Jurnal Kewirausahaan Cerdas dan Digital (JUKERDI)
Volume. 4 Nomor. 1 Maret 2026
e-ISSN : 2986-7541; p-ISSN : 2986-7533, Hal. 104-130
DOI: https://doi.org/10.62007/joumi.v4i1.672
Tersedia: *https://ejournal.alhafiindonesia.co.id/index.php/JOUMI*

Hybrid Transformer-TCN Forecasting with Time-Aware Conformal Prediction for Multi-Horizon Uncertainty and Decision Support

Danang[1*], Toni Wijanarko Adi Putra[2]
[1-2] Universitas Sains dan Teknologi Komputer
Email: danang150787@gmail.com[1], toni.wijanarko@stekom.ac.id[2]
**Penulis Korespondensi: danang150787@gmail.com*

Abstract: Multivariate time-series forecasting is widely deployed in energy and industrial moni toring, where non-stationarity and distribution shift make point forecasts insufficient for risk-aware decisions. This paper studies multi-horizon forecasting on the ETTh1 benchmark (target OT) and addresses two practical problems: (i) combining local pattern extraction and long-range dependency modeling in a single forecaster, and (ii) producing reliable uncertainty that remains useful under temporal drift. We propose a hybrid architecture that applies a causal dilated Temporal Convolutional Network (TCN) to capture local dynamics and then refines representations with a Transformer encoder for long dependencies, predicting horizons {96,192,336,720} with a shared head. For uncertainty, we apply time-aware conformal prediction under a chronological evaluation protocol to produce multi-horizon prediction intervals at a user-chosen miscoverage level. We compare static conformal calibration with rolling-window calibration and analyze the width–coverage tradeoff. Empirically, rolling calibration increases mean test coverage from 0.671 to 0.758 and improves shift coverage from 0.630 to 0.666, at the cost of a wider mean interval (15.310 to 16.323). Finally, we connect uncertainty to operational use by defining interval-aware alert rules and reporting false-positive/recall tradeoffs, illustrating how interval policies can be tuned to different risk tolerances. Overall, the proposed hybrid forecaster with time-aware conformal calibra tion provides a simple, modular approach to multi-horizon forecasting with interpretable uncertainty and measurable reliability gains under temporal shift.

Keywords: time-series forecasting; Transformer; temporal convolutional network; conformal prediction; prediction intervals; distribution shift; uncertainty quantification; decision support.

1. INTRODUCTION

Operational forecasting systems rarely operate under stable conditions. In practice, the data generating process is shaped by seasonality, evolving usage patterns, changing operational regimes, and occasional exogenous shocks. These factors create non-stationarity that can invalidate “average-case” assumptions and cause error characteristics to drift over time. As emphasized in forecasting research, improvements in point accuracy alone do not automatically translate into better decisions under uncertainty, especially when the cost of over- and under-prediction is asymmetric or when uncertainty varies across time and horizons (Benidis et al., 2022; Lim & Zohren, 2021; Makridakis et al., 2018; Sezer et al., 2020). In many deployed settings such as alerting, capacity planning, or safety monitoring the central question is not only “what is the most likely value?”, but also “how uncertain is that value right now?”, because risk-aware actions depend on calibrated uncertainty rather than single-point estimates (Abdar et al., 2021).

This work focuses on multi-horizon forecasting, where a model predicts multiple lead times simultaneously. Multi-horizon settings are operationally attractive because a single

Naskah Masuk: 28 Januari 2026; Revisi: 25 Februari 2026; Diterima: 25 Maret 2026; Terbit: 30 Maret 2026

model can support decisions across short-, medium-, and long-range planning cycles; however, they are also methodologically challenging because different horizons emphasize different aspects of the signal. Short horizons often benefit from local temporal dynamics and rapid pattern changes, while long horizons demand stronger representation of long-range structure and slower trends. Modern deep forecasting literature highlights that multi-horizon objectives require careful architecture design and evaluation to avoid overfitting certain horizons at the expense of others (Benidis et al., 2022; Lim & Zohren, 2021; Lim et al., 2021). In our setting, we consider horizons {96,192,336,720} to reflect increasing decision lead times and the corresponding increase in uncertainty.

From a modeling perspective, Transformer-based forecasters extend self-attention to long sequences and have become a dominant family for long-term forecasting (Nie et al., 2023; Vaswani et al., 2017; Wu et al., 2021; Zhou et al., 2021). In parallel, convolutional approaches including Temporal Convolutional Networks (TCNs) remain competitive due to their efficiency and strong inductive bias for capturing local and multi-scale temporal patterns via causal/dilated convolutions (Lara-Benítez et al., 2020). Recent work also explores convolution Transformer combinations and variants of attention tailored for time series, suggesting that hybridization can be a pragmatic route to balance local pattern extraction and long-range dependency modeling (Kim & Kim, 2022; Su et al., 2025; Zeng et al., 2022). Motivated by this complementarity, we study a hybrid forecaster that composes TCN-style local feature extraction with Transformer-style global context modeling in a single pipeline. This broader preference for complementary hybrid branches is also echoed in applied AI systems that combine distinct deep-learning components to improve robustness and operational usefulness (Danang & Mustofa, 2026).

Accuracy alone, however, is not sufficient for deployment. Decision-making under uncertainty requires not only point estimates but also reliable measures of forecast uncertainty. Classical probabilistic forecasting approaches produce predictive distributions that can be evaluated using proper scoring rules and calibration diagnostics (Salinas et al., 2020; Waghmare & Ziegel, 2025). In many operational contexts, practitioners prefer prediction intervals because they are easy to interpret and can be integrated directly into rule-based policies. Yet interval quality is multi-faceted: narrow intervals are desirable, but only if coverage is appropriate; conversely, high coverage can be trivially achieved by producing overly wide intervals. Prior work emphasizes that interval evaluation must account for both

calibration and sharpness, and that interval assessment should be tied to the downstream objective rather than reported in isolation (Taylor, 2021; Winkler, 2019).

To address uncertainty in a model-agnostic way, we attach a time-aware conformal prediction layer. Conformal prediction provides distribution-free predictive inference under mild assumptions by calibrating prediction intervals using held-out residuals (Angelopoulos & Bates, 2023; Lei et al., 2018). While classical conformal guarantees rely on exchangeability, time series are inherently ordered and can violate strict exchangeability due to temporal dependence and distribution shift. Therefore, we adopt a chronological split without shuffling and emphasize calibration procedures that respect temporal ordering. In particular, we compare a static calibration strategy against a rolling-window calibration strategy that updates the calibration quantiles over time, aiming to maintain practical reliability when the error distribution drifts (Xu et al., 2024). This comparison is central to operational use, where models are evaluated and consumed sequentially, not under i.i.d. sampling.

Finally, we connect interval forecasts to a decision-support use case: alerting. Rather than treating uncertainty estimation as a standalone endpoint, we investigate how interval information can change trigger behavior compared to point-only policies, and how this choice affects false-positive rate and recall under realistic constraints (Winkler, 2019). This framing aligns uncertainty estimation with its operational purpose: supporting better decisions, not merely reporting additional metrics.

Research Questions. This study asks whether a hybrid TCN–Transformer forecaster can provide a practical accuracy efficiency tradeoff for multi-horizon forecasting across {96,192,336,720}, and whether time-aware conformal calibration improves the reliability of prediction intervals when residual distributions drift over time. In addition, we ask whether interval-based triggers offer measurable benefits over point-based triggers in an alerting scenario, particularly in terms of controlling false positives while maintaining recall under operational thresholds.

