University of Crete
Department of Computer Science

# Conformal Prediction for Reliable Uncertainty Quantification in Chronos Time-Series Forecasting

**Author:** Kyriaki Kiouri Kyparisi
**Student ID:** 4906
**Supervisor:** Prof. Grigorios Tsagkatakis
**Academic Advisor:** Prof. Panagiotis Tsakalides

Irakleio, Crete, 29/10/2025

**Abstract**

Reliable uncertainty quantification is essential for time-series forecasting, where prediction intervals guide critical decisions in energy, retail, and healthcare. While pretrained models like Chronos-T5 achieve strong predictive accuracy, their probabilistic forecasts are often miscalibrated, meaning that the stated confidence levels do not reflect true empirical coverage. This thesis addresses that limitation by applying Conformalized Quantile Regression (CQR), a distribution-free, post-processing calibration method, to improve the reliability of Chronos' forecast intervals.

Using the Chronos-T5 Small model in zero-shot mode on the M4 Daily benchmark dataset, forecasts were generated through the `chronos-forecasting` and `Forecasting Evaluation Framework (FEV)` libraries and recalibrated via a manually implemented CQR procedure. Evaluation metrics, such as Prediction Interval Coverage Probability (PICP), Interval Calibration Error (ICE), Sharpness, Winkler Interval Score, and Pinball Loss, show that conformal calibration raises empirical coverage from 0.74 to 0.86 and reduces ICE from 0.115 to 0.036, while only modestly widening intervals and preserving median-forecast accuracy. These results demonstrate that a lightweight, model-agnostic conformal layer can transform pretrained time-series forecasters into sources of statistically valid, trustworthy uncertainty estimates, enabling more reliable decision-making without retraining the underlying model.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Over recent years, forecasting errors propagate to resource-allocation decisions in retail, energy, and health sectors. Reliable forecasts enable organizations to plan ahead, allocate resources, and make informed decisions. As data has become more abundant and complex, the need for models that can capture patterns across diverse domains has become stronger than ever.

Traditional statistical methods such as ARIMA and ETS remain widely used due to their interpretability and robustness in stable environments. However, they struggle to capture nonlinear dependencies and high-dimensional patterns. Advances in machine learning and deep learning have introduced more flexible models, and recent work has explored pretrained architectures that aim to perform well across many datasets without retraining. These developments open the door to more general and powerful forecasting systems.

This thesis focuses on Chronos[1], a pretrained model for time-series forecasting, and addresses one of its key limitations: the miscalibration of its probabilistic forecasts. While Chronos can produce point forecasts and quantile intervals, the reported coverage levels often deviate from their nominal values, reducing their reliability in practice. To address this issue, the thesis applies conformal prediction[6] to calibrate Chronos' quantile forecasts. The objective is to evaluate whether calibration can ensure valid coverage across multiple datasets, and to compare performance before and after calibration. By doing so, the study aims to demonstrate how pretrained forecasting models can provide more trustworthy and actionable uncertainty estimates.

# Chapter 2

# Literature Review

## 2.1 Intro to ML in Time-Series Forecasting

Time-series forecasting has traditionally been addressed with statistical models such as ARIMA, ETS, and the Theta method. These approaches remain important baselines due to their interpretability and robustness, but they are limited in handling nonlinear relationships, long-term dependencies, and cross-series patterns.

The availability of larger and more complex datasets has shifted attention toward machine learning and deep learning methods. Early neural approaches such as RNNs and LSTMs enabled models to learn sequential dependencies directly from data, while CNN-based and temporal convolutional architectures captured hierarchical structures. More recently, Transformer-based models have become central to forecasting, leveraging attention mechanisms to model long-range dependencies and heterogeneous covariates.

Despite these advances, most deep learning forecasters are trained on individual datasets, which restricts their portability across domains. By contrast, natural language processing has shown that large pretrained models trained on massive, diverse corpora (e.g., GPT) can be applied to new datasets with little or no extra training. Inspired by this, researchers have begun to explore pretrained approaches for time series that aim to perform well in both zero-shot and fine-tuned settings.

Against this backdrop, state-of-the-art forecasting methods can be broadly grouped into three categories: local statistical models, task-specific neural models, and pretrained models[3]. The following sections review the main characteristics of each category.

## 2.2 State-of-the-Art Time-Series Forecasting Methods

### 2.2.1 Local Models

Local statistical models remain an important reference point in time-series forecasting. They estimate parameters for each time series independently, relying on well-understood statistical theory. Among the most widely used are ARIMA, ETS, and the Theta method, which form the basis of many forecasting competitions and applications. These models are valued for their simplicity and transparency, as their assumptions about trend, seasonality, and residuals can be explicitly examined. They often provide robust short-term forecasts in stable contexts with limited data.

However, their inability to share information across multiple series restricts their usefulness in modern large-scale applications. Recent evaluations (Hyndman & Athanasopoulos, 2018) confirm that while these methods remain strong baselines, they struggle with complex nonlinear dynamics and high-dimensional covariates. For this reason, they are increasingly complemented—or replaced—by machine learning and deep learning approaches that can learn from diverse collections of time series.

