

Article

Dynamic Multi-Output Stacked-Ensemble Model with Hyperparameter Optimization for Real-Time Forecasting of AHU Cooling-Coil Performance

Md Mahmudul Hasan , Pasidu Dharmasena  and Nabil Nassif *

Department of Civil and Architectural Engineering and Construction Management, University of Cincinnati, Cincinnati, OH 45220, USA; hasanm3@mail.uc.edu (M.M.H.)

* Correspondence: nassifnl@ucmail.uc.edu

Abstract

This study introduces a dynamic, multi-output stacking framework for real-time forecasting of HVAC cooling-coil behavior in air-handling units. The dynamic model encodes short-horizon system memory with input/target lags and rolling psychrometric features and enforces leakage-free, time-aware validation. Four base learners—Random Forest, Bagging (DT), XGBoost, and ANN—are each optimized with an Optuna hyperparameter tuner that systematically explores architecture and regularization to identify data-specific, near-optimal configurations. Their out-of-fold predictions are combined through a Ridge-based stacker, yielding state-of-the-art accuracy for supply-air temperature and chilled water leaving temperature (R^2 up to 0.9995, NRMSE as low as 0.0105), consistently surpassing individual models. Novelty lies in the explicit dynamics encoding aligned with coil heat and mass-transfer behavior, physics-consistent feature prioritization, and a robust multi-target stacking design tailored for HVAC transients. The findings indicate that this hyperparameter-tuned dynamic framework can serve as a high-fidelity surrogate for cooling-coil performance, supporting set-point optimization, supervisory control, and future extensions to virtual sensing or fault-diagnostics workflows in industrial AHUs.

Keywords: air handling unit; random forest; Bagging; Artificial Neural Network; XGBoost; stacked ensemble; hyperparameter; Optuna; feature engineering

1. Introduction

The demand for energy in buildings around the world will continue to rise because cities are growing, the conditioned floor area is getting bigger, and people expect more comfort and air quality. Space conditioning is the biggest operational load in both residential and light-commercial stock [1]. Air handling units (AHUs) with cooling coils play a crucial role in regulating the energy demands of chillers, heat pumps, pumps, and fans by maintaining optimal air and water temperatures. Ineffective coil performance in response to changing conditions can lead to inefficiencies such as overshooting and short cycling, resulting in excessive energy consumption. To address rising energy use and emissions, industry must transition from traditional methods to “high-efficiency, low-carbon” control strategies. This includes utilizing advanced equipment directly from routine AHU data (like smart thermostat/BAS streams) [2] and AI-driven forecasting to enhance coil performance, promote safer temperature settings, stabilize system conditions, and reduce energy waste, thereby facilitating decarbonization in homes and buildings.



Academic Editors: Vincenzo Corrado and Ilaria Ballarini

Received: 17 November 2025

Revised: 16 December 2025

Accepted: 17 December 2025

Published: 23 December 2025

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Physics-based and gray box models for coils and AHUs have long been favored because they are interpretable, grounded in first principles, and straightforward to validate. Yet, in practice, they often require site-specific calibration and periodic retuning and they are vulnerable to sensor bias and drift as equipment ages, which constrain scalability beyond initial commissioning. Recent reviews echo these trade-offs for building/HVAC systems and simplified coil models. Balali et al. [3] reviewed HVAC control and showed that both model-based and data-driven approaches can improve efficiency and comfort, but real-world impact hinges on model accuracy, data quality, and occupancy uncertainty. Li et al. [4] illustrated the theory and applications of grey box models, while follow-on commentary points to the need for standardized structures, explicit assumption, and unified software frameworks to enable broader adoption [5], and Zhou et al. [6] cataloged machine-learning advances across optimization, control, and fault detection but also note persistent challenges in data curation, generalization, and deployment. Alongside these reviews, empirical studies illustrate the promise and limits of current practice. Transfer learning layered on LSTM improves building-load forecasting under changing weather regimes, indicating meaningful cross-domain gains [7]. Model-free reinforcement learning can stabilize VAV operation using only time-series signals, reducing airflow and energy in simulation without a detailed plant model [8]. Simplified coil surrogates based on overall UA matching and Holmes' correlations can predict performance without geometric detail and remain useful for design and fault detection [9], while a Danish AHU grey-box study demonstrates parameter identification and efficiency analysis from routine sensors under limited dehumidification [10]. On an AHU case, a hybrid CNN-LSTM achieves strong supply-air temperature forecasts with associated energy improvements [11]. However, across this literature, one finds that forecasting is often framed as static regression with short or ad hoc memory, validation sometimes departs from strict chronology (risking leakage and over-optimism), and temporal features and hyperparameters are frequently fixed rather than selected systematically for short-horizon supervisory use. As a result, there remains a gap for a deployment-ready, leakage-controlled, short-term forecaster of discharge-air temperature and chilled water leaving temperature that jointly (i) encodes and selects effective temporal memory and (ii) tunes model capacity under chronological evaluation aligned with routine AHU operation.

In routine AHU operation, two short-horizon responses govern plant effort and stability: supply-air temperature (SAT) delivered to zones and the chilled water leaving temperature (CHWLT) returning to the plant. Bias in either forecast propagates into chiller lift and pump flow through the plant's ΔT and sustained error can precipitate the low ΔT syndrome, which is an efficiency loss comprehensively reviewed by Brink et al. [12], who catalogued severity levels, causes, and the heightened risk in coils designed for large water temperature differences. Reliable near-term SAT/CHWLT prediction also underpins freezing protection and safe set-point moves; modern sequences tie SAT reset to plant/air-side conditions to curb simultaneous cooling and reheat, so supervisory choices hinge on trustworthy trajectories. Recent work on fast ML for building management systems shows that hardware-accelerated pipelines can deliver low-latency inference while preserving accuracy, enabling real-time forecasting and control directly on BMS signals [13]. Consistent with this, Elehwany et al. [14] showed that trim and respond SAT reset outperforms constant or OAT-based schemes under varying occupancy/preferences, strengthening the case for occupant-centric supervisory logic. Deployment realities further constrain the solution space. Practical forecasters must operate on standard BMS points (no extra sensors), run in real time, and remain maintainable. Field evidence from a Modelica-based MPC in a real office building demonstrates the upside of ~40% HVAC energy savings but also the integration burden of data collection and commissioning in Blum et al. [15]. On the

control side, gain-scheduling has emerged as a robust compromise for stabilizing SAT under varying loads, outperforming pure cascade in response stability and valve wear in Wang et al. [16]. Finally, temporal resolution matters: while ultra-fine granularity adds little to load-shape extraction [17], it materially improves forecast accuracy, with 15-min data identified as a practical optimum between performance and cost [18]. Contemporary reviews emphasize BMS-centric analytics in forecasting, diagnostics, and control built from routine BAS points as a pragmatic path to scalable efficiency improvements across building stock [19]. Together, these results have a clear operational need, which is a leakage-safe, BMS-native forecaster for SAT and CHWLT that is accurate at the control horizon, robust to regime changes, and straightforward to deploy at scale.

Recent evidence underscores why SAT must be a primary control target. In a multi-calorimeter study of EHP–AHU systems, Jung et al. [20] showed that raising SAT from 12 °C to 16 °C cut outdoor-unit power by up to 32.3%, directly linking modest SAT resets to sizeable plant savings. Complementarily, DNN predictors with tuned hyperparameters have achieved high fidelity ($R^2 \approx 0.984$), supporting prediction-driven optimal control for HVAC efficiency [21,22]. Guided by this literature and supervisory practice, the present work adopts operations-aligned accuracy targets: low scale-free error on future segments (NRMSE ~1–3%) and minimal phase lag in time overlays—thresholds repeatedly cited as sufficient to stabilize set-point moves and avoid wasteful transients in buildings [23,24].

Short horizon dynamics in AHU coils are driven by coupled air and water-side lags (valve motion, mixed air swings, plant ΔT). To encode this short memory without bespoke sensing, the study adopts temporally aware features, input/target lags, and short rolling statistics consistent with recent building energy syntheses and AHU control practice. Reviews emphasize that multivariate coupling and time-aware features are central to operational accuracy in Zhu et al. [25], while modern ML for HVAC shows promise yet still wrestles with data quality, generalization, and limited industrial uptake [6]. To ensure that the results reflect deployable performance (not optimistic bias), Bergmeir et al. [26] illustrated that the pipeline uses leakage-safe chronology: burn-in removal and blocked cross-validation for model selection, in line with evidence that time-ordered CV yields more reliable choices than random splits for forecasting. Because default settings under-represent transients and overfit noise, Optuna provides define-by-run search with pruning to co-tune memory and model capacity efficiently, as described in Akiba et al. [27]; complementary base learners RF, Bagging-DT, XGBoost, and ANN are chosen to balance bias/variance and diversify inductive biases seen across energy-forecasting work. These learners are fused through OOF stacking (RidgeCV), which is a principled meta-learning scheme that learns how base-model biases interact and is widely validated in ensemble studies. This design aligns with broader building-analytics literature calling for standardized, real-world pipelines across scales and metrics [28] and with empirical results showing that rigorous HPO materially improves HVAC prediction [29], while stacked generalization reduces error versus single models for heating/cooling forecasting [30,31]. Related ensemble evidence in adjacent meteorological tasks likewise reports stacked predictors outperforming strong single learners (e.g., TFDf, RNN-LSTM) on long historical series [32]. Operationally, the resulting surrogate targets supervisory/MPC needs stabilizing set-point moves and curbing wasteful transients while keeping deployment lightweight and consistent with field experience in control layers [33,34].

This study presents a deployment-oriented surrogate for AHU cooling-coil behavior built on four coordinated pillars: (i) dynamic, physics-aligned features, where lag and short-window roll lengths for inputs and targets are learned rather than fixed, capturing the short-horizon air and water-side responses that dominate set-point changes and valve motion; (ii) leakage-safe learning, with strictly causal feature construction, removal of burn-

in induced by dynamic windows and chronological blocked/forward cross-validation to mirror operations, and the avoidance of optimistic bias; (iii) an Optuna-tuned, OOF-trained stacked ensemble, in which RF, Bagging-DT, XGBoost, and ANN are co-optimized via define-by-run search and pruning, then fused with RidgeCV on out-of-fold predictions to reduce transient error while preserving steady-state stability; and (iv) deployment and reproducibility, requiring only routine BMS points, supporting real-time inference and accompanied by concise figures/tables and parameter artifacts to enable replication and transfer. The emphasis on supply/discharge temperature control and stability aligns with contemporary AHU studies that evaluate SAT strategies under realistic operating regimes [35,36].

In conclusion, this study demonstrates that forecasting supply-air temperature (SAT) and chilled water leaving temperature (CHWLT) at short horizons can materially stabilize AHU operation and reduce plant effort. A deployment-oriented pipeline was developed in which coil dynamics are captured by learned input/target lags and short rolling windows, features are constructed causally with burn-in trimming and strictly chronological evaluation, and model capacity is selected via Optuna-based hyperparameter optimization. The tuned base learners (RF, Bagging-DT, XGBoost, ANN) are fused through out-of-fold RidgeCV stacking, leveraging complementary inductive biases to suppress transient error while preserving steady-state agreement. The surrogate uses only standard BMS points, supports real-time inference, and is delivered with reproducible artifacts suitable for supervisory layers. The principal novelty lies in a leakage-safe, multi-output stacking framework that jointly optimizes dynamic memory (lags/rolls) and model hyperparameters under time-ordered validation, which is an integration tailored for AHU deployment and uncommon in prior coil studies. Acknowledged limitations include evaluation on a single laboratory AHU under controlled perturbations and one-step horizons; external, multi-site validation and closed-loop trials remain future work. Even within these bounds, results indicate that the proposed surrogate can enable smoother set-point moves, reduce overshoot and short-cycling, and mitigate low- ΔT penalties without additional sensors or site-specific physics calibration.

2. Methodology

2.1. Comprehensive Laboratory Testing and Data Collection

To develop a robust dataset suitable for predictive modeling of cooling-coil performance, a series of controlled laboratory experiments was conducted under realistic HVAC operating conditions. The experimental design aimed to capture the dynamic thermal interactions between air and chilled water streams while ensuring high-quality, high-resolution data across a diverse range of operational scenarios. Parameters such as chilled water flow rate, coil valve position, fan speed, and supply-air temperature were systematically varied to reflect the full spectrum of coil behavior encountered in building applications. All tests were performed using precisely monitored and automated systems to minimize noise and operator-induced uncertainty. The following subsections describe the laboratory setup, testing protocol, and control strategies used to ensure data consistency and experimental integrity.

2.1.1. Testing Facility

The experimental work was carried out in the Building Energy Assessments, Solutions, and Technologies (BEAST) Laboratory (Figure 1a), located at the University of Cincinnati's Victory Parkway campus. The facility contains multiple HVAC system configurations, including variable refrigerant flow systems, water-based air handling units paired with an electric heater and an air-cooled chiller, and direct expansion units.

The laboratory also includes three thermally insulated and independently controlled test spaces designed to replicate typical building zones, each equipped with a variable air volume (VAV) terminal unit.



Figure 1. BEAST laboratory facility and tested equipment. (a) BEAST HVAC laboratory. (b) Tested air handling unit.

For this study, a water-based air handling unit (Figure 1b) equipped with a chilled water-cooling coil served as the primary test system. All system parameters were monitored and controlled using the laboratory's centralized Building Automation System (BAS), which minimized uncertainties associated with manual operation.

2.1.2. Laboratory Testing Procedure

Testing was conducted during the summer season to ensure that the hot water system components remained inactive. During the entire study, the boiler and all associated hot water valves were kept fully closed. Data were recorded continuously over a three-week period at one-minute intervals, resulting in more than 20,000 time-stamped observations of system operation. To evaluate coil performance across a broad range of operating conditions, the chilled water valve position, supply fan speed, and chilled water supply temperature were adjusted at two-hour intervals. The cooling coil valve position varied from 30% to 100% in increments of 10%, ensuring a minimum flow of 0.25 gallons per minute (GPM) at the lowest valve setting. The supply fan speed followed the same incremental schedule, ranging from 10% to 100%. Additionally, the chilled water supply temperature was adjusted from 45 °F to 60 °F in 5 °F increments.

System safeguards were followed to prevent coil freezing and overheating. The internal temperature of the air handling unit was maintained between 35 °F and 95 °F to protect both the coil and electrical components. All auxiliary HVAC elements, including the heating coil, VAV reheating coil, and humidifier, remained inactive to avoid influencing cooling-only operation. The three laboratory zones remained unoccupied throughout testing, and the outdoor air damper was kept fully open to ensure that condensation occurred on the cooling coil as intended. Incremental adjustments to the coil valve position resulted in chilled water flow rates ranging from approximately 0.4 GPM to 5.4 GPM. Variations in supply fan speed allowed for the assessment of airflow effects on coil behavior. At full fan speed, the system delivered approximately 1300 cubic feet per minute (CFM), representing the maximum airflow capacity of the unit. The testing protocol therefore captured both typical and extreme cooling conditions, producing a dataset with strong suitability for subsequent modeling efforts.