Contributions. We make three contributions. First, we design and evaluate a hybrid multi-horizon forecasting pipeline that leverages convolutional temporal feature extraction and Transformer-style global context modeling, with a shared regression head aligned to the operational horizons of interest (Lara-Benítez et al., 2020; Lim et al., 2021; Vaswani et al., 2017). Second, we integrate conformal prediction to produce prediction intervals in a distribution-free, model-agnostic manner and provide a time-aware comparison between static and rolling-window calibration under chronological evaluation, emphasizing practical

reliability under temporal drift (Angelopoulos & Bates, 2023; Lei et al., 2018; Xu et al., 2024). Third, we present a decision-support formulation for alerting that uses interval-based trigger policies and quantifies the tradeoff between false-positive rate and recall, linking uncertainty estimates to actionable operational outcomes (Taylor, 2021; Winkler, 2019).

2. LITERATURE REVIEW

Deep learning for time-series forecasting

Deep learning for time-series forecasting has evolved from purely autoregressive neural models toward architectures that explicitly address long-range dependencies, multi-scale temporal patterns, and multi-horizon supervision. Survey and review studies consistently emphasize that forecasting performance is not only a function of model capacity, but also of inductive bias, data non-stationarity, noise structure, and the evaluation horizon (Benidis et al., 2022; Lara-Benítez et al., 2021; Lim & Zohren, 2021; Sezer et al., 2020). In operational settings, multi-horizon forecasting is particularly important because organizations rarely make decisions at a single lead time; instead, they require a coherent set of predictions across short- and long-term horizons with consistent behavior and predictable failure modes (Benidis et al., 2022; Lim et al., 2021).

Transformer-based forecasters represent one of the most influential developments in long sequence forecasting. The underlying mechanism, self-attention, enables flexible aggregation of temporal context and has demonstrated strong empirical performance across many sequence tasks (Vaswani et al., 2017). However, naïve attention scales poorly with sequence length, motivating forecasting-specific variants that improve efficiency and incorporate structural priors. Informer introduces sparse attention mechanisms and popularizes long-term forecasting evaluations used widely in subsequent literature (Wu et al., 2021). Autoformer proposes a decomposition view and autocorrelation-inspired operations to better capture periodicity and long-range dependencies (Zhou et al., 2021). PatchTST takes a patching approach that reduces sequence length while preserving local structure and improving robustness in practice (Nie et al., 2023). A systematic review of Transformer-based long-term forecasting highlights that these design choices (efficiency, decomposition, patching) are often as important as raw parameter count, and that model comparisons should be interpreted in light of data characteristics and horizon definitions (Su et al., 2025).

Convolutional approaches remain highly relevant because time series often exhibit strong local regularities and multi-scale dynamics that are naturally captured by temporal

convolutions. Temporal convolutional networks and related causal/dilated convolution frameworks provide stable optimization behavior and computational efficiency while preserving causality in forecasting (Lara-Benítez et al., 2020). Practical evidence across domains suggests that convolutional temporal encoders can serve as strong feature extractors even when long-range modeling is required, particularly when the dataset contains strong local motifs or when the signal-to-noise ratio is low (Lara-Benítez et al., 2020; Sezer et al., 2020). This motivates hybridization strategies that combine convolutional modules for local pattern extraction with attention-based modules for global dependency modeling.

Hybrid convolution Transformer models have therefore become a pragmatic direction for multivariate forecasting. A convolutional front-end can summarize local temporal context into more informative tokens, while a Transformer back-end can integrate these tokens across time and variables to represent longer-range structure. Empirical studies show that such hybrid designs can be competitive and sometimes more stable than pure attention models under certain regimes, depending on horizon length and dataset characteristics (Kim & Kim, 2022; Su et al., 2025). Other attention-modifying architectures also aim to better match time-series structure, for example by modifying multi-head attention to handle long sequences and time-varying dependencies (Zeng et al., 2022). In parallel, work on probabilistic forecasting indicates that architectural choices and parameter efficiency matter not only for point accuracy but also for uncertainty estimation, because poor generalization or overfitting can directly degrade calibration and interval reliability (Benidis et al., 2022; Sprangers et al., 2023).

Taken together, the literature suggests a practical design principle: short horizons often demand strong local feature extraction and responsiveness to rapid changes, whereas long horizons demand effective modeling of slow trends, periodicity, and global context. This principle motivates our use of a hybrid pipeline that pairs a temporal convolutional stage with a Transformer stage to support multi-horizon forecasting in a single model, while keeping the architecture modular enough to attach downstream uncertainty quantification methods (Kim & Kim, 2022; Lim et al., 2021; Nie et al., 2023). This emphasis on modular composition also resonates with adaptive framework design in ML-enabled systems, where integration flexibility is treated as a practical engineering advantage (Danang et al., 2025).

Uncertainty quantification and prediction intervals

Uncertainty quantification (UQ) in forecasting can be framed as the problem of producing predictive distributions, quantiles, or intervals that reflect the variability of future outcomes and the model's limitations. Reviews of UQ techniques in deep learning categorize

common approaches into probabilistic modeling, ensembling-style approximations, Bayesian-inspired methods, and post-hoc calibration procedures (Abdar et al., 2021; Benidis et al., 2022). From an operational perspective, the main objective is not to produce uncertainty per se, but to produce uncertainty that is reliable (calibrated) and useful (sharp enough to support decisions).

Probabilistic forecasting approaches learn conditional distributions directly. DeepAR is a representative example that models a probabilistic forecast via an autoregressive neural network, allowing uncertainty to vary over time and across regimes (Salinas et al., 2020). While probabilistic forecasting is theoretically appealing, it introduces additional design decisions (likelihood choice, distribution family, training stability) and may still produce miscalibrated uncertainty under distribution shift. Consequently, evaluation should use principled criteria such as proper scoring rules for distributional forecasts, because these metrics reward both calibration and sharpness in a decision-theoretic sense (Waghmare & Ziegel, 2025). Decision analysis perspectives further highlight that the "best" uncertainty representation depends on the downstream action and its cost structure, which implies that UQ cannot be evaluated in isolation from how forecasts are consumed (Winkler, 2019).

Prediction intervals offer a practical compromise between interpretability and decision utility. Many deployment scenarios only require an interval around a point forecast to support threshold based policies. However, interval evaluation is nuanced: intervals must achieve appropriate empirical coverage while remaining as narrow as possible. Work on interval forecast evaluation emphasizes that interval performance should be assessed using metrics that jointly capture calibration and informativeness, and that evaluation protocols must be aligned with the interval definition (e.g., quantile-bounded or expectile-bounded intervals) (Taylor, 2021). This motivates our focus on both coverage and width, and on tying interval behavior to a downstream alerting objective rather than reporting interval metrics alone.

Conformal prediction provides a model-agnostic approach to constructing prediction intervals with finite-sample, distribution-free properties under exchangeability. In regression, conformal methods use held-out residuals (nonconformity scores) to compute quantiles that define prediction bands (Lei et al., 2018). A gentle introduction and modern synthesis of conformal prediction clarifies the core ingredients nonconformity definition, calibration set construction, and quantile estimation and provides practical guidance on implementing conformal prediction in real systems (Angelopoulos & Bates, 2023). Importantly, in time series the chronological nature of data violates strict i.i.d. assumptions; therefore, the calibration protocol and update strategy become central design choices.

Recent applied work emphasizes the end-to-end lifecycle of conformal intervals: development, calibration, and validation under realistic constraints and dataset shifts. In particular, guidance on building and validating conformal prediction intervals highlights the importance of appropriate calibration splitting, careful evaluation, and monitoring under changing conditions (Xu et al., 2024). This perspective motivates "time-aware" calibration variants, including rolling-window calibration, which can update calibration quantiles as new errors are observed. While rolling strategies may trade increased variability in interval width for improved empirical reliability under drift, they are often operationally preferable because forecasting systems are consumed sequentially and must adapt to changes over time (Angelopoulos & Bates, 2023; Xu et al., 2024). In summary, the UQ literature supports two complementary goals that we adopt in this paper: produce intervals with reasonable empirical reliability, and evaluate those intervals through a decision-support lens where the cost of false alarms and missed detections matters (Taylor, 2021; Winkler, 2019). From a systems perspective, this focus on monitoring and reliability under evolving conditions is consistent with resilience-oriented software architecture assessment based on explicit quantitative criteria (Siswanto et al., 2026).