### 2.2.2 Task-Specific Neural Models

Task-specific neural models are trained or fine-tuned for individual datasets, enabling them to capture complex nonlinear dependencies and multi-scale temporal patterns. Examples include DeepAR, which employs recurrent neural networks to generate probabilistic forecasts, and N-BEATS and N-HiTS, which use feed-forward residual blocks to achieve strong accuracy in univariate tasks. Transformer-based methods such as the Temporal Fusion Transformer (TFT) and PatchTST leverage attention mechanisms to capture long-range dependencies and heterogeneous covariates.

A recent comparative study (Murray et al., 2023) evaluated architectures including LSTMs, GRUs, CNN-LSTMs, ConvLSTMs, and Transformers across five benchmark datasets. Their findings showed that convolution-enhanced recurrent models and Transformer variants consistently ranked among the top performers, especially for longer prediction horizons. The study also compared direct versus iterative forecasting strategies, concluding that iterative methods often yield lower errors in multi-step forecasts. While powerful, these models require dataset-specific training, limiting their portability across domains.

### 2.2.3 Pretrained Models

Pretrained models represent the newest direction in time-series forecasting. Instead of training on individual datasets, they leverage large and diverse corpora to learn general temporal representations that can be applied across tasks. Lag-Llama (Rasul et al., 2023) demonstrates that pretraining on multivariate time series enables strong transfer performance across domains. Moirai (Woo et al., 2024) introduces a hierarchical architecture that captures both short- and long-term dynamics and shows robust performance on multiple benchmarks. More recently, TimeGPT has been proposed as a commercial forecasting API, but its proprietary nature limits its accessibility for research and reproducibility.

Among the publicly available models, Chronos-T5 small (46 M parameters) is trained on 42 benchmark datasets, supports point and quantile forecasts, and balances accuracy with computational cost[5]. Given these qualities, Chronos is examined in detail in the following section.

### 2.2.4 Chronos models

Chronos is released in multiple sizes, ranging from Chronos-T5 Mini (20M parameters) to Chronos-T5 Large (710M). These variants provide a balance between accuracy and computational efficiency: larger models achieve the strongest results, while smaller ones are suitable for deployment in constrained environments. This thesis focuses on Chronos-T5 Small (46M parameters), which offers a favorable balance between forecasting accuracy and computational cost. Experiments in the Chronos paper[1]. show that Chronos-T5 Small, when fine-tuned, surpasses larger Chronos variants in zero-shot mode and outperforms the best task-specific baselines on Benchmark II (zero-shot evaluation) datasets .

Chronos models are trained on a mix of real datasets from domains such as finance, healthcare, energy, and climate, combined with synthetic augmentations such as TSMixup and KernelSynth. These augmentations diversify the training corpus and improve robustness on unseen datasets. Benchmark results confirm that performance scales with model size, but Chronos-T5 Small is particularly practical in research environments where efficiency is critical.

Chronos was chosen for this study because it provides native support for quantile forecasting, enabling direct estimation of predictive intervals that are essential for uncertainty quantification. Furthermore, unlike some pretrained alternatives such as TimeGPT, Chronos is fully open-source, with released model weights across all sizes, ensuring reproducibility and trans-

parency. These qualities (reproducibility, benchmark coverage, and compatibility with uncertainty calibration), make Chronos the most suitable choice for this thesis.

## 2.3 Conformal Prediction for Uncertainty Quantification

Conformal Prediction (CP) is a statistical framework that constructs finite-sample intervals that contain the future observation with probability 1-$\alpha$ under the exchangeability assumption. Unlike probabilistic models that depend on distributional assumptions, CP requires only that data points are exchangeable, making it a powerful, model-agnostic tool for uncertainty quantification. The core idea is to use past data to calibrate a model's residuals, thereby constructing intervals that contain future observations with a chosen confidence level.

Several variants of CP have been developed to address different settings. Split Conformal Prediction (SCP) divides the data into training and calibration sets, producing intervals around a model's predictions. Conformalized Quantile Regression (CQR, as proposed by Romano et al. [4]) extends the method to quantile forecasts by correcting model-estimated lower and upper quantiles. More advanced approaches, such as Ensemble Batch Prediction Intervals (EnbPI) and adaptive methods for time series, refine CP for sequential or dependent data. A growing body of tutorials, reviews, and libraries (e.g., MAPIE, TorchCP) makes these methods increasingly accessible.[6]

In the context of time-series forecasting, CP is particularly relevant because many models, including Chronos, can output quantiles that are not calibrated. A nominal 90% prediction interval may contain the true value far less often. By applying this to adjust these intervals, it is possible to achieve coverage that matches the desired confidence level while keeping intervals as sharp as possible. This thesis leverages CP to calibrate Chronos' forecasts, ensuring that its uncertainty estimates are statistically valid and more trustworthy for decision-making.

The theoretical guarantees of Conformal Prediction rely on the assumption that calibration and test samples are exchangeable; that is, their joint distribution is invariant under permutation. In practice, this assumption is often violated in sequential or temporal data, where observations exhibit autocorrelation. For time-series forecasting, exchangeability can be approximated by carefully selecting non-overlapping calibration and test windows, yet residual dependence may remain. Recent research ( Zaffran et al., 2022 [8] ) extends conformal methods to dependent or online settings, using weighted

or adaptive schemes that maintain approximate coverage over time. While this thesis assumes approximate exchangeability across the M4 Daily series, future work could investigate adaptive or sequential conformal approaches better suited to temporally dependent data.