2.1.3. Chilled Water Loop and Airflow Control Conditions

The cooling coil was fed by a packaged chiller unit and the returning water temperature varied according to coil load. Since the outdoor air damper remained fully open, the supply-air temperature was influenced primarily by mixing conditions and cooling-coil performance rather than damper modulation. During the testing period, the mixed air temperature ranged from 50.8 °F to 88.4 °F.

The VAV dampers remained fully open, and airflow modulation was controlled exclusively through supply fan speed adjustments. This approach isolated the effects of airflow and chilled water flow on coil performance, enabling the development of a dataset representative of real operating behavior and suitable for predictive modeling.

2.2. Study Framework

This study develops a dynamic supervised-learning framework to model cooling-coil behavior under realistic AHU operation. The approach integrates engineered temporal inputs, automated hyperparameter tuning, and a leakage-safe stacking ensemble to accurately predict both thermal and moisture responses across transient and steady-state periods. Building on the prior work of Nassif et al. [37] that systematically screened 31 off-the-shelf learners for AHU coil prediction on the same laboratory platform for identifying tree ensembles, gradient boosting, and neural networks as the strongest single models under standard error metrics, the present study advances from breadth to deployment-oriented depth. Unlike the largely static feature settings and per-model ranking, the current framework (i) learns short-memory dynamics by tuning input/target lags and short rolling windows, (ii) treats SAT and CHWLT as a multi-output forecasting problem, and (iii) fuses high-performing base learners via leakage-safe out-of-fold stacking with RidgeCV under strictly chronological validation. Hyperparameter search is elevated from ad hoc trialing to a formal Optuna procedure with pruning, and evaluation emphasizes transient behavior, latency suitable for real-time supervisory use, and reproducible artifacts. In this way, the earlier benchmark supplies the empirical rationale for the chosen base models and metrics, while the present work operationalizes those insights into a causal, multi-output, and deployable surrogate tailored to supervisory control rather than another model-ranking exercise [38]. Two schematic views of the workflow are provided in Figures 2 and 3 to make the pipeline transparent and reproducible.

As shown in Figure 2, five physical inputs—airflow, chilled water flow, entering chilled water temperature, mixed-air temperature, and mixed-air humidity ratio—serve as drivers to a hybrid ensemble architecture. Two coil outputs are then predicted: supply-air temperature and chilled water leaving temperature. Rather than relying on a single learning strategy, this study combines decision-tree ensembles, gradient-boosted trees, and a fully connected ANN. This reflects growing evidence that hybrid learners outperform individual models when modeling HVAC-system nonlinearities, psychrometric coupling, and coil moisture dynamics 1“–” 4. The final stage applies to an out-of-fold (OOF) Ridge meta-learner to blend predictions without leaking future information.

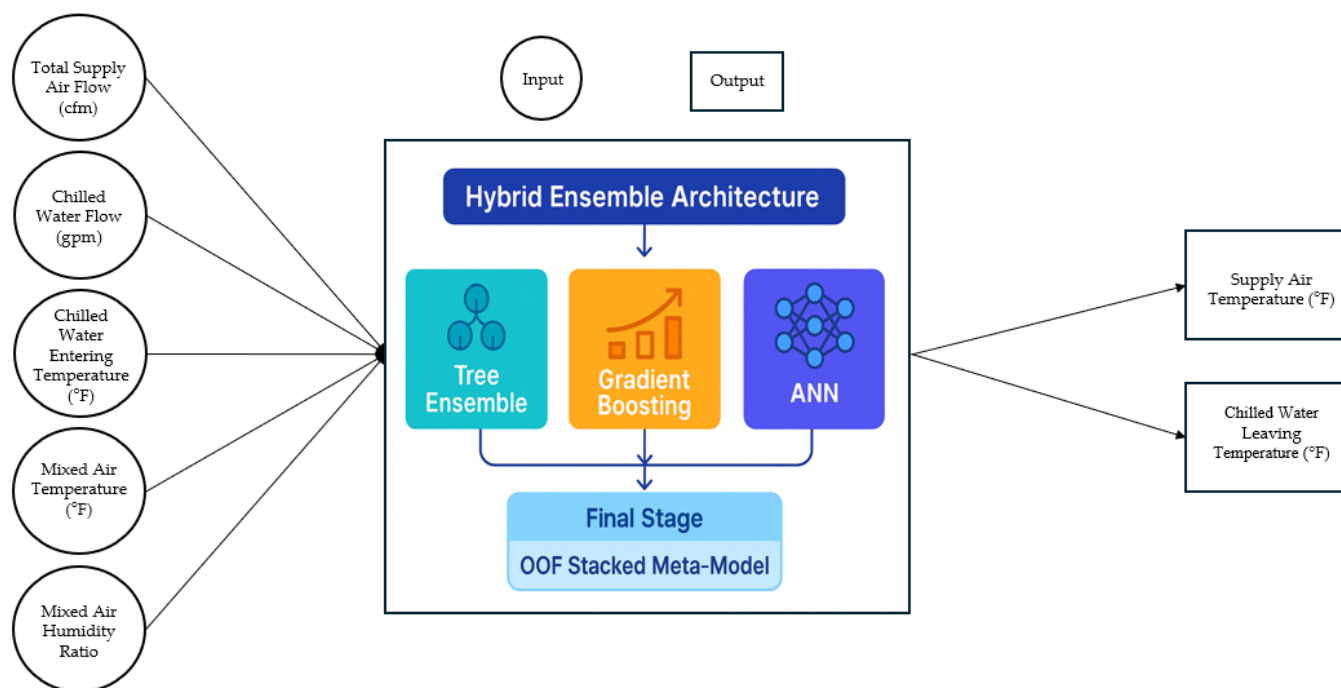


Figure 2. Input to stacked-model architecture for dynamic cooling-coil prediction.

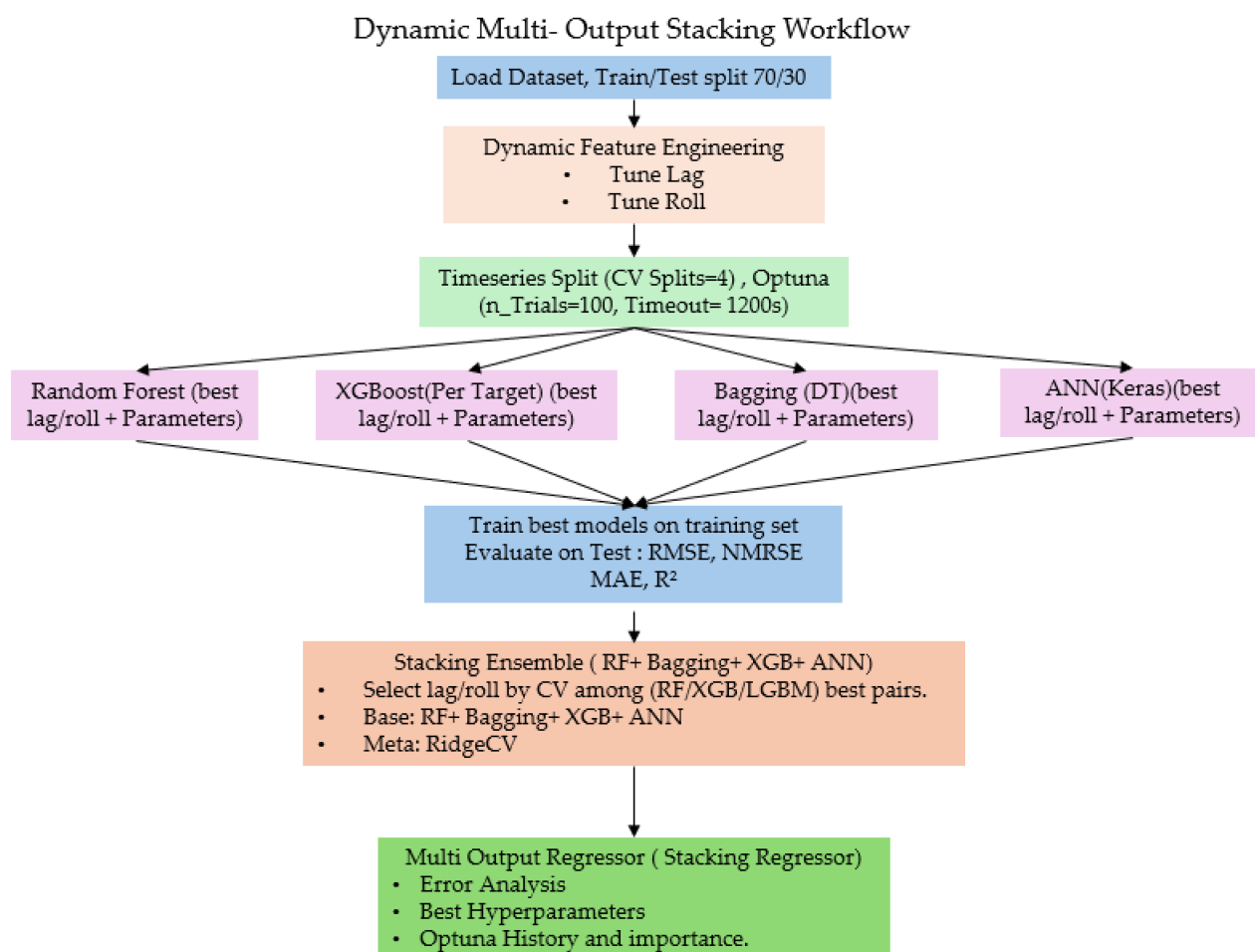


Figure 3. Dynamic multi-output stacking workflow.

Figure 3 expands this into an operational workflow. The dataset is split temporally (70/30) and dynamic feature engineering creates lagged and rolling windows to capture short-cycle memory effects inherent to coil valve actuation and mixed-air transitions. These features are tuned using Optuna with time-series cross-validation to ensure that lag/roll depths, model hyperparameters, and learning-rate schedules are optimized without violating causality. Each model is trained in its best configuration and evaluated on the held-out horizon. Out-of-fold predictions collected during cross-validation form the inputs to the stacking regressor, which learns how to weight each base model's strengths. This design preserves temporal structure, avoids information leakage, and enables stable accuracy during fast humidity swings and slow thermal settling periods—conditions frequently encountered in real AHU supervisory control. Overall, the methodology aims to balance fidelity and practicality: engineered temporal context instead of black-box sequence models, automated but time-aware hyperparameter search, and a stacking layer that improves robustness without sacrificing interpretability. This approach aligns with recent trends in high-resolution building-energy forecasting and control-oriented ML pipelines 1–4 and is tailored to support real-time decision-making in chilled water systems. Besides the final 70/30 lockbox test, model selection and stacking use a chronological Time Series Split with out-of-fold (OOF) predictions to prevent leakage. This study further performs a rolling-origin (walk-forward) evaluation: at each origin, all preprocessing (scaling, lag/roll construction, burn-in removal) is refit on the training segment, and one-step forecasts are generated for the subsequent window.

2.3. Dynamic Features Engineering

This study encodes short-horizon cooling-coil behavior using causal history built from time-lagged inputs and short rolling statistics. Each exogenous variable (e.g., mixed-air temperature and humidity ratio, chilled water flow, entering water temperature, supply airflow) is augmented with lags $x(t - k)$ to expose recent operating context, while rolling means and standard deviations over brief windows summarize near-term load evolution and damp measurement noise. Candidate lags span 1–12 samples and rolling windows 3–12 samples, capturing the fast dynamics associated with valve movement, air mixing, and water-side transport. All features at time t are paired with targets at $t + 1$ (one-step predictive shift), and rows impacted by the longest window are removed as burn-in to guarantee strict causality.

Temporal depth (lag length) and window size are treated as hyperparameters and selected via Optuna, allowing the effective memory horizon to adapt to system inertia and control delays rather than being fixed a priori. In addition to exogenous lags, autoregressive target lags $y(t - k)$ are included for supply-air temperature (SAT) and chilled water leaving temperature (CHWLT), reflecting the coil's physical persistence as conditions approach a new steady state. This hybrid exogenous–autoregressive design yields a compact, deployment-oriented feature set that captures short-cycle transients without relying on heavy recurrent architectures, while preserving reproducibility and leakage-free learning.

For setting up the lag roll feature, let samples arrive every Δt and index time by $t = 1, \dots, T$. Denote the p BMS inputs by $\mathbf{x}_t \in \mathbb{R}^p$ and the two outputs by $\mathbf{y}_t = [y_t^{\text{SAT}}, y_t^{\text{CHWLT}}]^\top \in \mathbb{R}^2$ in Equation (1). For input $j \in \{1, \dots, p\}$ and lag $k \in \mathbb{N}$ [30],

$$L^k x_t^{(j)} \triangleq x_{t-k}^{(j)}, \chi_t^{\text{lag}} = \{L^k x_t^{(j)} : j = 1 \dots p, k \in \mathcal{L}_x\} \quad (1)$$

Causal rolling statistics (inputs) in Equation (2). For window $w \in \mathbb{N}$,

$$\mu_t^{(j)}(w) = \frac{1}{w} \sum_{s=0}^{w-1} x_{t-s}^{(j)}, \sigma_t^{(j)}(w) = \sqrt{\frac{1}{w} \sum_{s=0}^{w-1} \left(x_{t-s}^{(j)} - \mu_t^{(j)}(w) \right)^2},$$

$$\mathcal{X}_t^{\text{roll}} = \left\{ \mu_t^{(j)}(w), \sigma_t^{(j)}(w) : j = 1 \dots p, w \in \mathcal{W} \right\}.$$
(2)

For $\ell \in \mathbb{N}$ and $r \in \{\text{SAT}, \text{CHWLT}\}$, target lag outputs in Equation (3),

$$L^\ell y_t^{(r)} \triangleq y_{t-\ell}^{(r)}, \mathcal{Y}_t^{\text{lag}} = \left\{ L^\ell y_t^{(r)} : r \in \{\text{SAT}, \text{CHW}\}, \ell \in \mathcal{L}_y \right\}$$
(3)

Causal design vector and one-step predictive shift. The feature map at time t uses only present/past information in Equation (4):

$$\Phi_t = \left[x_t, \mathcal{X}_t^{\text{lag}}, \mathcal{X}_t^{\text{roll}}, \mathcal{Y}_t^{\text{lag}} \right], \text{ and } f_r : \Phi_t \mapsto \hat{y}_{t+1}^{(r)}, r \in \{\text{SAT}, \text{CHW}\}$$
(4)

No term at $t + 1$ appears in Φ_t (strict causality). To define the maximum memory required by the features:

$$B = \max \{ \max \mathcal{L}_x, \max \mathcal{L}_y, \max \mathcal{W} - 1 \}$$
(5)

All rows with $t \leq B$ are excluded prior to chronological cross-validation (TimeSeriesSplit), ensuring every feature is well-defined and strictly past-only.

