Journal of Applied Automation Technologies

Research Paper

Record-Driven Calibration of Cooling-Tower Performance Digital Twins Using Hiera-UNet

Domantas Vasiliauskas[1, *], Benas Balčiūnas[1] and Marius Pakalniškis[1]

[1] Faculty of Technical Sciences, Šiauliai Academy of Vilnius University, Šiauliai, LT-76352, Lithuania
*Corresponding author: domantas.va@sa.vu.lt

Abstract. Although design equations and commissioning tests are typically used to start cooling-tower digital twins, variations in fill resistance, nozzle distribution, fan condition, instrumentation bias, and other factors cause their accuracy to decrease. Although the records are asynchronous, regime-dependent, and primarily comprise ordinary operating data rather than controlled test findings, plant historians have gathered the information required for recalibration. A record-driven Hiera-UNet model for ongoing performance calibration of wet cooling-tower digital twins is presented in this study. A low-order heat- and mass-transfer twin is contrasted with historian tags arranged into multi-resolution operational windows. Seasonal environmental variations, shift-scale loading, and rapid actuator changes are all represented by a hierarchical attention encoder. In order to calculate corrections for outlet water temperature, approach, efficacy, and the four identifying twin characteristics, a U-Net style temporal decoder uses encoder features and physical residual channels. The calibration is limited by energy-balance, monotonicity, smoothness, and uncertainty losses. Fourteen months of one-minute recordings from six tower cells at the two industrial sites served as the foundation for the technique. Hiera-UNet maintained a 90% interval coverage of 91.3% on the held-out station while lowering the approach error by 71.6% and the outlet-temperature mean absolute error for the uncalibrated twin to 0.29 °C from 1.18 °C. Additionally, fill-performance decrease at 4.1% and simulated sensor bias at 0.12 °C are found. The findings of the ablation experiment show that physical residual limitations and hierarchical encoding have added the most accuracy. The end product is a digital twin calibration layer that can be audited without changing the engineering framework.

Keywords: *Cooling Tower Calibration, Digital Twin, Hiera-UNet, Heat and Mass Transfer, Performance Monitoring*

Received on 30 December 2022, Accepted on 09 April 2023, Published on 16 April 2023

Introduction

The heat produced by power, chemical, steel, data center, and district energy plants is cooled by cooling towers. In order to guarantee the dependability of outlet-water temperature, approach to ambient wet-bulb temperature, thermal effectiveness, cell capacity, fan energy, and water loss, operators require the aforementioned parameters. These quantities are provided by design models in a clean and balanced condition, but plant operations progressively depart from this presumption. Natural-draft tower performance under measurable conditions has been modeled using neural networks in field experiments [1]. Additionally, delayed degradation can be deduced from the performance history rather than a single inspection, according to Fouling-oriented models [2]. A digital twin can maintain the structure of mass and heat transport while synchronizing parameters with the actual tower. Calibration is the real issue. Wet-bulb temperature, water flow, heat load, fan command, bypass position, and cell availability all fluctuate simultaneously in historical records, which are not controlled trials. A tower model could be utilized to preserve validity away from a single design point, according to an online cooling-water optimization study [3]. Additionally, coupled-zone research has demonstrated that there will be a material prediction error if spray, fill, and rain-zone interactions are ignored

[4]. Therefore, environmental excitation, control action, sensor bias, and actual equipment aging should all be kept apart in a good calibration layer.

Even when the average flow is the same, the distribution of air is altered by crosswind and fan arrangement [5]. While maintaining the physical meaning, hybrid modeling can lower the computational cost of intricate heat-transfer computations [6]. We suggest a record-driven architecture that can learn non-linear, multi-scale residual patterns without losing the base twin based on the aforementioned findings. In this research, fast- and slow-operating modes are represented by a Hiera-style hierarchical temporal encoding, and calibrated performance channels are reconstructed at the original time resolution using a UNet-style decoder. A constrained residual estimation problem is the suggested workflow. The canonical tower record is mapped to the raw station tags once they have been quality-checked and synchronized. The base-level state variables are generated by a low-order twin. After receiving observations and model residuals, Hiera-UNet outputs uncertainty, four parameter multipliers, and performance corrections. During training, enforce temporal smoothness, monotonic response, water-energy balance, and parameter constraints. This structure can continue to support what-if analysis of the calibrated twin while saving an engineering explanation for every update.

The following are the four. First, a historian-aligned approach makes a distinction between regime change, sensor replacement, and missing evidence. Second, instead of using pictures for multichannel process records, a hierarchical encoder and full-resolution decoder are employed. Third, under physical limitations, the network collaboratively calibrates interpretable twin characteristics and performance outcomes. Fourth, using particular operational data, the assessment reports address cross-station transfer, seasonal drift, sensor-bias recovery, degradation sensitivity, interval calibration, and online runtime. We can't alter the acceptance requirements if we don't add some routine temperature tests.

Related Work

Cooling-Tower Performance Models

Empirical characteristic curves and thorough computational explanations of multiphase heat and mass transfer are examples of cooling tower models. The effects of interior configurations and external variables on evaporation and thermal duty are demonstrated using combined heat-exchange techniques [7]. Additionally, microbial study has demonstrated that the tower system is vulnerable to operating safety hazards and water-system deterioration of heat-transfer surfaces [8]. Although most models have been proven through specialized experiments or limited-range operations, the aforementioned research support continuous condition monitoring.

Machine learning has been used in condition monitoring research to detect changes in sensor residuals [9]. The aforementioned techniques can identify the anomalies, but a drift score does not instruct the operator on how to fix the digital twin. Verification, validation, parameter estimation, and uncertainty reporting are all included in calibration research in energy simulation [10]. The two are not the same, because adjusting parameter errors that don't work in a new context could lead to a low prediction error.

Digital-Twin Calibration

Using measured or high-fidelity residuals, online digital-twin calibration has been developed as a quick, low-fidelity model correction [11]. Additionally, process-system twins have integrated optimized control, monitoring, and adaptive identification [12]. The aforementioned methods guarantee that the twin is always consistent with the physical asset and that any modifications are stored as part of the model's state rather than being concealed in an external dashboard.

Definitions of digital twins differ in terms of decision role, model scope, and synchronization frequency [13]. The equipment setup, measured boundary conditions, internal parameters, anticipated outputs, and confidence must all be specified by the calibration service for a cooling tower. Although a record-only predictor is a helpful soft sensor, its inability to allow counterfactual operation outside of the observed control sequence prevents it from being a calibrated twin on its own.

Multiscale Learning Gap

Plant record patterns at various scales. Water treatment and fouling change over weeks; ambient excitation varies seasonally; fan staging produces minute-scale transients; and process loading varies by shift. For this, an RNN typically contains just one hidden state. While hierarchical attention effectively accumulates long-range context, image-inspired encoder-decoder networks have skip links to preserve local information. However, their use in engineering calibration necessitates uncertainty, parameter constraints, physical residual channels, and explicit time masks.

A model that improves the digital twin without replacement, utilizes the lengthy operational history without smoothing short transitions, and is still auditable in the event of sensor malfunctions is the unresolved issue. All three demands are rarely taken into account at the station level in current research. To solve this issue, a Hierarchical U-Net calibration layer is added, and its output still matches the conventional cooling tower performance index.

Record Alignment and Calibration Formulation

Historian Record Construction

Six induced-draft counterflow cells at the two industrial sites provided the records. Water temperatures at the intake and exit, basin temperature, circulating water flow, fan speed, motor power, valve status, ambient dry bulb temperature, relative humidity, and station heat load are all measured in each cell. The Tag sample lasts for two to five minutes. Stable semantics and clear interfaces for the physical-digital state are the main goals of standardized digital twin models [14]. As a result, we preserve the source, engineering unit, timestamp quality, and replacement status while mapping local tags to a canonical schema.

Forward filling is not enough to maintain consistency. Parameter updates must adhere to the degradation mechanism and maintain model identity, according to research on machine-tool twins [15]. In this work, time-weighted aggregation is used to resample continuous tags to one minute. The most recent valid state is used by discrete commands. It is deemed interpolated if there is less than three minutes between neighboring rate limitations; otherwise, it is not. The redundant humidity measurements are used to calculate the wet-bulb temperature, which is discarded if the psychrometric closure is greater than 0.35 °C.

