

Article

Control-Aware Multi-Horizon PUE Forecasting for Coordinated Data Center Demand-Side Management and Microgrid Dispatch

Yingqi Liang ^{1,2,*} , Junjie Peng ², Guanyu Fu ² and Dipti Srinivasan ^{2,*} ¹ Department of Electrical Engineering, School of Automation, Guangdong University of Technology, Guangzhou 510006, China² Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077, Singapore; junjie.peng@u.nus.edu (J.P.); fuguanyu@u.nus.edu (G.F.)

* Correspondence: yingqi.liang@gdut.edu.cn (Y.L.); dipti@nus.edu.sg (D.S.)

Abstract

Data centers can provide demand-side flexibility by coordinating computing workloads, cooling systems, and on-site energy resources. However, facility demand varies with information technology (IT) load and cooling operation, making a fixed power usage effectiveness (PUE) or a forecast independent of planned controls inconsistent with dispatch. This paper proposes a control-aware, multi-horizon PUE forecasting framework for coordinated data center demand-side management (DSM) and microgrid dispatch. The key idea of this control-aware approach is to forecast PUE using planned workload and cooling schedules as inputs. A power-consistent Temporal Fusion Transformer (PC-TFT) predicts quantiles of non-IT overhead power and rack inlet temperature from telemetry, weather forecasts, admitted requests, and candidate workload and cooling schedules. Facility power and PUE are derived from the algebraic power balance, ensuring consistency among IT, overhead, and facility power and PUE values no lower than 1. Empirical split conformal calibration and temporally dependent scenarios characterize forecast uncertainty. A trajectory-conditioned piecewise-affine control response map with a recursive thermal state links the forecasts to a risk-informed model predictive controller that coordinates workloads, cooling, photovoltaic generation, battery storage, and grid exchange. The proposed framework is validated through EnergyPlus simulations of a Shenzhen data center, coupled with workload and microgrid simulations. Forecasting performance is compared with persistence and matched-input neural baselines, while dispatch is benchmarked against deterministic and oracle controllers. The results demonstrate improved multi-horizon PUE forecasting accuracy and empirical interval calibration, lower operating cost and peak grid demand, higher renewable energy utilization, and fewer service quality violations. These findings indicate that control-aware, power-balance-constrained probabilistic PUE forecasts can provide a reliable basis for coordinated data center DSM and microgrid dispatch.

Keywords: data center; power usage effectiveness; forecasting; demand-side management; microgrid dispatch; Temporal Fusion Transformer; model predictive control; uncertainty quantification; deep learning



Academic Editors: Qi Wang, Marcin Kaminski and Pavlos S. Georgilakis

Received: 21 July 2026

Revised: 8 August 2026

Accepted: 14 August 2026

Published: 16 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

1. Introduction

Data centers can support demand-side management (DSM) by shifting deferrable computing workloads and coordinating local energy resources. Batch requests may be delayed within service deadlines, and cooling setpoints may be adjusted within rack

temperature limits and monitored humidity conditions. These options must be coordinated with on-site photovoltaic (PV) generation, a battery energy storage system (BESS), and grid exchange, all under variable solar output and time-of-use (TOU) tariffs. Their value depends on an accurate facility power estimate: an underestimate may increase cost, create a new demand peak, or cause a thermal constraint violation. Thus, the scale and variability of data center electricity use make facility-level demand modeling highly relevant to grid operations [1,2].

Facility demand cannot generally be represented by a fixed power usage effectiveness (PUE). PUE is the ratio of total facility power to information technology (IT) power and is no less than 1 under a consistent measurement boundary [3]. Non-IT overhead power is the difference between facility and IT power and includes cooling and auxiliary loads. A workload change modifies the IT power denominator and may also change cooling demand, whereas a cooling setpoint changes overhead power without directly changing IT power. A PUE value obtained under one workload and cooling schedule may therefore be invalid for another schedule, and a lower PUE does not necessarily imply lower facility energy. This study consequently treats PUE as a derived efficiency indicator rather than a control objective. Dispatch is based on overhead and facility power, which directly determine grid exchange and operating cost.

Overhead power depends nonlinearly on weather, IT utilization, supply air conditions, airflow, chiller staging, and equipment state. The thermal dynamics of the data hall and the cooling system also introduce delays. The EnergyPlus simulation platform can represent these interactions, but site calibration and computational burden limit repeated evaluation within rolling optimization [4]. Recirculation, fan power, and economizer performance also depend on airflow management, climate, and cooling plant configuration [5,6]. A forecasting model should therefore include both environmental variables and planned controls and should be applied only within the site and operating range represented by its data.

Data-driven studies have used neural networks [7], belief rule-based systems [8], gated recurrent units (GRUs) [9], other deep learning models [10], and bidirectional gated recurrent units (BiGRUs) [11] to predict PUE. Predictive capacity management has also been studied for hyperscale facilities [12]. These works establish that telemetry can represent nonlinear relationships between operating conditions and PUE. Most reported formulations, however, are evaluated under the observed operating policy. They do not explicitly estimate how overhead power changes when a dispatcher modifies workload allocation or cooling setpoints. Hence, a model can achieve low test error yet become inconsistent when it is embedded in an optimizer that deliberately changes those controls.

Scheduling also requires probabilistic forecasts over the complete decision horizon. Here, multi-horizon forecasting produces a prediction for every 10 min interval up to 24 h ahead. The information set separates historical telemetry, known future information, and candidate control trajectories. Future measured responses are excluded. The Temporal Fusion Transformer (TFT) combines variable selection, recurrent processing, and temporal attention for this purpose [13,14]. Quantile regression represents asymmetric uncertainty, and split conformal calibration improves empirical marginal coverage [15]. Because finite-sample conformal guarantees rely on exchangeability, coverage for chronological operating data is interpreted within the studied regime rather than under arbitrary distributional shifts. Conditional value at risk (CVaR) provides a corresponding measure of costly tail outcomes in dispatch [16].

A separate literature has established several sources of data center flexibility. Workload shifting and local generation can reduce coincident peak demand [17]. Reviews identify temporal and geographical workload scheduling, cooling control, thermal storage, and

batteries as principal resources [18]. Other studies coordinate demand with on-site renewable generation, microgrid operation, and multiple energy carriers [19–22]. Related work examines deadline-constrained workloads [23], islanded microgrid optimization [24], flexibility-based demand management [25], scheduling with wind generation [26], and robust or model predictive microgrid control [27,28].

The forecasting and dispatch strands remain only partially connected. PUE studies generally assess prediction accuracy without using the forecast in grid scheduling. Data center dispatch models, in contrast, commonly use a constant PUE, an exogenous facility load forecast, or a deterministic conversion from IT power to facility power. Such representations may be adequate under a fixed operating policy, but they become inconsistent when the optimizer changes the workload or cooling trajectory that determines overhead demand. The resulting research gap is a tractable interface that carries forecasts of facility response, their dependence on control decisions, and their uncertainty into microgrid dispatch. Related data center flexibility and microgrid MPC studies [18,27,28] likewise rely on site-specific response models or exogenous load forecasts. However, existing methods do not provide probabilistic forecasts of overhead power and rack temperature that account for control actions and can be directly used for microgrid dispatch.

Closing this gap requires more than accurate marginal quantiles. The quantiles must be converted into trajectories that preserve temporal and intervariable dependence. A multistage formulation must also impose nonanticipativity, so scenarios with the same revealed history receive the same decision. Embedding the full forecasting network would make repeated optimization difficult, while an unconstrained surrogate could expose unsupported controls. These requirements motivate the use of a scenario tree and a bounded piecewise-affine response map.

This paper examines whether a control-aware, multi-horizon PUE forecast can improve data center DSM and microgrid dispatch. The proposed approach predicts nonnegative overhead power quantiles and rack inlet temperature quantiles conditional on candidate workload and cooling schedules. Facility power and PUE are then obtained from the algebraic power balance. Empirical split conformal calibration and a dependence model generate scenarios for rolling dispatch. A trajectory-conditioned piecewise-affine map transfers the control response into a scenario-tree MPC without embedding the complete network. The controller coordinates deferrable workload, cooling, PV generation, BESS operation, and grid exchange under service, thermal, storage, and nonanticipativity constraints.

The main contributions of this paper are as follows:

- A control-aware multi-horizon PUE prediction model subject to the algebraic facility power balance is developed. It conditions nonnegative overhead-power quantiles and rack-temperature quantiles on candidate workload and cooling schedules and analytically derives facility power and PUE. The outputs satisfy the power balance and $PUE \geq 1$ without clipping.
- An information structure suitable for online operation separates historical telemetry, known future inputs, and candidate workload and cooling schedules. Empirical split conformal calibration improves marginal coverage within the studied regime, while a dependence model produces temporally coherent scenarios.
- A risk-informed rolling MPC coordinates workload scheduling, cooling, PV generation, BESS operation, and grid exchange. A bounded piecewise-affine response map preserves the effects of candidate control histories through a recursively updated thermal memory state, and scenario-tree nonanticipativity ensures implementable decisions.
- Forecasting and dispatch are evaluated jointly against persistence, matched-input neural, constrained-output, deterministic, and oracle (i.e., with perfect information) baselines. Ablation and sensitivity studies assess the power-balance output, empirical

calibration, scenario dependence, control coverage, forecast horizon, risk preference, response-map accuracy, and rolling computational feasibility.

Section 2 formulates the scheduling problem. Section 3 presents the forecasting and dispatch methods. Section 4 describes the simulation and results, and Section 5 concludes the paper.

2. Problem Formulation

This section formulates the rolling scheduling problem for a data center microgrid. The formulation identifies the control variables, resource constraints, causal information set, and uncertain facility responses that must be considered in the forecasting and dispatch framework.

2.1. Data Center Microgrid and Rolling Operation

Figure 1 shows the data center microgrid and rolling control architecture informed by the proposed PUE forecasting framework. The studied site comprises an IT hall, a chilled water cooling plant, auxiliary electrical equipment, a rooftop or adjacent PV system, a BESS, and a bidirectional grid connection. The controller is updated at intervals of $\Delta t = 10$ min when used in energy equations. At update k , it receives the L most recent measurements, weather and PV forecasts, newly admitted computing requests, the current workload queue, and the battery state of charge, where L is the historical window length. It computes a schedule over H intervals, where H is the scheduling horizon length, implements only the first control action, and repeats the procedure when measurements for the next interval become available. This receding horizon process is causal: measurements after k are excluded, and future inputs are limited to forecasts, commitments, and candidate controls available at update k .

