Journal of Intelligent Information and Communication

Research Paper

Data Center Energy System Cooling Load Probabilistic Forecasting with a DeepAR-Quantile Model

Disha Malhotra[1, *], Preeti Soni[1], Shobha Razdan[1] and Kavya Reddy[1]

[1] Faculty of Information Technology, Indian Institute of Information Technology Allahabad, Prayagraj, 211012, India
*Corresponding author: disha.ma@iiita.ac.in

Abstract. For data center energy systems to operate in an energy-efficient and risk-aware manner, accurate probabilistic cooling load forecasting is necessary. The earlier point forecasting techniques are limited to calculating the average cooling demand and do not account for a number of sources of uncertainty, including variations in chiller performance, workload fluctuations, outside temperature changes, and delayed thermal response. DeepAR-Quantile model for multi-horizon probabilistic cooling load prediction. An autoregressive recurrent structure represents historical cooling demand, IT load, power system data, outdoor weather parameters, chilled-water/supply-air temperatures, and operational circumstances. The model enables prediction intervals and high-load risk estimates by producing conditional quantiles rather than a single deterministic result. The suggested approach will be compared with ARIMA, LightGBM, LSTM, GRU, Transformer, and deterministic DeepAR using experimental analysis of 18 months' worth of 5-minute operational data. In comparison to the strongest baseline, the suggested model achieved an anticipated P50 RMSE of 3.42 kW at the 1-hour horizon, enhanced the P90 interval coverage to 91.6%, and decreased the average pinball loss by 18.7%. The findings above demonstrate how cooling dispatch, reserve planning, and thermal risk control in data center energy systems may be analyzed using quantile-aware autoregressive models.

Keywords: *Data Center Cooling Load, DeepAR-Quantile, Probabilistic Forecasting, Prediction Interval, Thermal Risk*

Received on 10 October 2022, Accepted on 25 January 2023, Published on 06 February 2023

Introduction

Large-scale digital storage, cloud computing, artificial intelligence services, financial platforms, and industrial Internet applications are all provided by power-hungry data centers. Heat generated by high-peak demand must be dissipated by the cooling systems for the servers, power equipment, storage arrays, and network devices. The weather no longer determines cooling load in modern facilities; instead, high-power-density IT equipment, workload migration, airflow dispersion, chilled-water systems, humidity, and delayed thermal buildup in server rooms are the main causes [1]. Thus, chiller scheduling, cooling tower operation, pump control, thermal reserve allocation, and emergency protection against hot-spot development are all directly impacted by an accurate cooling demand forecast [2]. Unreliable forecasting results in either insufficient cooling capacity and a service failure or overcooling of the room and energy waste [3].

In order to improve the accuracy of point forecasting under nonlinear operating circumstances, existing load prediction research have progressively included machine learning, deep recurrent networks, and attention-based models [4]. Multiple anticipated values are also needed for cooling dispatch in data centers. Cooling systems may enter a partial-load condition with non-linear efficiency and reaction delays, external temperatures will fluctuate dramatically in a short period of time, and IT workloads may alter suddenly following virtual machine migration [5]. In these situations, a deterministic model will fail to capture the top tail of cooling demand while displaying a tiny average error. Thus, operational risk assessment has made use of quantile regression, probabilistic forecasting, and prediction interval estimation [6]. The conditional distribution of past sequences can be more easily learned by DeepAR and other autoregressive probabilistic models [7]. Additionally, without assuming a fixed Gaussian distribution of residuals, quantile-based loss functions can estimate different confidence levels for the forecaster [8].

A DeepAR-Quantile model for probabilistic cooling load forecasting in data center energy systems is examined in this research. In order to generate multi-horizon quantile projections, the model must first encode the history of cooling demand, IT power, outdoor meteorological parameters, chilled-water supply temperature, return-air temperature, humidity, and equipment status indicators. Reliably estimate the high-quantile risk and accurately predict the median at the same time. Slow thermal response, abrupt workload shift, probabilistic interval calibration, and peak-load alarm in uncertain operating settings are the main issues that need to be resolved.

Related Work

Cooling Load Forecasting in Data Center Energy Systems

Although it is a subset of building energy prediction, cooling load forecasting for data centers is more focused and time-sensitive. Data centers have a better correlation between IT workload and mechanical cooling demand [9], while general building studies concentrate on occupancy, weather, envelope reaction, and HVAC schedules [10]. According to data center energy studies, cooling contributes significantly to the facility's overall electricity usage, particularly when server density rises and airflow control is inadequate [11]. While shallow neural networks and conventional regression models are capable of learning daily patterns, they are likely to perform poorly in situations involving sudden equipment-state change and workload relocation [12]. By using high-dimensional sensor characteristics and recurrent memory, deep learning models enhance nonlinear representation [13]. Reserve-aware cooling dispatch is not appropriate for many of the existing cooling-load models as they only provide one predicted value [14].

Probabilistic Time-Series Forecasting Methods

Probabilistic forecasting techniques simultaneously display the range of uncertainty and the average future demand. Prediction intervals, bootstrap estimation, Bayesian regression, and quantile regression are the conventional techniques [15]. In energy forecasting, the coverage probability, interval width, pinball loss, etc. are frequently used to describe the reliability of an interval [16]. DeepAR is a popular tool for multi-series forecasting since it learns an autoregressive conditional distribution [17]. Although transformer-based forecasting models are better at modeling long-range dependencies, they often require more data and careful tweaking to estimate operational uncertainty [18]. Because they directly estimate P10, P50, P90, and other quantiles with a differentiable asymmetric loss function, quantile recurrent models are practical [19]. Additionally, data center cooling systems are quite straightforward; operators pay more attention to the upper-tail cooling need than to the overall demand [20].

Uncertainty-Aware Cooling Dispatch and Risk Indicators

Cooling dispatch must be both thermally safe and energy-efficient. A bigger cooling reserve is not designed to lower hot-spot hazards because a small forecast interval may be employed. The facility will be less energy-efficient and have a comparatively high chiller capacity if it is too big [21]. The quality of uncertainty is no longer treated as an ancillary outcome in energy forecasting research, which has started to take it into account alongside point accuracy [22]. Thermal inertia in racks and hallways, temperature swings, workload fluctuations, chiller cycles, and missing sensor data all contribute to increased uncertainty in data centers [23]. As a result, risk-averse forecasting should explicitly reflect the highest quantile of cooling demand and include a variety of factors [24]. In a similar manner, DeepAR-Quantile calibrates interval behavior under high-load settings and learns conditional quantiles from sequential operation data [25].

DeepAR-Quantile Probabilistic Forecasting Method

Multi-Source Cooling Load Feature Representation

Cooling load forecasting is viewed as a conditional probabilistic sequence issue in the suggested paradigm. As illustrated in Figure 1, cooling load, IT power, outside temperature, outdoor humidity, chilled-water supply temperature, chilled-water return temperature, air-side supply temperature, return-air temperature, cooling equipment status, and calendar indications are all monitored at each time step. The model learns delayed thermal effects based on operational experience by maintaining the order of these variables in time rather than compressing them into a fixed-size feature vector.