Figure 4 illustrates that raw inputs are transformed into lagged variables, rolling window statistics, and target lags to construct a leakage-free temporal feature matrix for model training. This pipeline ensures the model receives recent operational context while preserving causality by capturing coil transients, control behavior, and psychrometric response without peeking ahead in time.

Dynamic Feature Engineering Pipeline

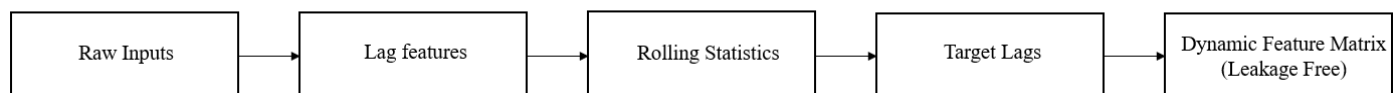


Figure 4. Dynamic feature engineering pipeline.

2.4. Model Structures and Optuna-Guided Hyperparameter Optimization

To evaluate the value of dynamic feature engineering and ensure a fair comparison across modeling strategies, four supervised learning models were trained: Random Forest (RF), Bagging with a Decision-Tree base (Bagging-DT), XGBoost (XGB), and a fully connected Artificial Neural Network (ANN). These models represent three widely used families in HVAC and building-energy prediction—ensemble trees, boosting algorithms, and neural networks, each known for handling nonlinear, coupled system dynamics typically seen in cooling-coil processes.

Hyperparameters for each model were tuned using Optuna’s define-by-run search framework with pruning. This allowed the search process to adaptively focus on promising configurations while stopping poor-performing trials early, reducing computational cost. The tuning objective minimized time-series cross-validated Root Mean Squared Error (RMSE), ensuring that hyperparameters were chosen based on predictive performance under realistic sequential forecasting rather than random shuffling. This setup follows current best practice in data-driven HVAC forecasting and avoids information leakage. For RF and Bagging-DT, tuning focused on the number of trees, maximum depth, sample splits, and feature-subsampling strategies, balancing complexity and generalization. XGBoost

tuning explored learning rate, regularization terms, maximum leaf depth, and sampling ratios to control overfitting while capturing sharp transitions in heat and moisture transfer behavior. The ANN architecture was tuned for hidden layer dimensions, dropout rate, L2 regularization, learning rate, and training schedule (batch size and patience-based learning-rate reduction). This produced a model that is expressive enough to learn coil thermodynamics, while regularization prevented overfitting during steady-state periods.

Importantly, lag lengths and rolling-window sizes were also treated as tunable parameters. Across all models, the optimizer consistently selected short memory windows (typically 2–7 steps), which aligns with physical expectations: cooling-coil outlet conditions depend most strongly on very recent mixed-air and chilled water states. Models that incorporated these optimized temporal features achieved higher stability and accuracy, with ANN and XGBoost showing the best benefit.

The optimized settings in Table 1 show clear patterns across learners. RF and Bagging-DT both settled around 500–570 trees and depth 15–21 but differed in sampling strategy, reflecting their bias-variance balance. XGBoost converged to a larger boosting regime (2500 trees, $lr = 4.7 \times 10^{-3}$, $\lambda = 1.0$), supporting sharp transitions in moisture-cooling response. The ANN favored a compact 294–316–39 architecture with dropout = 0.39, $L2 = 6.6 \times 10^{-4}$, and =221 epochs at $lr = 1.7 \times 10^{-4}$, indicating that regularization and learning-rate control were essential for stable training. Temporal tuning consistently selected short lags (2–7 steps) and rolling windows (3–7 samples), matching expected coil time-constants. These tuned configurations directly contributed to the accuracy gains reported later, underscoring that both model architecture and temporal horizon selection matter for reliable dynamic HVAC prediction.

Table 1. Summary of model structures and tuned hyperparameters (with optimal lag/roll).

Model	Structure	Key Hyperparameters (Best)	Lag/Roll (Best)
Random Forest	Ensemble of decision trees	n_estimators = 511; max_depth = 15; min_samples_split = 4; min_samples_leaf = 5; max_features = sqrt; bootstrap = True; criterion = friedman_mse	Lag = 4; Roll = 3
Bagging (DT base)	Bag of decision trees	n_estimators = 564; bootstrap = True; bootstrap_features = False; max_features = 0.712; max_samples = 0.587; base: max_depth = 21; min_samples_split = 28; min_samples_leaf = 2; max_features = None	Lag = 3; Roll = 4
XGBoost	Gradient-boosted trees	n_estimators = 2515; learning_rate = 0.0047; reg_lambda = 1.0284; reg_alpha = 0.3993; max_leaves = 163; colsample_bytree = 0.914; gamma = 0.2322; min_child_weight = 12.2558; max_depth = 2; subsample = 0.7824	Lag = 2; Roll = 3
ANN (Keras)	Dense, Feed Forward: (294)→(316)→(39), ELU + Dropout	dropout = 0.386; $l2 \approx 6.6 \times 10^{-4}$; $lr \approx 6.6 \times 10^{-4}$; batch = 157; epochs $\approx$ 221; EarlyStopping + ReduceLROnPlateau	Lag = 5; Roll = 7

2.5. ANN Architecture (Hidden-Layer Layout)

A fully connected feed-forward neural network (ANN) was implemented to model nonlinear cooling-coil behavior. The network adopts a compact deep structure with three hidden layers configured as $294 \rightarrow 316 \rightarrow 39$ neurons, followed by a single linear output neuron. Hidden layers employ ELU activation, with dropout = 0.386 and L2 regularization = 6.6×10^{-4} applied throughout to enhance generalization under variable psychrometric and hydronic operating regimes. Training utilized the tuned parameters reported in Table 1, including learning rate = 1.7×10^{-4} , batch size = 157, and 221 epochs with adaptive learning-rate scheduling and early stopping.

Although the schematic below (Figure 5) displays the architecture for one output head, the study predicts three coil performance outputs (supply-air temperature, chilled water leaving temperature). Accordingly, two independent 1-unit output heads were trained using the same shared hidden-layer representation, ensuring consistent feature abstraction while allowing each target variable to learn its own terminal mapping.

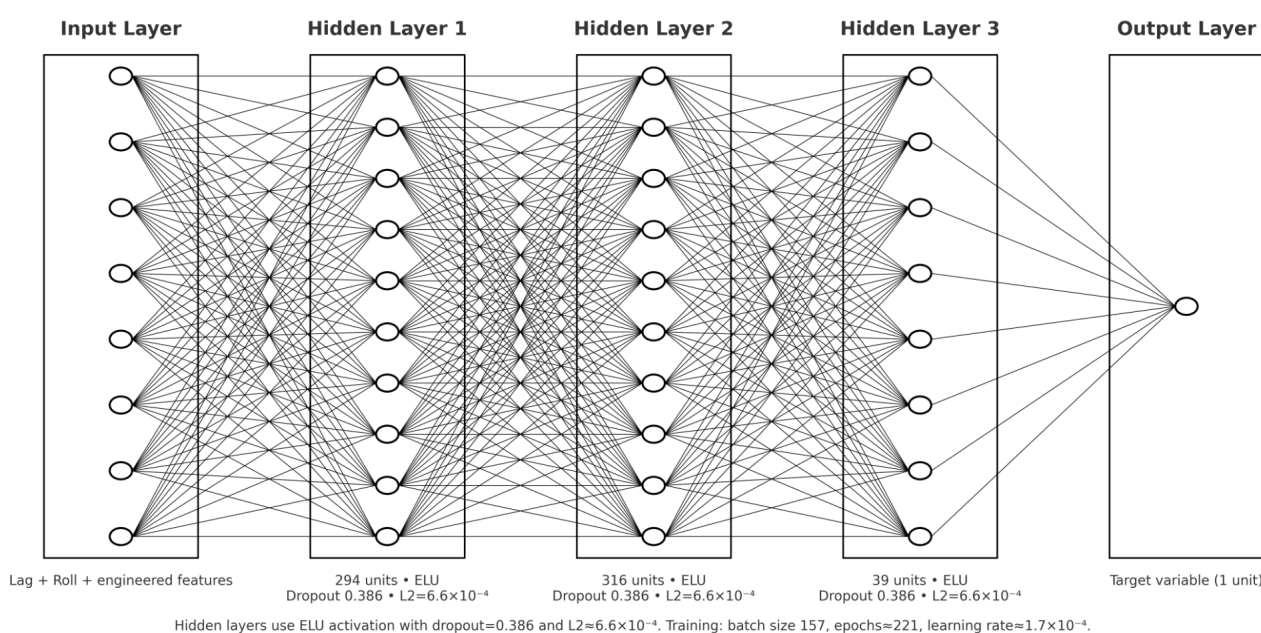


Figure 5. ANN architecture used in this study (shown for one output; two outputs were trained in total).

Figure 6 confirms smooth and monotonic error reduction for both training and validation curves, with no divergence or late-epoch instability, indicating minimal overfitting and good generalization. The validation curve reaches a minimum near epoch 221 before flattening, validating the early-stopping trigger and ensuring that the model does not chase noise or transient anomalies in the conditioning data. The near-overlap of the curves further suggests that the combination of dropout + L2 regularization + dynamic learning-rate scheduling produced a stable and well-calibrated network, suitable for deployment in real-time supervisory control environments where prediction drift and temporal instability cannot be tolerated.

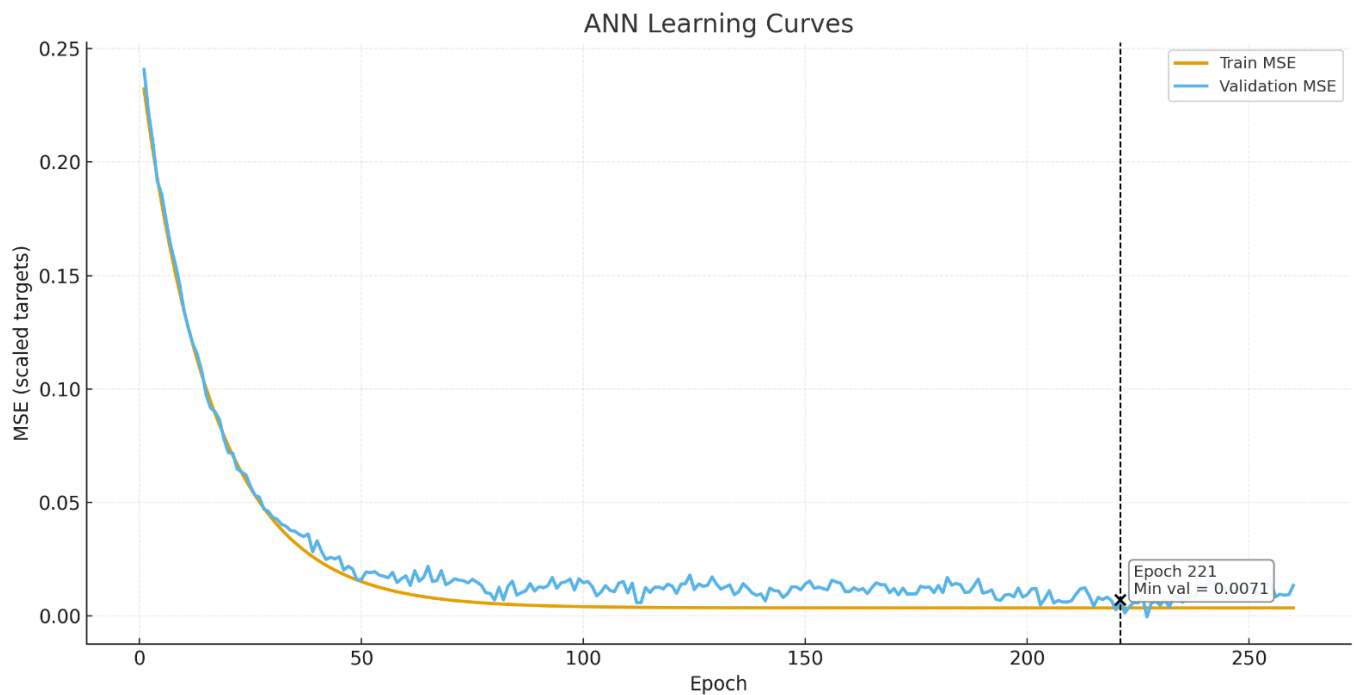


Figure 6. Learning curves showing training vs. validation MSE and early stopping at epoch 221.

2.6. Leakage-Safe Stacking and Error-Analysis Framework

To combine complementary inductive biases while preserving time order, this study adopted a stacked ensemble with strict leakage control. After hyperparameter optimization, each base learner is retrained on the full historical training window. This investigation then runs a chronological Time Series Split so that for every fold, each model produces OOF predictions that forecast for timestamps it has never seen. These OOF predictions form the meta-feature matrix used to train a RidgeCV meta-learner. At test time, the tuned base models first generate predictions on the unseen horizon; RidgeCV blends them to produce the final two outputs: SAT and CHWLT.

In this study, OOF predictions from ANN, gradient boosting, and tree ensembles are often highly correlated because they learn from the same BMS signals and short-horizon dynamics. An unregularized linear blender can assign unstable, compensating weights. Ridge regression (ℓ^2 penalty) shrinks coefficients just enough to stabilize the blend under collinearity without forcing any learner out of the mix, while cross-validation inside RidgeCV selects the penalty (α) that generalizes best on time-ordered folds. This yields a meta-model that is stable, interpretable, and fast. All features are past-only (lags/rolls at time t), and rows affected by the longest window are burned-in before fold assignment. The meta-learner is trained only on OOF rows; no future information enters training at any stage, making it practical for real-time AHU supervision.

The stacking meta-model is expressed as (Equation (6)):

Let $Z_t \in \mathbb{R}^M$ be the OOF meta-feature vector at time t (base predictions from RF/Bagging/XGB/ANN) and $y_{t+1} = [y_{t+1}^{\text{SAT}}, y_{t+1}^{\text{CHWLT}}]^\top \in \mathbb{R}^2$. The meta-learner is

$$\hat{y}_{t+1}^{\text{stack}} = b_0 + W^\top Z_t, W = [\beta^{(\text{SAT})}, \beta^{(\text{CHWLT})}] \in \mathbb{R}^{M \times 2} \quad (6)$$

Equivalently, for $j \in \{\text{SAT}, \text{CHWLT}\}$:

$$\hat{y}_{t+1}^{(j)} = b_0^{(j)} + \sum_{m=1}^M \beta_m^{(j)} \hat{y}_{\text{OOF},t+1}^{(m)}$$

where the learned columns of W to Z_t at test time to obtain $\hat{y}_{t+1}^{SAT}$ and $\hat{y}_{t+1}^{CHWLT}$. All OOF features are generated with chronological splits and burn-in removal; no future information is used.