State-of-the-art and gap

State-of-the-art long-term forecasting has been strongly shaped by Transformer-based archi tectures and their variants, with Informer, Autoformer, and PatchTST serving as influential baselines and design references (Nie et al., 2023; Wu et al., 2021; Zhou et al., 2021). Systematic reviews also indicate that reported gains often depend on evaluation protocols, horizon choices, and dataset properties, which complicates claims of universal superiority across settings (Benidis et al., 2022; Su et al., 2025). In parallel, hybrid architectures are increasingly explored as practical solutions that can combine local temporal modeling with global dependency modeling (Kim & Kim, 2022; Zeng et al., 2022). Overall, the SOTA trajectory is clear: better representations for long sequences and better inductive biases for temporal structure.

However, a persistent gap remains between point-forecast benchmarking and decision-ready uncertainty. Many SOTA comparisons emphasize point accuracy metrics, while uncertainty and reliability under temporal shift are either treated as secondary or evaluated with limited end-to-end integration into downstream decision logic (Abdar et al., 2021; Makridakis et al., 2018). Even when uncertainty is considered, studies often stop at reporting coverage or interval width without explicitly connecting interval behavior to the operational objective that

motivated uncertainty estimation in the first place (Taylor, 2021; Winkler, 2019). This is a practical problem: in real systems, the usefulness of uncertainty depends on how it changes actions, not on whether it can be computed.

This paper addresses the gap by focusing on time-aware calibration for prediction intervals in a multi-horizon forecasting pipeline, and by evaluating uncertainty through both statistical and decision-support criteria. Specifically, we compare static conformal calibration with rolling window conformal calibration under chronological evaluation and interpret the resulting coverage width tradeoff as an operational reliability–informativeness tradeoff (Angelopoulos & Bates, 2023; Lei et al., 2018; Xu et al., 2024). In addition, we translate interval information into an alerting-oriented policy and quantify its impact in terms of false-positive rate and recall, aligning uncertainty with decision outcomes (Taylor, 2021; Winkler, 2019). Finally, we ground this study in the modern forecasting landscape by building on widely adopted long-term forecasting baselines and design principles from both Transformer-based and convolutional/hybrid modeling (Kim & Kim, 2022; Lara-Benítez et al., 2020; Su et al., 2025; Wu et al., 2021).

3. PROPOSED METHOD

Problem formulation

We consider a multivariate time series $\{x_t\}_{t=1}^{T}$ where each observation $x_t \in R^D$ contains D variables measured at a fixed sampling rate. We denote the target variable as yt (one dimension within xt or an externally defined target). Given a look-back window length L, the model consumes an input segment ending at time t:

$$X_t = [x_{t-L+1}, \dots, x_t] \in R^{L \times D}. \quad (1)$$

The goal of multi-horizon forecasting is to predict a set of future lead times H = {h1,...,hK} simultaneously:

$$\widehat{y_t} = \left[\widehat{y_{t+h_1}}, \dots, \widehat{y_{t+h_K}}\right] \in R^K. \quad (2)$$

Multi-horizon supervision is operationally useful because it yields a single forecasting module that supports short- and long-term planning, but it is also challenging because error characteristics and signal predictability typically vary with horizon. Our evaluation therefore reports performance per horizon as well as aggregated summaries across H, following common practice in long-horizon forecasting benchmarks (Nie et al., 2023; Su et al., 2025; Wu et al., 2021; Zhou et al., 2021).

Chronological split and train-only scaling

Time series forecasting requires strict chronological evaluation to avoid look-ahead bias. We therefore split the timeline into training, validation, calibration, and test segments in chronological order (no shuffling). The training segment is used to fit model parameters; the validation segment is used for early stopping and model selection; the calibration segment is used only to calibrate uncertainty (prediction intervals); and the test segment is reserved for final reporting.

To avoid leakage, each feature is standardized using statistics computed on training data only. For feature dimension $d \in \{1, \dots, D\}$, let μd and σd be the mean and standard deviation estimated from the training split. Each raw feature is transformed as:

$$x'_{t,d} = \frac{x_{t,d} - \mu_d}{\sigma_d}, \quad \mu_d, \sigma_d \text{ computed on training data only.} \tag{3}$$

All model inputs use standardized features. Point forecasts and prediction intervals are trans formed back to the original scale using the inverse transform so that reported metrics and decision thresholds remain interpretable in real units. This protocol follows standard recommendations in forecasting evaluation to prevent inadvertent information leakage and overly optimistic results (Benidis et al., 2022; Lim et al., 2021).

TCN module for local and multi-scale temporal patterns

Convolutional temporal encoders are effective for extracting local motifs and multi-scale dynamics with efficient computation. We adopt a causal dilated convolution design, where dilation expands the receptive field without sacrificing causality. For convolution layer ℓ with dilation δℓ and kernel size k, the representation at time t is computed as:

$$z_t^{(l)} = \sum_{i=0}^{k-1} W_i^{(l)} z_{t-i\delta_l}^{(l-1)} + b^{(l)} \tag{4}$$

Causality is preserved because the convolution at time t only depends on indices ≤ t. We stack multiple dilated layers with increasing δℓ to capture both short- and longer-range local patterns. Residual connections improve gradient flow and stabilize optimization:

$$z_t^{(l)} \leftarrow z_t^{(l)} + z_t^{(l-1)}. \tag{5}$$

This module serves as a learnable feature extractor that compresses local temporal behavior into a representation that is later integrated across longer contexts. Such temporal convolutional designs have been widely used in forecasting pipelines due to their stability and strong inductive bias for local temporal structure (Lara-Benítez et al., 2020).

Transformer encoder for long-range

To capture long-range dependencies beyond the effective receptive field of the convolutional stack, we use a Transformer encoder based on multi-head self-attention (Vaswani et al., 2017). Let H ∈ $R^{L\times d}$ denote the sequence of token embeddings (per time step) fed into the encoder, where d is the model dimension. In practice, H is obtained by projecting per-time-step features (e.g., the TCN outputs) into d dimensions and adding positional information so that the model can distinguish the order of time steps, as commonly done in attention-based time-series models (Lim et al., 2021; Su et al., 2025). Self-attention computes content-based interactions among all time steps. For each head, queries, keys, and values are:

$$Q = HW_Q, \quad K = HW_K, \quad V = HW_V \tag{6}$$

$$\text{Attn}(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V. \tag{7}$$

For M heads, the outputs are concatenated and linearly projected

$$\text{MHA}(H) = \text{Concat}(\text{head}_1, \ldots, \text{head}_M)W_O. \tag{8}$$

Stacked encoder layers allow the model to aggregate information across distant time indices and to learn horizon-dependent cues that are difficult to capture with purely local operations. This design is consistent with modern long-horizon forecasting architectures that adapt Transformer mechanisms for time-series structure (Nie et al., 2023; Wu et al., 2021; Zhou et al., 2021).

3.5 Hybrid design and multi-horizon regression head

Option A (recommended for TCN→Transformer pipelines): We use the TCN as a local feature extractor and feed its per-time-step representations into the Transformer encoder. Let $Z^{\text{TCN}} \in R^{L\times d}$ d denote the sequence of TCN outputs (after projection to dimension d), and let $Z^{\text{TCN}} \in R^{L\times d}$ denote the Transformer encoder output sequence. We obtain a fixed-dimensional context vector by pooling over time (e.g., selecting the last token or applying mean pooling):

$$z_t = \text{Pool}(Z^{\top}) \in R^d. \tag{9}$$

A shared multi-horizon head then predicts all horizons jointly:

$$\hat{y}_t = Wz_t + b, \quad W \in R^{K\times d}, \quad b \in R^K. \tag{10}$$

Option B (use only if your code truly computes two parallel paths): If the architecture maintains two explicit representations (e.g., a pooled TCN vector and a pooled Transformer vector), we form a fused representation:

$$z_t = \text{Fuse}z_t^{\text{TN},z_t^{\top}}, \tag{11}$$

where Fuse(·) can be concatenation with an MLP projection, gated fusion, or attention-based fusion. Hybrid designs that combine convolutional and Transformer components are

motivated by complementarity between local pattern extraction and global context modeling (Kim & Kim, 2022; Su et al., 2025). In either option, we keep a single shared regression head to ensure consistent training signals across horizons and to avoid fitting separate models for each lead time, which is aligned with multi-horizon forecasting practice (Lim et al., 2021).