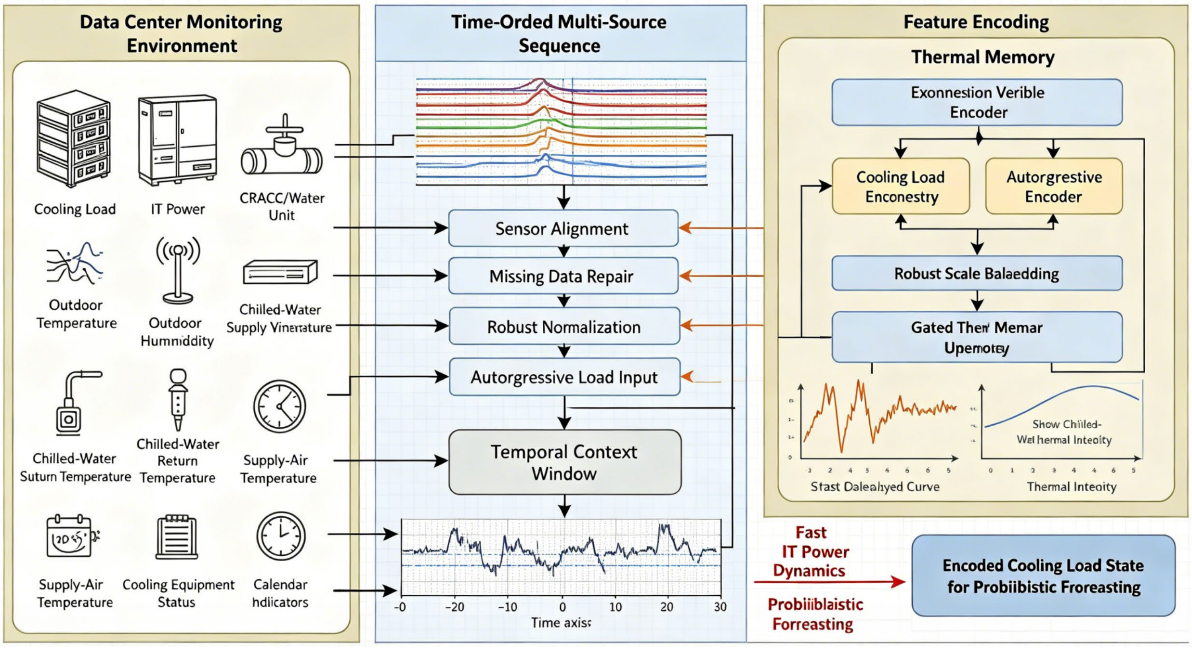

Figure 1. Multi-source cooling load feature representation framework

Let the multi-source observation at time to be represented by x_t, and let y_t denote the measured cooling load. The encoded sequence is formed as:

$$z_t = W_x x_t + W_y y_{t-1} + W_c c_t + b_z \qquad \text{Eq.(1)}$$

This representation separates exogenous operating variables from the autoregressive cooling load signal. The hidden state is then updated by a gated recurrent unit to retain delayed thermal memory:

$$h_t = G(h_{t-1}, z_t, r_t, u_t) \qquad \text{Eq.(2)}$$

The gate variables r_t and u_t regulate short-term workload impact and slower cooling-system inertia. To reduce scale imbalance among sensors, each physical variable is normalized by a robust deviation estimator:

$$\tilde{x}_{t,j} = \frac{x_{t,j} - \mathrm{median}(x_j)}{IQR(x_j) + \epsilon} \qquad \text{Eq.(3)}$$

The various physical ranges and noise characteristics of the operational variables need a rather reliable Normalization Strategy. Thermal inertia and control delay will cause the chilled water temperature to vary more slowly after the workload relocation, although the IT power will change very rapidly. The recurrent unit will give high-magnitude variables too much weight and suppress weaker but operationally meaningful signals if the aforementioned variables are directly concatenated without scaling. As a result, the model retains the original physical scale for assessment and inverse transformation while using normalized variables for training.

The prior L observations are stacked to provide a temporal context frame. Consequently, for a single predicting instance, the input tensor is:

$$X_t = [\tilde{x}_{t-L+1}, \tilde{x}_{t-L+2}, \dots, \tilde{x}_t] \quad \text{Eq.(4)}$$

Statistics are not used to determine the context's duration. Air mixing, chilled-water transit, and equipment control hysteresis can cause a data center's cooling system to react slowly to variations in server load. The daily work cycle is covered by a 24-hour context, although recurrence may still be employed to concentrate on a shorter time frame when the most recent IT power indication is noticeable. This design is more flexible in responding to fluctuating cooling needs and does not rely on a single set seasonal pattern.

DeepAR-Quantile Forecasting Architecture

An autoregressive hidden-state model that directly produces quantiles is called the DeepAR-Quantile network. As illustrated in Figure 2, during training, the prior ground-truth cooling load is used as the autoregressive input, and the projected median trajectory is fed forward recursively during inference. Future cooling demand is unclear, and this design represents the real conditions of online forecasting.

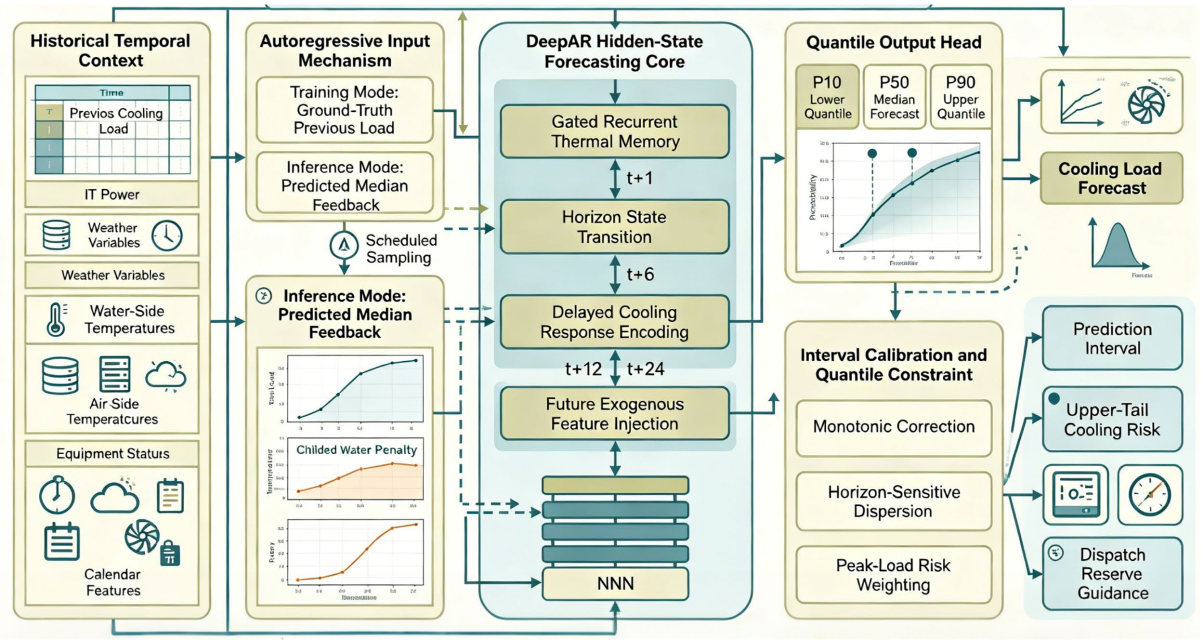

Figure 2. DeepAR-Quantile probabilistic forecasting architecture

For each future horizon tau, the hidden representation is projected into multiple quantile outputs:

$$\hat{y}^{q}_{t+\tau} = w_q^T h_{t,\tau} + \beta_q \quad \text{Eq.(5)}$$

The autoregressive horizon transition is calculated as:

$$h_{t,\tau} = G\left(h_{t,\tau-1}, \hat{y}^{0.5}_{t+\tau-1}, x^{ex}_{t+\tau}\right) \quad \text{Eq.(6)}$$

To represent uncertainty expansion over longer horizons, the model introduces a horizon-sensitive dispersion term:

$$d_{t,\tau} = \mathrm{softplus}\left(a^T h_{t,\tau} + \gamma \log(1+\tau)\right) \quad \text{Eq.(7)}$$

The upper and lower quantiles are constrained around the median by a monotonic correction:

$$\hat{y}^{q_u}_{t+\tau} = \hat{y}^{0.5}_{t+\tau} + d_{t,\tau}\eta_u \quad \text{Eq.(8)}$$

For lower quantiles, the correction is written as:

$$\hat{y}^{Q_t}_{t+\tau} = \hat{y}^{0.5}_{t+\tau} - d_{t,\tau}\eta_l \quad \text{Eq.(9)}$$

In noisy or infrequently observed operating circumstances, monotonic correction is employed to stop quantiles from crossing due to separate training of quantile heads. Such a cross makes it hard to read the interval and is both a mathematical blunder and a practical issue. As a result, direction-specific coefficients learn asymmetric uncertainty and the upper and lower limits share the dispersion term. The reason for this is that an abrupt increase in work often causes greater upper-tail uncertainty than lower-tail uncertainty.