Every training sample consists of a 120-minute full-resolution target interval and a 24-hour context window. The context has a scale of 1, 5, 30, and 120 minutes. Event channels are used in place of free text for maintenance records, fan trips, basin makeup events, and sensor calibration certificates. This design is able to distinguish between an unknown performance change and a known instrumentation intervention.

recording quality prior to the model being run. Sensor range, rate-of-change limits, redundant tag agreement, and the correlation between input, outlet, and basin data are all examined in temperature channels. Measured power and speed are harmonized with fan command. Header pressure and pump work are connected to flow. Instead of closing the interval, a quality failure adds a diagnostic property and changes the availability mask. As a result, documentation of recurring communication or instrument issues can be kept without tainting the calibration supervision.

Operating regimes are assigned by a deterministic state machine before learning. Regimes include shutdown, startup, stable partial load, stable full load, fan staging, flow transition, basin refill, and uncertain operation. Regime boundaries use commands and measured response, so a failed fan start is not labeled as ordinary running. The labels are inputs rather than targets and can be audited by operators. During training, the model may refine response patterns inside each regime but cannot silently redefine an equipment command.

$$x_t = [u_t, y_t, m_t, e_t] \quad \text{Eq.(1)}$$

The vector x_t combines operating inputs u_t, measured performance channels y_t, availability masks m_t, and maintenance or configuration events e_t. The mask is a model input as well as a loss weight. Consequently, a missing outlet-temperature value does not become an artificial zero or an apparently stable repeated value. Event channels remain active for a defined influence period determined by the maintenance class.

$$r_t = y_t - f_{\mathrm{DT}}(u_t, \theta_0) \quad \text{Eq.(2)}$$

The baseline residual r_t is the difference between the measured response and the digital-twin prediction obtained with commissioned parameters θ_0 . Residual channels include outlet temperature, approach, evaporation estimate, fan power, and water-energy imbalance. Using residuals focuses learning on the correction required by the twin instead of forcing the network to relearn psychrometric relationships from scratch.

Figure 1 shows how asynchronous plant evidence is converted into a calibration-ready record: quality flags and availability masks are retained through time alignment, so the baseline twin is corrected only where synchronized measurements and operating-regime information provide defensible support.

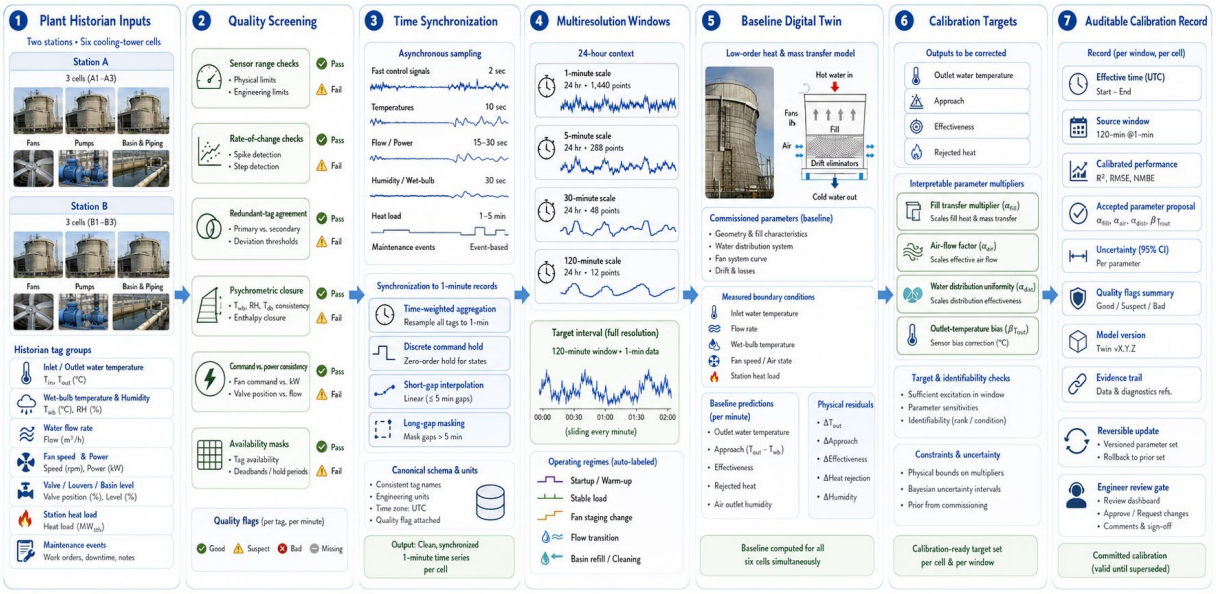

Figure 1. Historian ingestion, quality screening, multiresolution alignment, baseline digital-twin execution, and calibration targets.

Calibration Targets and Identifiability

Outlet water temperature, approach, efficacy, and rejected heat are the calibrated outputs. The effective fill transfer coefficient, air flow factor, water distribution uniformity, and outlet temperature sensor bias are the four parameter multipliers that are estimated. Model details must be modified based on the intended use, according to research on automated twin generation [16]. Only when a parameter exhibits a distinct response signature under the available stimulation are they chosen. Due to its strong correlation with the fill coefficient under typical operating conditions, the water-flow sensor's bias is not assessed simultaneously.

There should be a lot of excitement during the training window. If the range of water flow reaches 12%, the range of fan speed exceeds 15%, or the range of wet-bulb temperature exceeds 2.5°C, a window is chosen for parameter supervision. Windows without excitation do not update parameter loss, but they do train output correction and uncertainty. Identifiability is reported for each calibration suggestion and is based on the condition number of the local sensitivity matrix.

Points and intervals make up Parameter Supervision. An inspection might merely reveal that the fill performance is within a degradation band, but a portable airflow traverse offers a point estimate of the air-flow factor. If the forecast falls inside the permitted range, interval labels have a loss of zero; if not, there is a linear penalty. As a result, fake numerical precision won't represent the uncertainty of the routine examination. Additionally, it can use the same training pipeline for simulated fault campaigns, maintenance evidence, and commissioning tests.

The local sensitivity matrix is computed by perturbing each twin parameter around the proposed state while holding measured boundaries fixed. Columns are scaled by their engineering ranges before the condition number is evaluated. When two columns are nearly collinear, the service reports a combined performance deviation and freezes the less observable parameter. For example, fill and water-distribution factors are not separated during a narrow, constant-flow period. This rule is essential because a neural network can produce stable numbers even when the physical records cannot uniquely support them.

$$\theta_t = \theta_0 \odot (1 + \Delta\theta_t) \quad \text{Eq.(3)}$$

The calibrated parameter vector θ_t is formed by elementwise adjustment of the commissioned vector θ_0. Corrections are bounded by engineering ranges: ±25% for the fill coefficient, ±18% for air flow, ±15% for distribution uniformity, and ±1.0 °C for sensor bias. Smooth parameters cannot jump at fan staging, while sensor bias may change after an instrument event.

Data Splits and Acceptance Labels

Fourteen months of records yielded 512,640 valid cell-minutes after quality screening. Station A contributed four cells and Station B two cells. Training uses the first nine months of Station A; validation uses the next two months; mixed-station testing uses the final three months. Cross-station testing trains only on Station A and evaluates Station B. No overlapping context window crosses a split boundary.

Reference labels combine monthly thermal-test reconciliation, redundant calibrated sensors during three campaigns, maintenance inspections, and injected faults. Parameter labels are interval-valued where direct measurement is impossible. Output metrics are evaluated on all valid periods, whereas parameter recovery is evaluated on documented interventions and controlled simulations. The separation prevents precise-looking parameter scores from being inferred from uncertain routine records.

Hiera-UNet Calibration Model

Hierarchical Temporal Encoder

The encoder divides the one-minute record into nonoverlapping temporal patches and merges neighboring tokens across four stages. Digital-twin reviews emphasize that heterogeneous models and data require shared consistency rules [17]. Accordingly, each token carries station, cell, time-of-day, season, sensor-quality, and operating-regime embeddings. Masked channels use learned missing-value tokens, while numerical values are normalized by robust station statistics.

Hierarchical twin integration requires efficient access to local and global evidence [18]. Early Hiera stages use local attention over 16 min neighborhoods, capturing fan starts, valve motion, and sensor transients. Later stages attend across 2 h, 8 h, and 24 h contexts. Token dimensions are 96, 192, 384, and 512. Cross-scale residual links preserve trend information that might otherwise vanish after patch merging.

Practical digital-twin methodologies stress purpose-driven implementation rather than maximal model complexity [19]. The encoder therefore contains 18.6 million parameters, small enough for station deployment. Attention masks exclude invalid observations but retain timing. Relative-time bias distinguishes a true 30 min separation from thirty duplicated samples after a communication failure.

