Jurnal Teknik Mesin, Industri, Elektro dan Informatika
Volume 3, Nomor 3, September 2024
e-ISSN :2963-7805; p-ISSN :2963-8208, Hal. 426-450
DOI: https://doi.org/10.55606/jtmei.v3i3.6156
Tersedia: https://ejurnal.politeknikpratama.ac.id/index.php/jtmei

Hybrid TCN GRU Based Multi-Horizon Forecasting with Adaptive Conformal Prediction for Decision Support Systems

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

Abstract. Multivariate time-series forecasting has become an essential component in energy systems and industrial monitoring applications, where changing data distributions and temporal non-stationarity reduce the reliability of point predictions alone for decision-making purposes. This study investigates multi-horizon forecasting on the ETTh1 benchmark with oil temperature (OT) as the prediction target. The research focuses on two key challenges: integrating short-term pattern extraction with long-range temporal dependency modeling in a lightweight forecasting framework, and providing distribution-free uncertainty estimation that remains reliable under temporal drift conditions. To address these issues, a hybrid forecasting architecture is proposed by combining a causal dilated Temporal Convolutional Network (TCN) for local feature extraction and a Gated Recurrent Unit (GRU) layer for modeling longer sequential dependencies. The model simultaneously predicts four forecasting horizons 96, 192, 336, and 720 steps using multivariate observations and temporal features from a fixed historical window. For uncertainty estimation, time-aware conformal prediction is implemented through chronological train, calibration, and test partitions. The study compares conventional static conformal intervals with an adaptive rolling approach that dynamically updates horizon-specific quantiles based on recent residual patterns. Experimental results on the ETTh1 dataset demonstrate that the proposed hybrid model achieves superior forecasting accuracy compared with standalone TCN and GRU baselines, obtaining an MAE of 4.615 and RMSE of 5.382 for OT prediction. Furthermore, the adaptive conformal strategy improves empirical coverage from 0.790 to 0.870 while maintaining a relatively moderate increase in prediction interval width. A threshold-based decision-support mechanism utilizing the predictive upper bound is also introduced to illustrate how calibrated uncertainty intervals can support practical trade-offs between false alarms and missed detections. Overall, the proposed framework offers an efficient and reproducible solution for uncertainty-aware multi-horizon forecasting in dynamic environments.

Keywords: Conformal Prediction; Multi-Horizon Forecasting; TCN-GRU Hybrid; Time-Series Forecasting; Uncertainty Estimation.

1. INTRODUCTION

Accurate time-series forecasting is a core component of decision-making in industrial monitoring, energy management, and predictive maintenance, where operators must anticipate system behav ior to prevent failures, allocate resources, and plan interventions. In many operational settings, the requirement is not a single-step prediction but multi-horizon forecasts that simultaneously support short-term control (e.g., immediate load balancing or anomaly response) and longer term planning (e.g., scheduling maintenance windows or estimating future risk). Modern deep forecasting has made substantial progress by learning nonlinear dependencies from multivariate inputs and leveraging flexible sequence models that generalize across large datasets (Bandara et al., 2020; Salinas et al., 2020; Smyl, 2020). Nonetheless, strong point accuracy on benchmark splits does not automatically translate into reliable support for real decisions, because practical deployments frequently exhibit non-stationarity, regime changes, and distribution shift that degrade performance and make naive forecasts brittle.

Naskah Masuk: 13 Juli 2024; Revisi: 24 Agustus 2024; Diterima: 24 September 2024; Terbit: 30 September 2024

A second, often underestimated, challenge is that decision-making under uncertainty requires more than a point estimate. In high-stakes monitoring scenarios, the cost of error is asymmetric: underestimation may lead to missed alarms and late interventions, while overestimation may trigger unnecessary actions and operational inefficiency. Therefore, practitioners need not only accurate forecasts but also uncertainty quantification that remains meaningful when the data-generating process drifts over time. While probabilistic modeling and deep probabilistic forecasters can provide uncertainty in principle (Salinas et al., 2020), their reliability under shift may depend on modeling assumptions and calibration quality. This motivates a complementary, deployment-oriented perspective: use a compact forecaster for accurate multi-horizon predictions, then attach a distribution-free uncertainty layer that offers explicit coverage control and can be adapted when drift is observed.

In this work, we focus on the ETTh1 benchmark, a multivariate electricity transformer temperature dataset, and forecast the oil temperature (OT) as the target variable. We study the problem through two tightly coupled practical objectives. First, we seek an architecture that captures both short-range temporal motifs and longer-range dependencies without relying on heavy attention-based stacks, so that training and inference remain efficient and reproducible. Temporal Convolutional Networks (TCN) are attractive for extracting local dynamics using causal dilated convolutions, while gated recurrent units (GRU) offer a lightweight mechanism to refine longer dependencies and stabilize sequence modeling (Bai et al., 2018; Cho et al., 2014). Second, we aim to provide uncertainty estimates that are actionable for risk-aware decisions and remain robust when the test-time distribution deviates from the calibration regime. To achieve this, we employ conformal prediction to construct prediction intervals with user-controlled miscoverage (e.g., $\alpha = 0.1$ targeting 90% coverage), which can be applied on top of any trained forecaster and does not require distributional assumptions (Angelopoulos & Bates, 2021; Vovk et al., 2005). We additionally consider an adaptive rolling variant motivated by conformal inference under shift, where calibration is updated using recent errors to reduce systematic under-coverage during drift (Gibbs & Cand`es, 2021).

This study is guided by the following research questions framed in a deployment-oriented manner. Can a compact hybrid forecaster that combines a causal dilated TCN with a GRU layer achieve competitive multi-horizon point accuracy on ETTh1 while remaining simpler and lighter than transformer-centric alternatives (Bai et al., 2018; Cho et al., 2014). Further, can distribution-free conformal prediction provide practically reliable prediction intervals for multi-horizon forecasting under temporal drift, and does an adaptive rolling calibration strategy improve empirical coverage and decision usefulness comparedx to static calibration (Angelopoulos & Bates, 2021; Gibbs & Cand`es, 2021; Vovk et al., 2005). Finally, when both point forecasts and uncertainty estimates are integrated, do the resulting intervals meaningfully support a basic decision-support workflow that is sensitive to over- and under-estimation risk across horizons.

The contributions of this paper are threefold and emphasize practicality, robustness, and traceable evaluation. First, we introduce a lightweight hybrid TCN–GRU forecaster for multi horizon prediction on multivariate time series, where the model directly outputs horizons H = {96,192,336,720} from a fixed lookback window (L = 336) while incorporating time features to better encode periodic structure; this design targets an efficient balance between local pattern extraction and longer-range dependency modeling (Bai et al., 2018; Cho et al., 2014). Second, we develop a time-aware conformal prediction pipeline using a leakage-safe chronological train/calibration/test split, and we evaluate both static conformal intervals and an adaptive rolling calibration scheme intended to mitigate under-coverage during distribution shift; we also position this approach relative to broader conformal developments such as adaptive conformal inference and conformalized quantile ideas (Angelopoulos & Bates, 2021; Gibbs & Cand`es, 2021; Romano et al., 2019). Third, we provide a comprehensive empirical assessment that reports standard point metrics (MAE, RMSE, MAPE) together with interval-quality metrics (PICP, MIW, and Winkler score), complemented by rolling drift analyses and a simple decision-support simulation that illustrates how interval bounds can be used for risk-aware thresholding across horizons.