If upper quantiles are lower than lower quantiles, a penalty is imposed to avoid quantile crossover.

$$P_q = \sum_{i<j} \max(0, \hat{y}^{q_i} - \hat{y}^{q_j}) \quad \text{Eq.(10)}$$

The autoregressive decoder also distinguishes training and inference behavior. During training, the previous measured cooling load is available and stabilizes gradient estimation. During inference, the model must use its own forecasted median as the next autoregressive input. To reduce accumulated error, scheduled sampling is introduced in the last 40 training epochs. The probability of replacing the measured previous value with the predicted median gradually increases from 0.15 to 0.55, allowing the model to adapt to online forecasting conditions.

Quantile Loss and Prediction Interval Calibration

Quantile loss, interval calibration, smoothness control, and peak-load sensitivity are the optimization objectives. Because the error cost is asymmetric and relies on whether the forecast surpasses or falls short of the desired cooling load, quantile loss is employed.

The quantile q's Pinball Loss is determined by:

$$\rho_q(e) = \max(qe, (q-1)e) \quad \text{Eq.(11)}$$

The multi-horizon quantile loss is aggregated by horizon weights:

$$L_Q = \sum_{\tau} \omega_\tau \sum_{q} \rho_q\left(y_{t+\tau} - \hat{y}^q_{t+\tau}\right) \quad \text{Eq.(12)}$$

Because cooling load should vary smoothly unless workload or weather shifts occur, a trajectory regularizer is introduced:

$$L_S = \sum_{\tau} |\hat{y}^{0.5}_{t+\tau} - 2\hat{y}^{0.5}_{t+\tau-1} + \hat{y}^{0.5}_{t+\tau-2}| \quad \text{Eq.(13)}$$

Prediction interval coverage is calibrated by comparing the observed target with the predicted lower and upper bounds:

$$C_t = I(\hat{y}^{0.1}_t \le y_t \le \hat{y}^{0.9}_t) \quad \text{Eq.(14)}$$

The interval calibration loss is written as:

$$L_C = |\mathrm{mean}(C_t) - 0.8| + \lambda_w \mathrm{mean}(\hat{y}^{0.9}_t - \hat{y}^{0.1}_t) \quad \text{Eq.(15)}$$

Peak-load risk is emphasized by assigning higher weights to targets above the historical 85 th percentile:

$$\omega_t = 1 + \alpha I(y_t > Q_{0.85}(y)) \quad \text{Eq.(16)}$$

The final optimization objective is:

$$L = L_Q + \lambda_c L_C + \lambda_s L_S + \lambda_p P_q \quad \text{Eq.(17)}$$

A single loss function is inappropriate since design serves several functions. Smoothness loss reduces physically irrational oscillations, pinball loss enhances quantile estimation, calibration loss controls interval reliability, and crossing penalty maintains interval order. The model's sensitivity to a high-cooling-demand event is increased by peak-load weighting, but it still fairly accounts for typical operating circumstances because of a cap on the weighting coefficient. According to validation, alpha has been adjusted at 0.35 in the implemented environment in order to raise the maximum recall without appreciably increasing the frequency of false alarms.

https://doi.org/10.64972/jiic.2023v1.364p3s:27-41

Experimental Scheme and Quantitative Evaluation

Dataset Construction and Forecasting Settings

An experimental dataset has been created based on 18 months of data from a data center's energy system at a 5-minute sample frequency. Cooling load, IT power, outside temperature, humidity, chilled water supply and return temperatures, return air temperature, cooling tower status, chiller part-load ratio, and pump frequency are all included in the record. The dataset contains 151,200 valid time steps once the invalid sensor period and unduly brief maintenance pauses are eliminated. Ten percent of the sequence is utilized for validation, twenty percent is used for testing, and the first seventy percent is used for training. Forecast Timeframes: 15 minutes, 1 hour, 3 hours, and 6 hours. The related historical procedure is 24 hours long, and the context length is 288-time steps.

To avoid operational and seasonal similarities in the test set due to random shuffling, the dataset is split chronologically. Longer missing segments are masked and removed from the loss computation, while missing values shorter than three consecutive intervals are filled using local linear interpolation. Because they do not represent a typical cooling condition, cooling load values of less than 5 kW during a maintenance shutdown are not utilized for training. The greatest recorded value is 168.9 kW, the 90th percentile is 132.5 kW, and the average cooling load is 84.7 kW. For quantile calibration, the aforementioned statistics can produce a realistic high-load tail.

Accuracy, Interval, and Risk Evaluation Metrics

The three components of the system are peak-load risk sensitivity, probabilistic interval quality, and deterministic accuracy. The deterministic accuracy of the P50 forecast is displayed using RMSE, MAE, and MAPE. The average interval width, prediction interval coverage probability, and pinball loss at P10, P50, and P90 all demonstrate probabilistic performance. Peak-load recall, false alarm rate, and warning lead time are measures of operational risk performance. The aforementioned indices were selected because a data center cooling forecaster must be extremely precise under typical conditions and dependable in high-load danger situations. Both a vast interval with a low dispatch value and a narrow but inadequately covered period are unacceptable.

The actual tolerance band is defined as 78%–84%, whereas the intended coverage of the P10–P90 range is set at 80% for probabilistic evaluation. Larger coverage areas do not always translate into better models; if they do, the increased range will be excessively cautious in cooling reserve and waste energy. When the observed cooling loads in all future periods surpass the historical 90th percentile, the peak-load alarm is activated. A warning is only deemed legitimate if it is issued within 45 minutes of the incident. As a result, the late alarms will not be regarded as operational alerts.

Baseline Models and Training Configuration

ARIMA, support vector regression, random forest, LightGBM, LSTM, GRU, Transformer, deterministic DeepAR, and Temporal Fusion Transformer are compared with the approach shown here. Every neural network receives the same sensor input and has the same train-validation-test split. AdamW is used to train the DeepAR-Quantile model for 120 epochs. The initial learning rate is set to 0.0003, the batch size is 64, the weight decay is set to 0.01, and the gradient clipping is set to 1.0. Quantiles should be set at P10, P50, and P90. If the validation pinball loss does not improve after 12 epochs, early halting is used. In order to avoid the suggested model exhibiting too strong performance, the hardware setup for the RTX 3060 GPU and Intel i7 workstation CPU inference evaluations in fact are somewhat conservative. Lagged cooling load, rolling mean, rolling variance, IT power, weather factors, and calendar variables are all utilized in tree-based models. Both GRU and LSTM are recurrent layers with 64 hidden units. The Transformer baseline has a hidden size of 128 and four attention heads. The suggested method's recurrent backbone is used by the deterministic DeepAR baseline, which produces a single mean forecast rather than quantiles. Rolling statistics are computed using just the historical data accessible prior to the forecast origin in order to minimize data leaking, and this comparison separates the effect of direct quantile learning from the overall benefit of autoregressive sequence modeling. The operational prediction that the energy management system has access to represents weather variables at a future horizon; observed future values are not used. For a 6-hour horizon, include a realistic level of weather uncertainty. The test set is utilized once after finalization, and only the validation set is used for hyperparameter selection. Since this protocol complies with deployment procedures, streaming data must be processed using a cooling-load forecaster instead of retrospectively cleaned sequences.