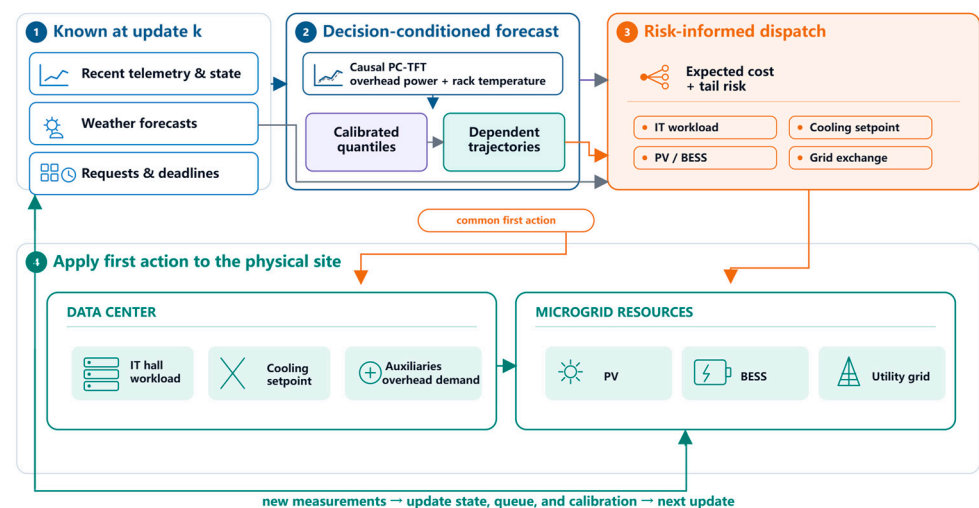


Figure 1. Data center microgrid and rolling control architecture informed by multi-horizon PUE forecasts.

Let k denote the dispatch interval index and h the lead time index. Define $\mathcal{T} = \{0, \dots, K\}$, $\mathcal{H} = \{0, \dots, H - 1\}$, $\mathcal{S}$, and $\mathcal{Q}$ as the study interval, prediction horizon, scenario, and quantile level sets, respectively; K is the final interval. Let superscript s denote a scenario unless otherwise stated. Within the adopted measurement boundary, facility demand is decomposed into IT power and non-IT overhead power. Overhead power includes chillers, pumps, fans, uninterruptible power supply losses, lighting, and other auxiliaries:

$$p_k^{\text{dc}} = p_k^{\text{it}} + p_k^{\text{aux}} \quad (1)$$

where P_k^{dc} , P_k^{it} , and P_k^{aux} are the facility, IT, and non-IT overhead powers in interval k , respectively. The scalar η_k denotes PUE and satisfies

$$\eta_k = \frac{P_k^{\text{dc}}}{P_k^{\text{it}}} = 1 + \frac{P_k^{\text{aux}}}{P_k^{\text{it}}}, \quad P_k^{\text{it}} > 0 \quad (2)$$

Equations (1) and (2) impose algebraic definitions of facility power and PUE. Predicting P_k^{aux} conditional on candidate controls and deriving η_k from (1) preserves consistency among IT, overhead, and facility power for each horizon and scenario. Thus, the optimizer uses overhead and facility power directly; PUE is a derived efficiency indicator because its denominator changes with the IT schedule.

2.2. Flexible Computing and Thermal Operation

For site-level scheduling, aggregate IT power is modeled as an affine function of IT utilization u_k :

$$P_k^{\text{it}} = P^{\text{idle}} + (P^{\text{rated}} - P^{\text{idle}})u_k, \quad 0 \leq u_k \leq 1 \quad (3)$$

where P^{idle} and P^{rated} are the aggregate idle and rated IT powers. Interactive requests are served immediately and impose a nondeferrable utilization floor. Flexible batch requests enter a queue and may be rescheduled within their delay limits. Let A_k , S_k , and B_k denote newly admitted work, served work, and the backlog at the beginning of interval k , respectively. The aggregate queue evolves as

$$B_{k+1} = B_k + A_k - S_k, \quad 0 \leq S_k \leq B_k + A_k \quad (4)$$

The work is measured in normalized server hours, and S_k maps linearly to the flexible component of u_k . For workload class i , let $A_{i,k}$ and $S_{i,k}$ denote arrival and service, and let d_i be the permitted service window length, including the arrival interval. An arrival at interval τ must be completed by $\tau + d_i - 1$. Therefore, for every $i \in \mathcal{I}^{\text{flex}}$ ($\mathcal{I}^{\text{flex}}$ is the set of flexible workload classes) and $k = d_i - 1, \dots, K$, cumulative service must satisfy

$$\sum_{\tau=0}^k S_{i,\tau} + V_{i,k} \geq \sum_{\tau=0}^{k-d_i+1} A_{i,\tau} \quad (5)$$

where τ is a summation index. $V_{i,k} \geq 0$ denotes the cumulative class- i service deficit at interval k ; V_k is the sum across flexible workload classes and is the slack penalized in Equation (22) and reported in the service-level agreement (SLA) metric. In a rolling operation, unserved work is carried forward as a carryover arrival with its remaining deadline. Capacity and nonnegativity constraints, together with any adopted ramp limits, complete the computing model. In the base case, deferrable work accounts for 25% of daily IT energy and may be delayed by at most 2 h. Within the prediction horizon, $A_{i,k+h}^s$ is fixed for requests already admitted at update k and is scenario-dependent for arrivals after k .

Cooling flexibility is represented by a bounded, ramp-limited supply air temperature setpoint. The independent control inputs are therefore flexible workload service and supply air temperature; airflow, valve positions, and equipment staging follow the embedded plant supervisory logic. Rack inlet temperature is imposed as an MPC scenario constraint. Rack inlet relative humidity is monitored as an operating condition rather than included as a PC-TFT output or MPC constraint, because humidification, dehumidification, and outdoor-air intake are handled by the internal building control logic. The forecast conditioned on control decisions predicts overhead power and rack inlet temperature. The constraint $\hat{T}_{k+h}^{\text{rack},s} \leq T^{\text{max}} - \varepsilon_T$ limits the predicted rack inlet temperature in scenario s at lead time h , where T^{max} is the adopted upper limit and the second term is the nonnegative response-

map safety margin defined in Section 3.5. Thermal limits are enforced directly and are not replaced by a PUE target.

2.3. PV, Storage, and Grid Exchange

Let E_k denote BESS stored energy, P_k^{ch} and P_k^{dis} the charging and discharging powers, and η^{ch} and η^{dis} the corresponding efficiencies. The energy state evolves as

$$E_{k+1} = E_k + \Delta t \left(\eta^{\text{ch}} P_k^{\text{ch}} - \frac{P_k^{\text{dis}}}{\eta^{\text{dis}}} \right) \quad (6)$$

The energy state satisfies $E^{\min} \leq E_k \leq E^{\max}$, and the charging and discharging powers are limited by the converter rating. Binary-mode variables prevent simultaneous charging and discharging in the mixed-integer linear program (MILP). The terminal constraint limits the end-of-horizon deviation from the stored energy at the forecast origin to $\pm 5\%$ of rated BESS energy, thereby limiting end-of-horizon energy borrowing.

Let $\bar{P}_k^{\text{pv}}$ denote available PV power and P_k^{curt} curtailed PV power. The utilized PV output is $P_k^{\text{pv,use}} = \bar{P}_k^{\text{pv}} - P_k^{\text{curt}}$. Using utilized PV power, the site balance is

$$P_k^{\text{imp}} + P_k^{\text{pv,use}} + P_k^{\text{dis}} = P_k^{\text{exp}} + P_k^{\text{it}} + P_k^{\text{aux}} + P_k^{\text{ch}} \quad (7)$$

where P_k^{imp} and P_k^{exp} are nonnegative grid import and export powers, respectively. Writing (7) in terms of $P_k^{\text{pv,use}}$ keeps available, curtailed, and utilized PV power distinct. The controller also carries the peak import already observed in the current accounting window as the state P_k^{hist} ; future decisions cannot reset this value.

2.4. Control-Aware Forecasting and Risk-Informed Scheduling

Overhead power and rack inlet temperature are uncertain because future weather, workload arrivals beyond requests already admitted at update k , and equipment response are imperfectly known. They are also dependent on control decisions: their predicted values change with the candidate IT service and cooling trajectories evaluated by the optimizer. Let x denote the collection of historical observations, z be exogenous forecasts and commitments available at update k , and v be candidate IT service and cooling controls. Define the response vector $\rho_{k+h}^s = [P_{k+h}^{\text{aux},s}, T_{k+h}^{\text{rack},s}]^T$, and let $q_{k,h}^s = [q_{k,h}^{P,s}, q_{k,h}^{T,s}]^T$ denote the separate copula-coordinate levels assigned to overhead power and rack temperature in scenario s at lead time h . The copula coordinates may take any value in $[0.1, 0.9]$ and are mapped by interpolation between adjacent trained quantile levels, as described in Section 3.4.

For $h \in \mathcal{H}$ and $s \in \mathcal{S}$, the forecast conditioned on control decisions is

$$\rho_{k+h}^s = \Phi_{\theta} \left(x_{k-L+1:k}, z_{k:k+H-1}, v_{k:k+H-1}; q_{k,h}^s \right) \quad (8)$$

The nonlinear map Φ_{θ} , parameterized by the learned vector θ , produces the predicted response at each lead time. Its inputs are restricted to historical observations, exogenous information available at update k , and explicitly supplied candidate controls. Let y be the full dispatch decision vector, of which v is the subvector of candidate controls; the remaining components describe workload service, storage, and grid exchange. Let $J(y)$ be the scenario-specific total cost, with realization $J_s(y)$ and probability π_s in scenario s . The scheduling objective minimizes expected cost and penalizes the upper tail:

$$\min_y \sum_{s \in \mathcal{S}} \pi_s J_s(y) + \lambda_r \text{CVaR}_{\alpha}(J(y)) \quad (9)$$

where the nonnegative scalar λ_r controls risk aversion. At confidence level α , the CVaR of $J(\mathbf{y})$ is represented by the auxiliary threshold ξ :

$$\text{CVaR}_\alpha(J(\mathbf{y})) = \min_{\xi} \left\{ \xi + \frac{1}{1-\alpha} \sum_{s \in \mathcal{S}} \pi_s [J_s(\mathbf{y}) - \xi]_+ \right\} \quad (10)$$

Scenario probabilities satisfy $\pi_s \geq 0$ and $\sum_{s \in \mathcal{S}} \pi_s = 1$. Each scenario cost includes imported energy, export revenue, battery degradation, PV curtailment, a daily peak import penalty, and service violation penalties.

The integrated problem is therefore to embed a forecast that depends on control decisions and satisfies the facility power balance within a tractable rolling dispatch model. Candidate controls must remain inside the range supported by training data; marginal quantiles must be converted into coherent trajectories; scenario decisions must satisfy nonanticipativity; and every facility power prediction must satisfy (1). Section 3 develops these components.

3. Proposed Methodology

3.1. Overall Framework and Causal Data Construction

The framework comprises six interconnected modules: causal data construction, multi-horizon forecasting with a power-balance output layer, split conformal calibration, dependent scenario generation, a piecewise-affine control response map, and risk-informed rolling optimization. The forecaster estimates the overhead power and rack inlet temperature for each candidate IT and cooling trajectory. Facility power and PUE are then obtained from the power balance. At each 10 min update, the controller implements the common first action and repeats the procedure after new measurements become available.

The historical window has $L = 144$ intervals, corresponding to 24 h. It contains measured IT and facility power, component-level cooling power, outdoor and indoor conditions, supervisory setpoints, valve and fan states, time encodings, and lagged forecast error statistics. A representative observed vector is

$$\mathbf{x}_k^{\text{obs}} = \left[P_k^{\text{pit}}, P_k^{\text{dc}}, T_k^{\text{out}}, \phi_k^{\text{out}}, T_k^{\text{sup}}, m_k, \dots \right]^T. \quad (11)$$

In Equation (11), T_k^{out} is outdoor dry bulb temperature, ϕ_k^{out} is outdoor relative humidity, T_k^{sup} is supply air temperature, and m_k is air mass flow rate. The ellipsis denotes additional channels retained after data quality screening. Future exogenous inputs comprise weather forecasts, calendar variables, and interactive and batch requests already admitted at update k . Work arriving after k is represented by workload forecasts and scenario residuals in the dispatcher. The candidate controls are planned IT utilization and the supply air temperature setpoint. PV forecasts and electricity prices are entered directly into the dispatcher because they do not affect the modeled thermal response. This separation keeps the physical controls distinct from the recursively updated thermal memory state and prevents future measurements from entering the forecast inputs.