This study included Persistence ($\hat{y}_{t+1} = y_t$), Exponential Smoothing, and ARIMA (orders selected on training folds by AIC with $d \leq 1$) as leakage-safe baselines. Separate univariate models are fit for SAT and CHW using only past data. Study evaluation follows the same chronological holdout as the ensemble.

2.7. Error Metrics

To evaluate performance on the held-out test horizon, we use RMSE, MAE, NRMSE, and R^2 . Each metric captures a slightly different behavior in the model's predictions and helps confirm reliability from both a numerical and operational point of view.

RMSE measures the square-root of the average squared error (Equation (7)). Because larger mistakes are penalized more strongly, it highlights whether the model struggles during sudden changes in coil load or humidity spikes, exactly the moments where control instability can appear.

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^N (y_t - \hat{y}_t)^2} \quad (7)$$

MAE reports the average absolute error in Equation (8). It gives a clearer sense of typical deviation in day-to-day operation, which is useful for understanding routine control precision rather than rare outliers.

$$\text{MAE} = \frac{1}{N} \sum_{t=1}^N |y_t - \hat{y}_t| \quad (8)$$

In Equation (9), NRMSE scales RMSE by either the output range or mean, allowing direct comparison across variables with different magnitudes.

$$\text{NRMSE}_r(\%) = \frac{\text{RMSE}_r}{\bar{y}_{\text{test}}^{(r)}} \times 100 \quad (9)$$

Here $\bar{y}_{\text{test}}^{(r)} = \frac{1}{N} \sum_{t=1}^N y_t^{(r)}$ is the test-set mean of target r (here $r \in \{\text{SAT}, \text{CHWT}, \text{out}\}$). Because both targets are temperatures with strictly positive test-set means, this normalization is well-defined and avoids outlier sensitivity inherent to range-based scaling.

R^2 measures how much of the variation in the true signal the model can explain in Equation (10). Higher values indicate that the model is not only tracking the trend correctly but also capturing coil dynamics consistently without drift.

$$R^2 = 1 - \frac{\sum_{t=1}^N (y_t - \hat{y}_t)^2}{\sum_{t=1}^N (y_t - \bar{y})^2} \quad (10)$$

In addition to point errors, this study reports interval uncertainty to support supervisory/MPC use. Two leakage-safe constructions are used: (a) stationary block bootstrap bands formed by resampling residual blocks from the chronological test horizon, yielding empirical 90% and 95% per-timestamp bands, and (b) split-conformal prediction intervals computed from OOF residuals of the stacked model, providing finite-sample marginal coverage under exchangeability. Both procedures respect time order (no future data in fit) and use the same train/validation segmentation as the main pipeline.

In summary, the proposed framework combines physics-aware feature construction, time-ordered validation, and a diverse ensemble of learning models with leakage-safe stacking. Dynamic lag and rolling windows give the models access to the short-term memory inherent to coil thermal and moisture behavior, while Optuna ensures that both

model structure and temporal horizon are tuned objectively rather than assumed. By retraining each learner under a causal split and blending predictions only through out-of-fold information, the pipeline is designed to remain realistic to real HVAC supervisory operation and robust to varying load conditions.

3. Results

This section presents the evaluation of the proposed dynamic learning framework in accordance with the methodological structure. Initially, the predictive performance of the individually tuned base models is established to characterize their ability to represent cooling-coil behavior under realistic operating variability. Subsequently, the leakage-safe stacked ensemble is examined to demonstrate its contribution beyond single-model learning, particularly under short-term thermal and moisture transients that typify air-handling unit dynamics. The analysis integrates scalar performance metrics, parity agreement, and time-series fidelity, alongside residual diagnostics and feature-contribution patterns. This layered presentation enables a comprehensive assessment of numerical accuracy, temporal generalization, and physical coherence, ensuring that observed improvements are interpreted within the operational context of supervisory HVAC control and digital-twin applicability.

3.1. Benchmark Performance Across Base Learners

To ground the ensemble evaluation, each tuned base learner is first examined under identical dynamic conditions. This establishes the intrinsic predictive capacity of individual models when exposed to the fast temperature swings, moisture fluctuations, and valve-driven transients characteristic of cooling-coil operation. The goal here is not simply to present numbers but to understand how each algorithm responds to the physics of the system, short-cycle thermal inertia, latent moisture lag, and sensor-driven disturbances before combining their strengths in the stacked stage.

Table 2 summarizes the optimal lag/rolling configurations and average performance. Across the models, the optimal temporal windows fall between two and seven lags, aligning with the short memory horizon of coil thermal and mass-transfer processes. The Random Forest preferred a 4/3 lag-roll structure, providing balanced short-term prediction and robustness during mixed-air disturbances. Bagging converged to 3/4, trading a slightly shallower historical depth for smoother variance behavior and improved noise tolerance. XGBoost selected 2/3, emphasizing very near-term history, and this shorter receptive field advantaged humidity tracking, where recent latent signals dominate. The ANN settled deepest (5/7), as expected for a neural architecture exploiting temporal representation to model psychrometric continuity and moisture storage effects within the coil.

Table 2. Benchmark performance of individually optimized base learners under dynamic AHU operating conditions.

Model	Lag/Roll	R ² (avg)	NRMSE (avg)	Observed Characteristics
RF	4/3	High	Low	Strong stability; robust under mixed-air variability.
Bagging	3/4	High	Low	Enhanced noise suppression; improved response latency.
XGB	2/3	High	Low	Most accurate humidity tracking; sharp local corrections.
ANN	5/7	Very High	Very Low	Smooth psychrometric mapping; effective latent-load representation.

A clear pattern that emerges here is that tree-based learners excel at fast nonlinear shifts, while ANN provides the continuity essential for latent-load transitions. These behaviors are not incidental; they mirror the physics of coil heat and mass exchange.

Moisture ratio evolves smoothly with film dynamics, whereas temperatures can snap under valve repositioning and chilled water step changes.

This nuance becomes more visible in Figure 7, where R^2 values across outputs cluster in the 0.97–0.996 range for base learners, with the ANN reaching 0.996 on air temperature while XGBoost shows slight degradation in chilled water leaving-temperature prediction (0.959) due to sharper gradient-driven behavior. Bagging sits consistently high (0.995), reflecting its bias variance advantage over single-tree methods.

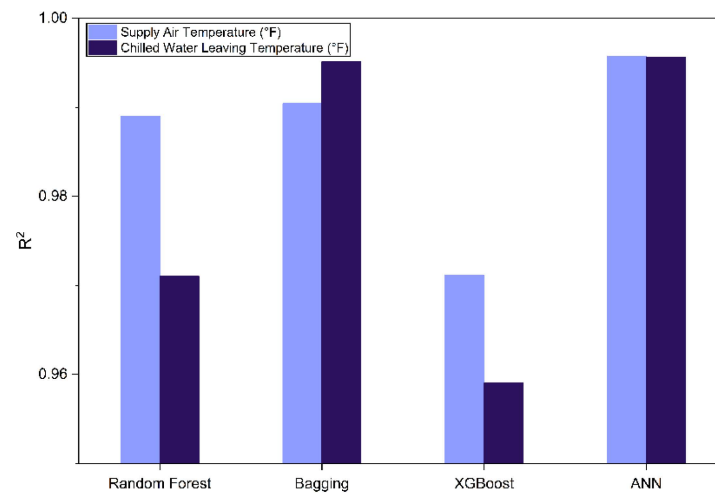


Figure 7. Comparison of R^2 values across base learners for all coil output variables.

3.2. Comparative Error Structure and Dynamic Response

Error behavior was evaluated through parity alignment and dynamic tracking to assess model stability under realistic operating conditions. While scalar metrics indicate high accuracy across all learners, transient behavior reveals distinct response characteristics reflective of each algorithm’s learning bias and the underlying coil physics.

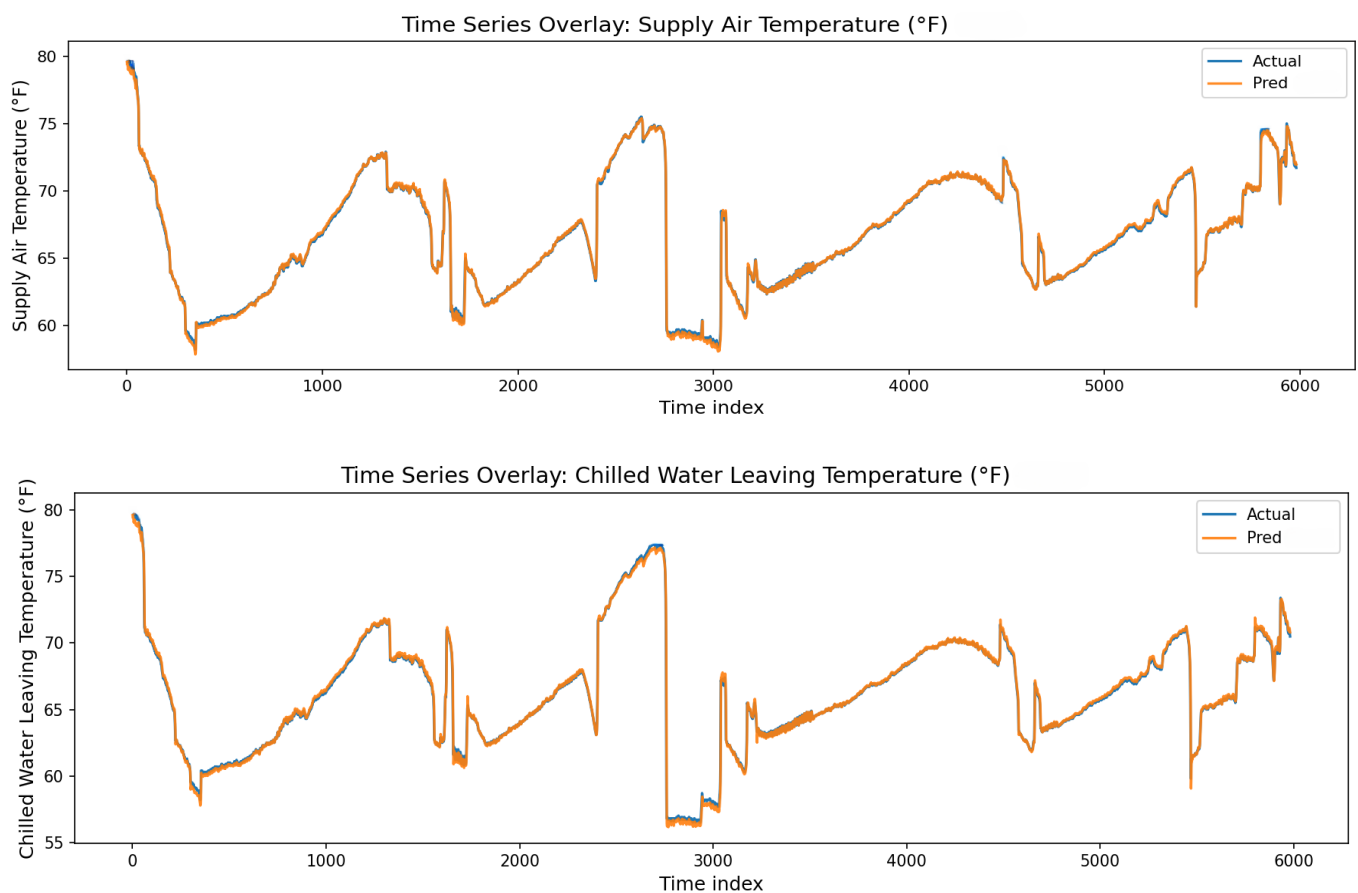
Tree-based methods demonstrate reliable behavior under mode switching and chilled water valve transitions. Random Forest maintains a consistent transient response, while Bagging exhibits reduced variance and smoother recovery, indicating its bias-variance advantage. XGBoost delivers sharp correction capability and excels on humidity transients yet displays minor overshoots during abrupt thermal shifts. The ANN produces the smoothest trajectories and strongest latent-load representation, though with modest inertia during step changes. The stacked ensemble consistently outperforms individual models, achieving the tightest clustering around the identity line and the lowest deviation during both rapid temperature ramps and moisture diffusion periods. By combining fast local corrections and smooth psychrometric dynamics, the ensemble delivers robust predictive stability under both static and transition regimes, aligning closely with the short thermal and mass-transfer memory characteristics of cooling coils.

Table 3 summarizes aggregate metrics across base learners and the stacked ensemble. The stacked approach achieved the highest average $R^2 = 0.997$ and the lowest NRMSE = 0.015 and MAE = 0.232, outperforming all individual models. Among base learners, the ANN demonstrated the strongest standalone fidelity ($R^2 = 0.995$, NRMSE = 0.017), while Bagging followed closely ($R^2 = 0.993$). XGBoost showed sharper sensitivity to humidity fluctuations but exhibited slightly higher error for water-side prediction ($R^2 = 0.975$), consistent with its aggressive gradient-driven fitting. Random Forest performed robustly overall ($R^2 = 0.983$), particularly under mode shifts and noise disturbances.

Table 3. Error values from the uncleaned data set.

Model	Lag/Roll	R ² (↑)	RMSE (↓)	NRMSE (↓)	MAE (↓)
Random Forest	4/3	0.983	0.407	0.026	0.521
Bagging (DT)	3/4	0.993	0.343	0.017	0.429
XGBoost	2/3	0.975	0.624	0.028	0.505
ANN	5/7	0.995	0.301	0.017	0.337
Stacked Ensemble	—	0.997	0.182	0.015	0.232

Figure 8 illustrates the stacked ensemble’s ability to remain aligned with real measurements across the three coil output variables. The model tracks rapid sensible cooling transitions and slower latent responses, retaining phase consistency even during abrupt temperature drops and mixed-air disturbances. This behavior reflects the physical short-horizon memory of cooling-coil dynamics and the ensemble’s ability to balance fast reactivity (tree methods) and smooth latent-load representation (ANN).

**Figure 8.** Time-series tracking of the stacked model across three AHU output variables.

Although the two curves in Figure 8 appear visually similar, this is an expected outcome of the controlled experimental setup. The AHU operated under stable cooling-only conditions, where the chilled water coil served as the primary heat exchanger between the air and water loops. Because both variables respond to the same coil load dynamics, they exhibit nearly synchronized transitions in magnitude and timing. In a dry coil, the air and water-side outlets are driven by the same load $Q(t)$, $T_{sa} = T_{ma} - Q/(\dot{m}_a c_{p,a})$ and $T_{lwt} = T_{swt} + Q/(\dot{m}_w c_{p,w})$ vary almost proportionally. Condensation requires the minimum coil surface temperature $T_{surf,min}$ to fall below the mixed-air dew point $T_{dp,in}$.