Result Interpretation and Comparative Analysis

Forecasting Accuracy Comparison

The following Figure 3 represents the overall predicted error. Deep sequence models are more suited for managing nonlinear temporal dependencies in the target, according to earlier studies on energy and load forecasting [26]. 15-minute forecast accuracy is displayed in Figure 3(a). The suggested model's RMSE and MAE are 1.86 kW and 1.21 kW, respectively; LSTM's RMSE is 2.34 kW and LightGBM's is 2.58 kW. The 1-hour horizon is depicted in Figure 3(b), and the suggested model has obtained 2.36 kW MAE and 3.42 kW RMSE. Autoregressive memory, which maintains the delayed effect of IT power on cooling demand, is primarily responsible for the aforementioned improvement. The 3-hour horizon is shown in Figure 3(c); the Transformer underestimates the abrupt high-load time even if it is superior to the recurring baseline. The 6-hour horizon is depicted in Figure 3(d), where the suggested model outperforms deterministic DeepAR at 7.6% and GRU at 8.4% by reducing MAPE to 5.8%.

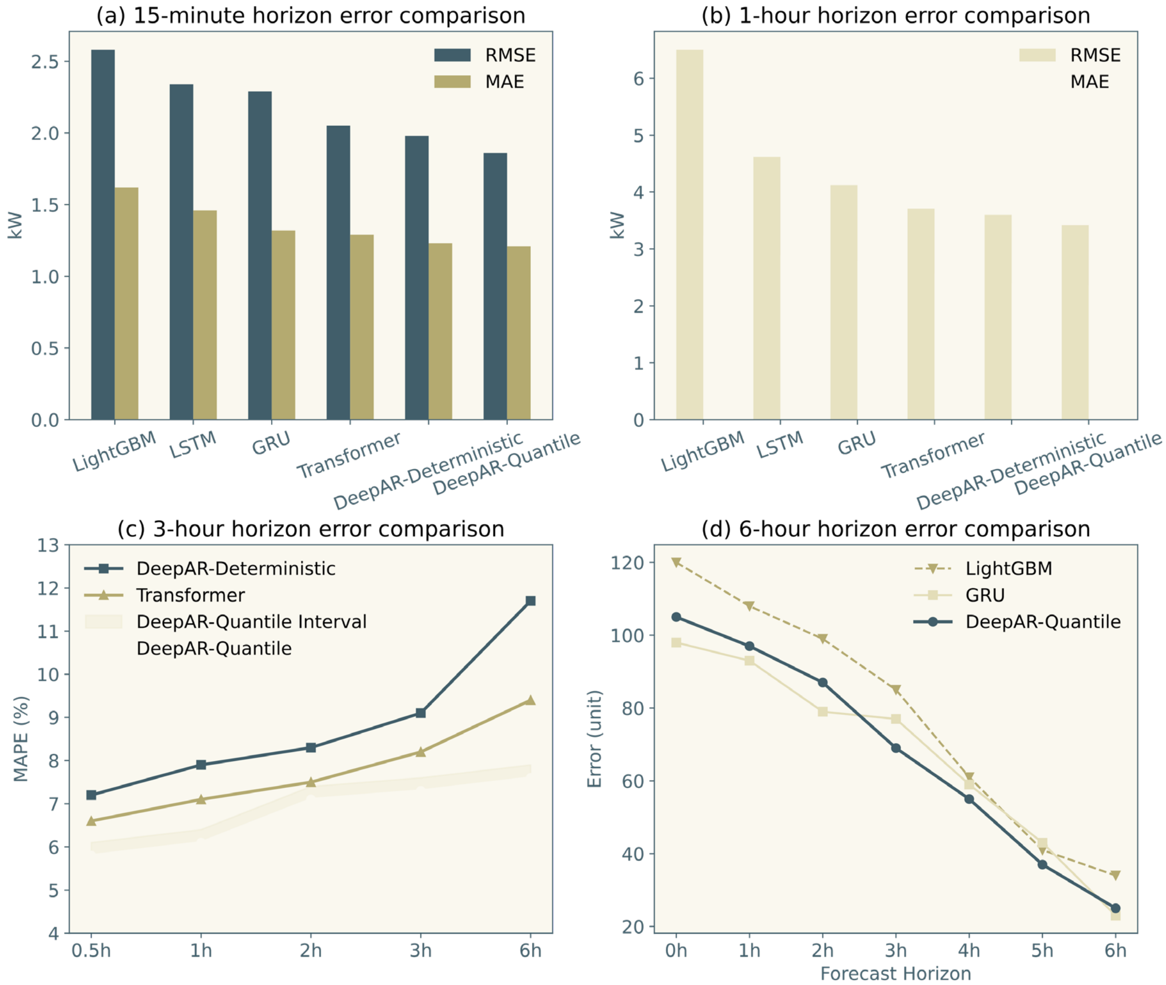

Figure 3. Cooling load forecasting accuracy comparison. (a) 15-minute horizon error comparison. (b) 1-hour horizon error comparison. (c) 3-hour horizon error comparison. (d) 6-hour horizon error comparison.

Error behavior is also influenced by the type of load condition. Because they require less cooling, most models perform similarly under steady nighttime operation conditions. However, shallow models are unable to keep up with the day's peak activity. By encouraging the hidden state to hold onto high-load evidence rather than smoothing it out, the suggested approach better addresses this situation. Additionally, studies on data center cooling reveal that airflow and chilled-water loops delay thermal response [27], making models based just on recent point values unreliable.

The research above indicates that the range between normal demand and high-cooling demand is where the largest benefit takes place. The model decreased the frequency of underprediction from 13.6% to 6.9% at the one-hour horizon. Under no circumstances can the range be underestimated for thermal control. The system will have to react more forcefully later if the chiller is turned on too late. Although the suggested method's median forecast is somewhat less smooth than the LightGBM forecast, cooling benefits from this.

A data center's load is not exactly periodic. It can react to changes in equipment control or a surge in workload that aren't displayed by the calendar function. The Transformer baseline is steady during the short run but good in the middle term, according to the tests mentioned above. Attention systems may be excessively sensitive to noise in a limited number of recent sensor readings, despite their ability to acquire long-range relationships. Because recurrent autoregression naturally preserves local continuity and updates the hidden state gradually, it is more stable in the near future. In addition to adding quantile outputs for uncertainty, the model will display the stability mentioned above. As a result, it has done well across all time horizons and enhances more than just one forecasting length.

Simultaneously optimize the cooling system's control. If the worst mistakes happen during the ramp-up phase, the point prediction might not be feasible even with a low RMSE. Within 30 minutes following a sudden boost in IT power, 61.3% of the biggest baseline faults in the test set manifest. This% is lowered to 38.7% by the aforementioned model, suggesting that the concealed state is more in line with transitory cooling behavior. The combined sequence of IT power, water temperature, and air temperature has taught the model a statistically stable delay pattern, but this does not imply that it has fully learnt the physical loop.

The outcome of the deterministic comparison is displayed in Figure 4. The residual error distribution for several models is shown in Figure 4(a). The new method features fewer severe underprediction locations and a lower middle error range. Working-day and weekend prediction behavior is depicted in Figure 4(b). Weekend cooling loads are more consistent, whereas working-day mistakes are greater owing to frequent fluctuations in the IT demand. The suggested model has consistently improved both groups, and forecasting tests on load series have demonstrated that model assessment should be divided by operating condition [28].

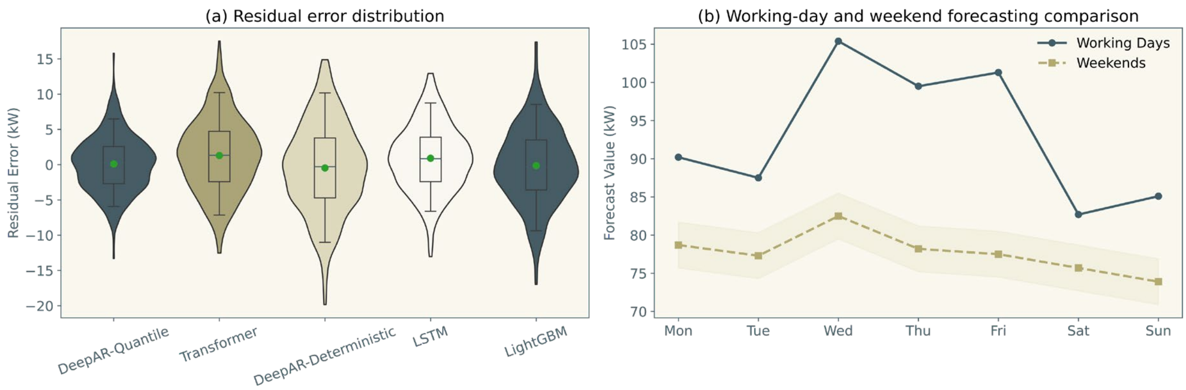

Figure 4. Error distribution under operating conditions. (a) Residual error distribution. (b) Working-day and weekend forecasting comparison.