Chronological data splits are used because adjacent building measurements are strongly correlated. Missing blocks shorter than 30 min are interpolated using only information available at the forecast origin; longer blocks are excluded. Continuous inputs are standardized with training statistics, and no forecast window crosses a training, calibration, or test boundary. Prediction under alternative controls also requires deliberate excitation. Airflow and equipment staging remain outputs of the plant supervisory logic. The chronological test period and the independent validation combinations are excluded

from the excitation data used for fitting. The dispatch envelope is then restricted to control combinations covered by the excitation campaign.

3.2. Multi-Horizon Temporal Fusion Model with an Explicit Power Balance

Figure 2 shows the information flow of the proposed temporal fusion model. The forecaster retains the principal components of the standard TFT: variable selection networks, gated residual networks, a recurrent encoder and decoder for local temporal order, interpretable multihead attention for longer dependencies, and output heads for each lead time [13,14]. The encoder processes only historical observations. The decoder receives exogenous future inputs and candidate controls as separate groups. No future measured response is supplied, and no bidirectional recurrent layer is used. Thus, the information flow is consistent with the online input set defined in Section 2.4. Causal masking also prevents the output at lead time h from using candidate controls after h .

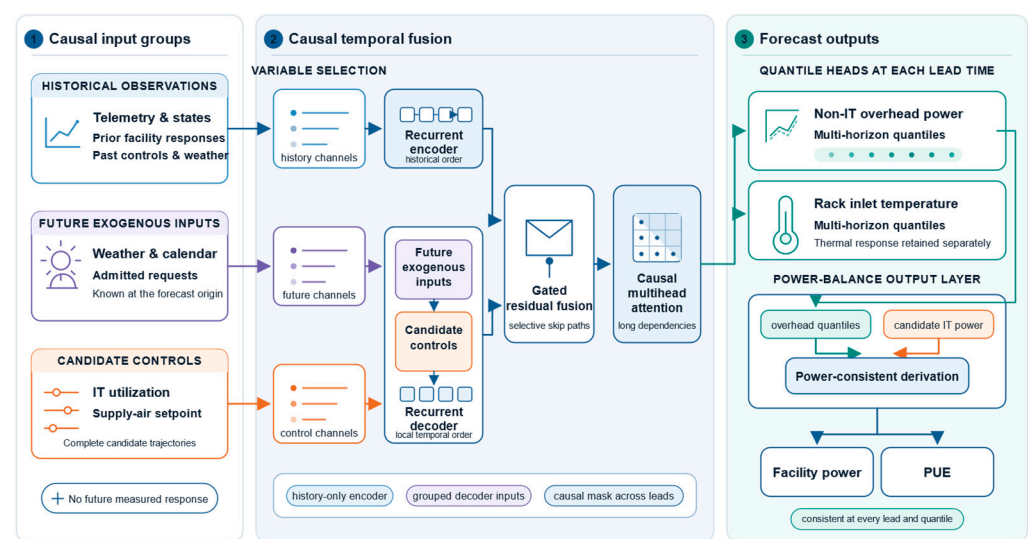


Figure 2. Information flow of the temporal fusion model. The network predicts quantiles of non-IT overhead power and rack inlet temperature; the power balance output layer derives facility power and PUE.

For each lead time h and quantile level q , the output head produces scalar activations $a_{k+h,q}$ and $b_{k+h,q}$. The first is mapped to nonnegative overhead power, while the second gives the rack temperature quantile. Facility power and PUE are derived from the candidate IT power:

$$\begin{aligned}\hat{P}_{k+h,q}^{\text{aux}} &= \text{softplus} \left(a_{k+h,q} \right) = \ln \left(1 + e^{a_{k+h,q}} \right) \geq 0, \\ \hat{T}_{k+h,q}^{\text{rack}} &= b_{k+h,q}, \\ \hat{P}_{k+h,q}^{\text{dc}} &= P_{k+h}^{\text{it}} + \hat{P}_{k+h,q}^{\text{aux}}, \\ \hat{\eta}_{k+h,q} &= \frac{\hat{P}_{k+h,q}^{\text{dc}}}{P_{k+h}^{\text{it}}} = 1 + \frac{\hat{P}_{k+h,q}^{\text{aux}}}{P_{k+h}^{\text{it}}} \geq 1, \quad P_{k+h}^{\text{it}} > 0.\end{aligned}\quad (12)$$

Equation (3) places a strictly positive lower bound on IT power through the idle power. Equation (12) therefore requires no denominator clipping and satisfies the facility power balance at every lead time and quantile. The algebraic output constraint does not replace thermal modeling; rack temperature and equipment limits are enforced separately.

The quantile set is $\mathcal{Q} = \{0.1, 0.2, \dots, 0.9\}$. For a scalar target y , prediction $\hat{y}$, and quantile level q , the pinball loss is

$$\uparrow_q(y, \hat{y}) = (q - \mathbb{I}\{y < \hat{y}\})(y - \hat{y}), \quad (13)$$

where $I\{\cdot\}$ is the indicator function. Let N_{tr} be the number of training forecast origins. The training objective combines the pinball losses for overhead power and rack temperature with an overhead power first difference loss:

$$\begin{aligned} L &= L_P + \lambda_T L_T + \lambda_\Delta L_\Delta, \\ L_P &= \frac{1}{N_{\text{tr}} H |\mathcal{Q}|} \sum_{n=1}^{N_{\text{tr}}} \sum_{h=0}^{H-1} \sum_{q \in \mathcal{Q}} \downarrow_q \left(p_{n,h}^{\text{aux}}, \hat{p}_{n,h,q}^{\text{aux}} \right), \\ L_T &= \frac{1}{N_{\text{tr}} H |\mathcal{Q}|} \sum_{n=1}^{N_{\text{tr}}} \sum_{h=0}^{H-1} \sum_{q \in \mathcal{Q}} \downarrow_q \left(T_{n,h}^{\text{rack}}, \hat{T}_{n,h,q}^{\text{rack}} \right), \\ \Delta p_{n,h}^{\text{aux}} &= p_{n,h}^{\text{aux}} - p_{n,h-1}^{\text{aux}}, \quad \Delta \hat{p}_{n,h,q}^{\text{aux}} = \hat{p}_{n,h,q}^{\text{aux}} - \hat{p}_{n,h-1,q}^{\text{aux}}, \\ L_\Delta &= \frac{1}{N_{\text{tr}} (H-1) |\mathcal{Q}|} \sum_{n=1}^{N_{\text{tr}}} \sum_{h=1}^{H-1} \sum_{q \in \mathcal{Q}} \left| \Delta \hat{p}_{n,h,q}^{\text{aux}} - \Delta p_{n,h}^{\text{aux}} \right|. \end{aligned} \quad (14)$$

The nonnegative weights λ_T and λ_Δ are selected from the validation set using prediction loss and interval coverage. A monotone rearrangement is applied across q before calibration and dispatch to remove residual quantile crossings. Training uses Adam optimization, early stopping, gradient clipping, and dropout.

Variable selection weights and temporal attention scores are retained as diagnostic quantities, not as estimates of causal effects. Persistent weight on outdoor conditions, recent overhead power, the IT schedule, and the supply air setpoint is physically plausible. A marked increase in the weight of unrelated channels is treated only as a possible indicator of data or regime drift. These diagnostic weights indicate model reliance on inputs rather than causality and should be used only to flag unexpected signals for further verification of data quality, system status, and underlying causes before any control action.

3.3. Empirical Split Conformal Calibration

Quantile regression does not by itself ensure nominal interval coverage. The trained forecaster is therefore frozen and evaluated on a chronological calibration block. For target coverage $1 - \beta$, let $q_l = \beta/2$ and $q_u = 1 - \beta/2$. For calibration sample i , the nonconformity score for overhead power is

$$r_i = \max \left\{ \hat{p}_{i,q_l}^{\text{aux}} - p_i^{\text{aux}}, p_i^{\text{aux}} - \hat{p}_{i,q_u}^{\text{aux}} \right\}. \quad (15)$$

Let N_{cal} be the number of scores in the relevant lead time band. The finite sample correction and calibrated nonnegative interval are

$$\begin{aligned} \gamma &= \min \left\{ 1, \frac{\text{ceil}((N_{\text{cal}} + 1)(1 - \beta))}{N_{\text{cal}}} \right\}, \\ \hat{r}_{1-\beta} &= \text{Quantile}_\gamma \left(\{r_i\}_{i=1}^{N_{\text{cal}}} \right), \\ \tilde{p}_{q_l}^{\text{aux}} &= \max \left\{ 0, \hat{p}_{q_l}^{\text{aux}} - \hat{r}_{1-\beta} \right\}, \\ \tilde{p}_{q_u}^{\text{aux}} &= \hat{p}_{q_u}^{\text{aux}} + \hat{r}_{1-\beta}. \end{aligned} \quad (16)$$

Equations (15) and (16) are shown for overhead power. The same nonconformity-score construction is applied independently to rack inlet temperature, without the nonnegativity truncation used for the lower overhead-power bound. Separate target- and lead-band-specific corrections are estimated for the 0–1, 1–4, 4–12, and 12–24 h bands. The overhead-power correction is propagated through Equation (12), ensuring that the calibrated facility-power and PUE intervals preserve power-balance consistency and $\eta \geq 1$; the independently

calibrated rack-temperature interval is used in the thermal constraints. During rolling operation, the corrections are recomputed at each 10 min interval using bounded calibration windows updated with newly verified observations, while the neural network parameters remain fixed. This rolling update allows the calibration to gradually track seasonal changes.

Split conformal prediction provides finite-sample marginal coverage under exchangeability [15]. Because chronological operating data are only approximately exchangeable within a stable regime, the calibrated intervals are treated as empirical rather than distribution-free guarantees for dependent time series. Accordingly, coverage is monitored across seasons and operating regimes, and the nominal level should be interpreted as applying to conditions adequately represented by the calibration data. Flash crowds that cause substantial distribution shifts may reduce prediction interval coverage probability (PICP). Persistent undercoverage requires rebuilding the calibration window using recent regime-specific data; if the underlying input–response relationship has changed, the forecasting model must also be updated.