# Chapter 3

# Methodology

## 3.1 Research Approach

The research follows a two-stage experimental design. We test whether CQR can raise the empirical PICP of Chronos-T5 small from its pre-calibration value to the nominal 0.90 while minimizing interval width (sharpness). The first stage focuses on establishing a baseline using Chronos, a pretrained Transformer model that provides probabilistic forecasts in the form of quantile predictions. In the second phase, conformal prediction is implemented using CQR to calibrate the quantile forecasts, thereby guaranteeing statistically valid coverage. The overall workflow consists of three main components:

1. **Forecast Model Version**

   Chronos-T5 Small is used in zero-shot mode on the M4 Daily dataset. Historical data from multiple time series are provided as input, and the model outputs quantile forecasts at levels 0.05, 0.50, and 0.95 over a 14-step prediction horizon. These quantiles represent the model's estimated uncertainty before any calibration is applied.

2. **Evaluation of Uncalibrated Forecasts**

   The raw Chronos forecasts are compared against ground truth values to assess their probabilistic accuracy. Evaluation metrics include the Prediction Interval Coverage Probability (PICP), Interval Calibration Error (ICE), Sharpness (mean interval width), Interval Score (Winkler), and Pinball loss at $\tau = 0.5$. These metrics collectively capture both calibration and sharpness, forming the baseline for later comparison.

3. **Calibration and Re-evaluation**

In the next stage, we will adjust Chronos' quantiles using conformalized quantile regression. The recalibrated forecasts will then be evaluated using the same metrics, allowing a direct before-and-after comparison of calibration quality.

## 3.2 Data Collection

The experiments in this study use the M4 Daily subset from the `autogluon/chronos_datasets` repository, a publicly available benchmark widely used for evaluating forecasting models. Each series represents daily observations, and the task involves predicting the next 14 time steps, corresponding to a forecast `horizon = 14`.

The dataset is accessed through the Forecasting Evaluation Framework (FEV), which provides standardized tools for dataset loading, temporal window generation, and ground-truth extraction. Using FEV ensures that all experimental runs are reproducible and aligned with the official Chronos benchmarks.

From this dataset, windows of past observations were extracted using the `fev.Task` interface. The Forecasting Evaluation Framework (FEV) was configured with a single evaluation window (`num_windows=1`), corresponding to one temporal split per time series. For this window, the framework provides two aligned datasets for each series: (i) `past_data`, containing the historical target values and covariates used as model input, and (ii) `future_data`, containing the timestamps and known features for the 14-step forecast horizon (without target values). The corresponding ground-truth values for the forecast horizon were retrieved separately using the `window.get_ground_truth()` function.

The evaluation uses all 4 227 time series available in the M4 Daily dataset, each contributing a single forecast window of 14 days. No additional preprocessing or normalization steps were applied beyond those internally handled by Chronos, preserving the integrity of the zero-shot evaluation setup and ensuring comparability across experiment.

## 3.3 Tools and Techniques

All experiments were conducted in Python 3.10 using the Google Colab High-RAM environment equipped with an NVIDIA L4 GPU, which provided the necessary computational capacity for large-scale forecasting and array operations. The implementation relied on the `chronos-forecasting` library

(Amazon Science)[1] to load and run the pretrained Chronos-T5 Small model in zero-shot mode, meaning that the model was applied directly to the M4 Daily dataset without any task-specific fine-tuning. The fev (Forecasting Evaluation Framework) package was employed to manage dataset access, window generation, and ground-truth retrieval, ensuring reproducibility and consistency with Chronos benchmarks.

Supporting libraries including `pandas` for data handling and `matplotlib` for visualization were used. `NumPy` played a crucial role in the implementation of the manual Conformalized Quantile Regression (CQR) procedure and related numerical calculations. The entire workflow was executed in Google Drive mounted directories to maintain version control and experiment reproducibility. The Chronos model weights were automatically retrieved from the Hugging Face repository under the identifier `amazon/chronos-t5-small`.

## 3.4 Uncertainty Quantification Strategy

To assess the reliability of the probabilistic forecasts produced by the Chronos-T5 Small model, we designed two complementary experimental stages aimed at quantifying and improving predictive uncertainty. Uncertainty was evaluated through the following methods:

- **Baseline Forecast Evaluation:** In the first stage, the uncalibrated forecasts generated by Chronos were analyzed to measure how accurately the model's native quantile predictions reflect true uncertainty. Using the 0.05, 0.50, and 0.95 quantiles produced for each 14-step forecast horizon, we compared the predicted intervals against the ground-truth values obtained from the M4 Daily dataset. This evaluation establishes the model's baseline calibration and serves as a reference for the subsequent calibration phase.

- **Conformal Calibration:** The second stage applies CQR to adjust Chronos' lower and upper quantiles. By introducing a separate calibration set, we modified the forecast intervals so that the empirical coverage of the predictions aligns with the desired nominal confidence level, thereby producing statistically valid and better calibrated uncertainty estimates.