The residual distribution demonstrates that by lowering the average error, the suggested model has not considerably raised the variance. The residuals' interquartile range is 4.18 kW, whereas Transformer's and deterministic DeepAR's are 5.72 kW and 6.41 kW, respectively. When compared to the best baseline, the model lowers MAE by 16.5% on working days. Over the weekend, there was a comparatively little decline of 11.9%. As a result, cooling loads with a high level of operational uncertainty are better suited for the model.

The suggested model also has a more uniform distribution of residual signals. Because they have optimized for the average loss across the whole distribution, some deterministic baselines consistently underestimate high-load times. The proportion of negative residuals with high IT use has decreased, but not all underpredictions have been eradicated by the model. Cooling reserve planning requires this behavior, and a biased model may

https://doi.org/10.64972/jiic.2023v1.364p3s:27-41

consistently underestimate the same class of operational events. This bias is seen in interval coverage and pinball loss in the quantile framework, and model failure is easier to spot.

The model has not learnt a general pattern for all days, according to a comparison of working-day and weekend outcomes. If this were the case, there would still be a lot of high-load weekday mistakes, but weekend performance would be far better than working-day performance. Instead, in both situations, the model has a comparable relative advantage. As a result, the prediction will take into account the dynamic factors of chilled water temperature and IT power. The function will be used to adjust to changes in the workload pattern following the implementation of maintenance procedures, cloud resource transfer, or service growth.

Probabilistic Interval and Uncertainty Analysis

Figure 5 is the outcomes of probabilistic forecasting. Sharpness and interval coverage are both necessary for probabilistic energy forecasting [29]. The P50 cooling-load curve is shown in Figure 5(a), and the suggested model closely matches the observed trajectory's high-load and low-load phases. The P10-P90 prediction interval is shown in Figure 5(b). A fixed-width interval is inappropriate as the interval grows at times of heavy demand and contracts during regular operation. Prediction interval coverage probability is compared in Figure 5(c). Deterministic residual-based intervals are only at 84.2%, whereas the new model has achieved 91.6% P90 coverage. Pinball loss is depicted in Figure 5(d), and when compared to Temporal Fusion Transformer, the suggested model lowers the average loss by 18.7%.

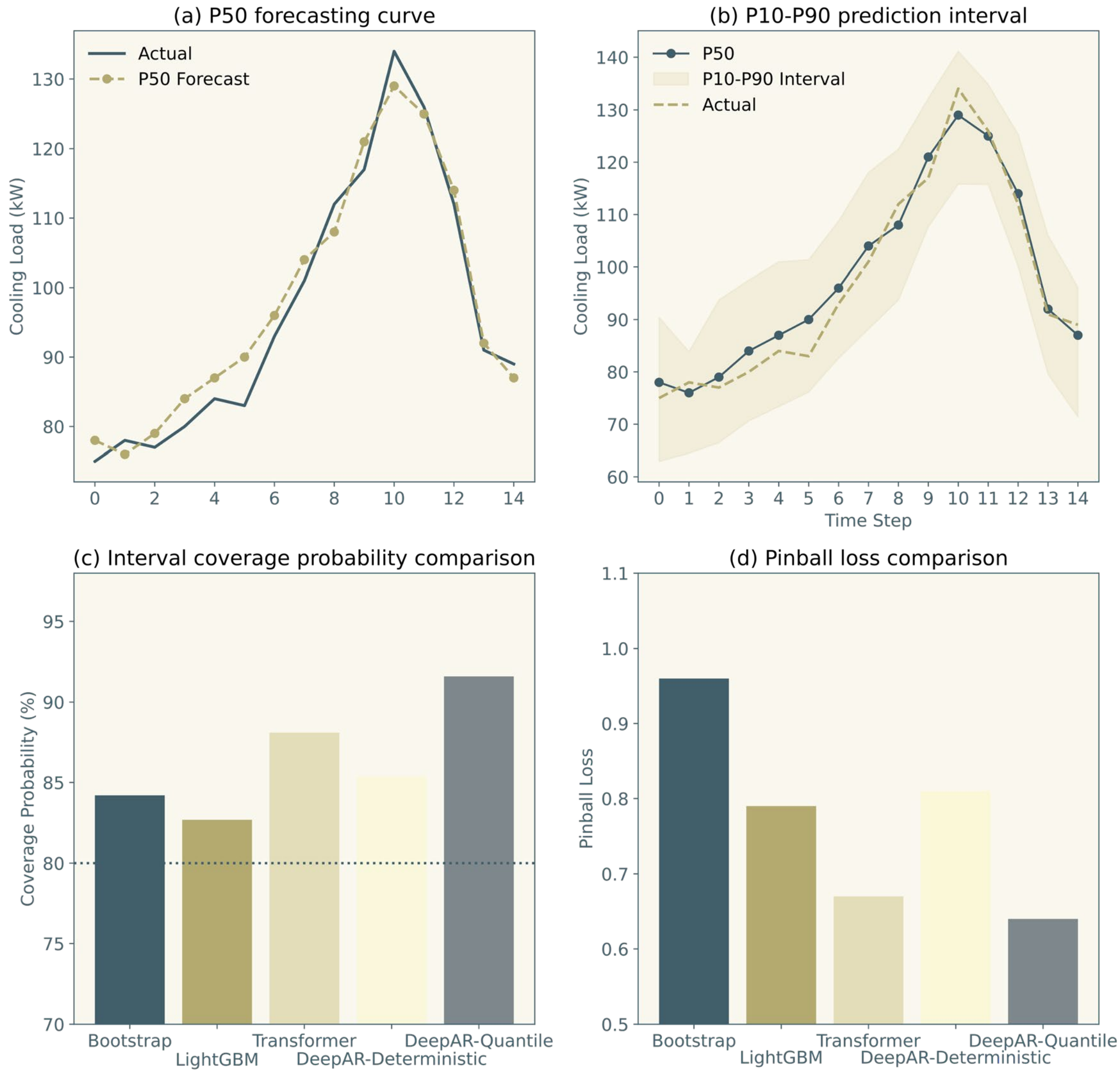

Figure 5. Quantile interval forecasting performance. (a) P50 forecasting curve. (b) P10-P90 prediction interval. (c) Interval coverage probability comparison. (d) Pinball loss comparison.

There is a strong technical advantage to the Interrupt Behavior. A data center operator must fulfill the upper quantile for chiller capacity planning in addition to the average cooling demand. Because the new method is based on a hidden temporal state rather than being computed after point prediction, it has a smaller interval than the bootstrap baseline. Direct quantile learning often provides better operational calibration than post-hoc residual fitting, according to quantile forecasting research [30]. The model modifies uncertainty based on the operating circumstances, as seen by the interval width increasing by 21.4% during the workload spike phase and decreasing by 16.8% during the steady night period of this dataset.

Calibration and Horizon Behavior Are Different. The P10-P90 range is narrow at the 15-minute horizon since the near-future condition is already determined by the recent cooling load and IT power. The delay lengthens at the 6-hour horizon because of equipment-state uncertainty and cumulative weather influences. The suggested approach has a comparatively small interval and good coverage of the target region. The average P10-P90 width is 14.2 kW at one hour and 25.8 kW at six hours; both are 9.6% less than the residual-bootstrap interval but still offer adequate coverage.