3.4. Temporally Dependent Scenario Tree Construction

Marginal intervals do not define dependence across lead times or uncertain variables. The calibrated quantile curves are therefore converted into a finite scenario tree. After monotone rearrangement, the overhead power curve is sampled at the interior levels

$$q_j = \frac{j}{|Q| + 1}, \quad j = 1, \dots, |Q|. \quad (17)$$

With $|Q| = 9$, (17) gives the trained quantile levels from 0.1 to 0.9; the number of retained scenario leaves is selected independently during subsequent scenario reduction. Residuals of overhead power, rack inlet temperature, PV power, weather variables, and forecast arrivals beyond admitted commitments are arranged as chronological 24 h blocks. Their empirical ranks are transformed to Gaussian scores. The covariance matrix is estimated using Ledoit–Wolf shrinkage, with its minimum eigenvalue set to 10^{-6} to ensure positive definiteness. A Gaussian copula then generates continuous dependent coordinates. Coordinates within $[0.1, 0.9]$ are mapped by piecewise-linear interpolation between adjacent trained quantiles; lower and upper tail coordinates are capped at the conformally adjusted 0.1 and 0.9 endpoints, respectively, so unsupported tail extrapolation is avoided. The other variables are mapped through their empirical residual distributions. We first generate 1000 trajectories using random seed 2026. Similar standardized trajectories are then grouped using stagewise k-medoids with weighted Euclidean distance, and representative scenarios are retained through forward reduction. Each retained scenario is assigned the proportion of original trajectories it represents. This construction preserves estimated temporal and intervariable dependence but does not convert marginal calibration into simultaneous trajectory coverage. For each scenario s , the Gaussian-copula coordinates for overhead power and rack temperature are denoted $q_{k,h}^{P,s}$ and $q_{k,h}^{T,s}$, respectively. They are generally unequal; intervariable dependence is imposed by their joint copula rather than by forcing identical marginal ranks.

The generated trajectories are clustered at the 1, 4, 12, and 24 h branching stages. Each reduced leaf s is assigned frequency π_s . Let $n_h(s)$ denote the clustered node occupied by leaf s immediately before the decision at lead time h . The node probability is $p_n = \sum_{s: n_h(s)=n} \pi_s$, and a child n' has conditional probability $p_{n'|n} = p_{n'}/p_n$. Implementable scenario decisions satisfy

$$y_{k+h}^s = y_{k+h}^{s'} \quad \text{if} \quad n_h(s) = n_h(s'), \quad s, s' \in \mathcal{S}. \quad (18)$$

Equation (18) imposes equality through the clustered node labels, rather than equality of continuous uncertainty trajectories. Nine leaves are a computational design choice rather than a coverage guarantee; sensitivity to the number of scenarios is examined separately. The nine-leaf tree approximates uncertainty within the calibrated empirical distribution, and therefore the CVaR results are conditional on the adopted scenario model. It does not guarantee protection against unmodeled extreme events; specified extremes must be represented explicitly as stress scenarios or addressed through deterministic reserve constraints.

3.5. Piecewise-Affine Control Response Map

The piecewise-affine map retains the dependence of facility response on the control path without embedding the full TFT in the optimizer. For each forecast origin and exogenous scenario, the frozen forecaster is evaluated on randomly sampled, ramp-feasible, complete 24 h candidate control trajectories rather than independent control points. Every trajectory combines the measured control history before k , a feasible candidate prefix $v_{k:k+h}$, and a feasible continuation to 24 h. Because the decoder is causally masked, the response at lead time h depends on the measured history and candidate prefix but not on controls after h . Figure 3 illustrates a two-dimensional slice of the resulting response map at a fixed thermal memory state.

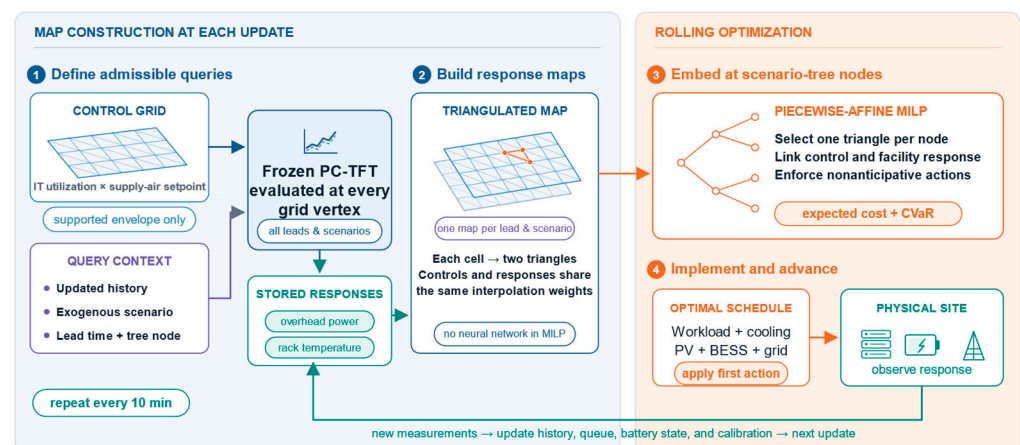


Figure 3. Construction and rolling application of the piecewise-affine control response map.

To transfer the prefix dependence into the MILP compactly, $\chi_{k+h}^s \in \mathbb{R}^2$ stores exponentially weighted histories of normalized IT utilization and supply air temperature. It is initialized from the last L measured controls and updated by the first relation in (19), where $A_\chi = \text{diag}(a_u, a_T)$, $0 \leq a_u, a_T < 1$, is selected on the chronological validation block. The matrix I in Equation (19) is the 2×2 identity matrix. The augmented four-dimensional coordinate $\tilde{v}_{k+h}^s$ combines the current two controls and the two memory states. Each vertex stores this augmented coordinate and the PC-TFT response obtained from a complete trajectory whose prefix produces the corresponding memory state. Let $\mathcal{C}$ be the set of candidate simplices and $\mathcal{M}_c$ the vertices of simplex c . The binary variable $z_{k,h,c}^s$ selects one simplex, and $\lambda_{k,h,c,m}^s$ is the nonnegative weight assigned to vertex m . The weights satisfy $\sum_{m \in \mathcal{M}_c} \lambda_{k,h,c,m}^s = z_{k,h,c}^s$ and $\sum_{c \in \mathcal{C}} z_{k,h,c}^s = 1$. The map is

$$\chi_{k+h+1}^s = A_\chi \chi_{k+h}^s + (I - A_\chi) v_{k+h}^s \quad (19)$$

$$\tilde{v}_{k+h}^s = \left[(v_{k+h}^s)^T, (\chi_{k+h}^s)^T \right]^T, \quad (20)$$

$$\begin{bmatrix} \tilde{v}_{k+h}^s \\ \rho_{k+h}^s \end{bmatrix} = \sum_{c \in \mathcal{C}} \sum_{m \in \mathcal{M}_c} \lambda_{k,h,c,m}^s \begin{bmatrix} \tilde{v}_{c,m} \\ \rho_{k,h,c,m}^s \end{bmatrix}. \quad (21)$$

Because (19) interpolates the augmented coordinate and response using identical weights while χ_{k+h}^s is recursively constrained, identical current IT utilization and supply air temperature reached through different histories can yield different responses. Simplex selection gives a piecewise-affine MILP formulation. The overhead power component enters the site balance in (7), while the rack inlet temperature is subject to the thermal limit in every scenario. The response map covers IT utilization from 45% to 100% in 5% increments and supply air temperature from 18 to 27 °C in 1 °C increments; its memory coordinates and admissible simplices are limited to those obtained from the feasible-trajectory campaign. These restrictions enforce the verified excitation envelope as a hard constraint in the MILP and prohibit extrapolation. As (19) uses a one-hot representation, it introduces many binary variables. This can be reduced by an equivalent logarithmic simplicial encoding and by sharing variables at common scenario-tree nodes, without changing the feasible mapping. If an emergency request cannot be fully met within this envelope, the controller uses only the admissible in-envelope actions and available microgrid resources. Any remaining shortfall is reported as unmet flexibility, and unsupported control schedules are neither evaluated nor implemented.

At each scenario-tree node, the response-map estimate of rack inlet temperature is constrained below the upper limit with a safety margin for response-map approximation error. The scenario tree represents forecast uncertainty, whereas the safety margin accounts for response-map approximation error. In this case study, the margin is 0.146 °C, equal to the maximum absolute temperature interpolation error obtained from the independent response-map tests.

3.6. Risk-Informed Rolling Microgrid Dispatch

The scenario-tree MPC combines the expected cost and CVaR objectives in Equations (9) and (10) with the response map and the resource constraints in Section 2. For scenario s , the power operating cost rate and total scenario cost are

$$\begin{aligned} C_{k+h}^s &= c_{k+h}^{\text{buy}} p_{k+h}^{\text{imp},s} - c_{k+h}^{\text{sell}} p_{k+h}^{\text{exp},s} + c^{\text{deg}} \left(p_{k+h}^{\text{ch},s} + p_{k+h}^{\text{dis},s} \right) + c^{\text{curt}} p_{k+h}^{\text{curt},s}, \\ J^s &= \sum_{h=0}^{H-1} \left(\Delta t C_{k+h}^s + c^{\text{sla}} V_{k+h}^s \right) + c^{\text{pk}} p^{\text{max},s}. \end{aligned} \quad (22)$$

The coefficients c^{buy} , c^{sell} , c^{deg} , and c^{curt} value grid purchases, exports, battery throughput, and PV curtailment, respectively. Their energy units are converted to CNY/MWh before optimization, so multiplication by power and Δt gives an interval cost. The service coefficient c^{sla} is expressed per server hour because V_{k+h}^s is already measured in server hours. The peak coefficient c^{pk} is expressed in CNY/(MW·day) and is used to value DSM; it is not presented as a verified Shenzhen utility demand tariff.

The daily peak variable retains the maximum import observed before the current update and bounds grid import in every future interval and scenario:

$$p^{\text{max},s} \geq p_k^{\text{hist}}, \quad p^{\text{max},s} \geq p_{k+h}^{\text{imp},s}, \quad h \in \mathcal{H}, s \in \mathcal{S}. \quad (23)$$

The historical peak state in Equation (23) is a reference for calculating the daily peak rather than a hard import limit. Exceeding it remains feasible but increases the peak demand penalty. If a contractual import limit is imposed, a safety buffer that only accounts for underestimation errors should be applied to the predicted import constraint.

After the scenario costs are defined, Equation (10) introduces the linear CVaR auxiliaries, and Equation (18) enforces nonanticipativity. The base case uses $\lambda_r = 0.25$ and $\alpha = 0.9$, selected from a validation frontier that balances expected cost and service violations. Setting $\lambda_r = 0$ gives expected cost stochastic MPC, whereas retaining only the median trajectory gives the deterministic baseline.

At each 10 min update, the controller validates and appends measurements; updates the workload queue, stored battery energy, and observed peak; and receives forecasts of weather, PV, price, and future workload arrivals together with admitted requests. It then updates the empirical calibration correction, constructs the scenario tree, evaluates the response grid within the supported envelope, and solves the nonanticipative MILP. The common first action is implemented. If required inputs are unavailable or the solver fails, a fallback policy prioritizes urgent work, maintains a verified safe thermal setpoint, and dispatches the battery based on its state of charge and the tariff period. When operating conditions fall outside the verified envelope, the controller switches to a fallback policy. This policy prioritizes critical loads, keeps cooling within safe setpoints, and enforces battery state of charge (SOC) and power limits, while supervisory protection systems retain ultimate control.