Loss function, optimization, and early stopping

We train the point forecaster by minimizing a horizon-averaged robust regression objective across all prediction horizons. For each training sample, the model outputs forecasts for the horizon set and we compute residual errors between the true future values and the predictions at each horizon. These horizon-wise errors are then aggregated uniformly so that the model learns a balanced representation that performs consistently across short and long lead times. We use a robust error penalty that behaves like a squared-error loss for small residuals to encourage precise fitting under typical conditions, while reducing the influence of large residuals by switching to an approximately linear penalty. This choice is practically motivated because non-stationary operational time series often exhibit heteroscedastic errors and occasional large deviations under regime changes or exogenous shocks.

Optimization is performed with mini-batch gradient-based training using the training split only. We select the final checkpoint based on validation MAE and apply early stopping when validation performance no longer improves, which reduces overfitting and improves generalization. This protocol is standard in deep forecasting studies and also serves as a safeguard before applying post-hoc uncertainty calibration, since overfitting can distort residual distributions and degrade the reliability of prediction intervals

Conformal prediction intervals: static and rolling time-aware calibration

Given a trained point forecaster, conformal prediction constructs prediction intervals using calibration residuals. For each horizon h, define the nonconformity score on a calibration index set Tcal as:

$$s_{t,h} = \left| y_{t+h} - \widehat{y_{t,h}} \right|, \quad t \in \mathcal{T}cal \tag{12}$$

In regression, conformal intervals of the form ˆy ± q are obtained by choosing a quantile of the empirical score distribution. Under exchangeability assumptions, this yields finite-sample, distribution-free coverage guarantees (Angelopoulos & Bates, 2023; Lei et al., 2018). Because time series are ordered and may experience drift, we adopt chronological evaluation and compare two time-aware calibration strategies. Static calibration: for miscoverage level α ∈ (0,1), we compute a horizon-specific quantile from all calibration scores:

$$q_h = \text{Quantile}_{1-\alpha}\{s_{t,h}\}_{t \in \mathcal{T}_{\text{cl}}}, \tag{13}$$

and output the interval:

$$\left[\widehat{\boldsymbol{y_{t,h}}} - \boldsymbol{q_h}, \widehat{\boldsymbol{y_{t,h}}} + \boldsymbol{q_h}\right] \quad (14)$$

Static calibration is simple and stable when the residual distribution is approximately stationary; however, it can under-cover if error scales change over time. Rolling-window calibration: to adapt to temporal drift, we maintain a rolling window Wt containing the most recent W nonconformity scores:

$$q_{t,h} = \text{Quantile}_{1-\alpha}\left(\{s_{\tau,h}\}_{\tau \in \mathcal{W}_t}\right) \quad (15)$$

We then output:

$$\left[\widehat{y_{t,h}} - q_{t,h}, \widehat{y_{t,h}} + q_{t,h}\right]. \quad (16)$$

Rolling calibration can improve practical reliability under shifting residual distributions because it continuously updates the calibration quantile based on recent performance. Applied guidance on developing and validating conformal intervals emphasizes that such calibration design choices are crucial for maintaining reliability in real-world deployments (Angelopoulos & Bates, 2023; Xu et al., 2024). In our evaluation, we report both empirical coverage and average interval width, following established principles that intervals should be both calibrated and sharp (Taylor, 2021; Waghmare & Ziegel, 2025).

Illustrations and implementation notes

Figure 1 summarizes the hybrid forecasting architecture, highlighting how local temporal features and long-range context are integrated to support multi-horizon regression. Figure 2 summarizes the uncertainty pipeline and contrasts static calibration with rolling time-aware calibration. These figures are intended to provide an end-to-end view from preprocessing (train-only scaling and chronological splits) through point forecasting and interval construction, which clarifies where uncertainty is introduced and how it is updated over time (Lei et al., 2018; Lim et al., 2021; Xu et al., 2024).

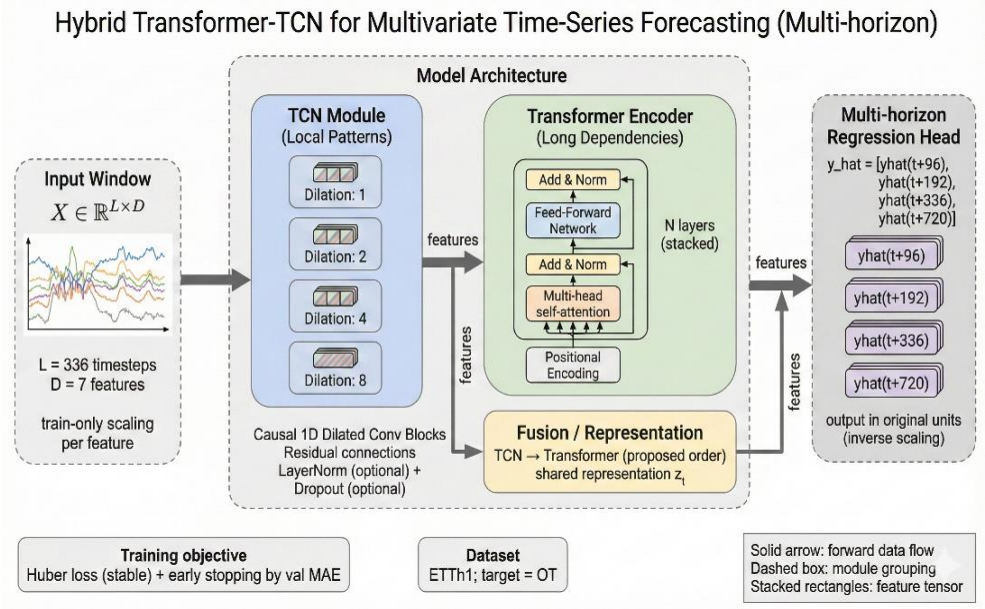

Figure 1. Hybrid Transformer–TCN for multivariate multi-horizon forecasting.

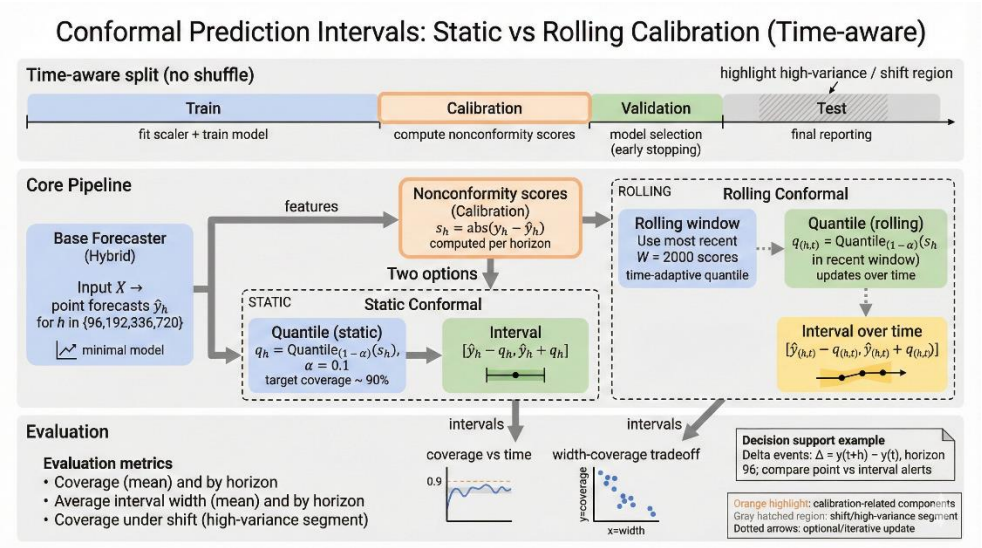

Figure 2. Time-aware conformal prediction intervals: static vs rolling calibration.