There has also been a rise in high-load risk detection. A cooling load that is higher than the 90th percentile in the past will be regarded as a dispatch-sensitive occurrence. The suggested model outperforms Transformer (88.1%) and LightGBM (82.7%) with a high-load recall of 93.8% at the 1-hour horizon. An early warning system for a high-cooling-demand location can prevent overheating, according to research on data center thermal management [31]. About 18.5 minutes before the high-load threshold is exceeded, the DeepAR-Quantile model will send out a warning, which is long enough for a phased chiller adjustment.

False alarm behavior is also permissible. Though the majority of the additional warnings happen close to the load threshold, the suggested approach has a false alarm rate of 4.6%, which is somewhat higher than the deterministic DeepAR baseline's 3.9%. This trade-off makes sense in real cooling control as a few cautious cautions are less expensive than a rapid thermal rise. As a result, the model's two levels are applied as follows: P90 is utilized for reserve preparation and supervisory alarms, whereas P50 is used for standard control.

P10 output serves an operational function as well. Even if the median estimate is rather low, it may be assumed that the cooling demand will not decline dramatically anytime soon if the lower quantile is high for a number of consecutive months. This sometimes happens as a result of extended periods of heavy server usage or comparatively high nighttime outside temperatures. As a result, early cooling decrease can be avoided by the lower quantile. However, operators may view the forecast as a risk-sensitive alert rather than an accurate load projection if the interval is comparatively big and biased towards the upper end.

The analysis of the interval curves demonstrates how different causes of uncertainty lead to varied interval forms under the model. The interval progressively rises as the load varies due to weather since the cooling loop is slowly impacted by the outside temperature. The upper interval expands more suddenly during workload-driven changes because the server's capacity to release heat varies quickly. Both can be employed for control; an abrupt upper-tail expansion necessitates quick reserve preparation, while a progressive interval growth can be managed by planned chiller staging. It would be challenging to get such behavior from a residual interval connected to a deterministic model.

Robustness and Operational Applicability Discussion

Figure 6 illustrates robustness to external perturbations. After 10% of the temperature sensor data was randomly masked, the predicted results are shown in Figure 6(a). The suggested model's LSTM is 5.02 kW and its RMSE is 3.91 kW. An exterior temperature-shift curve is shown in Figure 6(b). The suggested model still achieves an interval coverage of 89.4% when the test profile's outdoor temperature is raised by 3°C. We must guarantee the resilience of interval reliability under disturbances since sensor uncertainty is an issue in energy systems [32].

Figure 7 displays the deployment-oriented comparison. The inference delay is displayed in Figure 7(a), and the suggested model predicts one sequence in 14.6 ms on GPU and 46.3 ms on CPU. The model sizes are displayed in Figure 7(b), and after export, the suggested model is 18.4 MB. The scheduling support impact is depicted in Figure 7(c), and in the simulated dispatch situation, interval-aware cooling reserve decreases needless overcooling hours by 11.2%. Forecasting models must strike a compromise between accuracy and running costs, as demonstrated by edge and energy-system studies [33]. Therefore, the approach is suitable for online monitoring systems that update every five or fifteen minutes.

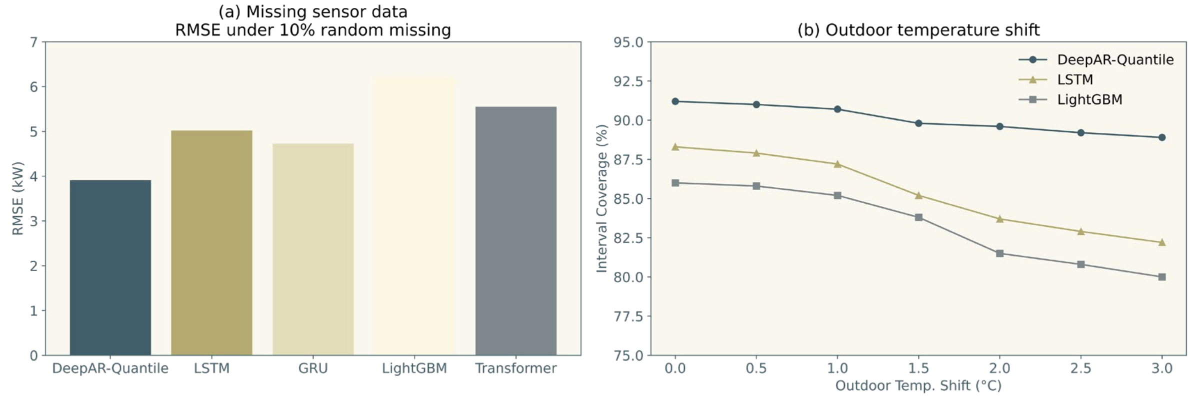

Figure 6. Robustness under missing and disturbed sensor conditions. (a) Missing sensor data. (b) Outdoor temperature shift.

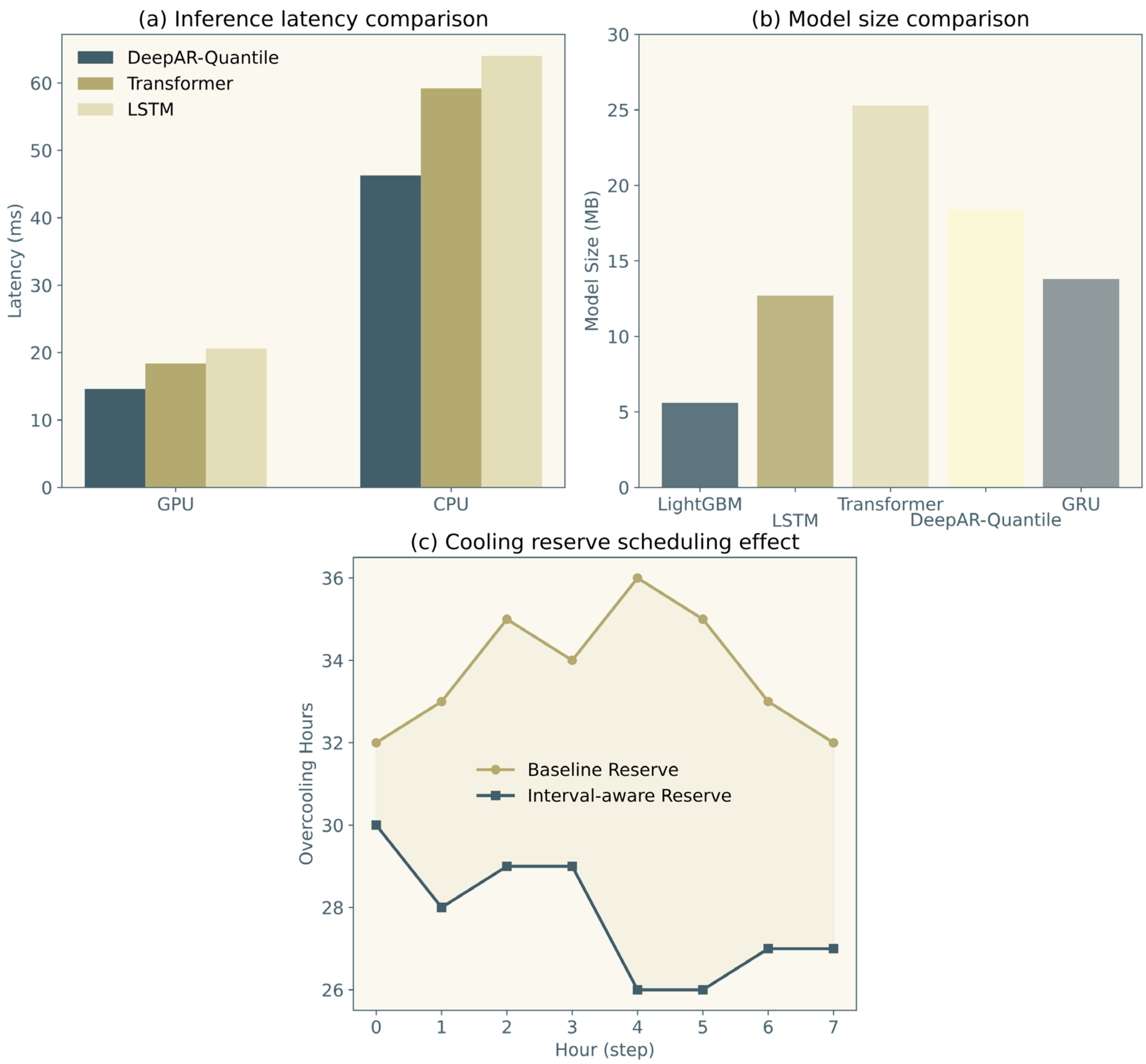

Figure 7. Deployment efficiency and cooling dispatch support. (a) Inference latency comparison. (b) Model size comparison. (c) Cooling reserve scheduling effect.