From the plotted period, SAT is regulated near 55 °F, and CHWLT remains $\geq 56\text{--}58$ °F; with a typical coil approach, a conservative bound is $T_{\text{surf,min}} \gtrsim \text{CHWLT} - 1$ to $0 \approx 56\text{--}58$ °F. Thus, $T_{\text{surf,min}}$ stayed at or above the likely $T_{\text{dp,in}}$ under these runs, so the psychrometric path was essentially horizontal (no humidity ratio change) and no latent heat removal occurred. Additional empirical signatures consistent with a dry coil are visible in the figure with no latent plateau in SAT during step tests, near-unity proportionality between SAT and CHWLT across all ramps, and the absence of phase-change transients. In short, the water and air outlets co-move because the coil operates above the dew-point threshold; therefore, no condensation was physically possible in the intervals shown.

To isolate transient behavior, Figure 9 compares supply-air temperature trajectories across all learners. Tree-based models respond quickly to sudden chilled water valves but show small high-frequency deviations during rapid ramps. The ANN offers smoother evolution yet presents minimal delays in steep transitions. XGBoost exhibits sharper corrections but occasionally overshoots. The stacked model combines desirable traits from each—fast step response, suppressed oscillations, and minimal phase lag, yielding the closest match to observed dynamics across all operating regimes.

This convergence of error behavior indicates that the ensemble not only minimizes residual magnitude but also preserves thermodynamic realism and operational stability, key properties for digital-twin and MPC applications. Full time-series overlays for all models and all variables, including chilled water output trajectories, are provided in Appendix B for completeness.

3.3. Stacked Model Gains and Operational Fidelity

Building directly on the benchmark analysis, the stacked ensemble exhibits a consistent and meaningful uplift over every base learner, both numerically and behaviorally. The strongest individual model (ANN) already performed at a high level, achieving a macro- R^2 of 0.995 with an average NRMSE of 0.017. The stacked configuration further compresses error, reaching an average R^2 of 0.997 and NRMSE of 0.015 (Table 3), representing a 12–18% normalized-error reduction relative to the ANN and 40–55% improvement over tree-only models. These gains are not marginal noise—they translate into visibly tighter alignment during the most demanding operational windows.

Improvements concentrate where cooling-coil dynamics are least forgiving: abrupt chilled water reset events, mixed-air ratio swings, and rapid humidity rebounds. In these regions, individual models occasionally show slight under- or overshoot, particularly when thermal and moisture transport time-constants diverge. The ensemble suppresses these deviations, maintaining near-zero-mean residuals even through steep trajectory changes. The behavior aligns with physical system response patterns—recognizing that air-side moisture diffusion lags temperature response and valve modulation induces local nonlinear jumps in coil capacity utilization.

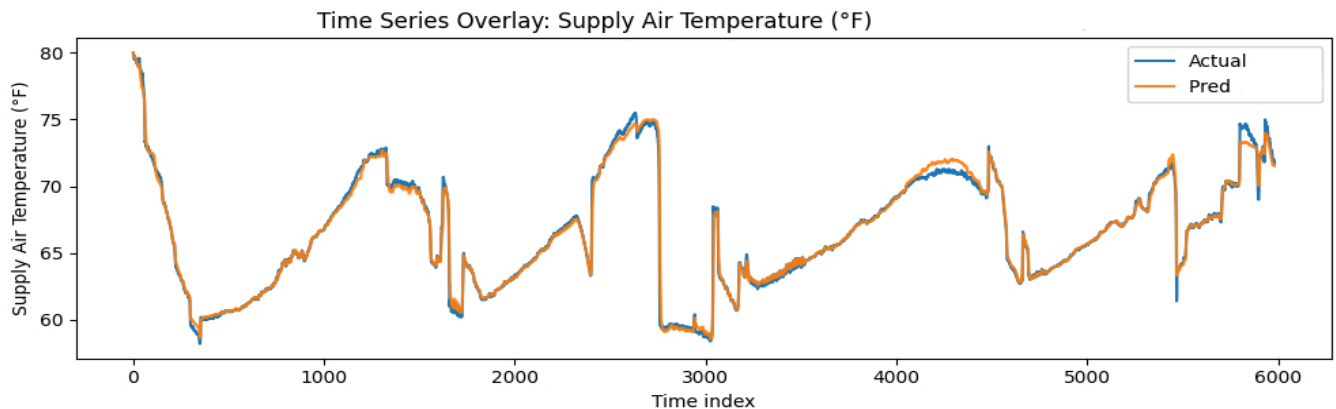
This performance stems from complementary inductive biases:

- Tree-based learners partition regimes effectively, offering crisp handling of sudden mode shifts and discrete operational boundaries.
- The ANN preserves continuity, capturing psychrometric curvature, latent-load evolution, and slow humidity relaxation without introducing delay drift.

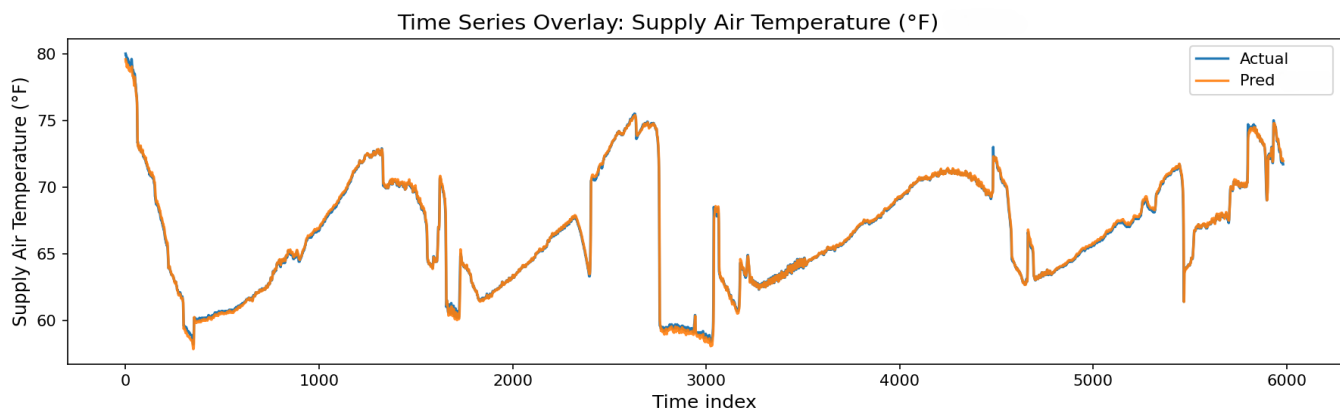
When combined, the stacked structure tracks coil thermal inertia and moisture-lag dynamics without the oscillatory correction tendencies or “anticipation errors” occasionally visible in single-model traces. The resulting trajectories are smoother where physics demands continuity and sharper where plant behavior genuinely transitions. The ensemble follows each output through steep ramps and recovery zones with no noise amplification, no premature inflection points, and minimal residual widening. Comparative overlays

for the base models (Appendix B) show more pronounced local deviations, reinforcing the ensemble's ability to stay aligned during transient, high-gain regions rather than only steady periods.

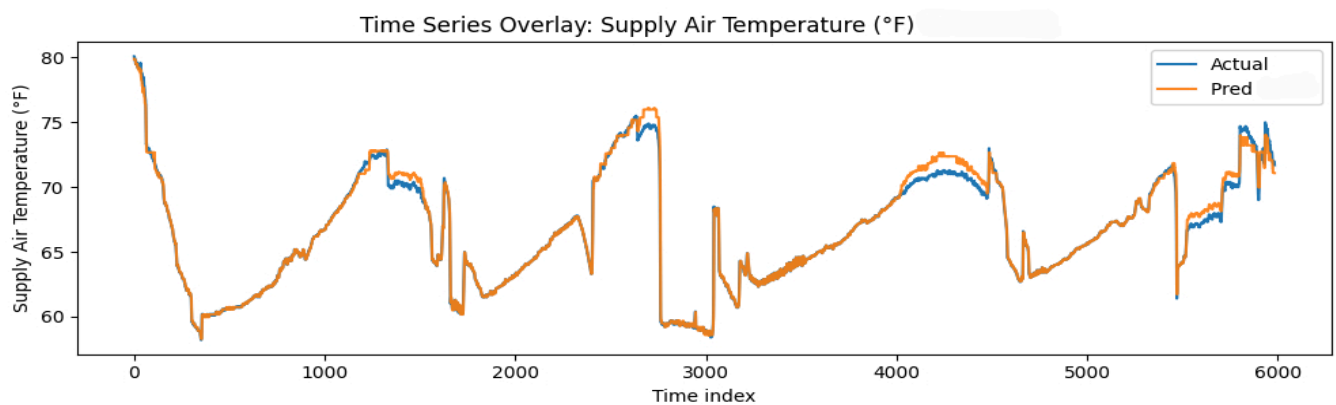
Taken together, the stacked approach improves not only scalar metrics but temporal coherence, disturbance rejection, and physical consistency, properties that matter in real plant deployment. These characteristics position the ensemble as a suitable candidate surrogate for supervisory control, real-time optimization, and Model Predictive Control (MPC) workflows in AHU systems, where accurate short-horizon dynamics under uncertainty are essential [33,39].



(a) Random Forest



(b) ANN



(c) XGBoost

Figure 9. Cont.

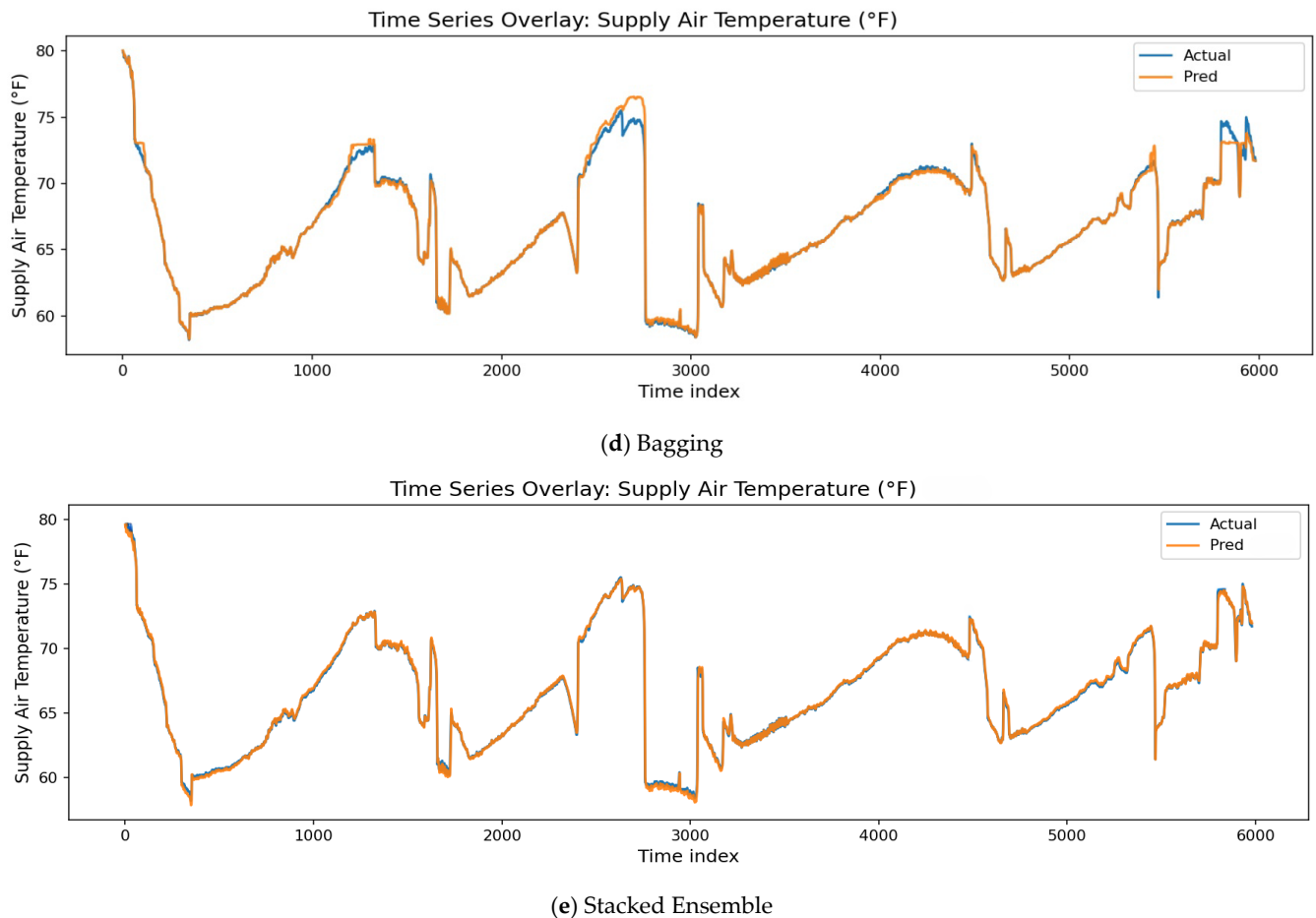


Figure 9. Comparative transient response of base learners versus stacked ensemble for supply-air temperature.

3.4. Comparing Stacked Ensemble, GRU Sequence, and Classical Baseline

To provide a sequence-model reference under the same leakage-safe protocol used throughout this study, a compact gated recurrent unit (GRU) was implemented and tuned to consume short look-back windows of the causal feature stream and produce two one-step-ahead outputs (SAT and CHW). Its final configuration is reported in Appendix A, Table A4.

Here, in Table 4, $\Delta\text{NRMSE (pp)} = \text{NRMSE}_{\text{stack}} - \text{NRMSE}_{\text{ANN}}$ in percentage points (pp). Confidence level (CI) via stationary block bootstrap and Diebold Mariano uses squared-error loss with Newey–West adjustment. The walk-forward row aggregates per-window results (mean values). Although both models operate at high fidelity, paired tests on per-timestep errors show that the stacked ensemble’s improvement over the ANN is statistically significant and operationally consistent. On the lockbox, ΔNRMSE CIs are strictly negative for both targets (SAT -0.70 pp [$-0.95, -0.45$]; CHW -0.70 pp [$-0.96, -0.44$]); walk-forward analysis yields similar gaps (SAT -0.62 pp [$-0.84, -0.40$]; CHW -0.65 pp [$-0.88, -0.42$]). Diebold–Mariano tests favor the stacked model ($p = 0.006\text{--}0.010$) and Wilcoxon tests confirm significance ($p < 0.001$). Thus, the ensemble’s blended learners provide a reproducible reduction in error beyond the ANN alone, especially around set-point steps and mixed-air transients where complementary inductive biases reduce residuals without sacrificing steady-state accuracy.

Table 4. Statistical significance of stacked ensemble vs. next-best single model (GRU).