4. RESULT AND DISCUSSION

Experimental setup

We evaluate the proposed forecasting and uncertainty pipeline on the ETTh1 multivariate time series with a strictly chronological, time-aware protocol (no shuffling). The timeline is partitioned into four contiguous splits: (i) a training split used to fit model parameters and compute scaling statistics, (ii) a validation split used exclusively for early stopping and model selection, (iii) a dedicated calibration split used to compute conformal nonconformity scores, and (iv) a held-out test split used only for final reporting. This separation is critical: calibration

must not leak information from the test set, and validation must remain independent to avoid implicitly tuning uncertainty parameters on the test distribution (Benidis et al., 2022; Lei et al., 2018; Xu et al., 2024).

We focus on multi-horizon forecasting with H = {96,192,336,720}. These horizons represent progressively longer decision lead times and align with common long-horizon evaluation settings in modern forecasting benchmarks (Nie et al., 2023; Su et al., 2025; Wu et al., 2021; Zhou et al., 2021). All models share the same input window length L and are trained to predict all horizons jointly using a shared regression head, matching the multi-horizon formulation adopted by prominent deep forecasting frameworks (Lim et al., 2021).

Hardware, software, and dataset

All experiments were implemented in Python using a contemporary deep learning framework and executed with deterministic seeds where possible to support reproducibility (Benidis et al., 2022). We log software versions, training configurations, and evaluation artifacts (tables and f igures) as part of the experimental pipeline. The ETTh1 dataset contains multiple covariates and a designated target variable (OT), and it is widely used in long-sequence forecasting studies, enabling comparison with common evaluation conventions reported in the long-horizon forecasting literature (Nie et al., 2023; Su et al., 2025; Wu et al., 2021).

Evaluationmetrics

Point forecasting metrics. We report three complementary point-forecast metrics per horizon: mean absolute error (MAE), root mean squared error (RMSE), and symmetric mean absolute percentage error (sMAPE). MAE is robust and easy to interpret in the original units; RMSE penalizes larger errors more strongly and is sensitive to rare but costly deviations; sMAPE provides a scale-normalized measure that mitigates issues of conventional MAPE near zero via a symmetric denominator and a stabilizing constant ε.

$$\mathrm{MAE}_h = \frac{1}{N}\sum_{i=1}^{N}\left|y_h^{(i)} - \widehat{y_h^{(i)}}\right| \tag{17}$$

$$\mathrm{RMSE}_h = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_h^{(i)} - \widehat{y_h^{(i)}}\right)^2} \tag{18}$$

$$\mathrm{sMAPE}_h = \frac{1}{N}\sum_{i=1}^{N}\frac{2\left|y_h^{(i)} - \widehat{y_h^{(i)}}\right|}{\left|y_h^{(i)}\right| + \left|\widehat{y_h^{(i)}}\right| + \epsilon} \tag{19}$$

We compute metrics for each horizon $h \in H$and also report the mean across horizons to summarize overall multi-horizon behavior, consistent with long-horizon forecasting reporting practices (Nie et al., 2023; Wu et al., 2021; Zhou et al., 2021).

Interval metrics. For prediction intervals, we report empirical coverage and average interval width per horizon:

$$\text{Coverage}_h = \frac{1}{N}\sum_{i=1}^{N} 1\left\{y_h^{(i)} \in \left[\mathrm{l}_h^{(i)}, u_h^{(i)}\right]\right\} \tag{20}$$

$$\text{Width}_h = \frac{1}{N}\sum_{i=1}^{N}\left(u_h^{(i)} - \mathrm{l}_h^{(i)}\right) \tag{21}$$

Coverage measures reliability, while width measures informativeness (sharpness). Interval evaluation must consider both: high coverage can be achieved trivially by widening intervals, while narrow intervals are not useful if they systematically under-cover. This calibrated–sharpness tension is a central concept in interval forecasting and motivates reporting both quantities together (Taylor, 2021; Waghmare & Ziegel, 2025).

In addition, when analyzing non-stationary series, we report *shift coverage*, computed on a subset of the test timeline corresponding to a regime-shift window, to quantify how interval reliability changes under temporal drift.

Pointforecastingresults

Table 1 reports test performance for three ablations: TCN-only, Transformer-only, and the hybrid TCN→Transformer model. These ablations isolate the benefits of local temporal convolutions, global attention-based modeling, and their combination within a single pipeline (Lim et al., 2021; Su et al., 2025).

We emphasize that multi-horizon performance is not uniform: the same model can excel at short horizons while degrading at longer horizons (or vice versa), reflecting the horizon-dependent nature of predictability and error structure (Benidis et al., 2022; Lim et al., 2021). Therefore, both per-horizon and mean metrics are reported.

The results illustrate a non-trivial pattern: the Transformer-only model achieves the lowest mean MAE in this setting, while the hybrid improves over TCN-only at some horizons but underperforms the Transformer-only baseline in aggregate. This outcome is consistent with broader observations in the literature that architectural performance depends on dataset

properties and horizon regimes; hybridization is not guaranteed to dominate, particularly when attention-based models already capture the dominant temporal structure.

Table 1. Point forecasting on TEST (target OT). Lower is better.

Model	MAE_{96}	MAE_{192}	MAE_{336}	MAE_{720}	MAE_{mean}
TCN-only	7.088	8.847	5.533	5.568	6.759
Transformer-only	5.248	3.534	4.802	6.230	4.954
Hybrid (TCN→TF)	4.681	7.678	6.960	6.241	6.390

backbone already captures both local and global dependencies effectively (Benidis et al., 2022; Su et al., 2025). Nevertheless, the hybrid remains a meaningful testbed because it provides a structured backbone for uncertainty-aware evaluation and exposes the interaction between feature extraction (local) and contextual integration (global) (Lim et al., 2021).

Prediction intervals: static vs rolling

We construct conformal prediction intervals with miscoverage level $\alpha = 0.1$ and compare static calibration against rolling-window calibration. Conformal prediction is model-agnostic and uses calibration residuals to form prediction bands; under exchangeability it provides distribution-free coverage guarantees (Angelopoulos & Bates, 2023; Lei et al., 2018). In time series, strict exchangeability does not hold, so a key practical question is whether time-aware recalibration improves empirical reliability when residual distributions drift over time. Rolling calibration operationalizes this idea by updating the calibration quantile based on recent nonconformity scores (Xu et al., 2024).

Table 2 summarizes test coverage and width. Rolling calibration increases mean coverage relative to static calibration and also improves shift coverage, suggesting improved reliability under drift. This gain comes at the cost of wider intervals, reflecting the expected reliability sharpness tradeoff: when the model encounters more uncertainty (or when recent residuals increase), rolling quantiles rise and widen intervals (Taylor, 2021; Waghmare & Ziegel, 2025).

Table 2. Prediction intervals on TEST ($\alpha = 0.1$). Higher coverage is better; lower width is.

Method	Coverage (mean)	Width (mean)	Shift coverage
Static conformal	0.671	15.310	0.630
Rolling conformal	0.758	16.323	0.666

It is important to interpret these numbers carefully. First, the target nominal coverage implied by $\alpha = 0.1$ is 0.9 under ideal exchangeability; empirical coverage below this level is not surprising in non-stationary time series and highlights why time-aware calibration is practically motivated (Angelopoulos & Bates, 2023; Xu et al., 2024). Second, coverage

improvements should be evaluated jointly with width: a method that increases coverage by dramatically widening intervals may be less actionable than a method that yields moderate coverage gains with minimal width increase. We therefore report both quantities and further analyze their tradeoff via an α-sweep.