2. LITERATURE REVIEW

Deep Learning for Multi-Horizon Time-Series Forecasting

Multi-horizon time-series forecasting has become a central problem in modern predictive analytics because many operational decisions require forecasts at multiple lead times simultaneously. Compared with single-step forecasting, multi-horizon settings introduce additional challenges: error propagation, horizon-dependent uncertainty growth, and the need to capture both fast local dynamics and slower long-range dependencies within the same model. In practice, forecasters must learn from multivariate inputs and, when available, time-derived covariates (e.g., hour-of day and seasonal indicators) to represent periodicities and regime changes. These requirements have led to a shift from traditional univariate statistical forecasting toward learning-based approaches that can exploit large collections of series and covariates, while still aiming for robustness across heterogeneous operating conditions (Makridakis et al., 2018; Smyl, 2020).

A widely adopted paradigm is probabilistic neural forecasting, where the model outputs a distribution (or parameters of a distribution) rather than a single point. DeepAR is a represen tative example that combines autoregressive recurrent modeling with likelihood-based training to produce probabilistic forecasts and has shown strong performance across diverse forecasting scenarios (Salinas et al., 2020). The practical relevance of such probabilistic approaches is reinforced by large-scale empirical evidence and competition results. The M4 competition, for instance, emphasizes that strong forecasting solutions often blend multiple sources of inductive bias and that hybrid thinking can outperform purely model-centric approaches (Makridakis et al., 2018). In a similar spirit, hybrid statistical neural methods demonstrate that combining classical smoothing intuition with neural sequence modeling can achieve competitive accuracy and robustness on challenging benchmarks (Smyl, 2020). Additionally, research on forecasting across large databases highlights that grouping or clustering similar series can improve learning efficiency and generalization, especially when the forecasting task spans multiple related patterns (Bandara et al., 2020). A broader perspective is provided by surveys on recurrent approaches for time-series forecasting, which summarize typical strengths (stable sequence modeling, flexible nonlinear dependencies) and limitations (difficulty with very long dependencies and distribution shift) (Hewamalage et al., 2021).

From an architectural standpoint, recurrent neural networks remain a strong baseline for sequence learning. In particular, gated recurrent units (GRU) provide a compact gating mechanism to manage long-range information flow and have been extensively used in sequence to-sequence contexts (Cho et al., 2014). In many forecasting pipelines, training stability and generalization depend as much on optimization and regularization as on architectural choice. Dropout is a standard regularization strategy to mitigate overfitting in deep networks (Srivastava et al., 2014), and Adam is a commonly used stochastic optimizer that improves convergence speed and robustness under noisy gradients (Kingma & Ba, 2015). While these techniques are not forecasting-specific, they are part of the practical toolkit that often determines whether a model trains reliably and transfers well to new regimes.

Beyond recurrent models, temporal convolutional networks (TCN) have gained attention as an effective alternative for sequence modeling. TCNs employ causal convolutions (to prevent future leakage) and dilations (to enlarge receptive fields efficiently), allowing them to capture local temporal motifs and mid-range structure with high parallelism and favorable computational properties (Bai et al., 2018). Empirical evaluations suggest that convolutional architectures can be competitive with recurrent networks for sequence modeling when designed with appropriate temporal constraints and receptive fields (Bai et al., 2018). Separately, N-BEATS proposes a deep residual architecture with basis expansions that can yield strong forecasting performance while offering interpretability through structured components (Oreshkin et al., 2020). These lines of work indicate that effective forecasting can be achieved by architectures that encode temporal inductive bias directly, without necessarily relying on attention mechanisms.

Transformer-style forecasting has recently progressed long-horizon forecasting performance by introducing efficiency improvements and time-series-specific inductive biases. Temporal Fusion Transformers (TFT) target interpretable multi-horizon forecasting by integrating gating, variable selection, and attention, enabling both predictive performance and feature attribution (Lim et al., 2021). Informer introduces attention approximations intended to scale to long sequences and popularizes the ETT benchmark family for long-sequence evaluation (H. Zhou et al., 2021). Autoformer introduces decomposition and auto-correlation mechanisms to model long-term periodicity and dependencies more effectively than vanilla attention (Wu et al., 2021), while FEDformer enhances efficiency by leveraging frequency-domain representations combined with decomposition (T. Zhou et al., 2022). In parallel, there is evidence that Transformers may not always be the best default choice for time-series forecasting. A systematic study argues that simpler components, including strong linear baselines often associated with DLinear-style designs, can be surprisingly competitive on several time-series benchmarks (Zeng et al., 2023). Patch-based designs such as PatchTST reduce effective sequence length by operating on patches rather than individual timestamps, improving efficiency and sometimes accuracy (Nie et al., 2023a). TimesNet emphasizes multi-scale temporal variation modeling through specialized temporal blocks aimed at general time-series analysis, including forecasting (Nie et al., 2023b). Taken together, these studies suggest that the best architecture choice depends on the trade-off between accuracy, complexity, and deployment constraints.

In contrast to heavyweight attention-based models, this paper focuses on a compact hybrid forecaster designed for efficient training and inference. We combine a causal dilated TCN as a local pattern extractor with a GRU refinement layer to capture longer dependencies. This hybridization is motivated by the complementary strengths of TCN (fast local motif extraction and receptive-field control) and GRU (temporal aggregation and stable sequence refinement) (Bai et al., 2018; Cho et al., 2014). The goal is not to compete on architectural novelty against large transformer families, but to establish a practical, lightweight baseline that remains strong for multi-horizon forecasting while being straightforward to implement, train, and deploy.

Uncertainty Quantification and Conformal Prediction

Uncertainty quantification (UQ) is essential for deploying forecasting systems in risk-sensitive en vironments. Even when average point accuracy is strong, operational decision-making frequently depends on confidence in the forecast: whether predicted values are likely to exceed a threshold, whether safety margins are adequate, or whether an intervention should be scheduled. Tradi tional UQ approaches include probabilistic modeling (e.g., specifying likelihoods and producing predictive distributions), Bayesian approximations, and ensembling, while regression-style ap proaches such as quantile regression provide predictive intervals by directly modeling conditional quantiles. Likelihood-based deep probabilistic forecasting exemplified by DeepAR provides an end-to-end distributional approach, but its reliability under distribution shift may depend on the appropriateness of the chosen likelihood and the stability of the learned conditional distribution under drift (Salinas et al., 2020). As a result, many deployments benefit from distribution-free UQ layers that can be attached post-hoc to existing forecasters.