https://doi.org/10.64972/jiic.2023v1.364p3s:27-41

Ablation research ascertain why a certain modification enhanced performance. The model is a deterministic autoregressive forecaster in the absence of quantile loss, and the P90 coverage decreases from 91.6% to 82.9%. After exogenous weather factors were eliminated, peak-load recall decreased by 6.8% points and 6-hour RMSE rose by 9.5%. Workload and temperature circumstances have a direct impact on data center load and cooling behavior [34]. Therefore, past cooling load should not be the exclusive basis for probabilistic cooling estimates.

Chilled-water variables are more suited for medium-horizon accuracy than for short-horizon precision, according to the ablation findings. The 3-hour RMSE is 8.7%, while the 15-minute RMSE is only 3.1% higher when the chilled-water supply and return temperatures are excluded. As a result, the IT power does not represent the water-side thermal information, which probably has slower dynamics. Because planned work behavior has a minor periodic component, calendar indicators increase weekend and nighttime stability while having the smallest individual impact.

There are flaws in this approach as well. The model will provide a wide range rather than a precise forecast when the cooling system is in an uncertain maintenance status. It is not very useful for precise dispatch, but it is appropriate for danger warning. Distribution shift is still a challenge for deep probabilistic models, according to forecasting literature [35]. Update the model on a regular basis using fresh data collected in practice. The primary residual errors arise when chiller changeover, workload migration, and an increase in the outside temperature all occur at the same time. The coupled phenomena indicate a non-linear reaction, according to research on thermal management and cooling technologies [36].

Although the model lacks an explicit thermodynamic equation, it is capable of learning statistical connections. It cannot ensure energy balance in the chilled-water loop; it can only ascertain the data's delayed reaction. The learned association will progressively alter if the sensor is still wandering after a few days. As a result, sensor quality checks and intervals will be part of a realistic deployment.

Periodic validation of the measured cooling equipment performance and drift monitoring. Since even a well-calibrated model would become inaccurate if the measurement layer changes, the following actions are essential in the mission-critical data center.

For practical reasons, it can be used in conjunction with a supervisory decision layer. It shouldn't immediately take the role of device protection logic or rule-based safety limitations. After giving the dispatch optimizer a probabilistic cooling demand, decide on chiller staging, pump frequency, and cooling tower operation while adhering to safety regulations. In this workflow, the interval width represents the uncertainty cost, P50 represents the economical operating point, and P90 represents the reserve required. The interpretation is more helpful to operators since it might assist them comprehend why the prediction has altered.

The forecast-to-energy-saving strategy relationship is now more transparent thanks to the model. Reducing the necessary cooling capacity, increasing the temperature of the chilled water supply, or postponing chiller start-up are many energy-saving measures. Only when the future load uncertainty can be measured are these procedures safe. Do not implement significant energy-saving measures based on a mild P50 projection if the P90 forecast is near the top limit of the present cooling capacity. On the other hand, the system can be safely lowered to reserve cooling if both P50 and P90 are low and the range is small. Adding a set safety margin to all point forecasts is not as instructive as using forecast distribution.

Lastly, gradual deployment is also supported by the approach suggested here. First, without automated operation, the model in advisory mode may be used to compare the actual cooling load with the projected interval. The P50 output may be utilized for scheduling suggestions, and the P90 output can be linked to alert thresholds. Interval reliability has been confirmed. As a result, the data center operator needs consistent behavior.

It will be used in a decision-support distribution system instead of a single-point estimating model since it has demonstrated good performance in practice. Because cautious thermal operation in high-density data centers requires transparency, the deployment objective strives to achieve a smaller forecast error and a more disciplined cooling reserve strategy that can be explained, audited, and altered by operators. It is also possible. From an operational perspective, the suggested projection may be combined with dashboards for energy management, chiller sequencing, and model predictive control. P90 output is used for reserve decisions, whereas P50 output is used for routine dispatch. Because the dangers of undercooling and overcooling are not equivalent, this separation is necessary for energy-efficient data center operation [37]. Both frameworks offer reasonable trade-offs between uncertainty-aware control and high-accuracy prediction. The cooling architecture, airflow design, and workload distribution are responsible for facility-level energy usage, according to earlier but still pertinent studies on data centers [38]. To illustrate the operator the likely cooling load and the necessary risk, include a probability interval. Temporal clustering and operating-condition separation have been suggested by other smart-meter and energy-behavior studies to improve deployment robustness [39]. It is possible to significantly minimize cooling energy while preserving thermal safety in the future by combining the suggested prediction and adaptive control approaches [40].

Conclusion

Here, DeepAR-Quantile is suggested as a solution to the cooling load forecasting problem in a data center energy system. In an autoregressive sequence forecasting framework, the model integrates IT power, cooling-load history, outdoor meteorological elements, chilled-water parameters, return-air temperature and humidity, and equipment-state indicators. For a non-deterministic forecasting model, the two are the upper-tail risk estimate and the median cooling demand, respectively.

According to the research above, quantile-aware autoregressive learning enhances interval reliability, high-load warning, and multi-horizon forecasting accuracy. The model is appropriate for a hot area, heats slowly, and performs well under heavy loads. Since a specific capacity must be reserved without excessive overcooling, prediction intervals are more practical than point predictions for the operational application of data center cooling dispatch.

The system may be expanded in further work by include airflow simulation functions, online model updates, physical cooling-loop limitations, and model predictive control. Incorporating the probabilistic cooling forecasts into a real-time data center energy management platform may also involve hardware-aware compression and edge deployment.

Author Contributions

Disha Malhotra contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, project administration, and funding acquisition. Preeti Soni contributes to conceptualization, methodology, software, validation. Shobha Razdan contributes to methodology, software. Kavya Reddy contributes to conceptualization. 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] Khalil, M., McGough, A. S., Pourmirza, Z., Pazhoohesh, M., & Walker, S. (2022). Machine learning, deep learning and statistical analysis for forecasting building energy consumption: A systematic review. Engineering Applications of Artificial Intelligence, 115, 105287. https://doi.org/10.1016/j.engappai.2022.105287

[2] Fathi, S., Srinivasan, R. S., Fenner, A. E., & Fathi, S. (2020). Machine learning applications in urban building energy performance forecasting: A systematic review. Renewable and Sustainable Energy Reviews, 133, 110287. https://doi.org/10.1016/j.rser.2020.110287

[3] Runge, J., & Zmeureanu, R. (2021). A review of deep learning techniques for forecasting energy use in buildings. Energies, 14(3), 608. https://doi.org/10.3390/en14030608

[4] Bedi, J., & Toshniwal, D. (2019). Deep learning framework to forecast electricity demand. Applied Energy, 238, 1312-1326. https://doi.org/10.1016/j.apenergy.2019.01.113

[5] Wang, Z., Hong, T., & Piette, M. A. (2020). Building thermal load prediction through shallow machine learning and deep learning. Applied Energy, 263, 114683. https://doi.org/10.1016/j.apenergy.2020.114683

[6] Gao, Y., Ruan, Y., Fang, C., & Yin, S. (2020). Deep learning and transfer learning models of energy consumption forecasting for a building with poor information data. Energy and Buildings, 223, 110156. https://doi.org/10.1016/j.enbuild.2020.110156