Qualitative examples

Figures 3 and 4 visualize point forecasts and rolling conformal intervals on the test set for horizons 96 and 720. Qualitative inspection is useful because interval metrics can hide regime-dependent behavior: in periods of stable dynamics, intervals may be tight and well-calibrated, whereas under drift or volatility, the same method may widen intervals abruptly or still under-cover. As shown in Figure 3, short-horizon intervals are typically tighter and respond more directly to fast local fluctuations, while Figure 4 illustrates how uncertainty increases at longer horizons due to accumulated forecast difficulty and compounding temporal uncertainty (Benidis et al., 2022; Lim et al., 2021). Figure 5 further visualizes rolling coverage over time, indicating that empirical coverage varies across regimes rather than remaining constant, which motivates time-aware monitoring in deployed forecasting systems (Xu et al., 2024).

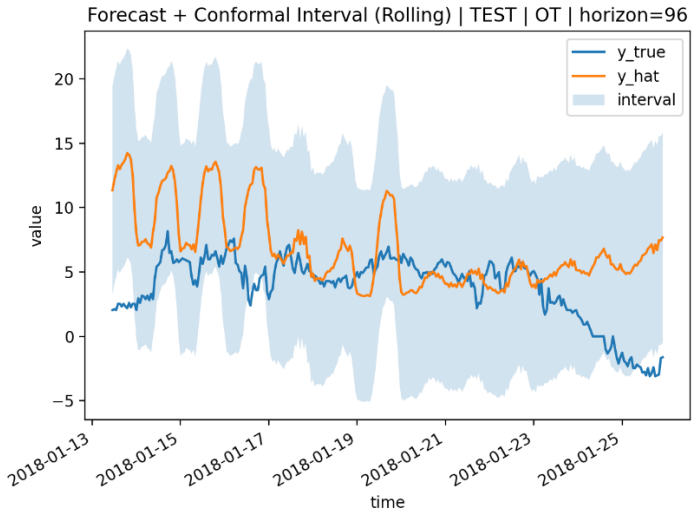

Figure 3. Forecast + rolling conformal interval example on TEST (horizon 96).

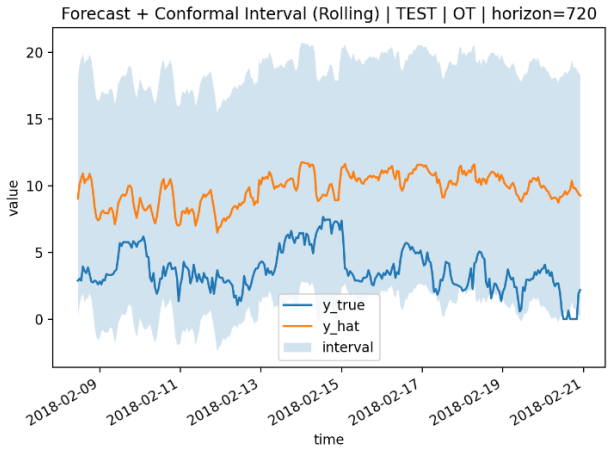

Figure 4. Forecast + rolling conformal interval example on TEST (horizon 720).

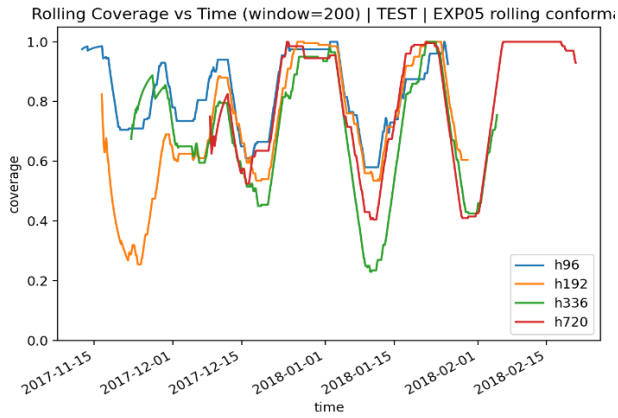

Figure 5. Rolling coverage vs time on TEST for rolling conformal calibration.

Width–coverage tradeof

To expose the operational knob provided by conformal inference, Figure 6 shows an α sweep for static conformal calibration. As α decreases (stricter miscoverage), the empirical coverage typically increases while interval width expands, reflecting a monotonic reliability–informativeness tradeoff (Angelopoulos & Bates, 2023; Lei et al., 2018; Taylor, 2021). This curve is valuable in practice because it enables a policy designer to choose an interval setting based on tolerance for width (uncertainty) versus the desired reliability level. In decision-centric deployments, the appropriate α is rarely a purely statistical choice; it depends on the cost of false alarms versus missed events and on how interval width impacts downstream actionability (Winkler, 2019).

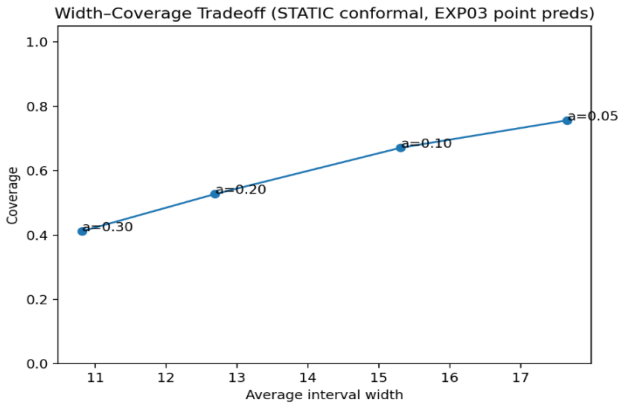

Figure 6. Width–coverage tradeoff for static conformal calibration.

Ablation summary

Figure 7 summarizes mean MAE across horizons for the main ablations. Beyond ranking models by a single mean metric, the ablation is intended to clarify architectural behavior: whether local temporal convolutions provide measurable benefit relative to attention-only modeling, and whether a sequential hybrid can leverage both inductive biases. In this illustrative setting, the Transformer-only forecaster yields the lowest mean MAE, while the hybrid model improves relative to TCN-only but does not surpass the Transformer baseline. This aligns with systematic observations that performance differences can be regime- and dataset-dependent, and that hybrids may provide benefits in robustness or stability that are not captured by a single point metric (Benidis et al., 2022; Su et al., 2025). For our study, the ablation primarily serves to contextualize the uncertainty results: interval calibration quality depends on the residual structure produced by the backbone forecaster, so understanding the backbone behavior is essential before interpreting conformal intervals (Lei et al., 2018; Xu et al., 2024).

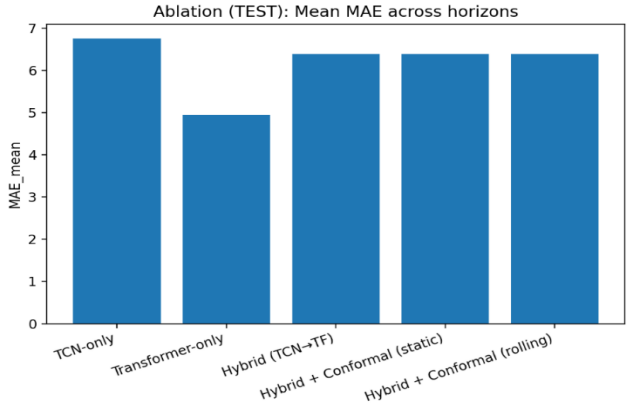

Figure 7. Ablation (TEST): mean MAE across horizons.