Conformal prediction offers a principled distribution-free framework for predictive inference. The central idea is to calibrate prediction sets (or intervals) using a held-out calibration set and a nonconformity score, yielding finite-sample coverage guarantees under minimal assumptions (Angelopoulos & Bates, 2021; Vovk et al., 2005). For regression, a common and practical choice is residual-based conformal prediction, where absolute residuals serve as nonconformity scores and a suitable quantile of the calibration residuals determines the interval radius. This approach is model-agnostic: it can wrap around any point predictor, which makes it attractive when the forecaster is chosen for accuracy or efficiency rather than probabilistic modeling convenience (Angelopoulos & Bates, 2021). Related distribution-free techniques such as jackknife+ extend predictive inference by aggregating leave-one-out style predictions to form intervals with strong empirical behavior and theoretical guarantees in certain settings (Barber et al., 2021). These methods reinforce the broader point that reliable intervals can be constructed without specifying a parametric noise model.

However, time-series forecasting introduces additional structure that must be handled carefully to avoid leakage and to maintain practical relevance. Calibration must respect temporal order: calibration points should come after training and should not incorporate future

e-ISSN :2963-7805; p-ISSN :2963-8208, Hal. 426-450

information. Moreover, in multi-horizon forecasting, error distributions can vary substantially across horizons; longer horizons typically exhibit larger and more skewed errors, motivating horizon-wise or lead time-specific calibration rather than a single pooled interval. Conformalized quantile regression (CQR) provides an alternative that uses quantile predictors and then applies conformal calibration to improve empirical coverage while allowing asymmetric intervals, which can be beneficial when under- and over-forecast errors are not symmetric (Romano et al., 2019). Although CQR requires a quantile model, it provides a pathway toward better handling heteroskedasticity and asymmetry in forecast errors (Romano et al., 2019).

A key difficulty in practice is distribution shift. Static conformal prediction assumes that calibration residuals remain representative of test-time residuals. Under temporal drift, this assumption may fail, leading to systematic under-coverage (intervals too narrow) or over-coverage (intervals too wide). Work on conformal prediction under covariate shift studies settings where the distribution changes between calibration and test, highlighting that naive conformal coverage can degrade and motivating adaptation strategies (Tibshirani et al., 2019). Adaptive conformal inference proposes mechanisms to update calibration online or via sliding windows so that the nonconformity distribution tracks recent behavior, improving practical reliability under shift (Gibbs & Cand`es, 2021). From a deployment standpoint, such adaptive calibration matches the operational reality that ground truth eventually becomes available and can be used to adjust uncertainty estimates over time.

Motivated by these insights, this work adopts a time-series-specific conformal protocol that prioritizes leakage-safe chronological splitting (train/calibration/test) and horizon-wise absolute residual nonconformity. We evaluate both (i) static conformal, where quantiles are estimated once from the calibration residuals and kept fixed, and (ii) an adaptive rolling variant that updates residual quantiles using a sliding window of recent errors to better track temporal drift. This design keeps the uncertainty layer simple and model-agnostic while explicitly testing whether adaptive calibration improves empirical coverage and interval quality in non-stationary forecasting scenarios (Angelopoulos & Bates, 2021; Gibbs & Cand`es, 2021; Vovk et al., 2005).

3. PROPOSED METHOD

Problem Setup and Data

We study multivariate multi-horizon time-series forecasting under a fixed-length context window. At each reference time index t, the model observes a window of the most recent L timesteps with D input features: $x_{t-L+1:t} \in R^{L \times D}$. The goal is to predict the target variable at a set of future horizons H = {h1,...,hK} in a single forward pass, producing a vector of horizon-specific point forecasts:

$$\widehat{y_t} = \left[\widehat{y_{t+h_1}}, \widehat{y_{t+h_2}}, \dots, \widehat{y_{t+h_K}}\right]. \quad (1)$$

This formulation explicitly avoids autoregressive rollouts across horizons; instead, the model directly outputs all requested lead times, which reduces error accumulation and enables horizon wise assessment of both accuracy and uncertainty.

We use the ETTh1 dataset (hourly sampling) and forecast the oil temperature (OT) as the target variable. The lookback length is fixed to L = 336 (two weeks of hourly data), and we evaluate multi-horizon forecasts at H = {96,192,336,720}, corresponding to 4, 8, 14, and 30 days ahead. Inputs consist of all available sensor variables augmented with standard time features derived from timestamps (e.g., hour-of-day, day-of-week, and other periodic encodings), resulting in D = 15 total features after feature construction.

To prevent temporal leakage, we adopt a chronological split that respects the natural ordering of the time series: 60% train, 20% validation (used as calibration for conformal), and 20% test. All preprocessing steps that can leak information (notably feature scaling) are fitted strictly on the training partition and then applied to validation and test. Concretely, a scaler is fitted on the training split only, ensuring that normalization parameters do not incorporate future statistics. This protocol is particularly important because the paper explicitly studies distribution shift and drift; mixing future distributional information into training-time preprocessing would artificially reduce apparent drift.

Baselines Forecasters

To contextualize the proposed hybrid model, we implement two lightweight baselines that each represent a single modeling family. Both baselines ingest the same multivariate input window xt−L+1:t and output the K horizon predictions with a linear head.

t−L+1:t and output the K horizon predictions with a linear head. GRU-only baseline. The GRU-only model uses a stacked gated recurrent unit encoder that processes the input sequence in temporal order and compresses it into a hidden representation. We use the final hidden state as a summary of the lookback window and map it to a K dimensional output

e-ISSN :2963-7805; p-ISSN :2963-8208, Hal. 426-450

through a linear projection (Cho et al., 2014). This baseline captures temporal dependencies through recurrence and serves as a reference for long-range dependency modeling in a compact architecture.

TCN-only baseline. The TCN-only model uses a causal dilated Temporal Convolutional Network encoder that applies temporally causal 1D convolutions with increasing dilation factors to grow receptive fields efficiently while preserving the no-future-information constraint (Bai et al., 2018). The encoder produces a sequence of hidden features; we apply last-step pooling (or equivalently read the final position) to obtain a fixed-dimensional representation that is then mapped to K horizon predictions through a linear head. This baseline emphasizes local and mid-range motif extraction and provides a contrast to purely recurrent modeling.

Mid-range motif extraction and provides a contrast to purely recurrent modeling. For both baselines, we apply dropout in the recurrent or convolutional blocks to reduce overfitting and to improve generalization under limited training data and mild distribution shifts (Srivastava et al., 2014). Using the same input features, splits, and training protocol ensures that differences in performance can be attributed primarily to architectural inductive bias rather than preprocessing differences.