- **Key Metrics for Quantifying Uncertainty:** To quantify predictive uncertainty and calibration across both stages, several complementary metrics are employed:

[1] `https://github.com/amazon-science/chronos-forecasting`

- *Prediction Interval Coverage Probability (PICP):* Measures the proportion of true observations falling within the nominal prediction interval (e.g., 90%), indicating empirical coverage. (Chronos provides quantile forecasts at levels 0.05, 0.50, and 0.95, corresponding to a nominal 90% prediction interval ($\alpha = 0.1$). The empirical coverage (PICP) therefore measures how often the ground-truth values fall within these 90% intervals.)
- *Sharpness:* Evaluates the mean width of the prediction intervals, where narrower intervals denote higher confidence.
- *Interval Calibration Error (ICE):* Captures the deviation between observed and nominal coverage, reflecting miscalibration.
- *Interval Score (Winkler):* Combines interval width and coverage penalties into a single proper scoring rule that balances calibration and sharpness.
- *Pinball Loss ($\tau = 0.5$):* Assesses the accuracy of the median forecast, corresponding to the model's central tendency.

These metrics allow for a comprehensive assessment of both the accuracy and the quality of Chronos' uncertainty quantification. The results of the uncalibrated baseline will later be compared with the conformally calibrated forecasts to demonstrate the effectiveness of the CQR adjustment.

## 3.5 Conformal Calibration Implementation

To improve the reliability of the quantile forecasts produced by the Chronos-T5 Small model, Conformalized Quantile Regression (CQR) was implemented as a post-processing calibration step. This procedure adjusts the model's lower and upper quantiles so that the empirical coverage of the prediction intervals matches the desired nominal confidence level of 90% ($\alpha = 0.10$).

**Dataset split:** The dataset was divided into calibration and test subsets based on unique series identifiers to ensure that no series appeared in both sets. An 80/20 split was applied, with 80% of the series allocated to the calibration set and the remaining 20% reserved for testing. This series-based partitioning enables evaluation of calibration performance on entirely unseen series.

Table 3.1: Dataset configuration for calibration and test subsets.

| | **Calibration** | **Test** |
|---|---|---|
| number of series | 3381 | 846 |
| Forecast horizon | 14 | 14 |
| Quantiles used | 0.05, 0.95 | 0.05, 0.50, 0.95 |

Each of the $N_{\text{cal}} = 3381$ calibration series produced one residual sample per forecast horizon ($h \in \{1, \ldots, 14\}$), resulting in 3 381 calibration scores for every horizon step.

**Nonconformity scores:** For each prediction sample $i$, the element-wise nonconformity score $s_i$ quantifies the extent to which the true observation $y_i$falls outside the predicted quantile interval $[\hat{q}_{\text{lo}}(x_i), \hat{q}_{\text{hi}}(x_i)]$. The score is defined as the maximum of 0 and the deviation of $y_i$ from the interval boundaries:

$$s_i = \max(0, \max(\hat{q}_{\text{lo}}(x_i) - y_i,\ y_i - \hat{q}_{\text{hi}}(x_i)))$$

where $\hat{q}_{\text{lo}}(x_i)$ and $\hat{q}_{\text{hi}}(x_i)$ are the Chronos-predicted lower (0.05) and upper (0.95) quantiles for input $x_i$. This ensures the nonconformity score is always non-negative, representing the distance from the interval when outside, and 0 when inside.

**Conformal quantile computation.** The conformal adjustment value $q_\alpha$ for a given miscoverage rate $\alpha$ is obtained as the empirical $(1-\alpha)$-quantile of the calibration scores $\{s_i\}$:

$$q_\alpha = \text{Quantile}_{1-\alpha}\big(\{s_i\}\big). \tag{3.1}$$

To ensure valid coverage in finite samples, the empirical $(1-\alpha)$-quantile is computed following the standard conformal correction rule. This involves calculating the index $k$ according to the formula:

$$k = \big\lceil (n+1)(1-\alpha) \big\rceil - 1$$

and then selecting the element at this index $k$ in the sorted list of calibration scores $\{s_i\}$ as the value for $q_\alpha$. Here, $n = N_{\text{cal}} = 3381$ denotes the number of calibration samples.

**Per-horizon calibration:** Calibration was performed separately for each forecast horizon $h$. For each $h$, the scores $s_{i,h}$ were pooled across all calibration series to compute a single padding value $q_{\alpha,h}$:

$$q_{\alpha,h} = \text{Quantile}_{1-\alpha}\big(\{s_{i,h}\}_{i=1}^{N_{\text{cal}}}\big). \tag{3.2}$$

This horizon-wise calibration allows the conformal intervals to adapt to the increasing uncertainty at longer prediction steps. The final calibrated intervals for a new input $x$ are obtained as

$$[L(x,h), U(x,h)] = \big[\hat{q}_{\text{lo}}(x,h) - q_{\alpha,h},\ \hat{q}_{\text{hi}}(x,h) + q_{\alpha,h}\big]. \tag{3.3}$$

## 3.6 Limitations

Although the experimental design provides a comprehensive framework for evaluating and calibrating uncertainty in pretrained time-series models, several limitations should be acknowledged.

First, the experiments were conducted using a single pretrained Chronos variant (Chronos-T5 Small) and a single benchmark dataset (M4 Daily). This configuration ensures reproducibility and manageable computational demand but may not fully capture the diversity of temporal behaviors observed across domains.