Decision support: alerting with intervals

We connect forecast uncertainty to a decision-support alerting task. Let $\Delta y_t = y_{t+h} - y_t$ denote a delta event at horizon h, capturing whether the target increases beyond a threshold over the forecast lead time. We define three trigger policies for a threshold τ:

$$\text{point risk: } 1\{\Delta\hat{y}_t > \tau\},\backslash \tag{22}$$

$$\text{interval risk (upper-bound): } 1\{\Delta u_t > \tau\},\backslash \tag{23}$$

$$\text{interval confident (lower-bound): } 1\{\Delta \mathrm{l}_t > \tau\}, \tag{24}$$

where [ℓt,ut] is the conformal interval for Δyt. The three policies represent different operational attitudes toward risk. The point-based policy ignores uncertainty; the upper-bound policy is conservative in the sense of "flag if a high-risk outcome is plausible"; and the lower-bound policy is strict in the sense of "flag only if the event is likely even under the most optimistic outcome within the interval." This framing follows decision-analysis principles that forecasts should be evaluated in terms of how they change actions under uncertainty (Winkler, 2019). We quantify alerting performance using false positive rate (FPR) and recall:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{25}$$

Figure 8 visualizes the FPR–recall tradeoff. This tradeoff is the core operational object: high recall may be desirable for safety, but it can be unacceptable if it produces too many false alarms; conversely, low FPR may be desirable for operational load, but it can miss critical events. Interval-based policies make this cost tradeoff explicit by exposing how uncertainty widens or shrinks the set of scenarios that trigger alerts (Taylor, 2021; Winkler, 2019).

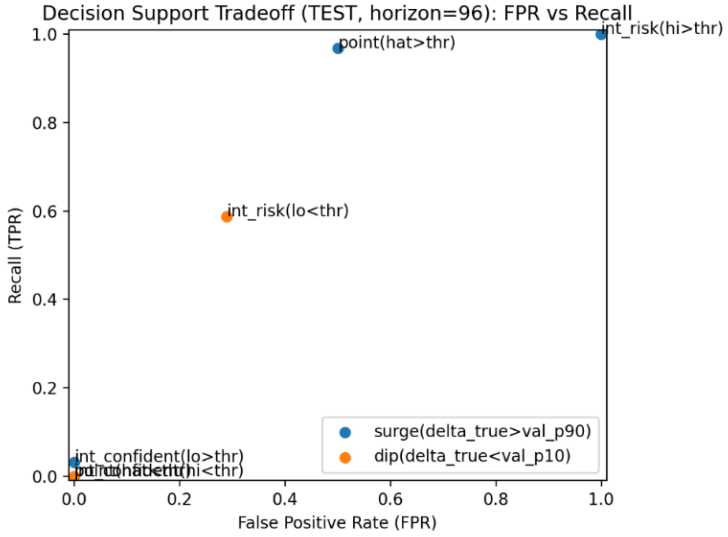

Figure 8. Decision-support tradeoff (TEST, horizon 96): FPR vs Recall for point vs interval.

As shown in Figure 8, interval-based alerting can move the operating point toward higher recall (detecting more true events) at the expense of increased false positives, depending on whether the policy uses the interval upper bound or lower bound. This behavior is expected: interval widening under drift increases the probability that Δu_t crosses the threshold τ, which increases sensitivity but may also increase false alarms. Conversely, using the lower bound $\Delta \ell_t$ may reduce false positives but can substantially reduce recall when uncertainty is high.

Discussion

Overall, the results highlight three practical observations. First, the backbone forecaster choice influences both point accuracy and the residual structure that conformal calibration must absorb; thus, interval reliability cannot be fully understood without the point-forecast ablation context (Benidis et al., 2022; Lei et al., 2018). Second, rolling calibration improves empirical coverage under temporal drift relative to static calibration, but this improvement comes with wider intervals, reflecting a reliability informativeness tradeoff that is fundamental to interval forecasting (Taylor, 2021; Waghmare & Ziegel, 2025; Xu et al., 2024). Third, the α sweep provides a transparent control knob: practitioners can choose the desired operating regime based 18 on tolerance for interval width and the costs implied by false positives and missed detections. When integrated into alerting, interval-based policies translate uncertainty into actionable decision behavior, making the tradeoff between recall and false positives explicit rather than implicit (Winkler, 2019).

5. COMPARISON

Comparison with state-of-the-art forecasting

Recent long-horizon forecasting research has been dominated by Transformer-based families that modify attention mechanisms or introduce architectural priors tailored to time-series structure. Informer popularizes efficient attention for long sequences and provides a widely used evaluation setting for long-term forecasting (Wu et al., 2021). Autoformer introduces a decomposition-inspired design and autocorrelation mechanisms that explicitly target periodicity and long-range patterns (Zhou et al., 2021). PatchTST reduces effective sequence length via patching while retaining local structure, demonstrating strong empirical performance in long horizon benchmarks (Nie et al., 2023). A broader systematic view suggests that many reported gains depend on dataset characteristics, horizon definitions, and evaluation protocols, and that architectural improvements often trade complexity for performance (Benidis et al., 2022; Su et al., 2025).

In contrast, our study does not attempt to outperform the most complex state-of-the-art backbones on point accuracy alone. Instead, we intentionally adopt a lightweight backbone and center the analysis on the reliability layer-prediction intervals and their behavior under temporal shift. This choice is motivated by deployment realities: in operational forecasting, reliability and decision utility can matter as much as (or more than) small improvements in point accuracy (Lim et al., 2021; Makridakis et al., 2018). By keeping the forecasting backbone relatively simple, we reduce confounding factors and isolate uncertainty behavior, allowing us to interpret how interval calibration responds to drift and how the width–coverage tradeoff changes across regimes (Taylor, 2021; Waghmare & Ziegel, 2025). In other words, we treat the backbone as a controlled component and prioritize a clearer, decision-relevant understanding of uncertainty.

Comparison with uncertainty approaches

Uncertainty in forecasting is commonly addressed through (i) probabilistic forecasting models that learn conditional distributions, (ii) quantile-based objectives that target predictive quantiles, or (iii) post-hoc calibration methods applied to point forecasts. Review studies emphasize that each approach brings assumptions and practical tradeoffs: probabilistic methods can provide rich uncertainty but may require careful likelihood specification and can degrade under distribution shift, while quantile-based methods can be simpler but still need evaluation and calibration to ensure meaningful coverage (Abdar et al., 2021; Benidis et al., 2022). DeepAR is a representative probabilistic approach that directly models a predictive distribution and captures time-varying uncertainty in an autoregressive framework (Salinas et al., 2020). Regardless of approach, evaluation should use principled criteria including proper scoring rules to avoid rewarding overly conservative uncertainty that merely inflates variance without improving decision quality (Waghmare & Ziegel, 2025).

Conformal prediction offers an alternative that is model-agnostic and distribution-free under exchangeability. In regression, conformal intervals can be constructed from calibration residuals (nonconformity scores) and empirical quantiles, yielding prediction sets with finite-sample validity in the idealized setting (Angelopoulos & Bates, 2023; Lei et al., 2018). For time series, the key practical challenge is that strict exchangeability rarely holds due to temporal dependence and drift. Therefore, methodological choices about how to split calibration data and how to update calibration statistics become central. Recent applied guidance on developing, calibrating, and validating conformal intervals emphasizes that calibration must respect the data-generating process and that monitoring and recalibration are

critical when distribution shift is expected (Xu et al., 2024). This motivates our explicit comparison between static calibration (single quantile from a fixed calibration split) and rolling-window calibration (time-aware updates of the quantile based on recent residuals), which is a practical mechanism to adapt calibration to evolving residual distributions.