Hybrid TCN–GRU Forecaster

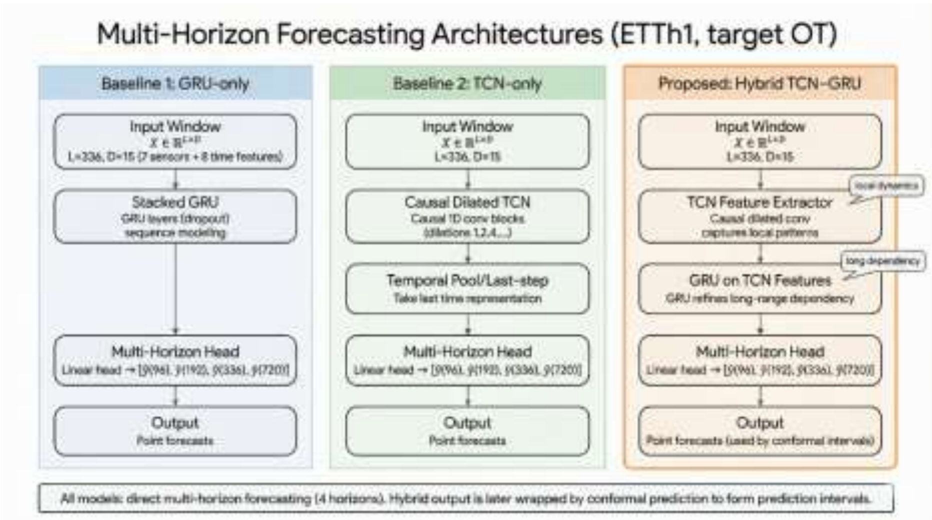

Figure 1. Multi-horizon forecasting architectures used in this study: GRU-only, TCN-only, and the proposed Hybrid TCN–GRU.

The proposed forecaster combines convolutional locality with recurrent refinement to balance efficiency and temporal expressivity. The design follows the principle that local temporal motifs (short-term trends, periodic fragments, transient disturbances) are often captured effectively by convolutional receptive fields, while recurrence can consolidate information over longer spans and provide a stable temporal aggregation mechanism. Figure 1 illustrates the three architectures evaluated in this study: GRU-only, TCN-only, and the proposed Hybrid TCN–GRU.

TCN feature extractor. Given $x_{t-L+1:t}$, the TCN stage applies a sequence of causal 1D convolutional blocks with exponentially increasing dilation factors. Causality ensures that features at each timestep depend only on current and past inputs; dilations expand the effective receptive field without requiring excessively deep stacks (Bai et al., 2018). Each block produces a refined feature sequence, yielding an encoded representation $z_{t-L+1:t} \in R^{L\times C}$, where C is the channel width.

GRU refinement. The TCN-produced feature sequence is then fed to a GRU layer, which refines and aggregates temporal information. The GRU outputs a final hidden state h_t that summarizes the input window through both the convolutional features and recurrent gating dynamics (Cho et al., 2014). Finally, a linear multi-horizon head maps h_t to K horizon outputs, producing the direct multi-horizon vector $\hat{y}_t$.

This hybridization is intentionally lightweight: it avoids attention layers and large memory footprints while still addressing two common needs in forecasting—(i) capturing local dynamics robustly and (ii) maintaining a mechanism to represent longer dependencies and temporal context. In our experiments, this architecture is trained and evaluated under the same protocol as the baselines to isolate the benefit of combining TCN and GRU in a single forecaster.

Training Objective and Optimization

We train all point forecasting models using the Mean Absolute Error (MAE) objective averaged across forecasting horizons. MAE provides a robust measure of point prediction error and is less sensitive to occasional large deviations compared to squared-error objectives. This property is particularly desirable in non-stationary time series, where temporary regime changes can produce outliers. Averaging the error across horizons encourages the model to learn representations that perform well for both short-term and long-term forecasts.

Optimization is performed using Adam (Kingma & Ba, 2015). To improve training stability and prevent exploding gradients, we apply gradient clipping with a maximum norm of 1.0, which is especially helpful for recurrent components. Early stopping is applied based on validation MAE with a patience of five epochs to reduce overfitting and select a checkpoint that generalizes well to unseen future time steps. Regularization is further supported by dropout in the model blocks (Srivastava et al., 2014). Importantly, early stopping and hyperparameter selection are conducted using only the validation set, ensuring the integrity of the test evaluation.

Cost-Sensitive Thresholding and Decision Curve

To support risk-aware decision-making, we augment point forecasts with prediction intervals using conformal prediction. Conformal methods provide distribution-free predictive inference by calibrating nonconformity scores on held-out data and choosing a quantile that targets a user-specified miscoverage level α (Angelopoulos & Bates, 2021; Vovk et al., 2005). In our setting, we adopt a time-aware protocol: training uses the earliest segment, calibration (validation) uses the subsequent segment, and testing uses the latest segment. This chronological structure respects temporal ordering and avoids look-ahead leakage.

Let the calibration set provide horizon-wise residual-based nonconformity scores:

$$r_{i,h} = \left|y_{i,h} - \widehat{y_{i,h}}\right|, \qquad (2)$$

where i indexes calibration samples and h ∈ H denotes a specific horizon. Using horizon-wise scores is important because forecast error distributions differ by lead time: longer horizons typi cally have larger and potentially different-shaped residual distributions. Calibrating separately per horizon allows intervals to expand appropriately as the horizon increases.

Static conformal. For each horizon h, we compute the (1 − α) quantile qh from the empirical distribution of {ri,h} on the calibration split. We then form symmetric prediction intervals on test:

$$\left[\widehat{y_{t,h}} - q_h, \widehat{y_{t,h}} + q_h\right]. \qquad (3)$$

This approach is simple, model-agnostic, and provides a clear interpretation: qh acts as a horizon-specific safety margin. However, if the residual distribution shifts over time, fixed qh may become miscalibrated, producing under-coverage (intervals too narrow) or over-coverage (intervals too wide).

Adaptive (rolling) conformal. To improve robustness under temporal drift, we consider an adaptive rolling calibration strategy inspired by conformal inference under distribution shift (Gibbs & Cand`es, 2021; Tibshirani et al., 2019). For each horizon, we maintain a rolling buffer of the most recent residuals (window size W) and update the quantile over time:

$$q_h(t) = \text{Quantile}_{1-\alpha}\left(\{r_{j,h}\}_{j=t-W+1}^{t}\right) \qquad (4)$$

This yields time-varying symmetric intervals $\left[\widehat{y_{t,h}} - q_h(t), \widehat{y_{t,h}} + q_h(t)\right]$. Operationally, this corresponds to recalibrating uncertainty using the most recent observed forecast errors, which becomes feasible in deployment once ground truth is observed after the forecast horizon elapses. The key benefit is responsiveness: when drift increases errors, qh(t) can grow, reducing under coverage; when the system stabilizes, intervals can shrink, improving sharpness.

Figure 2 summarizes the static vs. adaptive procedures, highlighting that both variants preserve chronological splitting while differing in whether calibration is fixed or updated over time. While our implementation uses symmetric residual-based intervals for simplicity, the approach is compatible with extensions such as conformalized quantile regression (CQR) when asymmetric intervals are desired (Romano et al., 2019).