[7] Zhang, X., Wen, J., Li, S., Chen, G., & Zhang, Z. (2021). A review of machine learning in building load prediction. Applied Energy, 285, 116452. https://doi.org/10.1016/j.apenergy.2021.116452

[8] Zhang, L., Li, J., Zhao, L., & Li, Y. (2020). A hybrid deep learning-based method for short-term building energy load prediction combined with an interpretation process. Energy and Buildings, 225, 110301. https://doi.org/10.1016/j.enbuild.2020.110301

[9] Liu, T., Tan, Z., Xu, C., Chen, H., & Li, Z. (2020). Study on deep reinforcement learning techniques for building energy consumption forecasting. Energy and Buildings, 208, 109675. https://doi.org/10.1016/j.enbuild.2019.109675

[10] Yang, J., Tan, H., Santamouris, M., & Lee, S. E. (2019). Building energy consumption raw data forecasting using data cleaning and deep recurrent neural networks. Buildings, 9(9), 204. https://doi.org/10.3390/buildings9090204

[11] Sha, H., Moujahed, M., & Qi, D. (2021). Machine learning-based cooling load prediction and optimal control for mechanical ventilative cooling in high-rise buildings. Energy and Buildings, 242, 110980. https://doi.org/10.1016/j.enbuild.2021.110980

[12] Lu, H., Tian, Z., Zhou, Y., & Liu, W. (2021). Multi-step-ahead prediction of thermal load in regional energy system using deep learning method. Energy and Buildings, 233, 110658. https://doi.org/10.1016/j.enbuild.2020.110658

[13] Gao, Y., Yu, X., Zhao, X., & Hu, W. (2022). A hybrid method of cooling load forecasting for large commercial building based on extreme learning machine. Energy, 238, 122073. https://doi.org/10.1016/j.energy.2021.122073

[14] Dong, B., Yu, Y., Quan, D., & Xiang, X. (2022). Short-term building cooling load prediction model based on DwdAdam-ILSTM algorithm: A case study of a commercial building. Energy and Buildings, 272, 112337. https://doi.org/10.1016/j.enbuild.2022.112337

[15] Chalapathy, R., Khoa, N. L. D., & Sethuvenkatraman, S. (2021). Comparing multi-step ahead building cooling load prediction using shallow machine learning and deep learning models. Sustainable Energy, Grids and Networks, 28, 100543. https://doi.org/10.1016/j.segan.2021.100543

[16] Kang, J., Wang, Y., An, J., & Yan, D. (2022). A novel approach of day-ahead cooling load prediction and optimal control for ice-based thermal energy storage system in commercial buildings. Energy and Buildings, 275, 112478. https://doi.org/10.1016/j.enbuild.2022.112478

[17] Wang, Z., Lee, W. L., Yuen, R. K. K., & Feng, W. (2019). Cooling load forecasting-based predictive optimisation for chiller plants. Energy and Buildings, 198, 261-274. https://doi.org/10.1016/j.enbuild.2019.06.016

[18] Sala-Cardoso, E., Delgado-Prieto, M., Kampouropoulos, K., & Romeral, L. (2020). Predictive chiller operation: A data-driven loading and scheduling approach. Energy and Buildings, 208, 109639. https://doi.org/10.1016/j.enbuild.2019.109639

[19] Chan, A. L. S., Wong, S. L., Yow, S., & Yuen, R. K. K. (2022). Development and performance evaluation of a chiller plant predictive operational control strategy by artificial intelligence. Energy and Buildings, 262, 112017. https://doi.org/10.1016/j.enbuild.2022.112017

[20] Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting, 36(3), 1181-1191. https://doi.org/10.1016/j.ijforecast.2019.07.001

[21] Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748-1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

[22] Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. International Journal of Forecasting, 37(1), 388-427. https://doi.org/10.1016/j.ijforecast.2020.06.008

[23] Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., et al. (2020). Criteria for classifying forecasting methods. International Journal of Forecasting, 36(1), 167-177. https://doi.org/10.1016/j.ijforecast.2019.05.008

[24] Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., et al. (2022). Forecasting: Theory and practice. International Journal of Forecasting, 38(3), 705-871. https://doi.org/10.1016/j.ijforecast.2021.11.001

[25] González Ordiano, J. Á., Gröll, L., Mikut, R., & Hagenmeyer, V. (2020). Probabilistic energy forecasting using the nearest neighbors quantile filter and quantile regression. International Journal of Forecasting, 36(2), 310-323. https://doi.org/10.1016/j.ijforecast.2019.06.003

[26] Huang, N., Li, R., & Zhu, J. (2020). An improved convolutional neural network with load range discretization for probabilistic load forecasting. Energy, 203, 117902. https://doi.org/10.1016/j.energy.2020.117902

[27] Brusaferri, A., Matteucci, M., Spinelli, S., & Vitali, A. (2022). Probabilistic electric load forecasting through Bayesian mixture density networks. Applied Energy, 309, 118341. https://doi.org/10.1016/j.apenergy.2021.118341

[28] He, Y., Cao, Y., Wang, Y., & Fu, L. (2022). Nonparametric probabilistic load forecasting based on quantile combination in electrical power systems. Applied Energy, 322, 119507. https://doi.org/10.1016/j.apenergy.2022.119507

[29] Hong, T., Xie, J., & Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. International Journal of Forecasting, 35(4), 1389-1399. https://doi.org/10.1016/j.ijforecast.2019.02.006

[30] Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. Science, 367(6481), 984-986. https://doi.org/10.1126/science.aba3758

[31] Yuan, X., Zhou, X., Pan, Y., Kosonen, R., Cai, H., Gao, Y., & Wang, Y. (2021). Phase change cooling in data centers: A review. Energy and Buildings, 236, 110764. https://doi.org/10.1016/j.enbuild.2021.110764

[32] Gupta, S., Asgari, S., Moazamigoodarzi, H., & Down, D. (2021). Energy, exergy and computing efficiency based data center workload and cooling management. Applied Energy, 299, 117050. https://doi.org/10.1016/j.apenergy.2021.117050

[33] Borkowski, P., & Piłat, A. (2022). Customized data center cooling system operating at significant outdoor temperature fluctuations. Applied Energy, 306, 117975. https://doi.org/10.1016/j.apenergy.2021.117975

[34] Keskin, M. E., & Soykan, G. (2022). Optimal cost management of the CCHP based data center with district heating and district cooling integration in the presence of different energy tariffs. Energy Conversion and Management, 254, 115211. https://doi.org/10.1016/j.enconman.2022.115211

[35] Hu, J., Zheng, Y., & Wang, Y. (2022). Packing computing servers into the vessel of an underwater data center considering cooling efficiency. Applied Energy, 314, 118986. https://doi.org/10.1016/j.apenergy.2022.118986

[36] Li, Y., & Li, X. (2020). Model-based optimization of free cooling switchover temperature and cooling tower approach temperature for data center cooling system with water-side economizer. Energy and Buildings, 227, 110407. https://doi.org/10.1016/j.enbuild.2020.110407

[37] Díaz, A. J., Cáceres, G., Torres, R., & Cardemil, J. M. (2020). Effect of climate conditions on the thermodynamic performance of a data center cooling system under water-side economization. Energy and Buildings, 208, 109634. https://doi.org/10.1016/j.enbuild.2019.109634

[38] Fu, Y., Zuo, W., Wetter, M., & VanGilder, J. W. (2019). Equation-based object-oriented modeling and simulation of data center cooling systems. Energy and Buildings, 198, 503-519. https://doi.org/10.1016/j.enbuild.2019.06.037

[39] Turek, T., & Radgen, P. (2021). Optimized data center site selection: Mesoclimatic effects on data center energy consumption and costs. Energy Efficiency, 14, 30. https://doi.org/10.1007/s12053-021-09947-y

[40] Abdelrahman, M. M., Zhan, S., Miller, C., & Chong, A. (2021). Data science for building energy efficiency: A comprehensive text-mining driven review of scientific literature. Energy and Buildings, 242, 110885. https://doi.org/10.1016/j.enbuild.2021.110885