Patch construction preserves channel groups. Ambient, hydraulic, thermal, electrical, quality, and event channels are projected separately and combined only after the first local block. This reduces the tendency of a high-variance electrical signal to dominate a small but important temperature residual. Group-specific normalization parameters are shared across cells, while design flow, rated fan power, and tower dimensions enter as fixed metadata tokens. The architecture can therefore compare dimensionless operating states without erasing the physical scale required by the digital twin.

The deepest encoder token represents an entire operating day, but it is not used as a single global summary. Four learned query tokens separately collect ambient exposure, hydraulic state, airflow state, and persistent degradation evidence. Their attention distributions are stored for audit. A proposed fill correction is expected to draw on degradation and hydraulic queries, whereas a short outlet-temperature correction should draw on the local thermal path. Inconsistent attribution widens uncertainty and prevents automatic persistence.

$$H_0 = W_p X + E_q + E_t \quad \text{Eq.(4)}$$

The initial token matrix H_0 is produced by patch projection W_p, sensor-quality embedding E_q, and temporal embedding E_t. Station identity is used only for normalization and uncertainty calibration; it is dropped in the final transfer experiment to test whether the model depends on a station label.

$$A = \mathrm{softmax}(\frac{QK^T}{\sqrt{d}} + B_\tau) \quad \text{Eq.(5)}$$

Attention A combines feature similarity with relative-time bias B_τ. The bias favors physically meaningful lags learned from the data, such as the delayed outlet response after a fan-speed change. It does not force a single residence time, which varies with water flow and basin inventory.

$$H_{s+1} = \mathrm{Merge}(H_s + \mathrm{Attn}(H_s)) \quad \text{Eq.(6)}$$

Each stage adds an attention residual before temporal merging. The merge operator concatenates adjacent valid tokens and projects them to the next channel width. Mask-weighted normalization prevents a partly missing patch from acquiring the same confidence as a complete patch.

UNet-Style Calibration Decoder

The decoder up samples the coarsest representation to one-minute resolution. Full-scale skip connections are motivated by encoder-decoder research that preserves fine detail while aggregating context [20]. Each decoder block receives the corresponding Hiera feature, baseline residual, operating input, and mask. The output comprises four performance corrections, four parameter corrections, and four log-variance channels.

Nested skip-connection research demonstrates that reducing semantic mismatch between encoder and decoder can improve reconstruction [21]. Here, skip features pass through regime gates conditioned on fan state, tower loading, and ambient class. A high-frequency fan transition can therefore use early encoder detail, while slow fill degradation relies more strongly on the deepest representation.

The model does not create a spatial temperature image. Its UNet geometry is applied along time and channel groups, producing a full-resolution calibration field. This distinction is important: every decoder location maps to a timestamp and each output channel maps to an engineering quantity. The structure supports trace-back to input tags and avoids implying unmeasured internal spatial resolution.

Decoder up sampling uses linear interpolation followed by a causal-symmetric convolution. The symmetric branch improves offline calibration by using the complete target interval; the causal branch supports online operation and is trained through shared weights. At service time, only observations available at the current timestamp are exposed to the causal output. Offline reports may use the symmetric output and are labeled accordingly, preventing future information from leaking into an online recommendation.

Parameter and residual heads share the first two decoder blocks but separate at 30 min resolution. The parameter branch pools evidence over the window and returns a slowly varying multiplier. The residual branch continues to one-minute resolution and responds to operating transitions. A leakage diagnostic measures how much a persistent parameter estimate changes when the final 30 min of the record are shuffled. Excessive change indicates that the parameter branch is reacting to short noise and disqualifies the checkpoint.

$$D_s = \mathrm{Up}(D_{s+1}) \oplus G_s(H_s) \quad \text{Eq.(7)}$$

Decoder feature D_s concatenates the upsampled deeper feature with gated skip feature $G_s(H_s)$. The gate suppresses poor-quality or regime-irrelevant channels. Concatenation is followed by two depthwise temporal convolutions and a pointwise projection, limiting parameter growth.

$$\hat{y}_t = f_{\mathrm{DT}}(u_t, \theta_t) + \delta_t \quad \text{Eq.(8)}$$

The calibrated prediction adds learned residual δ_t to the digital-twin output evaluated with calibrated parameters θ_t. Joint parameter and residual correction are deliberate: parameters capture persistent, interpretable change, while δ_t absorbs short unmodeled dynamics. A regularizer prevents the residual channel from explaining long-term degradation that should be assigned to parameters.

Figure 2 clarifies how short transients and slow degradation are separated before calibration: hierarchical encoder features preserve multiple temporal scales, regime-gated skip paths restore one-minute detail, and the output heads translate the shared representation into performance corrections, persistent parameter changes, and calibrated uncertainty.

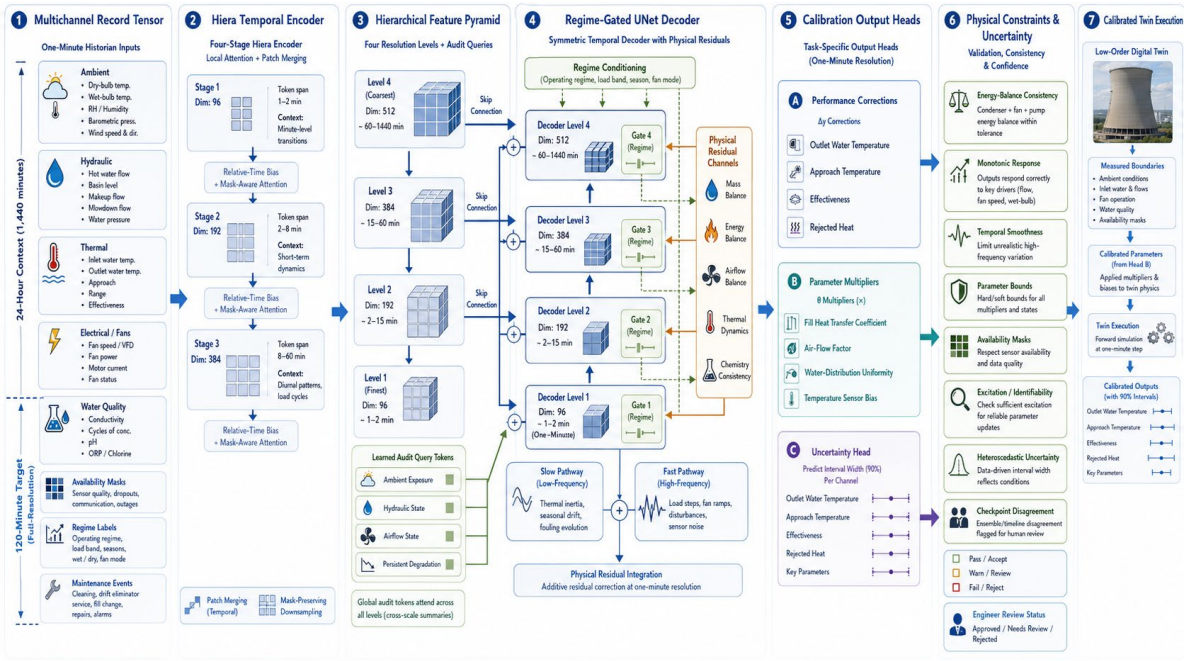

Figure 2. Hiera multiscale encoder, regime-gated UNet decoder, physical residual channels, parameter head, and uncertainty head.

Physical Constraints and Uncertainty

Physical constraints are applied to derived quantities rather than hidden features. The calibrated rejected heat must agree with the circulating-water energy decrease after accounting for measured makeup and blowdown. Outlet temperature must not exceed inlet temperature during valid cooling operation, and effectiveness must remain between zero and one. Parameter change is smooth except at a documented maintenance event.

The uncertainty head predicts heteroscedastic variance for each performance channel. Deep models for automated configuration show the value of adapting architecture and preprocessing to data properties [22], while self-configuring segmentation systems motivate reproducible pipeline choices rather than manual tuning [23]. In this work, uncertainty is calibrated on a held-out month using a single multiplicative temperature per output.

Constraint evaluation follows the same validity masks as the data. Energy balance is not enforced during confirmed basin overflow, unknown makeup flow, or a flow-meter calibration interval. Monotonic airflow response is evaluated only inside locally comparable wet-bulb and water-flow bins. This conditional enforcement avoids teaching the network a simplified physical rule outside its assumptions. Every disabled constraint writes a reason code, so a low total loss cannot be mistaken for complete physical validation.