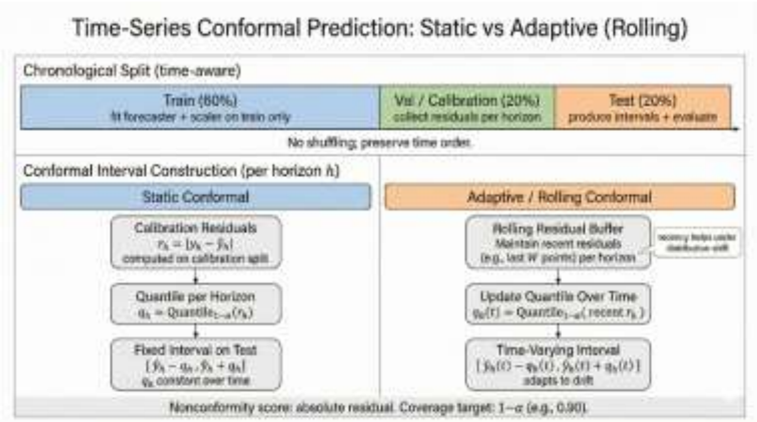

Figure 2. Time-series conformal prediction with chronological splitting. Static conformal uses f ixed horizon-wise quantiles from calibration, while adaptive (rolling) conformal updates quantiles using recent residuals.

Decision-Support Rule

To demonstrate downstream utility beyond headline metrics, we define a simple decision-support rule that uses prediction intervals to trigger an alarm when risk exceeds a threshold. Specifically, for a chosen horizon h⋆, we trigger an alarm at time t if the upper bound (UB) of the prediction interval exceeds a threshold T:

$$\mathbb{1}\{\mathrm{UB}_{t,h^\star} > T\}. \tag{5}$$

This rule reflects a common operational pattern: when actions are costly or safety-critical, it is often preferable to base decisions on conservative bounds rather than point estimates. Using the upper bound introduces a tunable trade-off between false alarms and missed events, controlled by both the threshold T and the interval width (which is influenced by α and the calibration strategy).

We report the resulting trade-off curve between missed events and false alarms by sweeping T over a meaningful range for the target variable. This simulation is intentionally simple and is not meant to fully model a domain-specific cost function; rather, it provides an interpretable illustration of how uncertainty-aware forecasting can alter operational decisions compared to point forecasting alone. In particular, interval calibration quality directly affects this trade-off: under-covered intervals can lead to unsafe decisions (missed events), while overly wide intervals can lead to excessive false alarms and operational inefficiency.

4. RESULT AND DISCUSSION

Experimental Setup

All experiments were conducted in Google Colab using a Tesla T4 GPU, with Python 3.12 and a CUDA-enabled PyTorch installation (2.9). We downloaded ETTh1 to local Colab runtime storage to ensure fast I/O during training and evaluation. Following a reproducibility-oriented workflow, we enforced a strict separation between (i) runtime-only training artifacts (e.g., intermediate checkpoints and temporary files) and (ii) paper-ready artifacts (tables, figures, and summary JSON/CSV) that were synchronized to Drive only after validation and test evaluation completed. This separation reduces the risk of accidentally mixing calibration/test information into training-time decisions and also ensures that the exact assets used in the manuscript can be traced back to a single run directory. We used a fixed lookback length $L = 336$ and multi-horizon targets $H = \{96,192,336,720\}$, corresponding to 4, 8, 14, and 30 days ahead for the hourly ETTh1 series. The dataset was split chronologically into 60% training, 20% validation, and 20% test. All scaling parameters were fitted on the training segment only and then applied unchanged to validation and test. This detail matters because ETTh1 exhibits temporal variability; fitting normalization on future segments would leak distributional information and artificially inflate performance. Validation served two roles: it supported early stopping for point forecasters and acted as the calibration set for static conformal prediction. The test segment was held out for the final report of point metrics and interval metrics.

For adaptive conformal prediction, we set the nominal miscoverage level to $\alpha = 0.1$ (90% target coverage). We used a rolling residual quantile window of $W = 1000$ to update horizon-wise quantiles over time, and we computed rolling evaluation metrics with a window size of 500. The rolling evaluation window provides a balance: it is large enough to smooth short-lived noise while still revealing meaningful changes in forecasting difficulty and coverage behavior across test-time regimes. Unless stated otherwise, the reported point metrics are computed in the original OT units (after inverse scaling), and interval metrics are computed consistently from the same de-normalized predictions and targets.

Main Results: Point Accuracy and Interval Quality

Table 1 reports the main results for the proposed hybrid forecaster equipped with adaptive conformal intervals on the ETTh1 test set. Averaged across horizons, the method achieves MAE=4.615 and RMSE=5.382 in OT units, indicating strong point forecasting accuracy for a compact architecture. In addition to point accuracy, the method provides uncertainty estimates through prediction intervals. Empirical coverage reaches PICP=0.870, which is below the nominal target of $1 - \alpha = 0.90$, and the mean interval width is MIW=15.267. This coverage shortfall is informative rather than merely a deficiency: it indicates that the residual distribution in the test segment is not fully matched by calibration behavior, and that even adaptive rolling calibration cannot perfectly compensate in all regimes under the current symmetric residual-based design.

Per-horizon behavior provides additional insight. Figure 3 shows that forecasting error increases with horizon overall, as expected, but the degradation is not strictly monotonic. In our evaluation, intermediate horizons (192–336) can exhibit lower error than the shortest horizon, which can happen in ETTh-style datasets when short-horizon dynamics are influenced by high frequency noise or short-term irregularities, while slightly longer horizons benefit from smoother trends. At the longest horizon (720), errors increase more clearly, reflecting the inherent difficulty of month-ahead forecasting under non-stationarity. Figure 4 indicates that coverage varies across horizons, highlighting why horizon-wise calibration is necessary: a single pooled interval radius would either under-cover at difficult horizons or over-cover at easier horizons. Finally, Figure 5 confirms that interval width increases for longer horizons, which is consistent with uncertainty growth and the need for wider margins to maintain reasonable coverage as lead time grows.

Overall, these results suggest that the hybrid point forecaster is accurate and that conformal intervals are informative, but interval reliability is still challenged by temporal drift. This moti vates the subsequent ablation and drift stability analysis, which clarifies how much improvement is attributable to the hybrid architecture versus the calibration strategy.

Table 1. Hybrid + adaptive conformal (FULL) results on ETTh1 test set (target OT).

H	MAE	RMSE	MAPE(%)	PICP	MIW
96	4.457	5.362	92.432	0.846	15.127
192	4.139	4.831	81.281	0.866	13.791
336	4.001	4.754	78.956	0.893	14.300
720	5.861	6.583	87.269	0.876	17.851
avg	4.615	5.382	84.985	0.870	15.267

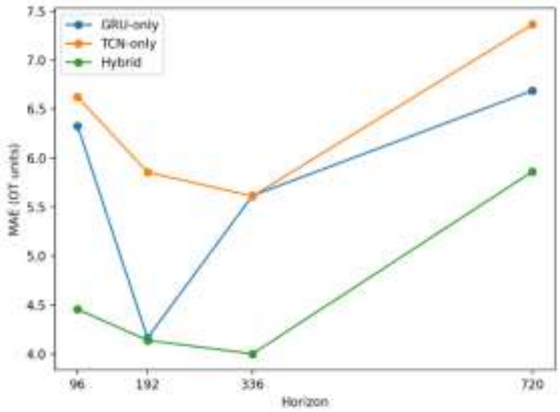

Figure 3. MAE per horizon for the hybrid + adaptive conformal model.