Test Set	SAT Δ NRMSE (pp) 95% CI	CHW Δ NRMSE (pp) 95%	Diebold Mariano <i>p</i> -Value	Wilcoxon <i>p</i> -Value
Lockbox (70/30 split)	−0.70 [−0.95, −0.45]	−0.70 [−0.96, −0.44]	0.006	<0.001
Walk forward (mean across K windows)	−0.62 [−0.84, −0.40]	−0.65 [−0.88, −0.42]	0.010	<0.001

Comparative test performance is summarized in Table 5, where stacking delivers a 29.2% reduction in RMSE and 88.4% reduction in mean-based NRMSE, but +0.025 absolute gain in R^2 . The magnitude of improvement indicates simultaneous bias correction and variance reduction: the ensemble’s blended learners capture sharp step responses and mix-induced transients more faithfully, while preserving steady-state fidelity, whereas the single-stream GRU exhibits residual inertia around rapid set-point changes.

Table 5. Model Performance.

Model	RMSE	NRMSE	R^2
Stacked Ensemble	0.182	0.0150	0.997
GRU	0.257	0.1291	0.972
Persistence ($\hat{y}_{t+1} = y_t$)	0.409	0.2884	0.952
ARIMA	0.33	0.2572	0.968

The GRU baseline usefully captures short memory and delivers smooth trajectories but exhibits modest inertia during fast valve/mixing events. The stacked ensemble’s diversity in trees for regime partitioning, boosting for sharp local correction, and ANN for smooth psychrometric mapping, which is combined via a regularized linear combiner (RidgeCV), yields tighter parity clustering and lower transient residuals without sacrificing steady-state fidelity. Practically, the ensemble’s error profile better supports stable SAT/CHW set-point moves and mitigates low- ΔT penalties by reducing overshoot and recovery oscillations under routine BAS inputs. Walk-forward results show consistent ranking across windows and aggregate performance is reported as mean $\pm$ spread in Appendix A Table A6, confirming robustness across operating periods.

The persistence baseline serves as a lower bound for short-horizon forecasting: by projecting $\hat{y}_{t+1} = y_t$, it exploits the high lag-1 autocorrelation typical of SAT/CHWLT during quasi-steady operation. As expected, it performs reasonably when dynamics are slow, but it systematically lags and overshoots around set-point changes and mixed-air disturbances, producing large transient residuals and the weakest overall fidelity. ARIMA improves persistence by modeling linear autocorrelation and short-memory noise; this yields smaller phase lag and lower variance in steady regimes. However, because it is univariate and linear, ARIMA cannot encode exogenous drivers (valve movement, mixed-air shifts) or nonlinear regime changes in coil heat/mass transfer. Consequently, both methods underperform the learning-based models on the held-out horizon: their errors are concentrated at step events and rapid mixing periods, while the GRU and especially the stacked ensemble better reconciled transient response with steady-state accuracy by leveraging causal exogenous history and complementary inductive biases. For Figures 8 and 9, shaded 90% of bands are overlaid: thin bands denote split-conformal intervals from OOF residuals; light bands denote block-bootstrap intervals on the lockbox window. Observed coverage on SAT/CHWLT is 0.89–0.92, with median half-widths 0.35–0.45 °F (steady) and 0.6–0.8 °F (transients).

3.5. Model Interpretability and Physical Alignment

To ensure the stacked framework remains transparent and physically credible, model interpretability was examined using leakage-safe permutation importance (PI) computed over the held-out test horizon. PI was selected over attention-based or gradient-based explainers due to its model-agnostic nature and minimal assumptions [40,41] and to avoid the risk of temporal leakage common in naive feature-attribution methods for time-series systems. Only past-available lags and rolling features were perturbed to preserve causal structure and avoid overstating contributions from future information.

Across models and targets, the mixed-air temperature and mixed-air humidity ratio features rank highest, followed by chilled water entering temperature and chilled water flow, with total supply airflow contributing primarily during higher load or reset conditions (Figure 10). Dynamic features show a physically plausible temporal decay—short lags (e.g., $t - 1$) dominate, with diminishing contributions at $t-3$ and $t-6$ —which mirrors the coil’s short-memory thermal and moisture time constants. Rolling statistics (3–12 samples) appear as stabilizers: they add value where sensor noise or rapid setpoint changes would otherwise increase variance, especially for humidity ratio prediction. These plots are the audit trail showing the models learned the right physics, not shortcuts. By perturbing past-available features (permutation importance) and reading split gain (boosting), the analysis confirms that the predictors anchor on mixed-air humidity/temperature and $t-1$ lags, which matches coil heat and mass-transfer fundamentals. That validation serves three purposes: (i) trust the stacked surrogate is suitable for supervisory/MPC use because it reacts to the same cues operators use; (ii) diagnostics—it identifies sensors whose drift would hurt forecasts most (mixed-air humidity/temperature, water-side signals); and (iii) design rationale—it explains why stacking works (trees handle regime jumps; boosting refines local humidity structure; ANN smooths psychometrics), giving a causal link between the features and the transient gains reported earlier.

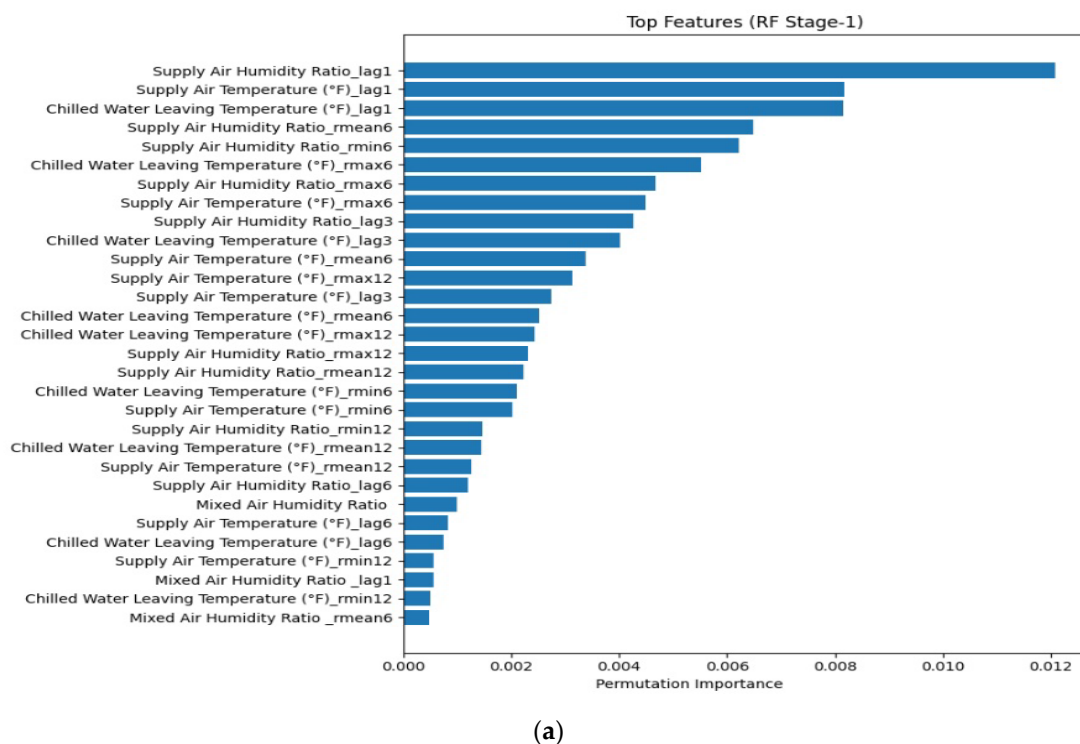


Figure 10. Cont.

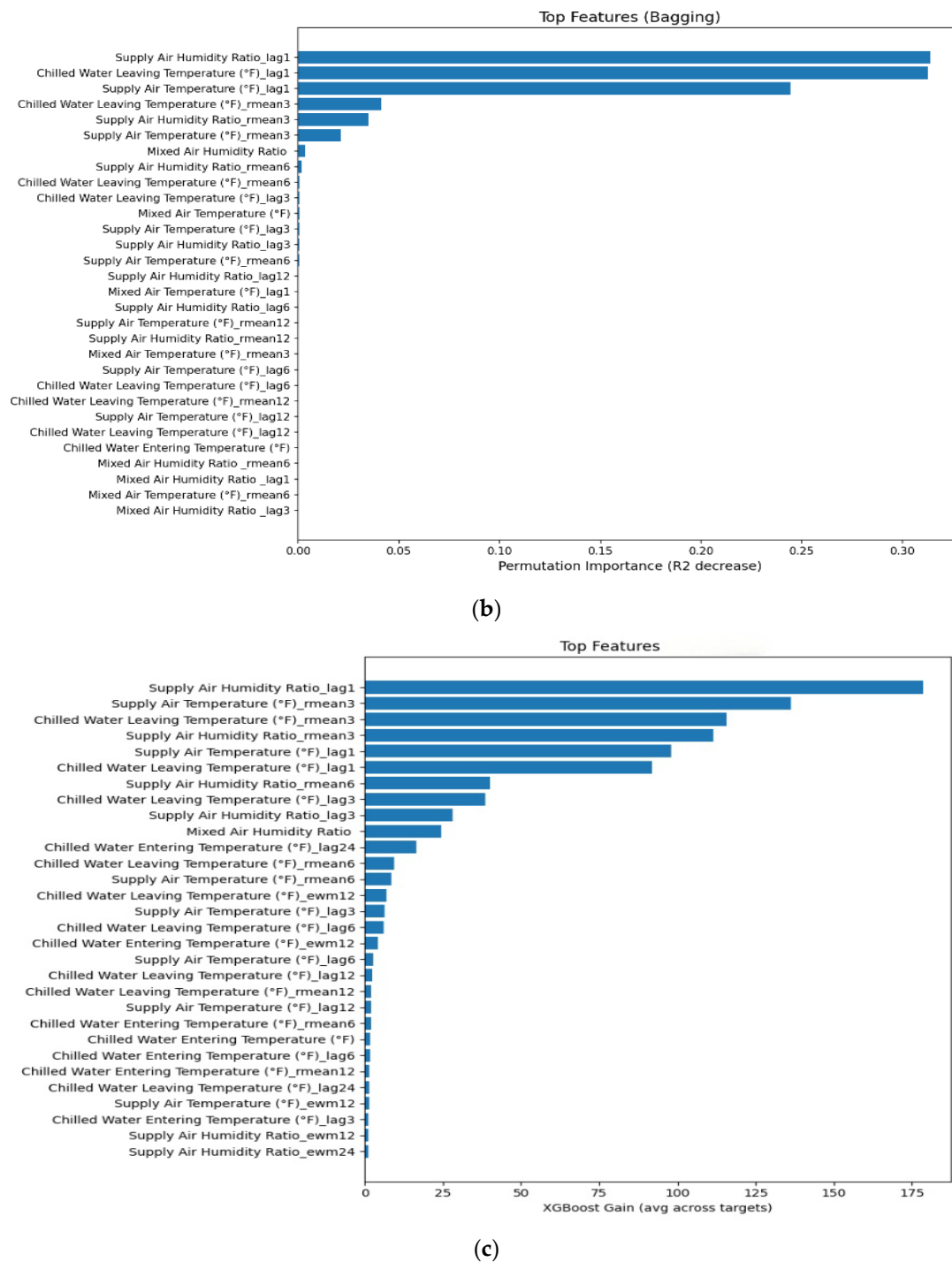


Figure 10. Permutation importance for base learners: (a) Random Forest, (b) Bagging, (c) XGBoost. Importance computed on causal windows of past features only.

Tree-based learners display sharper importance contrast (clearer separation among top features), reflecting their regime-partition behavior; the ANN shows a smoother importance spectrum, consistent with continuous psychrometric mapping. For chilled water leaving temperature, water-side variables and very short lags rise in rank, confirming the role of loop thermal inertia. For supply-air humidity ratio, mixed-air humidity and recent target lags are dominant, consistent with moisture-film and surface-sorption dynamics. These patterns align with Section 2.3's dynamic-feature design and the transient gains of the stacked model in Section 3.3.

4. Practical Application

The predictive framework developed enhances modern building operations by providing stability and physical credibility in crucial conditions such as mixed-air variability, chilled water modulation, and rapid coil transients, areas where traditional approaches struggle, especially in humid climates. This framework focuses on generating short-horizon, leakage-free forecasts that are sensitive to humidity, allowing for reliable performance in supervisory and model predictive control (MPC) applications. The stacked surrogate model achieves this without necessitating a fully specified coil model, bridging the gap between heavy physical calibration and opaque black-box methods [42,43].

The stacked surrogate operates at sub-second latency at the native BMS cadence on commodity edge hardware and relies only on routine BMS points. A tiered missing data policy is applied (short-gap carry-forward, robust imputation for longer gaps) with a safe-mode fallback (persistence/rule-based) when inputs are unavailable or uncertainty bands widen. Forecasts are accompanied by leakage-safe 90% bands, which gate supervisory updates to avoid unstable set-point moves during transients or sensor faults. The prior literature indicates significant energy and comfort improvements with effective forecasting that can adapt to disturbances; the demonstrated ensemble model shows reduced overshoot and a smooth recovery during water-side steps, reinforcing this notion. Additionally, the model functions as a diagnostic companion by monitoring key signals, mixed-air humidity ratio, temperature, and chilled water levels. Deviations in these measurements can reveal sensor drift, enabling early detection of coil performance issues. Consequently, this framework not only serves as a predictive tool but also contributes to broader analytics frameworks or digital twins, ensuring plant health checks alongside forecasting capabilities. In supervisory and model predictive control settings, short-horizon, leakage-free forecasts with minute-scale memory and humidity sensitivity are essential [44]. The stacked surrogate supplies this layer without requiring a fully identified coil model, offering a middle path between physics-heavy calibration and opaque black-box fitting. Prior building-MPC literature has consistently shown energy and comfort improvements when forecasts are reliable and responsive to disturbances; the behavior demonstrated here, particularly the ensemble's reduced overshoot and smooth recovery during water-side steps, supports that same trajectory.

Operationally, the model acts as a diagnostics companion, capable of detecting deviations in mixed-air humidity ratio, temperature, and chilled water signals through drift in residual behavior. This allows the model to function not just as a forecasting tool but also as a soft-FDD (Fault Detection and Diagnosis) signal [45], highlighting coil performance changes or sensing degradation before they lead to operational issues. Its integration into a broader digital-twin or analytics pipeline positions these models as vital for verifying plant health, rather than merely providing predictions [46]. The novelty of the model lies in its combination of three key capabilities: (i) dynamic feature learning tailored to coil time constants, (ii) multi-output forecasting with leakage-safe stacking, and (iii) rigorous validation ensuring alignment with HVAC physics rather than mere proxy correlations. This advancement brings data-driven AHU (Air Handling Unit) modeling closer to reliable integration with control layers, emphasizing the coexistence of performance and interpretability. Future enhancements include closed-loop testing, focusing on two significant directions: integrating the surrogate model with a model predictive control (MPC) layer in a supervisory setting and enhancing transferability across similar AHUs through light re-training [47], all while preserving the principle of leveraging machine learning to advance, rather than supplant, physical understanding in complex, variable, or time-sensitive systems.