4. Results and Discussion

4.1. Simulation Setup

The case study combines an EnergyPlus model of the data center building and cooling plant with simulations of workload arrivals, microgrid assets, tariff accounting, and rolling optimization. Shenzhen weather data drive the EnergyPlus model. The annual baseline contains 52,560 observations at 10 min resolution and 117 raw variables. A separate bounded excitation campaign uses randomly sampled, ramp-feasible 24 h IT-utilization and cooling-setpoint trajectories, including step, ramp, and reversal segments, to learn responses under alternative control histories. Facility and IT power are aggregated according to the electrical boundary in (1). In the case study, rack inlet temperature is limited to 27 °C, while rack inlet relative humidity is monitored against the 20–80% operating range; changes in the supply air temperature setpoint are limited to 1 °C per 10 min interval. The baseline PUE ranges from 1.17 to 1.43, with a mean of 1.29.

At each 10 min update, the MPC sets the flexible-workload service level and supply air temperature setpoint. EnergyPlus then uses its internal plant controllers to adjust air mass flow and chilled-water valve positions and to stage pumps and chillers in response to cooling demand, setpoint tracking, and equipment limits. These lower-level variables affect cooling power and rack inlet temperature but are internal plant responses rather than MPC decision variables or known future inputs. Their realized values are included in the historical input window at the next control update.

The simulated annual baseline spans 1 January 00:00 to 31 December 23:50. The chronological blocks are: training, 1 January–7 August (60%); hyperparameter selection, 8 August 00:00–13 September 11:50 (10%); conformal calibration, 13 September 12:00–19 October 23:50 (10%); and forward testing, 20 October–31 December (20%). No history or 24 h forecast window crosses a boundary. Excitation runs augment only the training block, and independent time-varying 24 h control trajectories are reserved for response-map testing. The rolling 30-day summer test (15 July–13 August) and seasonal day sets (15–17 January, 15–17 April, and 1–3 July) are separate EnergyPlus runs with independent workload and forecast error realizations. Their targets are excluded from model training, model selection, and initial offline calibration. During rolling evaluation, recalibration at forecast origin k uses only observations that have been verified and become available by k ; no future target is used.

The microgrid includes a 2.5 MW IT load, a 1.5 MW PV plant, and a 2.0 MWh/1.0 MW BESS. The BESS has charge and discharge efficiencies of 95% and operates within a SOC range of 10–90%. Valley, shoulder, and peak purchase prices are 0.38, 0.72, and 1.20 Chinese yuan (CNY)/kWh, respectively. Exported energy is valued at 0.28 CNY/kWh, and battery throughput is charged at 0.08 CNY/kWh as a simplified degradation proxy. The SLA penalty and daily peak import coefficient are set to 500 CNY per server-hour and 1.07 CNY/(kW·day), respectively. Flexible batch classes account for 25% of IT energy and have a 2 h deadline. Table 1 summarizes the key parameter settings.

Table 1. EnergyPlus data and microgrid parameters used in the case study.

Category	Parameter	Value
Data	Sampling interval/samples/raw variables	10 min/52,560/117
Forecast	History/horizon/quantiles	24 h/24 h/0.1–0.9
IT system	Rated/idle IT power	2.50/1.05 MW
Flexibility	Flexible share/deadline	25%/2 h
PV	Rated output	1.50 MW
BESS	Energy/power rating	2.0 MWh/1.0 MW
BESS	Efficiency/SOC limits	0.95/0.10–0.90
Tariff	TOU price/daily peak import penalty	0.38/0.72/1.20 CNY/kWh; 1.07 CNY/kW·day
Risk	CVaR confidence/risk weight	0.90/0.25

Asset ratings and tariffs are case study assumptions. The peak import coefficient represents a DSM penalty rather than a verified Shenzhen utility tariff.

The forecasting comparison includes daily seasonal persistence, GRU, BiGRU, and a standard TFT. Persistence repeats the most recent PUE profile at the corresponding time of day. All trainable baselines receive the same historical observations, future exogenous forecasts, and candidate workload and cooling controls available to PC-TFT; model sizes are matched within 5%. GRU and BiGRU produce deterministic forecasts for all lead times, while the standard TFT directly predicts PUE quantiles. A constrained direct-PUE TFT using $1 + \text{softplus}(\cdot)$ and an overhead-power TFT without the rack-temperature task are included in the ablation study. The proposed PC-TFT predicts the quantiles of overhead power and rack temperature, and then derives facility power and PUE. Dispatch is compared across no flexibility, deterministic MPC with constant PUE, deterministic MPC with persistence forecasts, PC-TFT median MPC, the proposed scenario-tree MPC, and a perfect-information oracle. The oracle receives the realized weather, PV output, workload arrivals, overhead power, and rack inlet temperature over the 24 h horizon but uses the same resource, service, thermal, and terminal constraints as the implementable controllers. Every learned model is queried only within the common excitation envelope. The hyperparameter setting of each model is presented in Table 2.

Let N_f denote the number of forecast origins and H the number of evaluated lead times. For forecast origin n and lead time h , $\eta_{n,h}$ and $\hat{\eta}_{n,h,0.5}$ denote the observed PUE and predicted median PUE, respectively. Point forecasts are evaluated using mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and the coefficient of determination. MAE and RMSE are defined as follows:

$$\text{MAE} = \frac{1}{N_f H} \sum_{n=1}^{N_f} \sum_{h=1}^H |\eta_{n,h} - \hat{\eta}_{n,h,0.5}|, \quad (24)$$

$$\text{RMSE} = \sqrt{\frac{1}{N_f H} \sum_{n=1}^{N_f} \sum_{h=1}^H (\eta_{n,h} - \hat{\eta}_{n,h,0.5})^2}. \quad (25)$$

Table 2. Forecasting models and selected hyperparameters.

Model	Inputs	Output	Parameters/Settings
Persistence	Last profile	Point PUE	Daily seasonal persistence
GRU	Past + exogenous forecasts + candidate controls	24 h point PUE	2×96 hidden units
BiGRU	Past + exogenous forecasts + candidate controls	24 h point PUE	2×64 per direction
TFT	Past + exogenous forecasts + candidate controls	PUE quantiles	128 hidden, 4 heads, dropout 0.15
Constrained PUE-TFT	Past + exogenous forecasts + candidate controls	PUE quantiles through $1 + \text{softplus}(\cdot)$	128 hidden, 4 heads, dropout 0.15
Proposed PC-TFT	Past + exogenous forecasts + candidate controls	Overhead-power and rack-temperature quantiles + derived PUE	128 hidden, 4 heads, dropout 0.15

All trainable models use Adam, gradient clipping at 1.0, and early stopping. Hyperparameters are selected using the chronological validation block.

MAE is used as the primary evaluation metric because it directly reflects the average prediction error in PUE values. In addition, the relative MAE reduction over the persistence model is reported to quantify the performance improvement in percentage terms.

Probabilistic forecasts are evaluated using pinball loss, PICP, and prediction interval normalized average width (PINAW). Let $[\hat{\eta}_{n,h}^l, \hat{\eta}_{n,h}^u]$ denote the lower and upper bounds of the predicted PUE interval and R_η the observed PUE range in the test set. The indicator in Equation (26) equals one when the observation lies within the predicted interval and zero otherwise:

$$\text{PICP} = \frac{1}{N_t H} \sum_{n=1}^{N_t} \sum_{h=1}^H \mathbb{I}\{\eta_{n,h} \in [\hat{\eta}_{n,h}^l, \hat{\eta}_{n,h}^u]\}, \quad (26)$$

$$\text{PINAW} = \frac{1}{N_t H R_\eta} \sum_{n=1}^{N_t} \sum_{h=1}^H (\hat{\eta}_{n,h}^u - \hat{\eta}_{n,h}^l), \quad R_\eta = \eta_{\max} - \eta_{\min}. \quad (27)$$

The nominal interval coverage is 80%. For stratified reliability checks, the Winkler interval score is normalized by the observed PUE range. Dispatch is evaluated based on daily operating costs, peak grid imports, PV utilization, BESS throughput, and SLA violations. SLA violations (%) denote deadline-expired service slack divided by total admitted flexible work. Every controller is tested ex post with the same workload arrivals and realized weather. Confidence intervals for daily performance differences are estimated using 1000 moving-block bootstrap replications with 24 h blocks to account for overlapping forecast windows.

4.2. Multi-Horizon Forecast Performance

Table 3 shows that the proposed PC-TFT achieves the lowest forecasting errors. Its PUE MAE is 0.0068, which is 17.1% lower than the standard TFT baseline, which directly predicts PUE using the same inputs, and 61.4% lower than the persistence baseline. Its RMSE is 0.0100. The MAEs of GRU and BiGRU are 0.0108 and 0.0099, respectively. All models use the same chronological test origins and information available at each origin; none receives future measured cooling or workload responses.

Table 3. Multi-horizon PUE forecasting performance on the chronological test set.

Model	MAE	RMSE	MAPE (%)	R ²	Relative MAE Reduction (%)
Persistence	0.0176	0.0248	1.36	0.832	—
GRU	0.0108	0.0152	0.84	0.934	38.6
BiGRU	0.0099	0.0142	0.77	0.943	43.8
TFT	0.0082	0.0119	0.64	0.960	53.4
Proposed PC-TFT	0.0068	0.0100	0.53	0.972	61.4

Relative MAE reduction is calculated against persistence. Bold entries indicate the best values.

The representative 48 h segment in Figure 4 shows that the PC-TFT median tracks both the daytime PUE increase and the overnight decline, without the phase lag associated with persistence. The calibrated interval widens with lead time and during afternoon transitions, when outdoor enthalpy and IT demand are high. The remaining misses cluster around rapid changes in equipment staging. Explicit chiller state inputs or regime-specific output heads may therefore improve calibration.

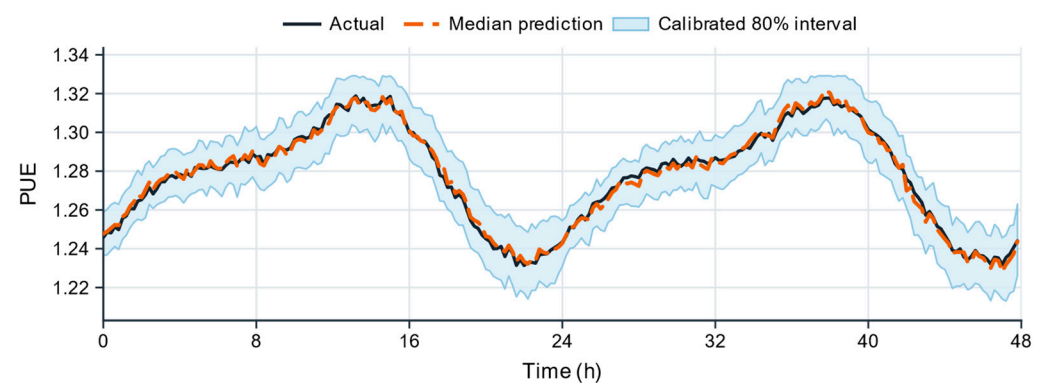
**Figure 4.** Representative PUE trajectory, PC-TFT median prediction, and calibrated 80% interval over 48 h.

Figure 5 shows that the advantage of PC-TFT increases with the forecast horizon. During the first hour, all learned models benefit from recent overhead power measurements. Beyond 4 h, weather forecasts, the planned IT schedule, and temporal attention become more influential. At 24 h, the MAEs are 0.0520 for persistence, 0.0310 for GRU, 0.0220 for TFT, and 0.0161 for PC-TFT. The improvement therefore persists through the final interval of the dispatch horizon.