Second, while Conformal Prediction was initially planned to be implemented through the MAPIE framework, practical issues related to package installation, version compatibility, and dependency conflicts in the Colab environment prevented a stable setup. MAPIE relies on specific versions of `scikit-learn` and `numpy`, which conflicted with those required by the `chronos-forecasting` and `torch` libraries. These incompatibilities made it difficult to execute CQR within the same runtime environment. As a result, the calibration procedure was implemented manually using the mathematical formulation proposed by Romano et al. (2019). This manual implementation follows the same theoretical principles as MAPIE's CQR estimator—computing nonconformity scores from the model's quantile forecasts, extracting their empirical $(1-\alpha)$-quantile as the conformal adjustment, and expanding the lower and upper prediction bounds accordingly. Therefore, the manual CQR used in this study is functionally equivalent to the MAPIE implementation but avoids versioning and compatibility issues, ensuring reproducibility and full control over the calibration process.

Third, the calibration process itself was constrained by the limited number of available quantiles and by computational resources. Chronos provides only three quantile levels (0.05, 0.50, 0.95), which restricts the resolution

of conformal adjustment and prevents the computation of dense-distribution metrics such as the Continuous Ranked Probability Score (CRPS) or the Expected Calibration Error (ECE). Moreover, calibration was performed on a representative subset of time-series windows to remain within GPU memory limits, which may influence the stability of the resulting coverage estimates.

In addition, differences in data-loading efficiency also affected runtime performance. Forecasts generated through the `fev.Task` interface completed in roughly one hour, whereas an equivalent experiment based on CSV files loaded from Google Drive via `pandas` required more than 3.5 hours. The additional time was primarily due to input/output latency and the computational overhead of manual preprocessing. This highlights the importance of using optimized dataset interfaces such as `fev` for large-scale forecasting experiments.

Finally, all experiments were executed within a Google Colab High-RAM environment equipped with an NVIDIA L4 GPU. While this setup was sufficient for model inference and calibration, memory constraints limited batch size, quantile density, and the exploration of more advanced calibration schemes such as adaptive or online conformal prediction.

Beyond these computational and methodological considerations, the overall scope of the study was limited by the use of a single benchmark dataset and a fixed evaluation configuration. While this approach ensured consistency across experiments, it may not fully reflect the model's behavior under different temporal patterns, seasonal variations, or domain-specific dynamics. These factors should be taken into account when interpreting the findings and considering their applicability to broader forecasting scenarios.

# Chapter 4

# Results

## 4.1 Overview

This chapter presents the experimental results obtained from evaluating the uncertainty estimates of the Chronos-T5 Small model before and after calibration with Conformal Prediction via CQR. Specifically, the analyses aim to:

- Quantify the probabilistic performance of Chronos' uncalibrated forecasts through key uncertainty metrics, including the Prediction Interval Coverage Probability (PICP), Interval Calibration Error (ICE), Sharpness, Interval Score (Winkler), and Pinball loss.
- Examine the degree of miscalibration in Chronos' native quantile forecasts, identifying whether the nominal 90% prediction intervals achieve their expected empirical coverage.
- Apply conformal calibration to adjust the lower and upper quantiles through Conformalized Quantile Regression (CQR), ensuring valid statistical coverage.
- Compare uncalibrated and calibrated forecasts to analyze the balance between coverage reliability and interval sharpness, highlighting improvements in reliability and interval efficiency.

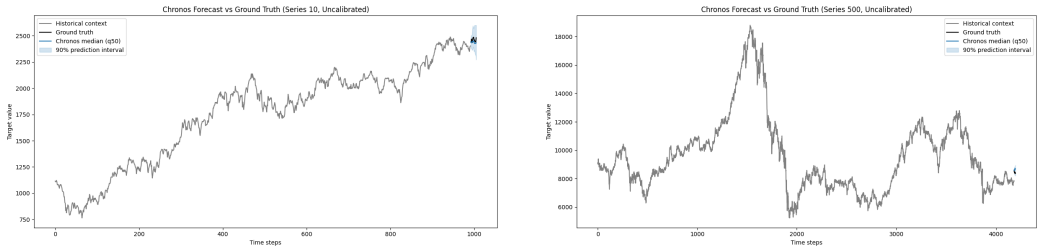
## 4.2 Comparison of Uncalibrated and Ground-Truth Forecasts

This section examines the probabilistic forecasts produced by the Chronos-T5 Small model before any conformal calibration is applied. The objective is to

establish a visual and quantitative baseline that reveals how well the model's native quantile estimates capture uncertainty on the `M4_Daily` dataset.

### 4.2.1 Forecasts and Visual Evaluation

Figure 4.1 presents representative examples of uncalibrated Chronos forecasts for several time series. The median predictions (blue) and 90% prediction intervals (shaded) are compared against the ground truth (black). The consistently narrow intervals reveal underestimation of uncertainty in the uncalibrated model.




(a) Series 10 (b) Series 500

Figure 4.1: Representative Chronos forecasts for selected time series before calibration.

Across all examined series, a consistent pattern that emerges is that the model accurately predicts central tendencies but fails to account for variability in more volatile segments. The underestimation observed here provides clear motivation for the subsequent conformal calibration stage, which aims to adjust interval widths such that the empirical coverage aligns with the desired nominal level of 90%.