Predictive uncertainty combines aleatoric variance from the decoder with epistemic variation across four independently initialized checkpoints. The online service uses the ensemble mean and adds the two variance components. Parameter intervals also include local sensitivity uncertainty from the digital twin. The combined interval expands under missing channels, out-of-range ambient conditions, weak excitation, or checkpoint disagreement. Operators therefore see both a calibrated value and a statement about how strongly the record supports it.

$$L_{\text{data}} = \sum_{t} m_t |\hat{y}_t - y_t| \qquad \text{Eq.(9)}$$

The masked data loss sums absolute error only where trusted observations exist. Channel weights scale errors by engineering tolerance: 0.3 °C for outlet temperature, 0.4 °C for approach, 2% for effectiveness, and 2.5% for rejected heat. This prevents the largest numerical channel from dominating optimization.

$$L_{\text{bal}} = |\hat{Q}_t - \dot{m}_w c_p (T_{\text{in}} - T_{\text{out}})| \qquad \text{Eq.(10)}$$

The balance loss compares predicted rejected heat $\hat{Q}_t$ with the water-side enthalpy decreases. It is down-weighted during basin refill, blowdown transitions, and known flow-meter maintenance. The term penalizes mutually inconsistent corrections even when individual sensor errors appear small.

$$L = L_{\text{data}} + \lambda_b L_{\text{bal}} + \lambda_s L_{\text{smooth}} + \lambda_u L_{\text{unc}} \quad \text{Eq.(11)}$$

The final objective combines data fit, energy balance, parameter smoothness, and uncertainty likelihood. Validation selected λ_b=0.30, λ_s=0.08, and λ_u=0.15. The weights are fixed for all stations. Parameter-bound penalties are implemented by bounded output transforms and therefore do not add another displayed equation.

Experimental Results

Optimization, Baselines, and Overall Evaluation

AdamW, batches of twelve 24-hour windows, a maximum learning rate of 3×10^{-2}, cosine decay, and gradient clipping at 1.0 were utilized for training across ninety epochs. When sending multi-resolution data into a dense decoder, full-scale UNet connectivity provides a useful benchmark [24]. 30% of the windows have fan staging, 20% have load ramps, 20% have high wet-bulb operations, 15% follow maintenance, and 15% are stable baselines. The station records are not evenly sampled per regime.

Reusing local leftover blocks enables efficient deeper multiscale processing, as demonstrated by nested U-structures [25]. Short Dropout Bursts, Timestamp Jitter, Calibration-Offset Injection, and Random Channel Masking are used. Since this would result in training samples that are physically impossible, augmentation does not alter command states independently of measured power.

Research on compound scaling indicates that a balance between depth, width, and resolution is necessary [26]. A large model with 34.2 million parameters only decreased the validation error by 0.01°C, whereas a small model with 9.7 million parameters underfit the seasonal context. The optimal cross-station accuracy per unit inference time was provided by the chosen 18.6 million-parameter setup.

On the mixed-station test set, the uncalibrated twin yielded an outlet-temperature MAE of 1.18 °C and an RMSE of 1.54 °C. Bidirectional LSTM was 0.43 °C, temporal UNet was 0.37 °C, flat transformer was 0.35 °C, random forest was 0.49 °C, and monthly least squares reduced MAE was 0.72 °C. The MAE and RMSE for Hiera-UNet were 0.24°C and 0.34°C, respectively. The effectiveness MAE fell by 3.5 percentage points (from 4.8 to 1.3), and the mean approach error decreased from 1.09 °C to 0.27 °C.

It was not possible to evaluate across stations. A 0.29°C outlet MAE, a 0.31°C approach MAE, and a 1.6-point effectiveness MAE were attained during training at Station A and testing at Station B. 0.46 °C, 0.51 °C, and 2.4 points were generated via the Temporal U-Net. The low-order twin assumption was broken by basin mixing, and the biggest Hiera-UNet mistake happened within the first 20 minutes following the two cells' simultaneous start-up.

In the steady state, error distributions were almost symmetrical, but during starting, they were positively skewed. The 90th absolute percentile was 0.48 °C, the 99th percentile was 1.07 °C, and the median outlet error was 0.03 °C. The reported performance is insensitive to removing the challenging period because eliminating the first 15 minutes after startup only decreased the overall MAE by 0.02°C. The MAE at the cell level varied from 0.21 to 0.32 °C, and the cell with the highest value had the least accurate flow measurement.

The calibrated twin had a 95th-percentile value of 4.9% and an average rejection rate of 2.2% when compared to the monthly thermal-test period. Over the same time periods, the commissioned twin displayed a mean inaccuracy of 7.6%. A fixed bias correction did not improve performance; in moderate ambient circumstances, the median correction was almost zero, and under humid, high-load operation, it rose. The tower mechanism depicted in the hierarchical record is consistent with this regime dependence.

Figure 3 distinguishes improvements in three complementary performance quantities. Figure 3(a) shows that the largest gain occurs in outlet-temperature calibration relative to both the commissioned twin and data-driven baselines, Figure 3(b) demonstrates that the reduction also extends to approach temperature rather than reflecting a single-output bias correction, and Figure 3(c) indicates that effectiveness errors contract across both typical and tail behavior.

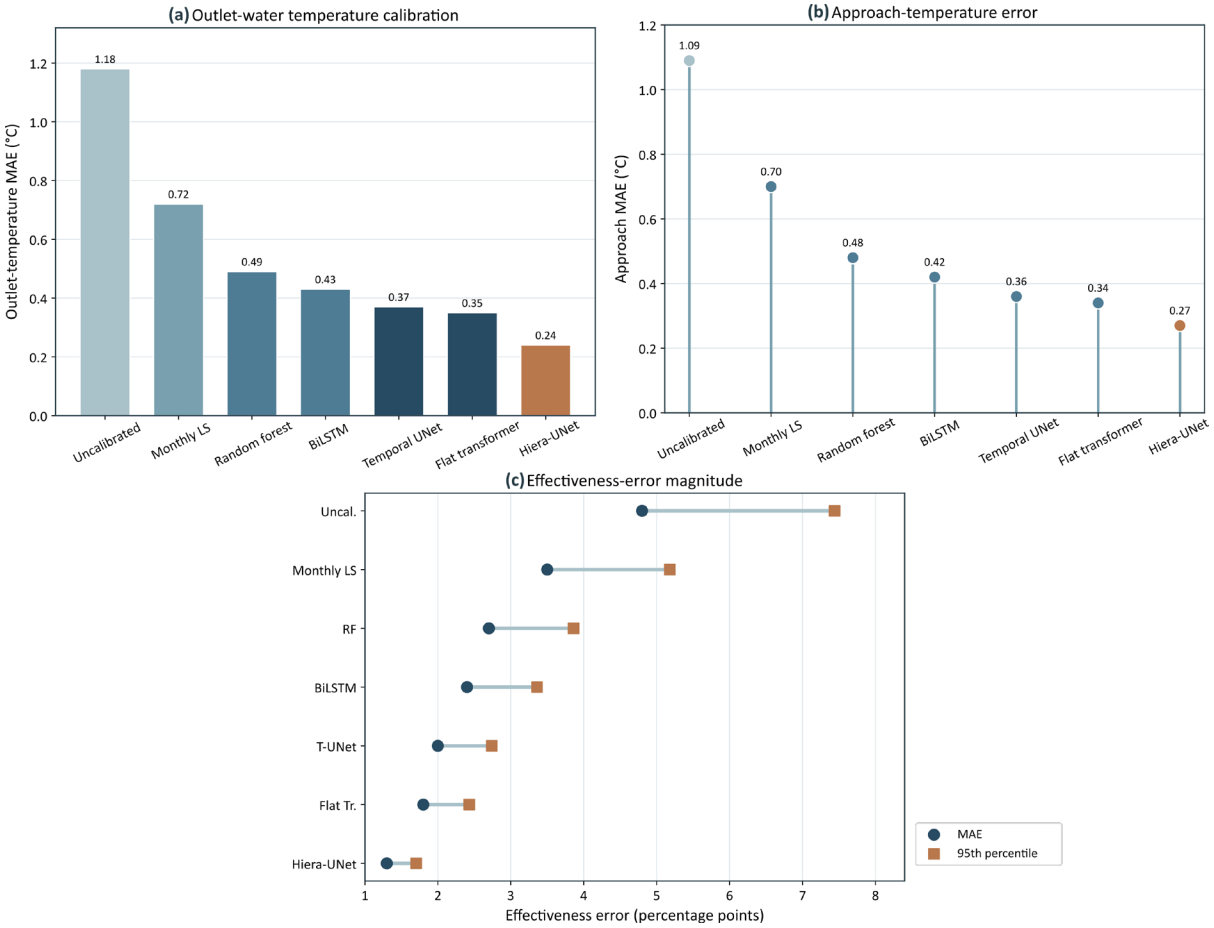

Figure 3. Overall calibration performance: (a) outlet-temperature MAE; (b) approach MAE; (c) effectiveness error across baseline models.