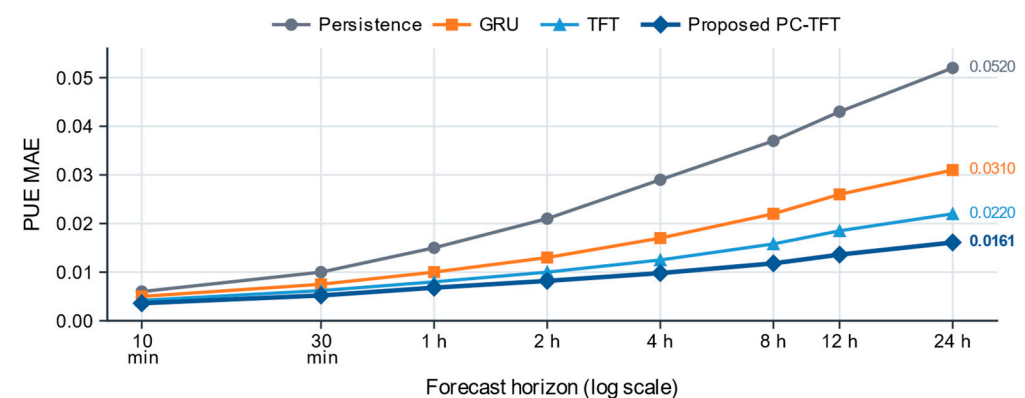
**Figure 5.** PUE MAE versus forecast horizon for the principal forecasting baselines.

Table 4 shows the controlled ablation of future control inputs and the native-output forecast performance. In the ablated model, the planned IT-utilization and supply-air-temperature trajectories are removed from the PC-TFT inputs; candidate IT power is retained only in the algebraic calculation of facility power and PUE. All other settings remain unchanged. Incorporating the future control inputs reduces the MAEs of PUE, overhead power, and rack inlet temperature by 24.4%, 34.6%, and 33.9%, respectively.

Table 4. Effect of future control inputs and native-output forecast performance.

Target	MAE Without Future Controls	MAE with Future Controls	MAE Reduction (%)	Proposed RMSE	Proposed 80% PICP (%)
PUE	0.0090	0.0068	24.4	0.0100	81.3
Overhead power (kW)	16.2	10.6	34.6	15.7	81.1
Rack inlet temperature (°C)	0.112	0.074	33.9	0.108	80.9

Both models use the same architecture, capacity, training protocol, and chronological test origins; only the planned future control trajectories are removed. Native-output metrics are evaluated on the chronological test set.

4.3. Power Consistency, Calibration, and Ablation

The ablation results in Table 5 distinguish the roles of the proposed components. The action-conditioned direct-PUE TFT produces 1.8% invalid values near the low-IT boundary. Applying $1 + \text{softplus}(\cdot)$ also eliminates PUE values below 1, confirming that this inequality alone is not unique to the power-balance output. However, its counterfactual facility power MAE remains 32.8 kW. Predicting overhead power through the algebraic balance reduces this error to 26.7 kW, and adding the rack-temperature multitask quantile loss further reduces PUE MAE to 0.0068 and facility power MAE to 25.9 kW. Thus, the principal advantages are the consistent facility power decomposition and improved response accuracy, rather than merely enforcing $\text{PUE} \geq 1$. All configurations use identical forecast origins, inputs, model capacity, and control bounds.

Table 5. Ablation and interval calibration results.

Configuration	MAE	80% PICP (%)	PINAW	Invalid PUE (%)	Counterfactual Facility-Power MAE (kW)
Action-conditioned direct PUE TFT	0.0082	74.9	0.084	1.8	34.1
Constrained direct PUE TFT, $1 + \text{softplus}(\cdot)$	0.0080	75.1	0.085	0	32.8
Overhead-power TFT (single task)	0.0074	75.6	0.085	0	26.7
+ rack-temperature multitask quantile loss	0.0068	76.2	0.088	0	25.9
+ empirical split conformal calibration	0.0068	81.3	0.095	0	25.9

The counterfactual facility-power MAE is evaluated on a separate set of independent, random, ramp-feasible, time-varying 24 h IT-load and cooling-setpoint trajectories within the excitation envelope; the remaining metrics use the chronological test set. Bold entries indicate the best value for each metric.

Empirical split conformal calibration leaves the median MAE unchanged. Raw quantile regression gives 76.2% pooled coverage for a nominal 80% interval; calibration increases coverage to 81.3%, and PINAW from 0.088 to 0.095. For the 0–1, 1–4, 4–12, and

12–24 h bands, calibrated PICP/PINAW/normalized interval scores are 80.8%/0.061/0.072, 81.7%/0.076/0.086, 81.4%/0.096/0.111, and 80.9%/0.113/0.132, respectively. A 1000-replication 24 h block bootstrap yields a 95% confidence interval of 80.1–82.4% for pooled PICP. Across seasons, IT-utilization quartiles, and supply-temperature bands, PICP ranges from 80.0% to 82.6%. These are empirical diagnostics, not conditional or simultaneous trajectory guarantees. For scenario trajectories, the dependent copula sampling reduces the normalized energy score from 0.751 to 0.682 and the variogram score from 0.164 to 0.118 relative to independent marginal sampling.

The response map introduces a tradeoff between approximation accuracy and optimization complexity. Across 1000 independent, ramp-feasible, time-varying 24 h control trajectories (144,000 interval responses), its mean, 95th-percentile, and maximum interpolation errors relative to PC-TFT are 3.8, 10.9, and 18.6 kW for overhead power and 0.028, 0.083, and 0.146 °C for rack inlet temperature, respectively. Compared with EnergyPlus, PC-TFT yields MAE values of 10.6 kW and 0.074 °C, while the final response map yields 11.4 kW and 0.081 °C. Removing the thermal memory state increases the mean map errors from 3.8 to 12.7 kW and from 0.028 to 0.116 °C, confirming that a static two-dimensional map does not retain the control-path effect.

EnergyPlus excitation simulations, PC-TFT training and model selection, initial conformational calibration and dependence estimation, and construction of the simplex topology and fixed MILP structure are performed offline. At each online update, the controller validates the received data; updates the system states, workload queues, rolling calibration corrections, scenarios, and response-map coefficients; evaluates the forecasts; warm starts and solves the MILP; and generates the first control command. Gurobi shared-memory multithreading is used. In practical deployment, independent data acquisition, preprocessing, and forecast evaluation tasks can run in parallel with the optimization where data dependencies permit.

With nine leaves and 144 intervals, the node-shared logarithmic MILP contains 118,436 continuous variables, 8672 binary variables, and 231,904 linear constraints. Gurobi 12.0.1 is run with warm starts, a 1% relative optimality gap target, and a 600 s time limit on a ThinkStation P3 Tower Gen 2 equipped with an Intel Core Ultra 9 285K processor. The computational performance of the complete rolling online pipeline is summarized in Table 6. The measured wall-clock pipeline begins on receipt of the simulator outputs and ends when the first control command is generated. The mean, 95th-percentile, and maximum final optimality gaps were 0.62%, 0.96%, and 1.00%, respectively, and all 4320 updates produced feasible schedules within 600 s. Physical sensing, communication, and actuation delays are excluded and require site-specific measurement; sensor-noise effects were not modeled and require site-specific evaluation. A field deployment should set its computation deadline to reserve the physical delays. At the computation deadline, the best feasible solution is applied; if none is available, the predefined fallback policy is invoked.

Table 6. Computational time of the rolling online pipeline over 4320 updates.

Timing Component	Mean (s)	95th Percentile (s)	Maximum (s)
MILP solution	472.0	536.0	582.0
Non-solver online processing	5.4	7.9	10.8
Complete software pipeline	477.4	542.9	586.9

The complete software-pipeline time includes MILP optimization but excludes physical sensing, communication, and actuation delays.

Variable selection and attention diagnostics are consistent with the modeled thermal behavior but are not causal estimates. Outdoor dry bulb temperature and enthalpy receive the largest future input weights in summer, whereas recent overhead power and chilled

water supply temperature dominate the historical inputs. Planned IT utilization becomes more influential after 3 h, as current telemetry provides less information about future server heat. Attention increases near daily weather extrema and tariff transitions. These patterns can support sensor and regime monitoring, but they do not identify causal effects.

4.4. Operational Value of Forecast-Informed Dispatch

Closed-loop rack temperature performance and monitored humidity are summarized in Table 7. Across 4320 control intervals, the 95th-percentile and maximum rack inlet temperatures were 26.61 °C and 26.86 °C, respectively, with no violation of the 27 °C limit.

Table 7. Closed-loop rack temperature and monitored humidity performance.

Metric	Statistic	Simulated Value	Operating Limit	Assessment
Rack inlet temperature	95th percentile	26.61 °C	≤27 °C	Within limit
Rack inlet temperature	Maximum	26.86 °C	≤27 °C	Within limit
Rack inlet temperature	Exceedance	0/4320 (0%)	0	No violation
Rack inlet relative humidity	Minimum	42.7%	20–80%	Within range
Rack inlet relative humidity	Maximum	68.4%	20–80%	Within range

Temperature statistics cover the rolling 30-day summer closed-loop test; humidity values cover the baseline, excitation, and closed-loop runs.

The rolling 30-day summer experiment shows that coordinated resource flexibility and forecast-informed control reduce cost and peak demand while increasing PV utilization (Table 8). The no-flexibility case costs 39,820 CNY/day and peaks at 3.12 MW. Deterministic MPC with constant PUE or persistence forecasts captures part of the value of workload shifting and battery arbitrage but underestimates the afternoon increase in cooling demand. The PC-TFT median MPC yields the lowest cost among implementable controllers: 34,960 CNY/day. The proposed risk-informed controller yields the lowest peak (2.49 MW), the highest PV utilization (94.3%), and the lowest SLA violation rate (0.1%) at a cost of 35,240 CNY/day.

Table 8. Dispatch performance for the rolling 30-day summer experiment.

Method	Cost (CNY/day)	Peak (MW)	PV Utilization (%)	BESS Throughput (MWh/day)	SLA Violations (%)
No flexibility	39,820	3.12	78.4	1.74	1.4
Constant PUE deterministic MPC	36,750	2.83	87.1	2.31	0.9
Persistence deterministic MPC	35,910	2.71	90.2	2.45	0.8
PC-TFT median MPC	34,960	2.54	93.7	2.62	0.6
Proposed scenario-tree MPC	35,240	2.49	94.3	2.55	0.1
Oracle MPC	34,620	2.46	94.8	2.59	0

Bold entries distinguish the lowest deterministic cost from the best risk-related values because the controllers optimize different objectives.

Compared with the no-flexibility case, the proposed strategy, together with available resource flexibility, reduces operating cost by 11.5% (95% CI: 10.7–12.3%) and peak grid import by 20.2% (95% CI: 18.9–21.4%). These system-level improvements include the inherent benefits of workload shifting, PV utilization, and battery arbitrage and therefore should not be attributed solely to PC-TFT. Compared with constant-PUE MPC, the proposed controller reduces cost and peak grid import by 4.1% and 12.0%, respectively, increases PV utilization by 7.2 percentage points, and reduces the SLA violation rate by 0.8 percentage

points. Compared with persistence-based MPC, it reduces cost and peak grid import by 1.9% and 8.1%, respectively. Its cost is 1.8% above the oracle result. PC-TFT median MPC is 0.8% less expensive, but its SLA violation rate is six times higher, and its peak import is 50 kW higher. This tradeoff follows from the CVaR term.