## 4.3 Uncertainty Metrics and Calibration Curves

This section presents an analysis of the uncertainty estimates produced by the Chronos-T5 Small model, both before and after conformal calibration. It is structured to first examine the uncalibrated results; through calibration curves, coverage behavior, and key uncertainty metrics, to diagnose how the model's nominal 90% prediction intervals perform in practice. The following part then introduces the calibrated results, describing the conformal calibration procedure and evaluating its impact on coverage reliability and interval width. By organizing the analysis in this sequence, the section provides a clear before and after comparison that highlights how conformal calibration

corrects deviations and improves the reliability of Chronos' probabilistic forecasts.

### 4.3.1 Uncalibrated Metrics and Calibration Curves

#### Calibration Curves and Coverage Behavior

Figure 4.2 shows the calibration curve for Chronos' uncalibrated forecasts. The red dashed line represents perfect calibration, where the empirical coverage would exactly match the nominal coverage across all confidence levels (i.e., a model whose predicted intervals contain the true values in precisely the expected proportion of cases). The blue line corresponds to the empirical coverage achieved by Chronos for each nominal level between 0 and 1.

For a well-calibrated model, the blue curve should closely follow the red diagonal. In this case, the blue line lies near the red line at low nominal coverages but diverges below it at higher levels. This means that while Chronos' lower-quantile intervals behave reasonably, its wider intervals (such as the nominal 90% ones) fail to include the true values as often as expected. Hence, the model tends to be slightly under-confident at small coverages and increasingly over-confident at higher coverages, revealing systematic miscalibration.

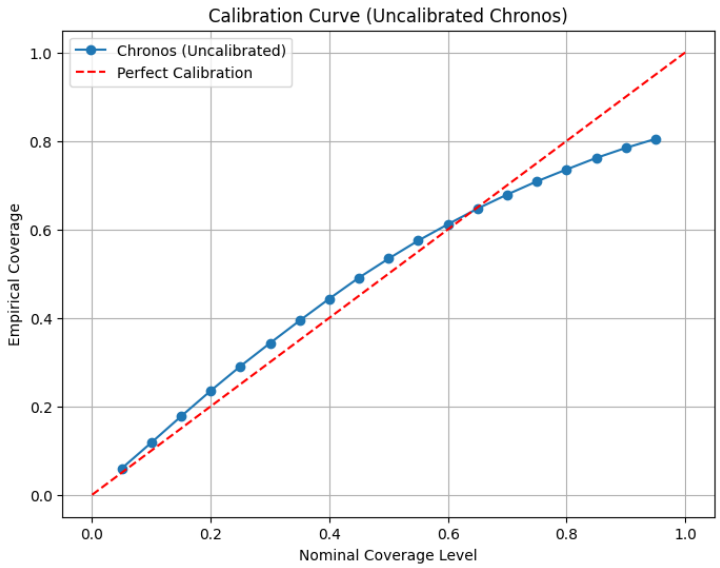



Figure 4.2: Calibration Curve for Chronos before calibration.

In Figure 4.3 the red dashed line represents the nominal 90% coverage target, while the blue line shows the empirical coverage obtained from the

model at each forecast step (1 to 14 days ahead). The empirical coverage fluctuates between approximately 0.74 and 0.82 but never reaches the nominal 0.9 level. This indicates that Chronos' prediction intervals are too narrow across all horizons (i.e., the model underestimates uncertainty, regardless of how far ahead the forecast is made).

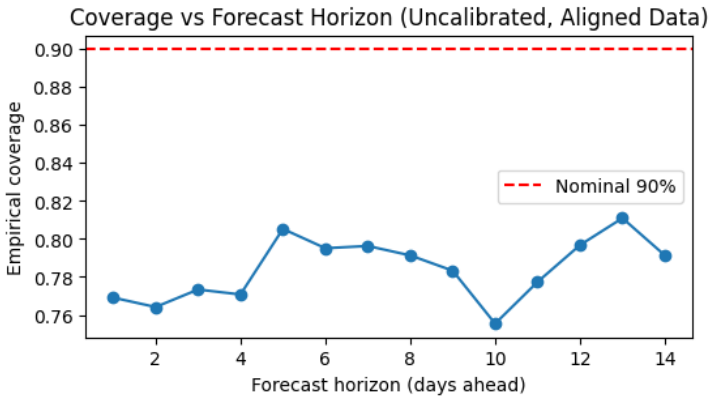



Figure 4.3: Coverage vs Forecast Horizon for Chronos before calibration.

Finally, Figure 4.4 illustrates the distribution of empirical coverage across all individual time series. Most series exhibit empirical coverage between 0.9 and 1.0, indicating that their prediction intervals are wider than necessary (over-coverage). However, a non-negligible subset of series lies well below 0.9, producing under-coverage that lowers the overall average PICP to 0.78 (see Table 4.1). This heterogeneity reveals that Chronos' uncertainty estimates are not uniformly calibrated across series; some forecasts are overly conservative, while others remain over-confident.