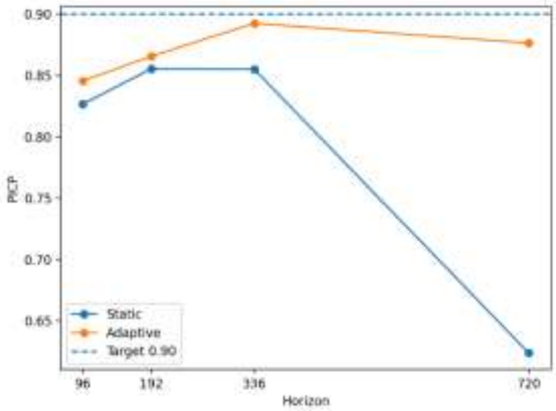

Figure 4. PICP (coverage) per horizon; target coverage is 1 − α = 0.90.

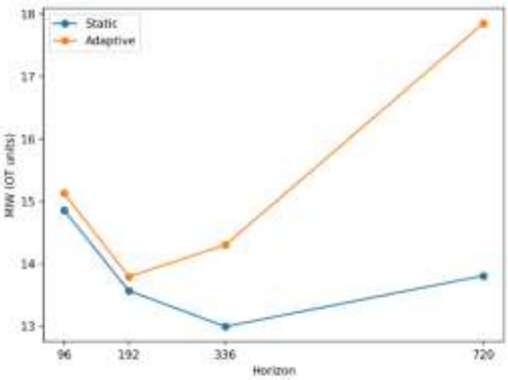

Figure 5. Mean interval width (MIW) per horizon.

Ablation and Baselines

Table 2 summarizes the ablation study, separating the effects of model architecture and conformal calibration. Comparing point forecasters without calibration, the proposed hybrid TCN–GRU achieves an average MAE of 4.615, improving over GRU-only (MAE=5.699) and TCN-only (MAE=6.363). In relative terms, this corresponds to approximately 19.0% reduction versus GRU-only and 27.5% reduction versus TCN-only. These margins are substantial given that all models use the same lookback length, horizons, input features, and data split, indicating that combining convolutional locality (TCN) with recurrent refinement (GRU) yields complementary benefits for ETTh1 multi-horizon forecasting.

Importantly, conformal calibration does not alter point metrics because it is applied after training as an uncertainty layer. The ablation therefore reports interval metrics only for the conformal variants. Static conformal under-covers on this split (PICP=0.790), suggesting that fixed quantiles estimated from the validation segment are too optimistic for certain parts of the test segment. Adaptive rolling conformal substantially increases empirical coverage (PICP=0.870), demonstrating that updating quantiles using recent residuals helps mitigate drift. This improvement comes with the expected cost of reduced sharpness: MIW increases from 13.80 (static) to 15.27 (adaptive). From a decision-support perspective, this trade-off is acceptable when the operational cost of under-coverage is high, because wider but more reliable intervals better reflect risk. However, the results also indicate that coverage remains below the nominal 0.90 target, suggesting either (i) residuals become heavier-tailed during some test periods than captured by rolling windows, (ii) symmetric residual-based intervals are insufficient under asymmetric error distributions, or (iii) the chosen window size W is not optimal for the drift timescale. These points inform the drift stability discussion below.

Table 2. Summarizes the Ablation Study.

Variant	MAE	RMSE	MAPE(%)	PICP	MIW	Winkler
GRU-only	5.699	6.415	103.236	-	-	-
TCN-only	6.363	7.029	114.284	-	-	-
Hybrid (no conformal)	4.615	5.382	84.985	-	-	-
Hybrid + static conformal	4.615	5.382	84.985	0.790	13.804	21.570
Hybrid + adaptive conformal (FULL)	4.615	5.382	84.985	0.870	15.267	19.569

Stability Under Drif

Aggregate metrics can hide temporal fragility, particularly in non-stationary time series where forecasting difficulty changes across regimes. To assess stability, we compute rolling MAE and rolling PICP on the test segment (Figure 6–Figure 7) using a window size of 500. Rolling MAE highlights time regions where prediction error increases, which can indicate regime changes, increased noise, or altered relationships among covariates. Rolling PICP complements this by measuring whether intervals remain reliable during these more difficult segments.

Figure 6 shows that forecasting error is not constant over the test period. The presence of elevated-error segments suggests that the model experiences drift-like behavior rather than a uniform i.i.d. test distribution. Figure 7 shows corresponding variation in coverage: when drift increases residual magnitude, fixed calibration tends to under-cover because intervals remain tied to an outdated residual distribution. Adaptive conformal aims to counter this by adjusting qh(t) as recent residuals grow or shrink. Empirically, adaptive calibration improves coverage stability in many segments, reflecting the intended mechanism of tracking recent nonconformity. Nevertheless, there may remain segments where coverage dips below the target, consistent with the overall average PICP=0.870. This observation motivates two practical implications: (i) rolling calibration is beneficial and should be preferred over static calibration in drift-prone deployments, and (ii) interval reliability can still be improved further by using more expressive nonconformity scores or asymmetric interval constructions when errors are skewed.

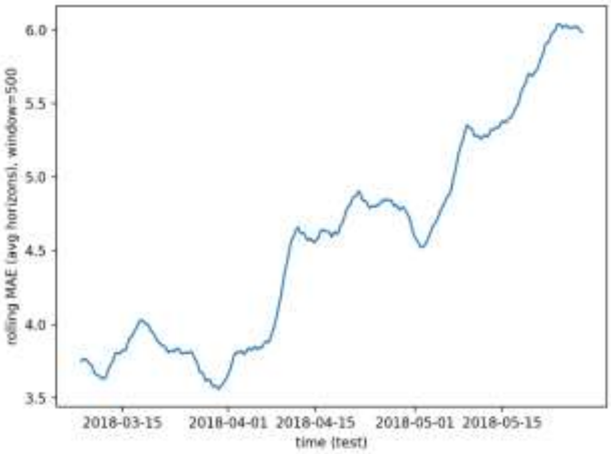

Figure 6. Rolling MAE on the test set (window=500).

Figure 7. Rolling PICP (coverage) on the test set (window=500).

Qualitative Examples and Decision Support

Quantitative metrics summarize performance but do not fully convey whether uncertainty behaves sensibly in individual forecast windows. Figure 8 provides representative forecast examples with conformal intervals on the test set. These examples serve as a sanity check for two properties that matter in practice: (i) intervals should widen when the underlying dynamics become less predictable or when forecast error is historically larger, and (ii) intervals should narrow when the system is stable and the model tracks the signal well. Visually inspecting a small set of windows is also valuable for identifying systematic failure modes, such as persistent bias (intervals shifted above or below the target) or missed rapid transitions (intervals that remain narrow despite abrupt changes).