Figure 6 illustrates the operating mechanism. Midday PV supplies facility demand and charges the BESS, which discharges during the evening price peak. Flexible work moves to intervals with lower prices or higher PV output while meeting deadlines and the rack temperature limit. Facility power does not scale with IT power by a constant factor because overhead demand varies with weather and control decisions. The controller therefore uses the resulting overhead and facility power response, while PUE remains a reporting metric.

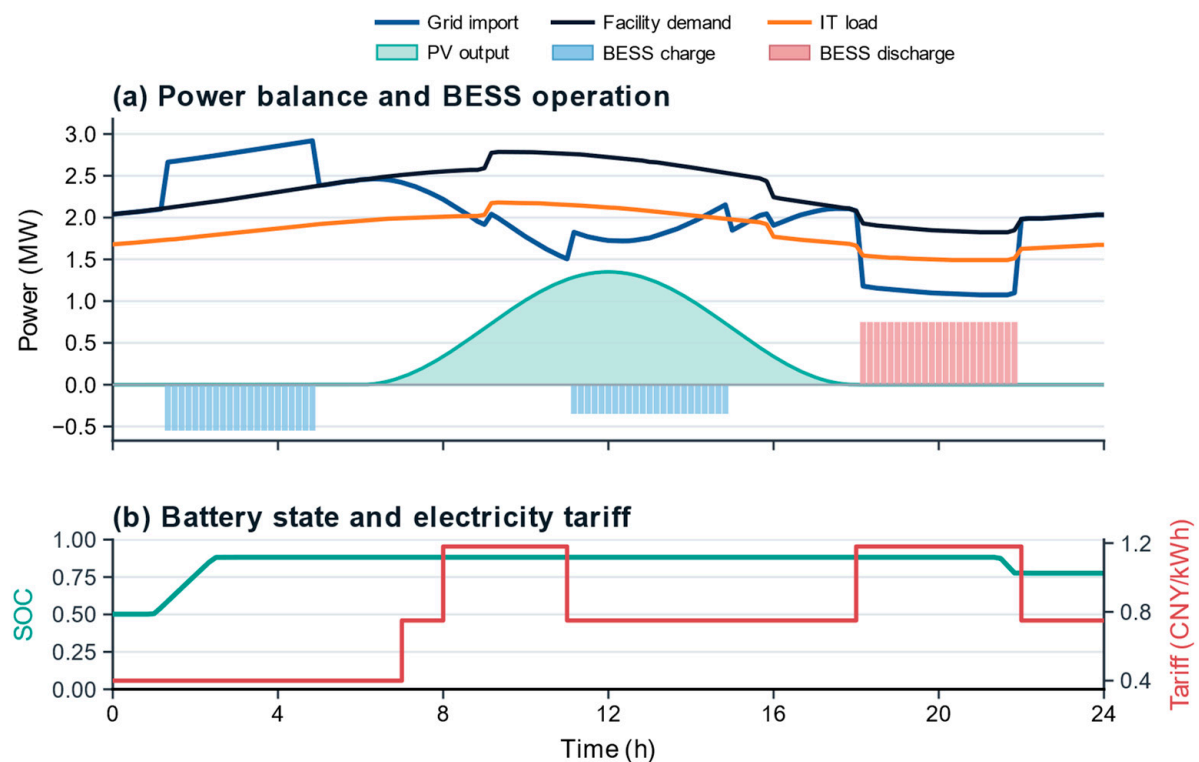


Figure 6. Representative 24 h dispatch profile under the proposed controller. Positive battery power denotes discharge.

Figure 7 separates the economic and reliability effects. Deterministic MPC with constant PUE or persistence forecasts captures much of the tariff arbitrage value, but its peak reduction and PV utilization remain below those of the PC-TFT controllers. The small cost gap relative to the oracle and the larger difference in SLA violations indicate that calibrated uncertainty primarily affects tail-risk decisions. The nonzero violation rate in the no-flexibility case arises from the server capacity limit; without this limit, the baseline would serve all admitted work at the cost of a higher peak.

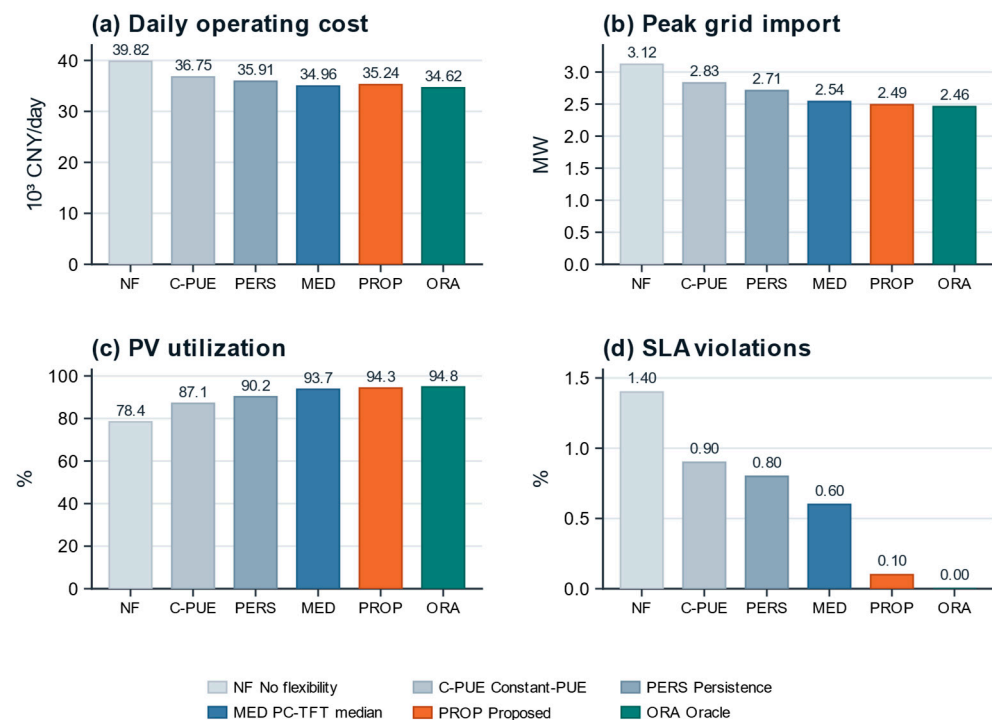


Figure 7. Comparison of operating cost, peak import, PV utilization, and service quality violations.

4.5. Seasonal, Risk Parameter, and Forecast Error Sensitivity

Table 9 shows that the benefits are not limited to one hot period. Cost reductions are 12.7%, 10.1%, and 8.4% for the summer, shoulder season, and winter representative day sets, respectively. The larger summer benefit is consistent with greater and more variable cooling demand. In winter, workload flexibility, PV generation, storage, and tariff response remain valuable, but forecasting overhead power as a function of control decisions contributes less because cooling demand is lower. The reported seasonal savings compare the full coordinated controller with the no-flexibility case. The value of control-aware forecasting depends on site-specific conditions (e.g., variability, forecast error, prices, and flexibility). Thus, in cold climates, it should only be adopted if it clearly outperforms simpler baselines to justify added complexity. Accordingly, no universal PUE floor can be inferred from this study; adoption requires a site-specific comparison that also accounts for computational and maintenance costs.

Table 9. Seasonal performance and sensitivity to the risk parameter.

Case	Cost Reduction vs. no Flexibility (%)	Peak Reduction (%)	PV Utilization (%)	SLA Violations (%)
Summer, proposed	12.7	20.9	94.0	0.12
Shoulder season, proposed	10.1	17.6	95.2	0.08
Winter, proposed	8.4	15.2	91.5	0.05
Risk weight $\lambda_r = 0$	12.2	18.6	94.4	0.61
Risk weight $\lambda_r = 0.25$	11.5	20.2	94.3	0.10
Risk weight $\lambda_r = 0.75$	10.4	20.8	93.6	0.03

Seasonal rows summarize the representative day sets at the base risk parameter. Risk parameter rows refer to the rolling 30-day summer experiment.

Risk parameter sensitivity quantifies the cost of improved reliability. Increasing the risk parameter from $\lambda_r = 0$ to $\lambda_r = 0.25$ reduces the SLA violation rate from 0.61% to

0.10% and improves peak reduction by 1.6 percentage points, while cost savings decline by 0.7 percentage points. Increasing the weight to $\lambda_r = 0.75$ yields only an additional 0.07% reduction in violations and lowers PV utilization because more battery headroom is reserved for adverse scenarios. Scenario-count sensitivity is limited: using 5, 9, and 15 leaves gives cost reductions of 11.2%, 11.5%, and 11.6%, peak reductions of 19.4%, 20.2%, and 20.3%, and SLA violation rates of 0.18%, 0.10%, and 0.09%, respectively. Nine leaves are retained because 15 leaves provide negligible operational improvement at a higher computational cost.

In Figure 8, scaling forecast errors clarifies when calibrated uncertainty has operational value. At twice the base error, deterministic MPC has a 4.10% cost gap to the oracle and 3.10% SLA violations; the risk-informed controller limits these values to 3.15% and 0.58%, respectively. When forecast errors are small, deterministic MPC is slightly less expensive because a CVaR reserve provides little benefit. As error magnitude and tail persistence increase, the risk-informed formulation provides stronger protection.

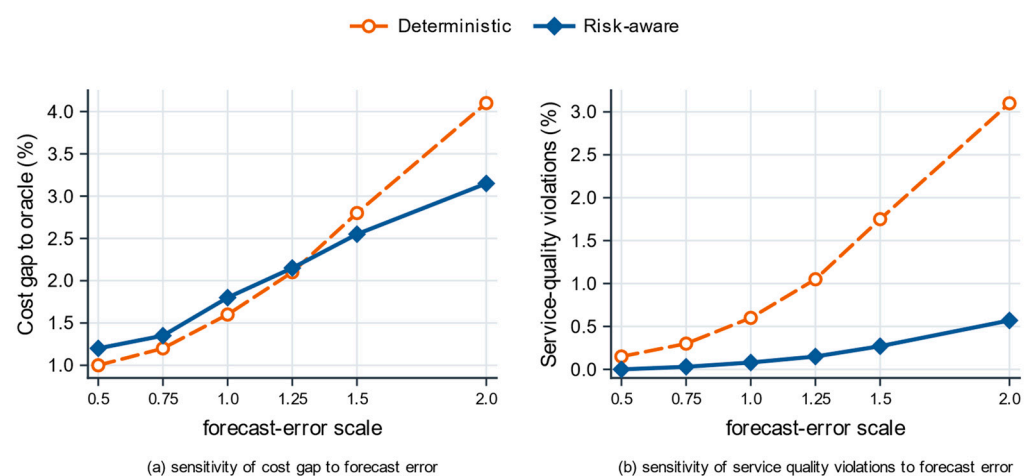


Figure 8. Sensitivity of the cost gap and service quality violations to scaled forecast error.