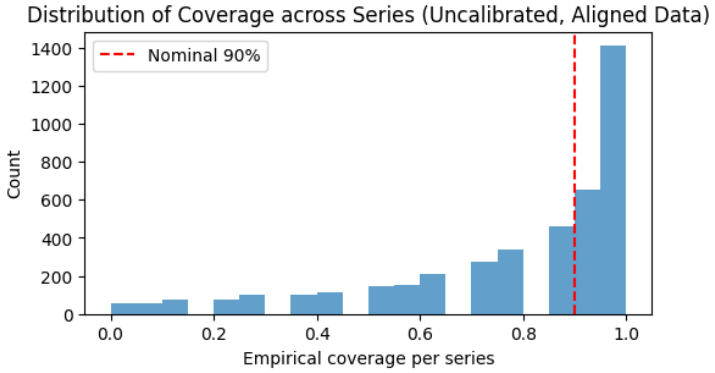



Figure 4.4: Distribution of empirical coverage across time series for Chronos before calibration.

### Quantitative Metrics

Table 4.1: Uncertainty metrics for Chronos before conformal calibration.

| **Metric** | **Value** |
|---|---|
| PICP | 0.744 |
| ICE | 0.155 |
| Sharpness | 449.41 |
| Interval Score (Winkler) | 1052.92 |
| Pinball Loss @ $\tau = 0.5$ | 78.49 |

As shown in Table 4.1, Chronos achieves a PICP of 0.744, meaning that only about 74.4% of the true values are contained within the predicted bounds. This under-coverage indicates that the model underestimates its forecast uncertainty and produces intervals that are too narrow.

Also the model exhibits an ICE of 0.155, corresponding to an 15 % shortfall from the desired 90 % coverage. Lower ICE values indicate better calibration; therefore, this result confirms that the uncalibrated forecasts are substantially miscalibrated.

The mean interval width (Sharpness $\approx$ 449.41) represents the average distance between the 0.05 and 0.95 predicted quantiles across all forecast steps and series. Because this value is computed in the original data scale, it reflects the typical magnitude of the series rather than absolute narrowness or width. Given that many M4 Daily series reach values in the thousands, an average width of roughly 449.41 indicates that Chronos produces relatively tight prediction intervals in proportion to the overall signal level. However, since coverage remains below the nominal 90%, these intervals are slightly too narrow to capture the full variability of the observations.

The Winkler Interval Score ($\approx$ 1052.92) combines interval width and coverage accuracy into a single metric. Penalizes both excessively wide intervals and cases where the true value falls outside the predicted range. Since lower values indicate better calibration–sharpness trade-offs, the relatively high score here reflects the penalty incurred by Chronos' systematic under-coverage, consistent with the narrow but over-confident intervals observed earlier.

The Pinball Loss at $\tau = 0.5$ ($\approx$ 78.49) evaluates the accuracy of the median (central) forecast and is equivalent to the Mean Absolute Error at that quantile. Although the value appears numerically large due to the scale of the M4 Daily data, it provides a consistent baseline for comparison.

Overall, the uncalibrated Chronos model provides accurate median forecasts and relatively sharp intervals but fails to achieve the desired 90% empirical coverage. These findings justify the next step of applying *Conformalized Quantile Regression* (CQR), which will adjust the lower and upper quantiles to ensure statistically valid coverage without overly inflating interval width.

### 4.3.2 Calibrated Metrics and Calibration Curves

#### Calibration Procedure Summary

To address the systematic undercoverage identified in the previous subsection, CQR was applied to the Chronos forecasts to adjusts the nominal lower and upper quantiles predicted by Chronos by adding an empirical offset derived from calibration residuals. For each calibration example, conformity scores $S_i = \max(\hat{q}_{0.05}(x_i) - y_i,\, y_i - \hat{q}_{0.95}(x_i))$ quantify how far the true value lies outside the predicted interval. The $(1-\alpha)$ quantile of these scores, denoted $q_{1-\alpha}(S)$, is then used to expand the forecast intervals as $[\hat{q}_{0.05}(x_t) - q_{1-\alpha}(S),\, \hat{q}_{0.95}(x_t) + q_{1-\alpha}(S)]$.

**Calibration statistics:** The vector $q_{\alpha,h}$ represents the conformal correction values per-horizon applied to the Chronos quantile intervals, which progressively increase with the forecast horizon as predictive uncertainty accumulates.

$$q_{\alpha,h} = \begin{matrix} 25.92, & 30.52, & 43.88, & 49.59, & 38.36, & 46.60, & 51.43, & 63.07, \\ 69.13, & 94.46, & 90.37, & 72.05, & 57.86, & 65.63 \end{matrix}$$

Summary statistics of the adjustment values: Mean = 57.06, Median = 54.64, Max = 94.46, Min = 25.92, Standard Deviation = 19.61

#### Calibration Curves and Coverage Behavior

After applying Conformalized Quantile Regression (green curve), the calibrated calibration line moves much closer to the diagonal reference, demonstrating improved probabilistic calibration across all nominal levels. The empirical coverage now matches the nominal coverage for most confidence levels, confirming that the conformal correction successfully restores statistical validity.