The MAE of outlet-temperature inaccuracy was 0.21°C during cool-dry periods, 0.23°C during mild periods, and 0.31°C during hot-humid operation, according to seasonal analysis. Ninety-eight percent of the predictions were within 0.5°C when the wet-bulb temperature exceeded 26°C. Because there was minimal additional airflow excitation in the record, error rose when all fans ran at saturation.

All six cells had the same relative ranking of the baseline models; the only cell with a greater airflow imbalance had an increase in absolute inaccuracy. Because its hierarchical branches distinguished between short fan-staging disruptions and slower ambient and hydraulic variations, Hiera-UNet maintained the lowest inaccuracy. A single time scale proved unable to account for both basin mixing and a progressive approach-temperature drift, and the benefit was most noticeable above the 80% rated load.

Calibration quality was more susceptible to stimulation than to calendar season alone, according to the stratification of errors by operating mode. Fill and airflow corrections could be distinguished by windows with moderate fan speed or water-flow change. A modest residual was deemed insufficient evidence to establish that a parameter had stabilized because nearly steady windows had broader uncertainty intervals even when their point forecasts were accurate.

This state of affairs was maintained by cross-station transfer. A sensor-quality mask stopped missing tags from being perceived as low process values, while normalization by Design water flow and rated fan power decreased systematic station bias. The seasonal and station-specific breakdown displayed below is the result of the residual transfer mistake that happened during the simultaneous cell start-up and fan-saturation period.

Figure 4 explains where transfer performance remains sensitive to operating context. Figure 4(a) separates station and cell effects after design-based normalization, Figure 4(b) shows how error rises as hot-humid wet-bulb conditions reduce the available thermal margin, and Figure 4(c) links fan saturation to weak airflow excitation and wider uncertainty even when point predictions remain usable.

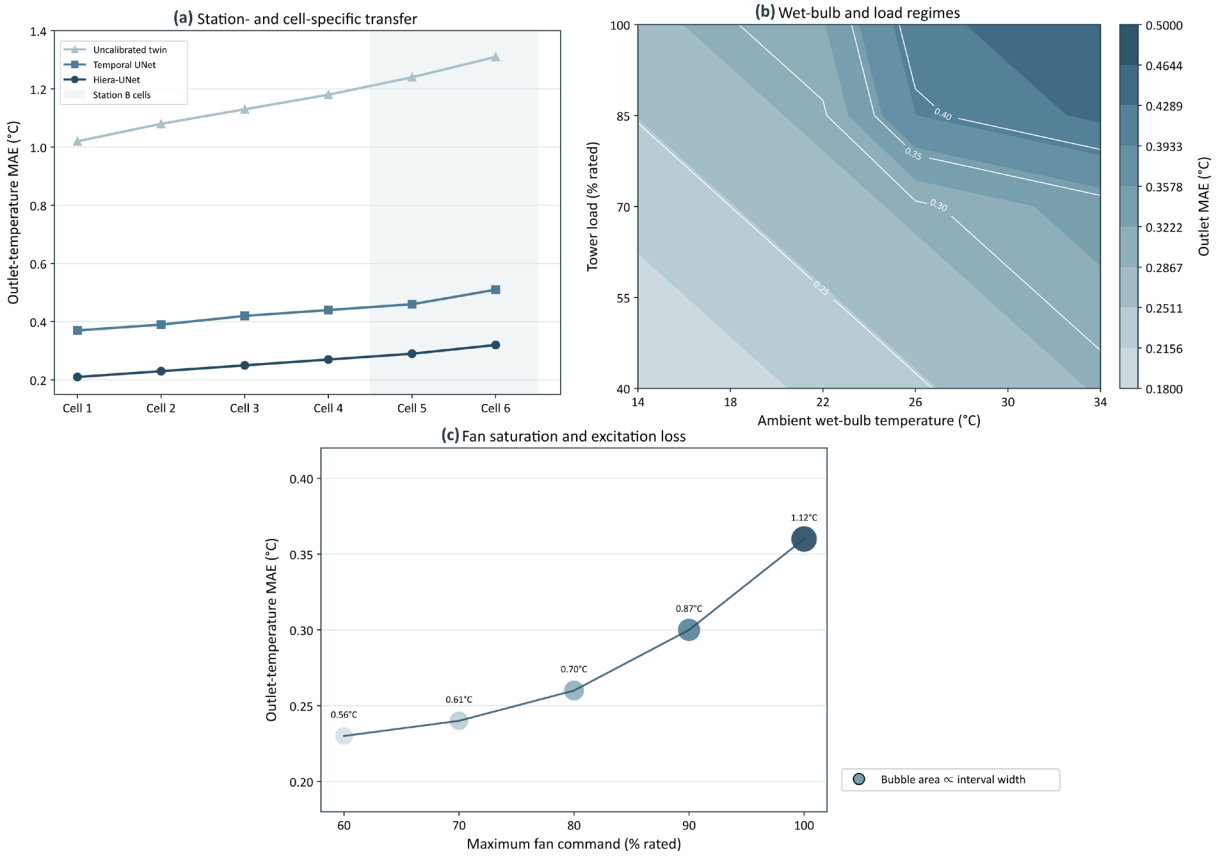

Figure 4. Cross-station and seasonal generalization: (a) station-specific error; (b) wet-bulb bins; (c) fan-saturation regime.

Reproducibility, Ablation, and Robustness

Long-range token interaction is valuable, according to Transformer study [27]. The commissioned digital twin, monthly least-squares calibration, temporal convolution, bidirectional LSTM, random forest residual correction, standard temporal UNet, and a flat transformer with the same parameter budget serve as baselines. The same historical channels, masks, training windows, and station splits are used by all data-driven models.

Observations and governing residuals can be integrated by physics-informed learning [28]. The regular UNet and Hiera-UNet are trained with and without energy-balance and smoothness parameters in order to separate that contribution. At Station A, only hyperparameters are chosen. There will be no post-test recalibration, and Station B remains unaltered until the cross-station test.

Realistic noise and diverse sensors are taken into account in industrial fault diagnostic research [29]. For every model, three seeds are run. Store model hashes, software versions, tag dictionaries, split manifests, normalization statistics, and event mappings. The experiment packages provide generated tensors and timestamp hashes, while the station environment contains raw historian values.

The risk is seen in features that encode source identity rather than equipment state, as demonstrated by multi-source fault learning [30]. As a result, the transfer experiment maps all units to SI, normalized by design flow and rated fan power, and excludes the station embedding. If the local sensitivity condition number is higher than 80 or more than 25% of the necessary channels are hidden, the calibration request is refused.

By adding offsets of -0.8 °C and +0.8 °C to the output sensor, you can create a regulated bias. The model had the right sign in 98.6% of windows and a bias of 0.12 °C MAE. With an absolute inaccuracy of 4.1%, the simulated fill-coefficient decline from 5% to 25% was recovered. Due to partial confounding with fill performance, the distribution-uniformity error was 5.4% and the air-flow multiplier error was 3.6%.

The cross-station outlet MAE increased from 0.29°C to 0.38°C in the absence of hierarchical merging. It increased to 0.41 °C after the physical remnant channels were removed. The 95th-percentile heat-rejection inconsistency

increased from 4.7% to 9.6% when the energy balance was eliminated, although the average MAE was only 0.32 °C. The inaccuracy rose to 0.34 °C and the maintenance transition became noisier when direct concatenation was used in place of the gated skips.

Every random seed had the same parameter recovery. The fill-multiplier error's standard deviation was 0.8%, while the flat transformers was 2.7%. The outlet accuracy decreased by less than 0.02°C after eliminating parameter supervision and solely optimizing the output error, however the fill estimations differed by more than 18% between the seeds. As a result, in this experiment, the exact reconstruction of the output does not produce an actual calibration parameter.