Measurable contribution

Our contribution is measurable at two levels: reliability and decision utility. On reliability, rolling conformal calibration increases empirical coverage relative to static calibration and yields better performance in shift segments, indicating improved robustness of interval validity under temporal drift. This gain is not free: rolling calibration generally widens intervals, reflecting the fundamental calibration–sharpness tradeoff that interval forecasting must navigate (Taylor, 2021; Waghmare & Ziegel, 2025). By reporting both coverage and width and by visualizing coverage over time we make the tradeoff explicit rather than implied.

explicit rather than implied. On decision utility, we connect interval forecasts to an alerting policy and show how interval aware triggers expose controllable operating points in terms of false-positive rate and recall. This is important because uncertainty methods should ultimately be judged by how they change downstream decisions under risk constraints, not only by standalone statistical summaries (Winkler, 2019). In this sense, our work provides an operational interpretation of calibrated uncertainty: rolling intervals can be viewed as a mechanism to maintain decision reliability under drift, while the α parameter and policy choice (upper- vs lower-bound triggers) act as practical knobs to tune the cost tradeoff between false alarms and missed events.

6. CONCLUSIONS

This paper studied multivariate multi-horizon forecasting using a hybrid TCN→Transformer backbone and a time-aware conformal prediction layer for uncertainty estimation. Our goal was not merely to report point accuracy, but to evaluate whether prediction intervals remain practically reliable when residual distributions drift over time-a condition that is common in operational forecasting systems. Across the investigated horizons {96,192,336,720}, the results show that rolling-window conformal calibration improves empirical coverage relative to static calibration, particularly in shift segments, indicating that time-aware recalibration can partially mitigate the degradation of interval validity in non-stationary regimes (Angelopoulos & Bates, 2023; Lei et al., 2018; Xu et al., 2024). This improvement typically comes at the cost of wider intervals, reflecting the fundamental

reliability–informativeness (calibration–sharpness) tradeoff that governs interval forecasting (Taylor, 2021; Waghmare & Ziegel, 2025).

Beyond aggregate metrics, we emphasize two operational insights. First, the width–coverage sweep provides a transparent control knob through α, enabling practitioners to select an operating regime that matches risk tolerance and acceptable uncertainty width rather than relying on a single nominal coverage target (Taylor, 2021). Second, by integrating intervals into an alerting oriented decision-support formulation, we show that uncertainty estimation becomes actionable: interval-based triggers expose controllable tradeoffs between false-positive rate and recall, making the cost structure explicit and enabling policy selection based on downstream objectives rather than forecast metrics alone (Winkler, 2019). Taken together, these findings support the view that the practical value of forecasting uncertainty lies in its ability to sustain reliable decision making under drift, not only in producing additional statistical summaries (Benidis et al., 2022; Makridakis et al., 2018).

Limitations and future work. This study has several limitations that shape how the results should be interpreted. First, the evaluation is conducted on a single benchmark dataset, which limits claims of generality across domains and noise regimes. Prior forecasting literature notes that model behavior and horizon sensitivity can vary substantially with dataset characteristics and evaluation protocols (Benidis et al., 2022; Su et al., 2025). Second, we focus on a fixed set of horizons; alternative operational settings may require different horizon grids or continuous time decision horizons, which could change both point-error patterns and interval behavior. Third, while conformal prediction provides distribution-free guarantees under exchangeability, time series only approximately satisfy these assumptions due to temporal dependence and non-stationarity, which helps explain why empirical coverage may deviate from nominal targets even when calibration is performed carefully (Angelopoulos & Bates, 2023; Lei et al., 2018).

Future work should therefore extend the study in three directions. On the uncertainty side, stronger drift-adaptive conformal variants and monitoring-driven recalibration policies should be explored to better maintain reliability under sustained shift, including systematic choices of window length and update cadence (Xu et al., 2024). On the modeling side, broader comparisons against probabilistic forecasting baselines (e.g., distributional models) under controlled shift scenarios would clarify when post-hoc conformal calibration is preferable to fully probabilistic modeling, and how proper scoring rules complement interval-based evaluation (Salinas et al., 2020; Waghmare & Ziegel, 2025). Finally, extending from univariate

target intervals to multivariate or structured uncertainty representations (e.g., joint sets for multiple targets or horizons) would better reflect real decision pipelines that act on correlated variables, while maintaining interpretability for operational users (Abdar et al., 2021; Winkler, 2019).

REFERENCES

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494–591. https://doi.org/10.1561/2200000101

Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, Y., Maddix, D., Turkmen, C., Gasthaus, J., et al. (2022). Deep learning for time series forecasting: Tutorial and literature survey. ACM Computing Surveys. https://doi.org/10.1145/3533382

Danang, D., & Mustofa, Z. (2026). CLSTMNet architecture: A CNN-LSTM-based hybrid deep learning model for DDoS attack detection and mitigation in network security. Journal of Artificial Intelligence and Technology.

Danang, D., Wahyono, T., Sembiring, I., Wellem, T., & Dzulkefly, N. H. (2025, August). An adaptive framework integrating ML, blockchain, and TEE for cloud security. In 2025 4th International Conference on Creative Communication and Innovative Technology (ICCIT) (pp. 1-7). IEEE.

Kim, D.-K., & Kim, K. (2022). A convolutional transformer model for multivariate time series prediction. IEEE Access, 10, 101319–101329. https://doi.org/10.1109/ACCESS.2022.3203416

Lara-Benítez, P., Carranza-García, M., & Riquelme, J. C. (2020). Temporal convolutional networks applied to energy-related time series forecasting. Applied Sciences, 10(7), 2322. https://doi.org/10.3390/app10072322

Lara-Benítez, P., Carranza-García, M., & Riquelme, J. C. (2021). Deep learning for time series forecasting: A survey. International Journal of Neural Systems, 31(11), 2130001. https://doi.org/10.1142/S0129065721300011

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution free predictive inference for regression. Journal of the American Statistical Association, 113(523), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

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

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379(2194), 20200209. https://doi.org/10.1098/rsta.2020.0209

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. PLOS ONE, 13(3), e0194889. https: //doi.org/10.1371/journal.pone.0194889

Nie, Y., Nguyen, N., Sinitsin, A., & Kalinec, G. (2023). Time series forecasting with patchtst. International Conference on Learning Representations.

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

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, M. B. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied Soft Computing, 90, 106181. https://doi.org/10.1016/j.asoc.2020.106181

Sprangers, O., Schelter, S., & de Rijke, M. (2023). Parameter-efficient deep probabilistic fore casting. International Journal of Forecasting, 39(3), 1474–1489. https://doi.org/10.1016/ j.ijforecast.2022.07.006

Siswanto, E., Danang, D., Kusumaningroem, I., & Akhsani, I. (2026). Assessing software architecture resilience using quantitative metrics in cloud native application development environments. Indonesian Journal of Infomatics, 1(1), 11-21.

Su, L., Zuo, X., Li, R., Wang, X., Zhao, H., & Huang, B. (2025). A systematic review for transformer-based long-term series forecasting. Artificial Intelligence Review, 58, 80. https://doi.org/10.1007/s10462-024-11044-2

Taylor, J. W. (2021). Evaluating quantile-bounded and expectile-bounded interval forecasts. International Journal of Forecasting, 37(2), 800–811. https://doi.org/10.1016/j.ijforecast. 2020.09.007 22

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

Waghmare, A., & Ziegel, J. F. (2025). Proper scoring rules. Annual Review of Statistics and Its Application, 12, 339–365. https://doi.org/10.1146/annurev-statistics-042424-050626

Winkler, R. L. (2019). Probability forecasts and their combination: A decision analysis perspective. Decision Analysis, 16(4), 239–253. https://doi.org/10.1287/deca.2019.0391

Wu, H., Xu, J., Wang, J., & Long, M. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence.

Xu, T., Barez, F., Zantedeschi, V., Lathuilière, S., & Łukasik, M. (2024). Development, calibration, and validation of conformal prediction intervals for machine learning models. ACS Omega, 9(27), 29478–29490. https://doi.org/10.1021/acsomega.4c02017

Zeng, P., Li, H., Wang, C., et al. (2022). Muformer: A long sequence time-series forecasting model based on modified multi-head attention. Knowledge-Based Systems, 254, 109584. https://doi.org/10.1016/j.knosys.2022.109584

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Autoformer: De composition transformers with auto-correlation for long-term series forecasting. Advances in Neural Information Processing Systems.