4.6. Discussion

Three findings emerge. First, the power-balance output improves the facility power response accuracy within the verified control envelope; enforcing $PUE \geq 1$ alone does not account for the gain. Second, empirical calibration, temporal dependence, and CVaR mainly improve tail-risk decisions. Staged ablation attributes 7.7 percentage points of cost savings to resource flexibility and a further 4.5 points to the action-conditioned median forecast. Adding calibration, dependent scenarios, and CVaR changes cost savings from 12.2% to 12.0%, 11.8%, and 11.5%, while reducing SLA violations from 0.60% to 0.34%, 0.22%, and 0.10%. Across five scenario-generation seeds, cost varies by less than 0.2% and peak import by less than 0.03 MW. Third, the proposed PC-TFT enables PUE-related facility responses to vary consistently with the control decisions considered in dispatch; a predictor fitted under a fixed cooling policy cannot be assumed valid when the optimizer changes that policy.

To facilitate practical deployment, we further discuss the implementation procedure in real data center environments. Data from the data center infrastructure management system, building management system, IT scheduler, and weather service are time-aligned to train and calibrate the facility-specific PC-TFT and to construct its response map. During online operation, the framework uses current facility states, pending computing requests, weather forecasts, and candidate workload and cooling schedules as inputs. At each control interval, the MPC determines the flexible workload service schedule and supply air temperature setpoint and sends them to existing supervisory control systems. Before full closed-loop

deployment, the framework is validated via historical replay and shadow-mode operation, followed by gradual activation within safety and operating limits. Prediction errors and interval coverage are monitored online. Note that minor changes that do not alter the relationship between control actions and system responses can be handled by updating the calibration window and monitoring performance; persistent errors require fine-tuning with new data; major upgrades or reconfiguration require fine-tuning or retraining and regeneration of the response map.

5. Conclusions

This paper developed a control-aware PUE forecasting framework for data center DSM and microgrid operation. PC-TFT predicts quantiles of overhead power and rack inlet temperature using telemetry, exogenous forecasts, admitted requests, and candidate workload and cooling controls. Facility power and PUE are derived from the algebraic power balance. Empirical conformal calibration and dependent scenarios represent uncertainty, while a trajectory-conditioned piecewise-affine map with recursive thermal memory embeds the control-path response in risk-informed MPC. The controller coordinates workloads, cooling, PV generation, battery storage, and grid exchange under service, thermal, storage, and nonanticipativity constraints.

Coupled EnergyPlus, workload, and microgrid simulations for a Shenzhen data center demonstrate the effectiveness of the proposed framework. PC-TFT reduces PUE MAE by 17.1% compared with the standard TFT that directly predicts PUE from the same inputs, and by 61.4% compared with persistence. Empirical calibration increases the coverage of the nominal 80% prediction interval from 76.2% to 81.3%. The maximum response-map errors are 18.6 kW for overhead power and 0.146 °C for rack inlet temperature. Compared with the no-flexibility case, the proposed controller, together with available resource flexibility, reduces operating cost by 11.5% and peak grid import by 20.2%, while increasing PV utilization from 78.4% to 94.3%. These system-level gains reflect both resource flexibility and forecast-informed control. Compared with constant-PUE MPC, the proposed controller provides additional cost and peak reductions of 4.1% and 12.0%, respectively. Although median PC-TFT MPC is 0.8% less expensive than the proposed risk-informed controller, the latter achieves a lower peak grid import and reduces the SLA violation rate from 0.6% to 0.1%, demonstrating a better balance between economy and reliability. Overall, these results show that, on the tested simulation platform, the proposed control-aware multi-horizon PUE forecasting provides an accurate and computationally tractable basis for coordinated data center DSM and microgrid dispatch. The rolling horizon implementation can accommodate unplanned workload arrivals or changes in cooling schedules by updating system states and re-optimizing at each control interval, provided that the updated operating state and candidate controls remain within the verified envelope; otherwise, the predefined fallback policy is invoked. Future work will further evaluate its performance under such deviations and investigate adaptive model updating under sustained changes in operating conditions.

Author Contributions: Conceptualization, Y.L. and D.S.; methodology, Y.L. and J.P.; software, Y.L. and J.P.; validation, G.F.; formal analysis, Y.L.; investigation, Y.L. and J.P.; resources, D.S.; data curation, J.P.; writing—original draft preparation, Y.L.; writing—review and editing, J.P., G.F., and D.S.; visualization, Y.L.; supervision, D.S.; project administration, D.S.; funding acquisition, D.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Key Research and Development Program of China, Ministry of Science and Technology of the People's Republic of China (2022YFB2403500).

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT Plus (5.6 Sol) to assist with the graphical design of Figures 1–3. The authors reviewed and edited the output and take full responsibility for the content of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

BESS	Battery Energy Storage System
BiGRU	Bidirectional Gated Recurrent Unit
CNY	Chinese Yuan
CVaR	Conditional Value-at-Risk
DSM	Demand Side Management
GRU	Gated Recurrent Unit
IT	Information Technology
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MILP	Mixed-Integer Linear Program
MPC	Model Predictive Control
PC-TFT	Power-Consistent Temporal Fusion Transformer
PICP	Prediction Interval Coverage Probability
PINAW	Prediction Interval Normalized Average Width
PUE	Power Usage Effectiveness
PV	Photovoltaic
RMSE	Root Mean Square Error
SLA	Service-Level Agreement
SOC	State of Charge
TFT	Temporal Fusion Transformer
TOU	Time-of-Use

References

1. Masanet, E.; Shehabi, A.; Lei, N.; Smith, S.; Koomey, J. Recalibrating global data center energy-use estimates. *Science* **2020**, *367*, 984–986. [[CrossRef](#)] [[PubMed](#)]
2. Dayarathna, M.; Wen, Y.; Fan, R. Data Center Energy Consumption Modeling: A Survey. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 732–794. [[CrossRef](#)]
3. The Green Grid. *PUE™: A Comprehensive Examination of the Metric*; White Paper No. 49; The Green Grid: Beaverton, OR, USA, 2012.
4. Crawley, D.B.; Lawrie, L.K.; Winkelmann, F.C.; Buhl, W.F.; Huang, Y.J.; Pedersen, C.O.; Strand, R.K.; Liesen, R.J.; Fisher, D.E.; Witte, M.J.; et al. EnergyPlus: Creating a new-generation building energy simulation program. *Energy Build.* **2001**, *33*, 319–331. [[CrossRef](#)]
5. Chu, W.-X.; Wang, C.-C. A review on airflow management in data centers. *Appl. Energy* **2019**, *240*, 84–119. [[CrossRef](#)]
6. Khalaj, A.H.; Scherer, T.; Halgamuge, S.K. Energy, environmental and economical saving potential of data centers with various economizers across Australia. *Appl. Energy* **2016**, *183*, 1528–1549. [[CrossRef](#)]
7. Gao, J. *Machine Learning Applications for Data Center Optimization*; Google White Paper; Google: Mountain View, CA, USA, 2014.
8. Hossain, M.S.; Rahaman, S.; Mustafa, R.; Andersson, K. A belief rule based expert system for datacenter PUE prediction under uncertainty. *IEEE Trans. Sustain. Comput.* **2017**, *2*, 140–153. [[CrossRef](#)]
9. Zhao, Y.; Liang, J.; Su, Q.; Liu, X.; Han, Z.; Li, F. On Predicting the PUE with Gated Recurrent Unit in Data Centers. In Proceedings of the 2019 IEEE/CIC International Conference on Communications in China, Changchun, China, 11–13 August 2019; pp. 1664–1670.
10. Ounifi, H.A.; Ben Abdallah, H.; Chalghoumi, H.; Aissa, S. Deep machine learning-based power usage effectiveness prediction for sustainable cloud infrastructures. *Sustain. Energy Technol. Assess.* **2022**, *52*, 101967. [[CrossRef](#)]

11. Kannan, P.; Trivedi, A.; Srinivasan, D. A Bidirectional Gated Recurrent Unit Model for PUE Prediction in Data Centers. In Proceedings of the 2025 International Joint Conference on Neural Networks, Rome, Italy, 30 June–5 July 2025; pp. 1–8.
12. Islam, M.A.; Ren, S.; Quan, G.; Shakir, M.Z.; Vasilakos, A.V. Capacity management of hyperscale data centers using predictive modelling. *Energies* **2019**, *12*, 3438. [\[CrossRef\]](#)
13. Lim, B.; Arik, S.Ö.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [\[CrossRef\]](#)
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2017; Volume 30.
15. Romano, Y.; Patterson, E.; Candès, E.J. Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2019; Volume 32.
16. Rockafellar, R.T.; Uryasev, S. Optimization of Conditional Value-at-Risk. *J. Risk* **2000**, *2*, 21–41. [\[CrossRef\]](#)
17. Liu, Z.; Chen, Y.; Bash, C.; Wierman, A.; Gmach, D.; Wang, Z.; Marwah, M.; Hyser, C. Data center demand response: Avoiding the coincident peak via workload shifting and local generation. *Perform. Eval.* **2013**, *70*, 770–791. [\[CrossRef\]](#)
18. Chen, M.; Gao, C.; Song, M.; Chen, S.; Li, D.; Liu, Q. Internet data centers participating in demand response: A comprehensive review. *Renew. Sustain. Energy Rev.* **2020**, *117*, 109466. [\[CrossRef\]](#)
19. Cao, X.; Wang, J.; Zeng, B. Data center demand response with on-site renewable generation: A bargaining approach. *IEEE/ACM Trans. Netw.* **2018**, *26*, 2707–2720. [\[CrossRef\]](#)
20. Yang, Y.; Chen, W.; Liu, J.; Wen, F.; Shahnia, F. Data center holistic demand response algorithm to smooth microgrid tie-line power fluctuation. *Appl. Energy* **2018**, *231*, 277–287. [\[CrossRef\]](#)
21. Yu, L.; Jiang, T.; Zou, Y. Real-Time Energy Management for Cloud Data Centers in Smart Microgrids. *IEEE Access* **2016**, *4*, 941–950. [\[CrossRef\]](#)
22. Wang, C.; Cao, Y.; Ding, Y. Flexible Multi-Energy Scheduling Scheme for Data Center to Facilitate Wind Power Integration. *IEEE Access* **2020**, *8*, 88876–88891. [\[CrossRef\]](#)
23. Cao, X.; Zhang, J.; Wang, J.; Zeng, B. Data-driven flexibility assessment for internet data center towards periodic batch workloads. *Appl. Energy* **2022**, *324*, 119665. [\[CrossRef\]](#)
24. Lian, H.; Li, C.; Zhang, Y.; Wu, Q. Robust multi-objective optimization for islanded data center microgrid operations. *Appl. Energy* **2023**, *330*, 120344. [\[CrossRef\]](#)
25. Basmadjian, R. Flexibility-based energy and demand management in data centers: A case study for cloud computing. *Energies* **2019**, *12*, 3301. [\[CrossRef\]](#)
26. Osnes, J.; Al-Gabalawy, M.; Olsen, R.L. Harnessing Task usage prediction and latency sensitivity for scheduling workloads in wind-powered data centers. *Energies* **2022**, *15*, 4469. [\[CrossRef\]](#)
27. Marín, L.G.; Sumner, M.; Muñoz-Carpintero, D.; Köbrich, D. Hierarchical energy management system for microgrid operation based on robust model predictive control. *Energies* **2019**, *12*, 4453. [\[CrossRef\]](#)
28. Garcia-Torres, F.; Bordons, C.; Ridao, M.A. Model predictive control for microgrid functionalities: Review and future challenges. *Energies* **2021**, *14*, 1296. [\[CrossRef\]](#)

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