In a counterfactual test, the twin's fan speed was altered while the recorded hydraulic and ambient inputs remained unchanged. In 99.1% of cases, Hiera-UNet and 94.7% of cases, the unconstrained temporal UNet maintained the expected direction of the outlet-temperature response. Although the Hiera-UNet parameter correction stayed under the limit, both models displayed an increase in uncertainty at the maximum extension. The calibrated twin inside the observed operational envelope can be subjected to limited what-if analysis using the outcome.

Figure 5 tests whether the learned corrections recover identifiable physical changes rather than merely reducing output error. Figure 5(a) compares imposed and recovered outlet-sensor offsets, Figure 5(b) traces the response to progressively stronger fill degradation, and Figure 5(c) contrasts the recovery uncertainty of airflow and distribution multipliers, exposing the greater confounding of water distribution with fill performance.

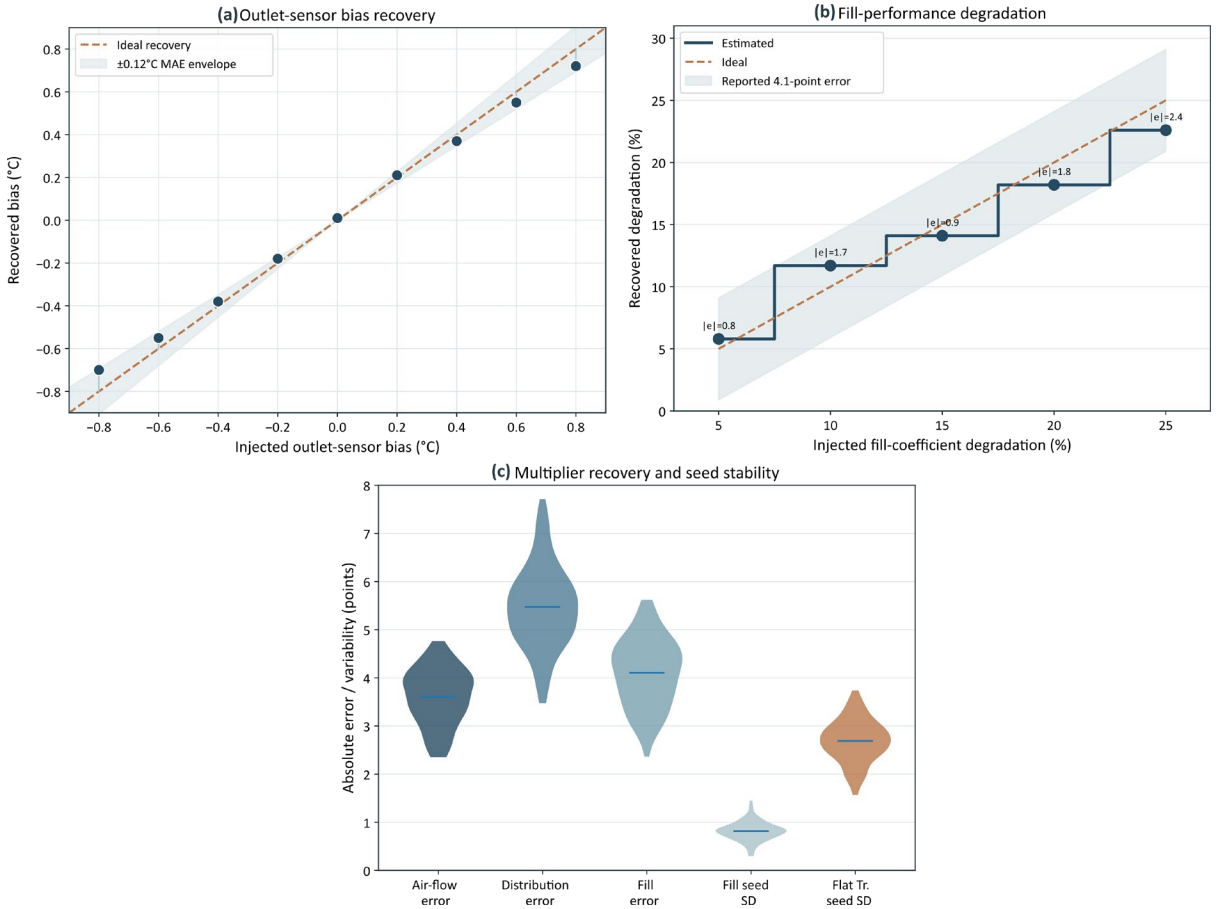

Figure 5. Calibration-parameter recovery: (a) outlet-sensor bias; (b) fill-coefficient degradation; (c) airflow and distribution multipliers.

The ablation experiment indicates that the output accuracy and parameter recovery are different. The inferred fill multiplier was unstable across random seeds, while removing parameter supervision only marginally altered the outlet-temperature MAE. Consequently, the model may match the measured outlet temperature and explain the discrepancy to a physical mechanism flaw by incorporating compensatory corrections.

By penalizing corrections that broke the water-side heat rejection condition or led to irrational short-term parameter changes, physical limitations decreased the ambiguity. The latter primarily decreased the 95th-percentile inconsistency and inhibited sudden parameter jumps close to maintenance transitions by increasing the tail of the error distribution.

The similar engineering trade-off was found in context-length tests. A 48-hour context increased computing cost without appreciably increasing accuracy, while short windows have reacted rapidly but show no signs of progressive deterioration. A one-minute decoder was employed to record the transient reaction following fan staging, and a 24-hour context-retained shift-scale and diurnal data were chosen.

Figure 6 separates architectural accuracy gains from the safeguards required for credible calibration. Figure 6(a) shows the incremental penalty from removing hierarchy, physical residuals, and gated skip paths, Figure 6(b) demonstrates that balance and smoothness constraints chiefly suppress tail inconsistency and unstable parameter motion, and Figure 6(c) identifies the 24-hour context with one-minute reconstruction as the best trade-off between degradation evidence, transient fidelity, and runtime.

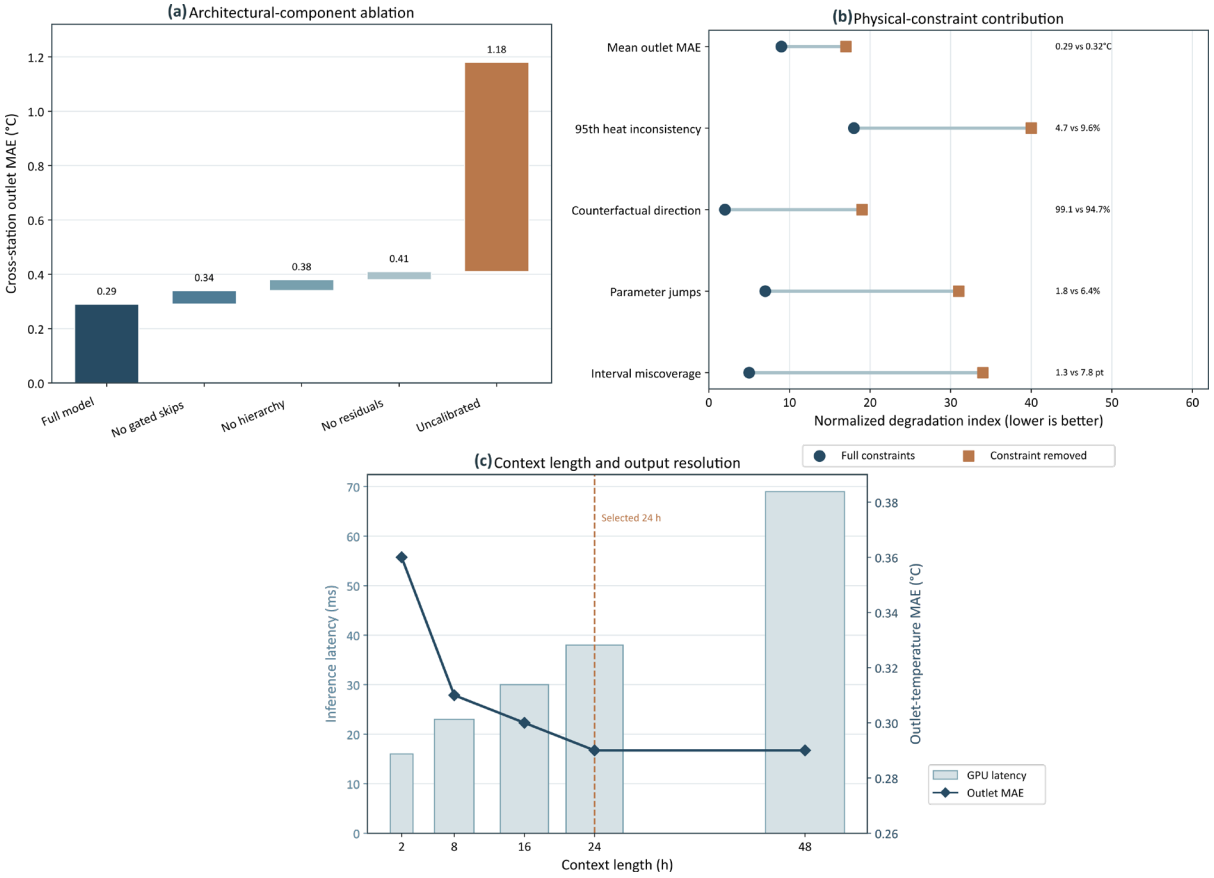

Figure 6. Ablation and context sensitivity: (a) architectural components; (b) physical constraints; (c) context length and output resolution.