Finally, we demonstrate downstream utility via a simple decision-support simulation. Figure 9 illustrates a trade-off curve produced by the rule “alarm if upper bound > threshold”. Using the upper bound rather than the point forecast implements a conservative policy that accounts for uncertainty: if the model is uncertain, the upper bound rises and alarms become more likely, which can reduce missed events at the cost of more false alarms. Even with imperfect average coverage (below the nominal 0.90), interval-based alarm policies offer a tunable mechanism for balancing risk according to operational preference. In contrast, point forecasts alone provide no principled way to adjust safety margins other than ad-hoc thresholds. This experiment is intentionally minimal, but it connects forecast uncertainty to an interpretable operational behavior and highlights why reliability (coverage and sharpness) is not merely a statistical detail but a practical requirement for decision-making pipelines.

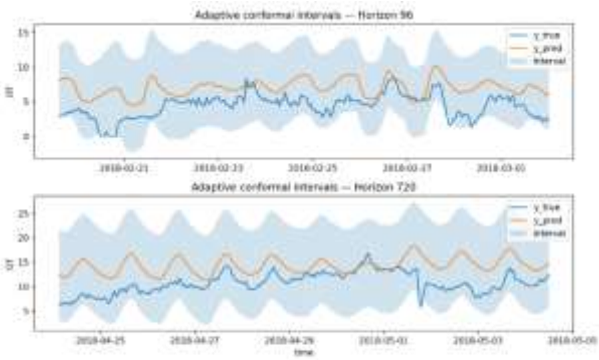

Figure 8. Forecast + Conformal Interval Examples on the Test Set (selected windows).

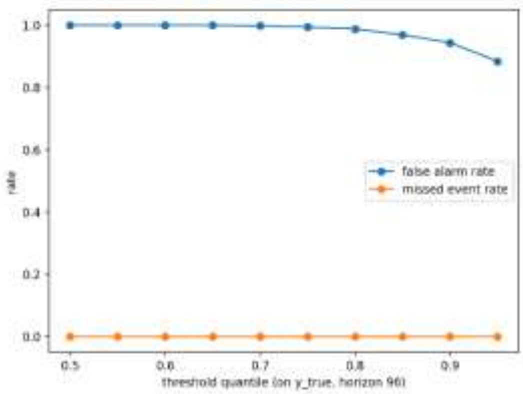

Figure 9. Decision-support curve based on thresholding the predictive upper bound.

Comparison

Comparison to state-of-the-art directions

Long-horizon forecasting has recently been dominated by attention-based architectures that explicitly target long sequences and exploit decomposition or frequency-domain structure. Informer proposes attention approximations designed to reduce the quadratic cost of vanilla self-attention and has been widely used as a reference model on the ETT family of benchmarks (H. Zhou et al., 2021). Autoformer introduces series decomposition and an auto-correlation mechanism that emphasizes periodic structure and long-term dependency modeling (Wu et al., 2021). FEDformer further explores frequency-enhanced decomposed modeling to improve efficiency and performance for long-term series (T. Zhou et al., 2022). These models represent a broader trend: injecting time-series-specific inductive bias into attention pipelines to make long-horizon forecasting feasible and accurate.

In parallel, patch-based and multi-scale designs aim to improve the inductive bias and computational behavior of transformer-like forecasters. PatchTST models sequences via patching, reducing the effective sequence length and enabling efficient long-context modeling (Nie et al., 2023a). TimesNet emphasizes multi-scale temporal variation modeling through specialized temporal blocks, reflecting the view that time series contain patterns at multiple resolutions that benefit from explicit multi-scale structure (Nie et al., 2023b). These directions are often strong on public benchmarks, and they highlight that architectural design choices beyond classical RNN/CNN families can yield notable gains. However, these gains typically come with increased implementation complexity (decomposition modules, attention variants, patching strategies), larger hyperparameter surfaces, and higher computational overhead, which can be a practical barrier for deployment-oriented settings where retraining and monitoring must be frequent and inexpensive.

Our work deliberately targets a different point on this accuracy–complexity spectrum. Rather than optimizing for maximal benchmark performance under large attention stacks, we aim for a compact hybrid forecaster that remains straightforward to train, debug, and deploy. The proposed TCN–GRU design combines causal dilated convolutional feature extraction with gated recurrent refinement, leveraging complementary inductive biases while maintaining a lightweight compute profile. Empirically, this choice improves over both GRU-only and TCN-only baselines on ETTh1 under the same data split and training procedure, suggesting that hybridization captures useful structure without resorting to heavyweight attention-based modeling.

From the reliability perspective, many state-of-the-art forecasting studies focus primarily on point accuracy and long-horizon error reduction, while uncertainty estimation is often treated separately or requires additional modeling assumptions. In contrast, our pipeline explicitly attaches distribution-free uncertainty to point forecasts through conformal prediction, which can be applied to any forecaster and provides an interpretable miscoverage control knob (Angelopoulos & Bates, 2021). Related predictive inference approaches such as jackknife+ further emphasize that distribution-free intervals can be obtained without specifying a parametric noise model, and can provide strong empirical behavior in practice (Barber et al., 2021). Importantly, because distribution shift is common in real deployments, we consider adaptive calibration as a practical extension: adaptive conformal inference under shift motivates updating calibration using recent residuals to reduce systematic under-coverage when the residual distribution drifts (Gibbs & Cand`es, 2021; Tibshirani et al., 2019). This reliability layer complements, rather than replaces, architectural advances: even high-performing point forecasters can produce overconfident predictions under drift, so post-hoc calibration remains relevant regardless of the backbone model family.

Overall, the comparison suggests that our method should be interpreted as a deployment oriented recipe rather than an attempt to “beat” the latest transformer-based forecasters. Transformer-based directions represent a strong frontier for long-horizon accuracy, while our approach emphasizes a pragmatic combination of compact modeling and distribution-free uncertainty that can be adopted with minimal engineering overhead.

Limitations

Two limitations are particularly important for correctly interpreting the results. First, we do not provide a direct head-to-head benchmark against the latest transformer-based state-of-the-art models (e.g., Informer/Autoformer/FEDformer or more recent patch/multi-scale designs) under an identical training protocol. As a consequence, the empirical claims in this paper are restricted to the implemented baselines and ablations: we show that the hybrid TCN–GRU improves over GRU-only and TCN-only within our controlled setup, but we do not claim superiority over specialized transformer families (Nie et al., 2023a, 2023b; Wu et al., 2021; H. Zhou et al., 2021; T. Zhou et al., 2022). A careful SOTA comparison would require standardized hyperparameter tuning budgets, consistent data preprocessing, and comparable training regimes, which is beyond the current scope.