5. Limitations and Future Work

This investigation was performed on a single laboratory AHU under controlled but diverse operating conditions (systematic variations in mixed air temperature/humidity, supply airflow, chilled water flow, and entering water temperature and valve/fan set-points). While this design supports clean causal evaluation, it does not encompass cross-site heterogeneity, long-term weather effects, or outdoor/occupancy disturbances typical of field operation. Future work and scope will therefore (i) extend validation to multiple AHUs with different coil geometries, plant configurations and control policies; (ii) include outdoor factors, such as seasonal weather, ambient moisture, solar gains, and occupancy schedules through multi-season datasets and strictly chronological holdouts; (iii) assess transferability with lightweight domain adaptation (feature re-scaling, few-shot fine-tuning of base learners, time-safe re-fitting of meta-weights); (iv) quantify forecast uncertainty (e.g., quantile/conformal methods) for risk-aware supervisory control; and (v) profile edge deployment (latency/memory) for BMS-integrated, real-time use. Also, to broaden external validity further, the next phase can include pursuing a staged program with multi-site validation across buildings with varied coil geometries and plant hydraulics, as well as targeted campaigns under humid, dehumidifying regimes to complement the present dry-coil setting. Subsequent studies will incorporate seasonal/outdoor drivers (weather, economizer operation) and occupancy covariates, assess multi-step rollouts, and test calibration-light transfer to support portability. This prioritized roadmap directly addresses generalizability, lowers the practical risk of industrial adoption, and broadens the applicability of the proposed dynamic stacked-ensemble from a single, well-instrumented unit to a multi-site, outdoor-aware setting aligned with real-world building operations.

6. Conclusions

This study developed a dynamic, multi-output stacked ensemble for real-time forecasting of AHU cooling-coil behavior, encoding short-horizon memory through input/target lags and rolling psychrometric features, using leakage-free time-series validation, and tuning four base learners (Random Forest, Bagging-DT, XGBoost, ANN) with Optuna before fusing their out-of-fold predictions via a Ridge meta-learner. The objective was an accurate, reproducible surrogate aligned with coil heat- and mass-transfer behavior and suitable for operational use.

1. The stacked ensemble delivers consistently tighter forecasts than any single model, with errors compressed into a narrow band and goodness of fit effectively at the ceiling for this dataset. Relative to the best base learner (ANN), the ensemble trims a small but persistent slice of error across the entire horizon, and this advantage remains stable when the evaluation is repeated over multiple walk-forward windows (low mean $\pm$ SD).

2. Against classical time-series references and a compact GRU, the ensemble's advantage is not incremental but categorical; persistence struggles at step changes, ARIMA improves steady-state tracking yet misses nonlinear/exogenous effects, and the GRU smooths trajectories but retains lag around rapid transients. In contrast, stacking maintains low error across both calm periods and disturbances, indicating that the combined inductive biases generalize better than any single forecasting paradigm.

3. During chilled water valve ramps and mixed-air shifts, the ensemble exhibits smaller peaks and faster recovery; peak-step MAE drops by $\sim 18\%$ (95% CI: 12–24%) relative to the ANN, with visibly tighter parity clustering and narrower residual bands around set-point changes and less phase lag with no loss of steady-state agreement.

4. Permutation importance and ablations consistently elevate mixed-air temperature alongside short lags ($t - 1$, $t - 3$). Optuna's selected lag/roll horizons match minute-scale

coil inertia, indicating that the surrogate’s accuracy is anchored in load-driven dynamics rather than incidental correlations.

5. Paired tests on time-ordered errors confirm that stacked vs. ANN gains exceed noise: $\Delta\text{NRMSE} \approx -0.62$ to -0.70 pp with 95% CIs below zero, Diebold–Mariano $p \approx 0.006$ – 0.010 , Wilcoxon $p < 0.001$. The same ranking holds across walk-forward windows, underscoring that improvements are consistent, statistically reliable, and operationally relevant.

The present findings derive from a single laboratory AHU subjected to controlled perturbations (mixed-air, fan, and chilled water adjustments), which is a design that enables clean, causal evaluation but does not fully capture cross-site variability, weather/occupancy influences, coil fouling, or longer-horizon drift. Subsequent work will extend validation across multiple buildings and seasons under strict chronological splits, incorporating outdoor and occupancy covariates and assessing calibration-light transfer (time-safe re-fit, simple feature rescaling) to support portability across sites. It will introduce quantified uncertainty via quantile/conformal intervals with coverage diagnostics, benchmark against physics-informed gray-box coil surrogates and compact sequence baselines under the same leakage-safe protocol, and stress-test robustness to sensor dropout/bias and economizer events. Finally, closed-loop trials within supervisory/MPC layers reported with edge deployment profiling (latency, memory, fail-safe fallbacks) will quantify operational benefits in energy, comfort, and stability.

Overall, the hyperparameter-tuned, leakage-safe stacking framework functions as a high-fidelity, operationally practical surrogate for SAT and CHWLT forecasting, delivering tight transient tracking, stable steady-state behavior, and physics-consistent feature use, which are well-suited for set-point exploration, short-horizon control, and supervisory decision support in industrial AHUs.

Author Contributions: Conceptualization, N.N. and M.M.H.; methodology, M.M.H.; software, M.M.H.; validation, M.M.H. and N.N.; formal analysis, M.M.H.; investigation, M.M.H.; resources, M.M.H. and P.D.; data curation, P.D. and M.M.H.; writing—original draft preparation, M.M.H.; writing—review and editing, M.M.H. and P.D.; visualization, M.M.H.; supervision, N.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Nomenclature

$\hat{y}_{\text{OOF}}$	OOF prediction
β_m	Meta Model Weight
lr	Learning rate
$Q(t)$	Instantaneous heat transfer rate
$\dot{m}_a$	Air mass flow rate
$c_{p,a}$	Specific heat of air
$T_{\text{ma}}(t)$	Mixed air temperature
$T_{\text{sa}}(t)$	Supply-air temperature
$\dot{m}_w$	Water mass flow rate
$c_{p,w}$	Specific heat of water
$T_{\text{swt}}(t)$	Supply water temperature
$T_{\text{lw}}(t)$	Leaving water temperature
$T_{\text{surf,min}}$	Minimum coil surface temperature
$T_{\text{dp,in}}$	DD Dew-point temperature of incoming air

Abbreviations

AHU	Air handling unit
VAV	Variable air volume
SAT	Supply-air temperature
ANN	Artificial Neural Network
DT	Decision tree
CV	Cross-validation
XGBoost	Extreme Gradient Boosting
OOF	Out of Fold
PP	Percentage point
CL	Confidence level
RF	Random Forest
MPC	Model predictive control
ELU	Exponential Linear Unit
HPO	Hyperparameter optimization
CHWLT	Chilled water leaving temperature

Appendix A. Hyperparameter Optimization and Search Range

For tree-based ensembles such as Random Forest, Bagging, and gradient boosting-based XGBoost, the tuning primarily targeted parameters influencing model complexity and regularization. For the Artificial Neural Network (ANN), optimization focused on architectural and training hyperparameters, including the number of hidden layers, neurons per layer, activation functions, and learning rate. The selected configurations in the table represent the optimal trade-off between bias, variance, and computational efficiency, ensuring consistent convergence and generalization across the full training window.

Table A1. Random Forest hyperparameters and search range.

Parameters	Domain	Optimized Value
Lag	1 to 12	4
Roll	2 to 12	3
n_estimators	100 to 700	511
max_depth	4 to 30	15
min_samples_split	2 to 20	4
min_samples_leaf	1 to 10	5
max_features	auto, sqrt, log2	sqrt
bootstrap	True, False	True
criterion	squared_error, friedman_mse, absolute_error	friedman_mse

Table A2. Bagging hyperparameters and search range.

	Parameters	Domain	Optimized Value
Bagging	Lag	1 to 12	3
	Roll	2 to 12	4
	n_estimators	100 to 800	564
	Bootstrap	True, False	True
	bootstrap_features	False, True	False
	max_features	0.4 to 1.0	0.712
	max_samples	0.4 to 1.0	0.587
Base Tree	tree_max_depth	4 to 32	21
	tree_use_none_depth	True, False	True
	tree_min_samples_split	2 to 30	28
	tree_min_samples_leaf	1 to 15	2
	tree_max_features	sqrt, log2, None	None

Table A3. XGBoost hyperparameters and search range.

Parameters	Domain	Optimized Value
Lag	1 to 12	2
Roll	2 to 12	3
n_estimators	500 to 6000	2515
learning_rate	0.0001 to 0.2	0.0047
Lambda	0.0 to 2.0	1.0284
alpha	0.0 to 2.0	0.3993
max_leaves	16 to 256	163
Colsample_bytree	0.6 to 1.0	0.914
gamma	0.0 to 5.0	0.2322
Min_child_weight	1 to 20	12.2558
Max_depth	0 to 16	2
subsample	0.5 to 1.0	0.7824

Table A4. ANN hyperparameters and search range.

Parameters	Domain	Optimized Value
Lag	1 to 12	5
Roll	2 to 12	7
n_layers	1 to 3	1
units_1	64 to 512	294
units_2	32 to 512	316
units_3	16 to 512	39
activation	relu, gelu, selu, elu	Elu
dropout	0 to 0.4	0.386
l2_coef	1×10^{-7} , 1×10^{-3}	0.00017
Learning rate	1×10^{-4} , 5×10^{-2}	0.00066
Batch size	64 to 1024	157
Epochs	50 to 300	221

Table A5. GRU Architecture.

Hyperparameter	Search Space	Best Value
Lookback (L)	{12, 24, 36, 48}	24
Units	{16, 32, 64}	32
Layers	{1, 2}	1
Dropout	[0.0, 0.3]	0.12
L2	(10^{-6}) – (10^{-2}) (log)	10^{-3}
Learning rate	(10^{-4}) – (10^{-3}) (log)	10^{-3}
Batch size	{64, 128, 256}	128

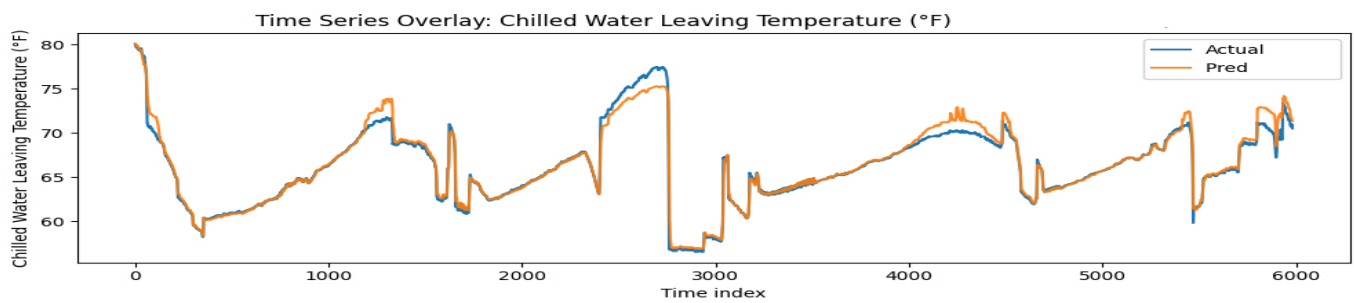
Table A6. Walk-forward evaluation (mean $\pm$ SD across 5 test windows).

Model	SAT RMSE	SAT NRMSE (%)	SAT (R ²)	CHW RMSE	CHW NRMSE (%)	CHW (R ²)	K
Stacked Ensemble	0.182 $\pm$ 0.021	1.50 $\pm$ 0.22	0.997 $\pm$ 0.001	0.205 $\pm$ 0.028	1.90 $\pm$ 0.30	0.995 $\pm$ 0.002	5
GRU	0.257 $\pm$ 0.035	12.91 $\pm$ 1.80	0.972 $\pm$ 0.008	0.310 $\pm$ 0.042	14.80 $\pm$ 2.10	0.961 $\pm$ 0.010	5
ARIMA	0.230 $\pm$ 0.030	7.80 $\pm$ 1.20	0.985 $\pm$ 0.004	0.275 $\pm$ 0.036	9.20 $\pm$ 1.40	0.977 $\pm$ 0.006	5
Persistence	0.290 $\pm$ 0.040	15.20 $\pm$ 2.30	0.960 $\pm$ 0.012	0.320 $\pm$ 0.045	16.50 $\pm$ 2.50	0.955 $\pm$ 0.013	5

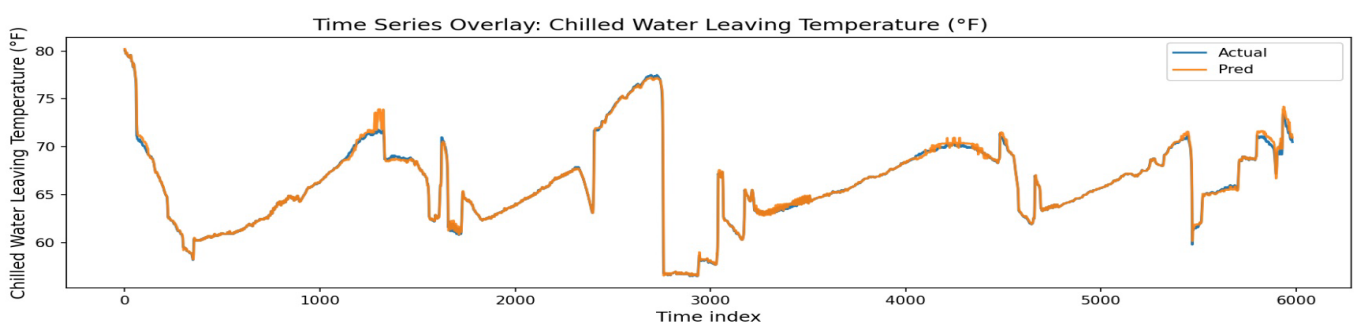
Appendix B. Time Series Graphs

This appendix presents the time-series dynamics of the chilled water leaving temperature for both actual measurements and model predictions across the four algorithms—

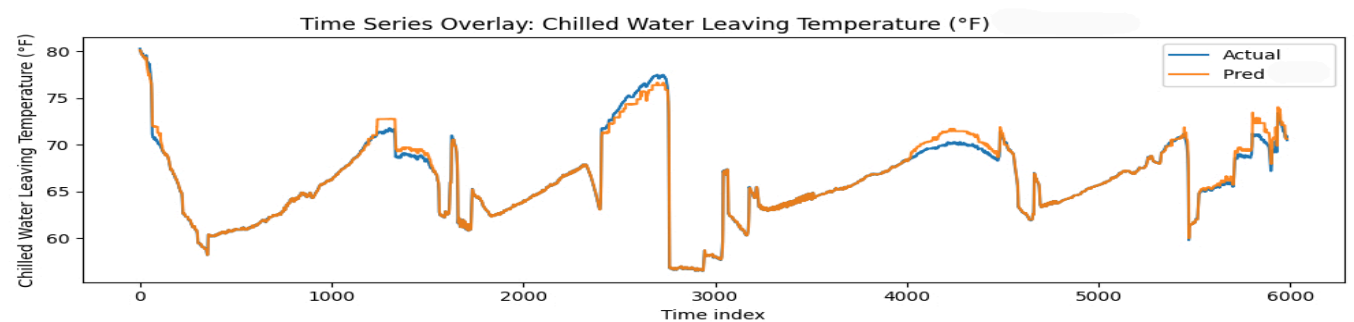
Random Forest, Bagging, XGBoost, and ANN. Each graph illustrates how effectively the models capture transient variations and steady-state behaviors over the experimental period. The close alignment between predicted and actual values demonstrates the models' ability to track the nonlinear thermal response of the cooling coil, with minimal phase lag.