Online Service, Drift, and Error Analysis

Deep flaw detection in industrial forging demonstrates that an operational activity must be linked to a relevant model [31]. The suggested calibration record, which includes the effective time, source window, parameter changes, performance adjustment, interval width, quality flags, and model version, is written by the online service every five minutes. The set of commissioned parameters is not replaced by it.

Multiscale anomaly detection is not a competing predictor, but rather a protective measure [32]. If a record window's feature distance is more than the training envelope, if a sudden residual cannot be explained by a logged event, or if the outlet temperature uncertainty is greater than 0.8 °C, the record window is not stored. Withheld windows do not cause automatic parameter persistence; instead, they go into an engineer review queue.

Explicit asset, model, data, and service linkages are the main focus of industrial digital twin systematization [33]. Thus, governance makes a distinction between equipment depreciation and sensor rectification. Field tests or a redundant sensor can be used to determine a temperature offset for a projected temperature. Persistence over three or more stimulated windows and one inspection notice are necessary for a decline in fill coefficient. Reversible updates maintain the previous calibration condition.

Shadow, advisory, and controlled-update modes are used for deployment. Shadow mode creates a baseline for data quality and computes the outcome without subjecting it to the operation. The performance and parameter proposal is displayed in advisory mode, but manual confirmation is necessary. Only when the quality, sensitivity, uncertainty, and evidence conditions are all met will the parameter be saved in controlled-update mode. The architecture does not require automatic write-back, therefore a plant can stay in advisory mode indefinitely. Areas with distinct cybersecurity and change management requirements may need to be covered by both.

The metrics are tracked by Cell and Regime. Residual mean, interval coverage, withheld-window rate, missing-channel pattern, feature distance, and operator correction rate are examples of service metrics. To prevent false alerts, a drift alarm is set off after two or more indicators have been above the threshold for a long time. Retraining maintains the frozen commissioning and cross-station benchmark while utilizing the recently added records. In addition to improving accuracy, the new model will preserve fairness and uncertainty coverage.

Following scalar calibration, 89.7% of the approach measurements and 91.3% of the valid outlet-temperature observations fell within the nominal 90%-interval. In the familiar condition, the mean interval width was 0.74 °C, but in the withheld high-distance window, it was 1.21 °C. 0.026 is the anticipated calibration error. The coverage was 78.4% and the raw variance estimate was overstated because there was no calibration.

Over the course of 30 days, a synthetic 12% fill deterioration was added; by day 30, it had achieved an estimated value of 11.3% after 4.6 days at the 5% threshold. Within eighteen minutes, a sharp 0.5 °C change in the sensor was detected; however, because the event signature did not support energy-balance, it was not regarded as equipment degradation. The achieved fill multiplier increased from 0.86 to 0.95 during three extended windows in an actual nozzle-cleaning procedure.

End-to-end inference ran at 214 ms/cell-window on an eight-core industrial CPU and 38 ms/cell-window on an RTX 3060. There was a significant buffer in the five-minute service cycle because historian extraction and quality screening took 0.9 seconds for the six cells. The CPU's peak memory was 2.7GB, while the GPUs was 1.1GB. For each cell-year, the calibration record added roughly 46 MB.

6.8% of all test windows were excluded by the review gate. There were 2.1% of missing essential temperatures, 1.9% of poor sensitivity, 1.4% of out-of-envelope operation, 0.9% of severe uncertainty, and 0.5% of contradictory event data. After retrospective analysis of the prediction, the withheld windows had an error of 0.71 °C, while the accepted-window outlet's MAE was 0.25 °C at the cross-station test. As a result, the gate concentrated the danger without appreciably reducing the quantity of regular operation.

Record the maintenance outcomes and simulate operator reviews. Three of the 42 persistent parameter recommendations were false positives because of unrecorded bypass operation, four were inconclusive, and 35 were validated by further inspection or calibrated-sensor evidence. The two situations were eliminated by adding a bypass-state consistency rule. Because the service offered contributing intervals, tag quality, balance residual, sensitivity, and the previous calibration condition, the median evidence-review time was 3.8 minutes.

Figure 7 connects statistical calibration with deployment behavior. Figure 7(a) shows that scalar calibration restores interval coverage close to its nominal level, Figure 7(b) distinguishes gradual fill deterioration from an abrupt sensor shift through their different temporal and balance signatures, and Figure 7(c) shows that historian preparation dominates the service cycle while neural inference occupies only a small fraction of the five-minute update budget. The results indicate that the model is accurate enough for monitoring and optimization support, but not a substitute for contractual thermal testing. Persistent parameter changes were more reliable than single-window corrections. The most consequential failure mode was poor excitation: when fan speed, flow, and ambient condition remained nearly constant, several parameter combinations explained the same residual. The service correctly widened uncertainty, but engineers still need to schedule an informative operating test before accepting a parameter update.

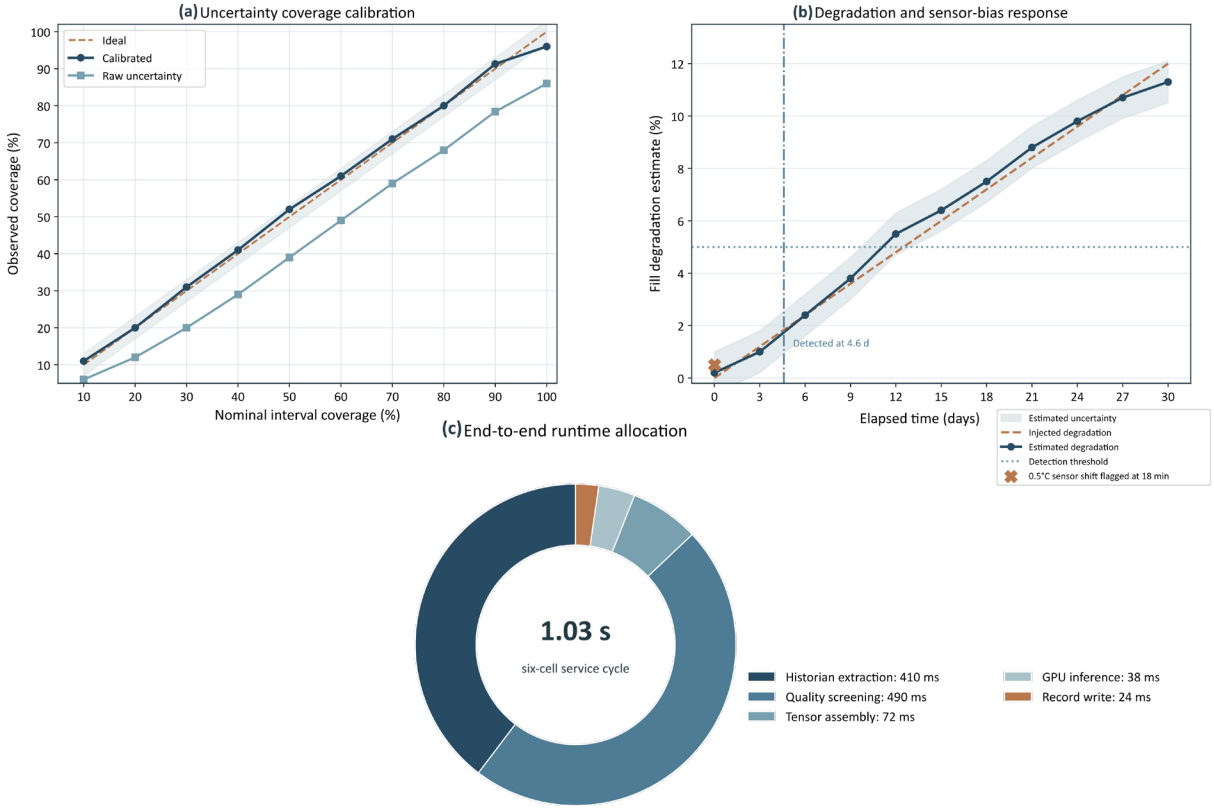

Figure 7. Deployment and error attribution: (a) uncertainty coverage; (b) degradation and sensor-bias response; (c) end-to-end runtime allocation.