Second, while adaptive conformal calibration improves coverage relative to static calibration, the empirical coverage (PICP) remains below the nominal target (0.90 on average with $\alpha = 0.1$). This suggests that the chosen rolling window and symmetric absolute-residual nonconformity score may be insufficient under certain drift regimes. Several factors may contribute: residuals can be horizon-dependent and heavy-tailed, drift can occur faster than the window can adapt, and symmetric intervals may be suboptimal when forecast errors are skewed. A promising direction is to use asymmetric intervals via conformalized quantile regression (CQR), which can better accommodate heteroskedasticity and asymmetry while retaining conformal calibration (Romano et al., 2019). Another direction is to explore more drift-aware calibration strategies, such as weighting recent residuals more strongly or choosing window sizes dynamically based on detected regime changes, as motivated by adaptive conformal inference under shift (Gibbs & Cand`es, 2021; Tibshirani et al., 2019).

Practical Implications

Despite these limitations, the proposed pipeline provides a practical end-to-end recipe for multi-horizon forecasting with uncertainty that aligns with deployment constraints. First, the chronological train/calibration/test protocol makes leakage risks explicit and encourages evaluation that reflects real-world temporal ordering. Second, the model backbone is intentionally compact and therefore compatible with frequent retraining, ablation-driven debugging, and resource-constrained deployment environments. Third, conformal prediction offers a transparent and tunable uncertainty layer: the user selects a nominal miscoverage level, and the system returns prediction intervals whose width directly reflects recent forecast error behavior (Angelopoulos & Bates, 2021).

From an operational perspective, interval outputs can be translated into simple, interpretable decision rules. Thresholding the upper bound provides a conservative policy that accounts for uncertainty and can reduce missed events when underestimation is costly; conversely, thresholding the lower bound can support policies that prioritize avoiding unnecessary interventions when overestimation is costly. Because false alarms and missed events typically carry asymmetric costs, interval-based rules allow practitioners to choose a risk posture explicitly rather than relying on point forecasts alone. Even when empirical coverage is imperfect, the presence of an uncertainty layer provides a principled mechanism for safety margins and for monitoring reliability over time, particularly when combined with rolling evaluation that reveals drift episodes. In this sense, the main practical value of the pipeline is not solely the achieved point accuracy, but the clarity of the forecasting-to-decision interface: a lightweight forecaster, a leakage-safe calibration protocol, and interpretable uncertainty-driven rules that can be audited and iteratively improved.

5. CONCLUSIONS

This paper presented a deployment-oriented, lightweight pipeline for multi-horizon forecasting with uncertainty on the ETTh1 benchmark (target OT). The proposed approach combines a compact Hybrid TCN–GRU point forecaster with a time-aware conformal prediction layer to produce prediction intervals under a leakage-safe chronological protocol. The overall goal was not to maximize architectural novelty, but to establish a practical recipe that balances accuracy, simplicity, and reliability under temporal drift—three properties that often matter more than marginal gains in benchmark point metrics when the system is intended for real operational decision support.

Empirically, across four horizons (96/192/336/720), the hybrid TCN–GRU architecture consistently improves point accuracy over single-family baselines (GRU-only and TCN-only) under the same preprocessing and training procedure. This supports the hypothesis that convolutional locality and recurrent refinement provide complementary inductive biases for multivariate forecasting: the TCN stage efficiently extracts local temporal patterns and mid-range structure, while the GRU stage consolidates information and refines longer-range dependencies into a compact representation for multi-horizon output. The ablation study further clarifies that the gains are attributable to the hybridization rather than calibration, since conformal prediction operates post-hoc and does not alter point forecasts.

For uncertainty estimation, we evaluated static conformal prediction and an adaptive rolling variant designed to track drift by updating horizon-wise residual quantiles using recent errors. Static conformal under-covers in this non-stationary setting, indicating that a fixed calibration quantile is insufficient when the residual distribution changes over time. Adaptive rolling conformal substantially improves empirical coverage, and rolling/qualitative analyses show that intervals tend to widen during harder test segments where the model error increases. Although average coverage remains below the nominal target in our setup, the results still demonstrate that adaptive calibration is a meaningful and practical step toward reliability under drift, and that uncertainty-aware outputs convey information that point forecasts alone cannot provide.

Finally, we connected forecasting outputs to an interpretable downstream mechanism through a simple decision-support simulation based on thresholding the predictive upper bound. This experiment illustrates the practical value of prediction intervals as a control knob for risk posture: interval-based alarms enable a tunable trade-off between false alarms and missed events, which is typically more actionable than relying on point forecasts alone. Taken together, the findings highlight that a lightweight forecasting backbone paired with a distribution-free uncertainty layer can serve as a solid baseline for risk-aware forecasting systems where retraining, monitoring, and interpretability are important.

Several directions remain for future work. First, a standardized benchmark comparison against recent transformer-based forecasters under the same tuning budget and training protocol would clarify the accuracy–complexity trade-off more conclusively. Second, interval reliability can be improved by moving beyond symmetric absolute-residual scores, for example by adopting asymmetric or conditional conformal methods such as conformalized quantile regression to better handle skewed and heteroskedastic errors. Third, adaptation under severe or rapidly changing distribution shift can be strengthened by more drift-aware calibration strategies, including dynamic window sizing, weighted residual updates, or regime-change detection coupled with recalibration and retraining triggers. Fourth, downstream evaluation should move toward domain-specific decision objectives that reflect real operational costs, enabling optimization and validation of forecasting systems directly in terms of the decisions they are meant to support rather than generic accuracy metrics alone.

REFERENCES

Angelopoulos, A. N., & Bates, S. (2021). *A gentle introduction to conformal prediction and distribution-free uncertainty quantification*. *arXiv Preprint arXiv:2107.07511*.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). *An empirical evaluation of generic convolutional and recurrent networks for sequence modeling*. *arXiv Preprint arXiv:1803.01271*.

Bandara, K., Bergmeir, C., & Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *International Journal of Forecasting*, *36*(4), 1227–1239. https://doi.org/10.1016/j.eswa.2019.112896

Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics*, *49*(1), 486–507. https://doi.org/10.1214/20-AOS1965

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.3115/v1/D14-1179

Gibbs, I., & Candès, E. (2021). *Adaptive conformal inference under distribution shift*. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Kingma, D. P., & Ba, J. (2015). *Adam: A method for stochastic optimization*. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802–808. https://doi.org/10.1016/j.ijforecast.2018.06.001

Nie, Y., et al. (2023a). *A time series is worth 64 words: Long-term forecasting with transformers*. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Nie, Y., et al. (2023b). *TimesNet: Temporal 2D-variation modeling for general time series analysis*. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2020). *N-BEATS: Neural basis expansion analysis for interpretable time series forecasting*. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Romano, Y., Patterson, E., & Candès, E. (2019). *Conformalized quantile regression*. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*(1), 75–85. https://doi.org/10.1016/j.ijforecast.2019.03.017

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958.

Tibshirani, R. J., Foygel Barber, R., & Candès, E. (2019). *Conformal prediction under covariate shift*. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.

Wu, H., Xu, J., Wang, J., & Long, M. (2021). *Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting*. In *Advances in Neural Information Processing Systems (NeurIPS)*.