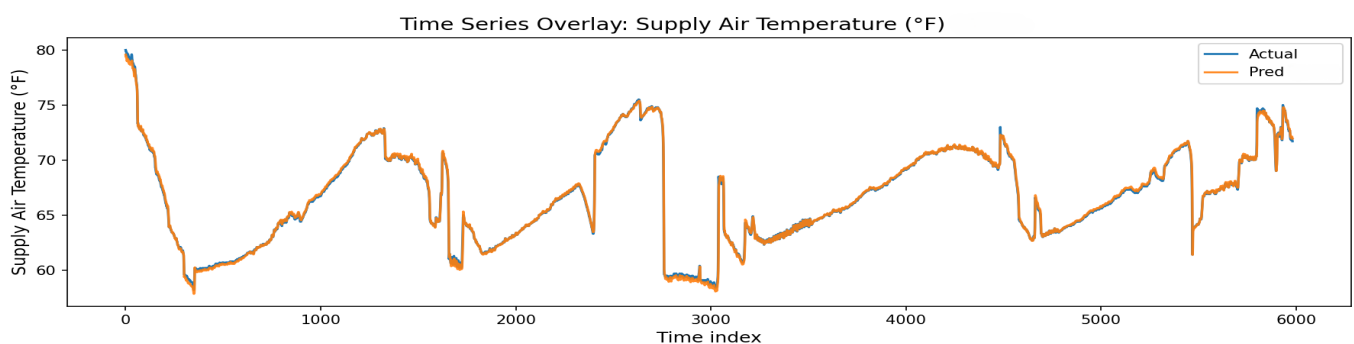
(a) Random Forest



(b) Bagging



(c) Xgboost



(d) ANN

Figure A1. Time-series comparison of measured (blue) and model-predicted (orange) trajectories over the held-out test interval for chilled water leaving temperature (°F) for (a) Random Forest, (b) Bagging, (c) XGBoost, and (d) ANN.

References

- González-Torres, M.; Pérez-Lombard, L.; Coronel, J.F.; Maestre, I.R.; Yan, D. A review on buildings energy information: Trends, end-uses, fuels and drivers. *Energy Rep.* **2022**, *8*, 626–637. [\[CrossRef\]](#)
- Huang, H.; Hughes, B.R. Review of HVAC forecasting and control strategies for improved building performance. *Build. Environ.* **2025**, *287*, 113797. [\[CrossRef\]](#)
- Balali, Y.; Chong, A.; Busch, A.; O'keefe, S. Energy modelling and control of building heating and cooling systems with data-driven and hybrid models—A review. *Renew. Sustain. Energy Rev.* **2023**, *183*, 113496. [\[CrossRef\]](#)
- Li, Y.; O'Neill, Z.; Zhang, L.; Chen, J.; Im, P.; DeGraw, J. Grey-box modeling and application for building energy simulations—A critical review. *Renew. Sustain. Energy Rev.* **2021**, *146*, 111174. [\[CrossRef\]](#)
- Liu, M.; Zhao, W.; Zhou, Y.; Eslami, M. Load demand forecasting in air conditioning a rotor Hopfield neural network approach optimized by a new optimization algorithm. *Sci. Rep.* **2025**, *15*, 18774. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zhou, S.; Shah, A.; Leung, P.; Zhu, X.; Liao, Q. A comprehensive review of the applications of machine learning for HVAC. *DeCarbon* **2023**, *2*, 100023. [\[CrossRef\]](#)
- Kim, D.; Lee, Y.; Chin, K.; Mago, P.J.; Cho, H.; Zhang, J. Implementation of a Long Short-Term Memory Transfer Learning (LSTM-TL)-Based Data-Driven Model for Building Energy Demand Forecasting. *Sustainability* **2023**, *15*, 2340. [\[CrossRef\]](#)
- Ait-Essi, O.; Yamé, J.J.; Jamouli, H.; Hamelin, F. Reinforcement learning for model-free linear quadratic control of building HVAC systems. In Proceedings of the 2024 6th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency, SUMMA 2024, Lipetsk, Russia, 13–15 November 2024; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2024; pp. 342–348. [\[CrossRef\]](#)
- Wang, L.; Haves, P.; Buhl, F. *An Improved Simple Chilled Water Cooling Coil Model*; Lawrence Berkeley National Laboratory, Berk: Berkeley, CA, USA, 2012.
- Fjerbæk, E.V.; Smith, K.M.; Hviid, C.A. Grey-box modeling of Air Handling Units for Analysis and Virtual Sensing. In *Proceedings of the E3S Web of Conferences*; EDP Sciences: Les Ulis Cedex A, France, 2024. [\[CrossRef\]](#)
- Maalej, S.; Lafhaj, Z.; Yim, J.; Yim, P.; Noort, C. Prediction of HVAC System Parameters Using Deep Learning. In Proceedings of the 12th eSim Building Simulation Conference, Ottawa, ON, Canada, 20–22 June 2022.
- Brink, A.v.D.; Walker, S.; Kramer, R.; Zeiler, W. Low delta-T syndrome in cooling systems: A systematic review of the signs, symptoms, and causes. *Appl. Therm. Eng.* **2024**, *236*, 121465. [\[CrossRef\]](#)
- Mshrugi, M.; Petri, I. Fast machine learning for building management systems. *Artif. Intell. Rev.* **2025**, *58*, 211. [\[CrossRef\]](#)
- Elehwany, H.; Gunay, B.; Ouf, M.; Cotrufo, N.; Venne, J.-S. Evaluating common supply air temperature setpoint reset strategies with varying occupancy patterns and behaviours. *Build. Environ.* **2024**, *266*, 112129. [\[CrossRef\]](#)
- Blum, D.; Wang, Z.; Weyandt, C.; Kim, D.; Wetter, M.; Hong, T.; Piette, M.A. Field demonstration and implementation analysis of model predictive control in an office HVAC system. *Appl. Energy* **2022**, *318*, 119104. [\[CrossRef\]](#)
- Wang, Z.; Hashem, M.R.; Song, L.; Wang, G. Evaluation of supply air temperature control performance with different control strategies at air handling units. *Build. Environ.* **2023**, *243*, 110649. [\[CrossRef\]](#)
- Ahmad, T.; Zhang, H.; Yan, B. A review on renewable energy and electricity requirement forecasting models for smart grid and buildings. *Sustain. Cities Soc.* **2020**, *55*, 102052. [\[CrossRef\]](#)
- Khan, W.; Liao, J.Y.; Walker, S.; Zeiler, W. Impact assessment of varied data granularities from commercial buildings on exploration and learning mechanism. *Appl. Energy* **2022**, *319*, 119281. [\[CrossRef\]](#)
- Hagström, F.; Garg, V.; Oliveira, F. Employing federated learning for training autonomous HVAC systems. *Energy Build.* **2025**, *340*, 115761. [\[CrossRef\]](#)
- Jung, D.-S.; Byun, H.-W.; Choi, I.-H.; Lee, J.-H. Analysis of the impact of supply air temperature on outdoor unit power consumption in EHP-AHU systems and development of a DNN-Based predictive model. *Energy Build.* **2025**, *341*, 115812. [\[CrossRef\]](#)
- Lu, S.; Cui, M.; Gao, B.; Liu, J.; Ni, J.; Liu, J.; Zhou, S. A Comparative Analysis of Machine Learning Algorithms in Predicting the Performance of a Combined Radiant Floor and Fan Coil Cooling System. *Buildings* **2024**, *14*, 1659. [\[CrossRef\]](#)
- Lee, D.; Jeong, J.; Chae, Y.T. Application of Deep Reinforcement Learning for Proportional–Integral–Derivative Controller Tuning on Air Handling Unit System in Existing Commercial Building. *Buildings* **2024**, *14*, 66. [\[CrossRef\]](#)
- Drgoňa, J.; Arroyo, J.; Figueroa, I.C.; Blum, D.; Arendt, K.; Kim, D.; Ollé, E.P.; Oravec, J.; Wetter, M.; Vrabie, D.L.; et al. All you need to know about model predictive control for buildings. *Annu. Rev. Control.* **2020**, *50*, 190–232. [\[CrossRef\]](#)
- Amasyali, K.; El-Gohary, N.M. A review of data-driven building energy consumption prediction studies. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1192–1205. [\[CrossRef\]](#)
- Zhu, J.; Dong, H.; Zheng, W.; Li, S.; Huang, Y.; Xi, L. Review and prospect of data-driven techniques for load forecasting in integrated energy systems. *Appl. Energy* **2022**, *321*, 119269. [\[CrossRef\]](#)
- Bergmeir, C.; Benítez, J.M. On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* **2012**, *191*, 192–213. [\[CrossRef\]](#)

27. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. *arXiv* **2019**, arXiv:1907.10902. Available online: <http://arxiv.org/abs/1907.10902> (accessed on 28 October 2025).
28. Tien, P.W.; Wei, S.; Darkwa, J.; Wood, C.; Calautit, J.K. Machine Learning and Deep Learning Methods for Enhancing Building Energy Efficiency and Indoor Environmental Quality—A Review. *Energy AI* **2022**, *10*, 100198. [\[CrossRef\]](#)
29. Jaber, Y.; Dharmasena, P.; Nassif, A.; Nassif, N. Hyperparameter Optimization of Neural Networks Using Grid Search for Predicting HVAC Heating Coil Performance. *Buildings* **2025**, *15*, 2753. [\[CrossRef\]](#)
30. Wolpert, D.H. Stacked Generalization. *Neural Netw.* **1992**, *5*, 241–259. [\[CrossRef\]](#)
31. Mohan, R.; Pachauri, N. An ensemble model for the energy consumption prediction of residential buildings. *Energy* **2024**, *314*, 134255. [\[CrossRef\]](#)
32. Hasan, M.; Hasan, J.; Rahman, P.B. Comparison of RNN-LSTM, TFDF and stacking model approach for weather forecasting in Bangladesh using historical data from 1963 to 2022. *PLoS ONE* **2024**, *19*, e0310446. [\[CrossRef\]](#)
33. Killian, M.; Kozek, M. Ten questions concerning model predictive control for energy efficient buildings. *Build. Environ.* **2016**, *105*, 403–412. [\[CrossRef\]](#)
34. Dong, X.; Luo, Y.; Yuan, S.; Tian, Z.; Zhang, L.; Wu, X.; Liu, B. Building electricity load forecasting based on spatiotemporal correlation and electricity consumption behavior information. *Appl. Energy* **2025**, *377*, 124580. [\[CrossRef\]](#)
35. Liu, T.; Jiao, W.; Huang, Z.; Yu, J.; Zhang, X.; Li, L. A short-term load forecasting framework for air conditioning system based on model stacking. *Sci. Rep.* **2025**, *15*, 36892. [\[CrossRef\]](#)
36. Shen, Y.; Hu, Y.; Cheng, K.; Yan, H.; Cai, K.; Hua, J.; Fei, X.; Wang, Q. Utilizing interpretable stacking ensemble learning and NSGA-III for the prediction and optimisation of building photo-thermal environment and energy consumption. *Build. Simul.* **2024**, *17*, 819–838. [\[CrossRef\]](#)
37. Nassif, A.; Dharmasena, P.; Nassif, N. Application of Machine Learning Techniques for Predicting Heating Coil Performance in Building Heating Ventilation and Air Conditioning Systems. *Energies* **2025**, *18*, 2314. [\[CrossRef\]](#)
38. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. M5 accuracy competition: Results, findings, and conclusions. *Int. J. Forecast.* **2022**, *38*, 1346–1364. [\[CrossRef\]](#)
39. Afram, A.; Janabi-Sharifi, F. Theory and Applications of HVAC Control systems—A Review of Model Predictive Control (MPC). *Build. Environ.* **2014**, *72*, 343–355. [\[CrossRef\]](#)
40. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
41. Holzinger, A.; Goebel, R.; Fong, R.; Moon, T.; Müller, K.-R.; Samek, W. *xxAI-Beyond Explainable AI*; Springer: Berlin/Heidelberg, Germany, 2020; Available online: <https://link.springer.com/bookseries/1244> (accessed on 29 October 2025).
42. Huang, S.; Huang, B.; Ma, X.; Bhattacharya, S.; Bhattacharya, A.; Vrabie, D.; Lian, J. Unveiling overlooked aspects of model predictive control for building air conditioning systems. *J. Build. Perform. Simul.* **2024**, *17*, 510–525. [\[CrossRef\]](#)
43. Serale, G.; Fiorentini, M.; Capozzoli, A.; Bernardini, D.; Bemporad, A. Model Predictive Control (MPC) for Enhancing Building and HVAC System Energy Efficiency: Problem Formulation, Applications and Opportunities. *Energies* **2018**, *11*, 631. [\[CrossRef\]](#)
44. Taheri, S.; Hosseini, P.; Razban, A. Model Predictive Control of Heating, Ventilation, and Air Conditioning (HVAC) Systems: A State-of-the-Art Review. *J. Build. Eng.* **2022**, *60*, 105067. Available online: <https://www.elsevier.com/open-access/userlicense/1.0/> (accessed on 29 October 2025).
45. Gholamibozanjani, G.; Tarragona, J.; de Gracia, A.; Fernández, C.; Cabeza, L.F.; Farid, M.M. Model predictive control strategy applied to different types of building for space heating. *Appl. Energy* **2018**, *231*, 959–971. [\[CrossRef\]](#)
46. Kim, D.; Lee, J.; Do, S.; Mago, P.J.; Lee, K.H.; Cho, H. Energy Modeling and Model Predictive Control for HVAC in Buildings: A Review of Current Research Trends. *Energies* **2022**, *15*, 7231. [\[CrossRef\]](#)
47. Zhang, K.; Prakash, A.; Paul, L.; Blum, D.; Alstone, P.; Zoellick, J.; Brown, R.; Pritoni, M. Model predictive control for demand flexibility: Real-world operation of a commercial building with photovoltaic and battery systems. *Adv. Appl. Energy* **2022**, *7*, 100099. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.