Conclusion

Hiera-UNet for record-driven calibration of cooling-tower performance digital twins is presented in this study. The technique encodes multiscale residual background, executes a low-order twin, aligns asynchronous plant records, and reconstructs calibrated outputs and identifiable parameters at a resolution of one minute. A twin's engineering function must be smooth, bounded, unpredictable, and physically balanced.

The uncalibrated outlet-temperature MAE for the mixed test was reduced from 1.18 °C to 0.24 °C in six cells and two stations, and 0.29 °C was attained in cross-station transfer. Calibration intervals attained near-nominal coverage, and parameter recovery tests separated sensor bias from gradual fill/airflow variations. Physical residuals and hierarchical encoding contributed more than just a bigger model.

Low-excitation times, basin-mixing transients, and parameter confusion are still limitations. Future research will include calibrated airflow measurements, additional water-quality states, and controlled excitation scheduling. Network outputs should continue to be treated by deployment as reversible suggestions that support maintenance evidence, record quality, sensitivity, and uncertainty.

Author Contributions

Domantas Vasiliauskas contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Benas Balčiūnas contributes to validation, analysis, investigation, data collection, draft preparation. Marius Pakalniškis contributes to investigation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Song, J., Ruan, S., Chen, Y., Wu, X., & Li, X. (2019). Research on Thermal Performance of Natural Draft Counterflow Wet Cooling Tower Based on Artificial Neural Network. Fluid Machinery. https://doi.org/10.3969/j.issn.1005-0329.2019.03.016

[2] Guo, Q., Qi, X., Wei, Z., Yin, Q., Sun, P., & Guo, P. (2019). Modeling and characteristic analysis of fouling in a wet cooling tower based on wavelet neural networks. Applied Thermal Engineering. https://doi.org/10.1016/j.applthermaleng.2019.02.041

[3] Ma, K., Liu, M., & Zhang, J. (2021). Online optimization method of cooling water system based on the heat transfer model for cooling tower. Energy. https://doi.org/10.1016/j.energy.2021.120896

[4] Yu, J. H., Qu, Z. G., Zhang, J. F., Hu, S. J., & Guan, J. (2022). Comprehensive coupling model of counter-flow wet cooling tower and its thermal performance analysis. Energy. https://doi.org/10.1016/j.energy.2021.121726

[5] Wang, W., et al. (2021). Crosswind influence on heat and mass transfer performance for wet cooling tower equipped with an axial fan. Case Studies in Thermal Engineering. https://doi.org/10.1016/j.csite.2021.101259

[6] Hosseini, S. S., et al. (2022). Development of a hybrid model for reliably predicting the thermal performance of direct contact countercurrent cooling towers. International Journal of Heat and Mass Transfer. https://doi.org/10.1016/j.ijheatmasstransfer.2022.123336

[7] Chen, X., et al. (2021). Analysis of a novel combined heat exchange strategy applied for cooling towers. International Journal of Heat and Mass Transfer. https://doi.org/10.1016/j.ijheatmasstransfer.2021.120910

[8] Toman, S., Kiilerich, B., Marshall, I. P. G., & Koren, K. (2020). Bacterial Respiration Used as a Proxy to Evaluate the Bacterial Load in Cooling Towers. Sensors. https://doi.org/10.3390/s20216398

[9] Bender, B., et al. (2022). Enhancing cooling tower performance with condition monitoring and machine learning based drift detection. Procedia CIRP. https://doi.org/10.1016/j.procir.2022.09.063

[10] Chong, A., Gu, Y., & Jia, H. (2021). Calibrating building energy simulation models: A review of the basics to guide future work. Energy and Buildings. https://doi.org/10.1016/j.enbuild.2021.111533

[11] Lin, Y., et al. (2022). Online autonomous calibration of digital twins using machine learning with application to nuclear power plants. Applied Energy. https://doi.org/10.1016/j.apenergy.2022.119995

[12] Min, Q., Lu, Y., Liu, Z., Su, C., & Wang, B. (2019). Data-driven digital twin technology for optimized control in process systems. ISA Transactions. https://doi.org/10.1016/j.isatra.2019.05.011

[13] VanDerHorn, E., & Mahadevan, S. (2021). Digital Twin: Generalization, characterization and implementation. Decision Support Systems. https://doi.org/10.1016/j.dss.2021.113524

[14] Liu, M., et al. (2020). The development of standardized models of digital twin. IFAC-PapersOnLine. https://doi.org/10.1016/j.ifacol.2021.04.164

[15] Wei, Y., Hu, T., Zhou, T., Ye, Y., & Luo, W. (2021). Consistency retention method for CNC machine tool digital twin model. Journal of Manufacturing Systems. https://doi.org/10.1016/j.jmsy.2020.06.002

[16] Lugaresi, G., & Matta, A. (2021). Automated manufacturing system discovery and digital twin generation. Journal of Manufacturing Systems. https://doi.org/10.1016/j.jmsy.2021.01.005

[17] Liu, M., Fang, S., Dong, H., & Xu, C. (2021). Review of digital twin about concepts, technologies, and industrial applications. Journal of Manufacturing Systems. https://doi.org/10.1016/j.jmsy.2020.06.017

[18] Cimino, C., Negri, E., & Fumagalli, L. (2021). Harmonising and integrating the Digital Twins multiverse: A paradigm and a toolset proposal. Computers in Industry. https://doi.org/10.1016/j.compind.2021.103501

[19] Qamsane, Y., Chen, C. Y., Balta, E. C., Kao, B. C., Mohan, S., Moyne, J., Tilbury, D. M., & Barton, K. (2021). A Methodology to Develop and Implement Practical Digital Twin Solutions for Manufacturing Systems. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3065971

[20] Tao, F., Zhang, H., Liu, A., & Nee, A. Y. C. (2020). Digital Twin-driven smart manufacturing: Connotation, reference model, applications and research issues. Robotics and Computer-Integrated Manufacturing. https://doi.org/10.1016/j.rcim.2019.101837

[21] Semeraro, C., et al. (2020). Digital Twin Models in Industrial Operations: A Systematic Literature Review. Procedia Manufacturing. https://doi.org/10.1016/j.promfg.2020.02.084

[22] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2020). UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. IEEE Transactions on Medical Imaging. https://doi.org/10.1109/TMI.2019.2959609

[23] Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods. https://doi.org/10.1038/s41592-020-01008-z

[24] Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y. W., & Wu, J. (2020). UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. ICASSP 2020. https://doi.org/10.1109/ICASSP40776.2020.9053405

[25] Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: Going Deeper with Nested U-Structure for Salient Object Detection. Pattern Recognition. https://doi.org/10.1016/j.patcog.2020.107404

[26] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv. https://doi.org/10.48550/arXiv.1905.11946

[27] Dosovitskiy, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv. https://doi.org/10.48550/arXiv.2010.11929

[28] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics. https://doi.org/10.1016/j.jcp.2018.10.045

[29] Zhao, R., et al. (2021). Deep learning for diagnosis and classification of faults in industrial rotating machinery. Computers & Industrial Engineering. https://doi.org/10.1016/j.cie.2020.107060

[30] Zhou, F., Yang, S., He, Y., Chen, D., & Wen, C. (2021). Fault diagnosis based on deep learning by extracting inherent common feature of multi-source heterogeneous data. Proceedings of the Institution of Mechanical Engineers, Part I. https://doi.org/10.1177/0959651820933380

[31] Lee, J., et al. (2021). Application of Deep Learning for Fault Detection in Industrial Cold Forging. International Journal of Production Research. https://doi.org/10.1080/00207543.2021.1891318

[32] Ren, H., et al. (2022). Multi-scale Anomaly Detection for Big Time Series of Industrial Sensors. arXiv. https://doi.org/10.48550/arXiv.2204.08159

[33] Trauer, J., Schweigert-Recksiek, S., Engel, C., Spreitzer, K., & Zimmermann, M. (2020). The Digital Twin Concept in Industry - A Review and Systematization. IEEE ETFA. https://doi.org/10.1109/ETFA46521.2020.9212089