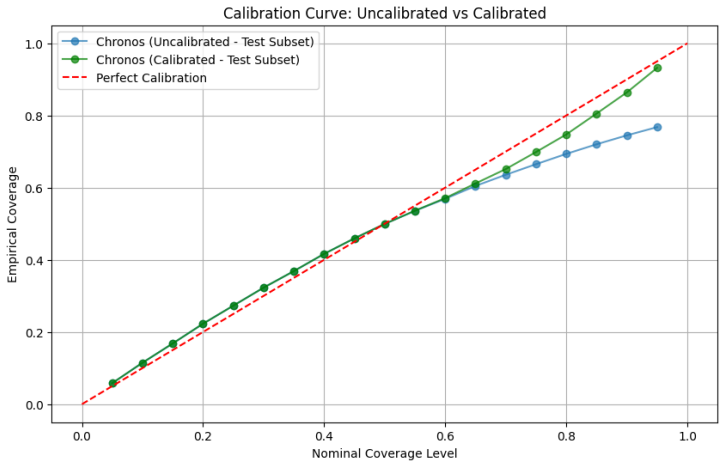



Figure 4.5: Comparison of calibration curves for Chronos before and after conformal calibration.

The per-horizon coverage plot further supports this improvement. In the uncalibrated case, coverage fluctuates between 0.72 and 0.77 across horizons, consistently remaining below the 0.9 target. Following calibration, the coverage curves for all 14 forecast steps increase toward the nominal line, achieving an average PICP of 0.863 (an increase of +0.119). This suggests that calibration effectively compensates for Chronos' horizon-dependent uncertainty underestimation.

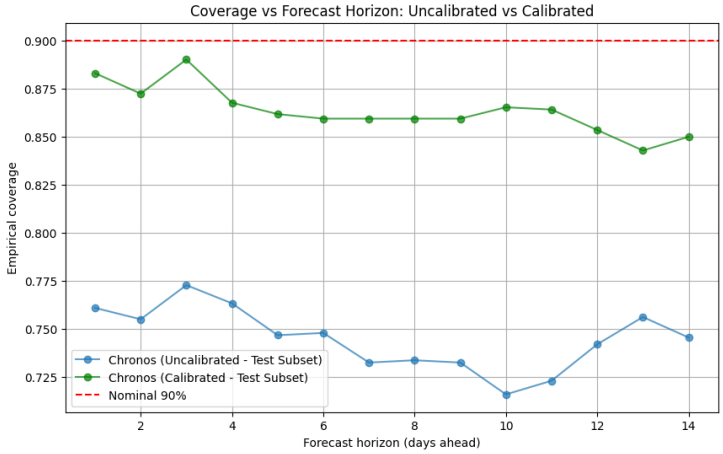



Figure 4.6: Empirical coverage as a function of forecast horizon for Chronos before and after calibration.

Finally, Figure 4.7 compares the distribution of empirical coverage across all series before and after conformal calibration. In the uncalibrated case,

the coverage values are widely dispersed, with many series falling well below the nominal 90% target, indicating heterogeneous and unreliable uncertainty estimates. After calibration, the distribution becomes sharply concentrated near the nominal level, demonstrating that the conformal adjustment effectively equalizes coverage across series. The calibrated model achieves both higher reliability and greater consistency, ensuring that the prediction intervals provide statistically valid uncertainty estimates across the entire dataset.

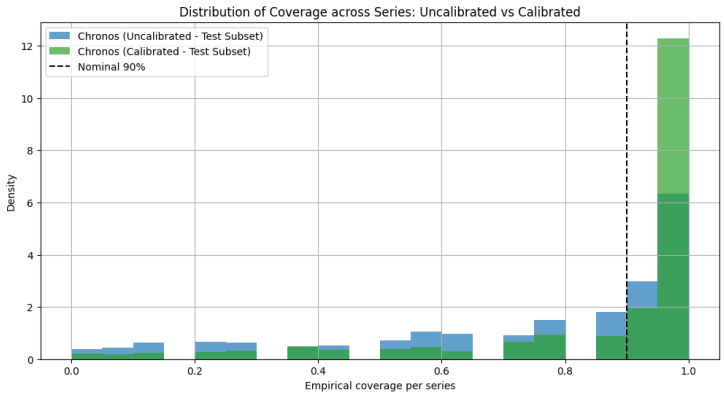



Figure 4.7: Distribution of empirical coverage across time series for Chronos before and after calibration.

Overall, these visual comparisons demonstrate that conformal calibration substantially improves both the reliability and uniformity of the Chronos probabilistic forecasts, aligning empirical coverage with nominal expectations while maintaining stable performance across horizons.

### Quantitative Metrics and Comparison

Table 4.2: Uncertainty metrics for Chronos before and after conformal calibration.

| **Metric** | **Uncalibrated** | **Calibrated** |
|---|---|---|
| PICP ↑ | 0.744 | 0.863 |
| ICE ↓ | 0.155 | 0.036 |
| Interval Width (Sharpness) ↑ | 449.41 | 555.46 |
| Interval Score (Winkler) ↓ | 1052.92 | 964.48 |
| Pinball Loss @ $\tau = 0.5$ ↓ | 78.49 | 78.49 |

Calibration increases empirical coverage (PICP) from 0.744 to 0.863, approaching the nominal target of 90% , while reducing the ICE by the same magnitude. This reliability gain comes with a modest widening of the prediction intervals, a necessary cost for attaining improved statistical validity. Winkler decreases substantially, confirming that the conformal adjustment enhances the calibration and interval efficiency. Pinball Loss remained unchanged after calibration, as the CQR adjustment modifies only the outer quantiles (0.05 and 0.95), leaving the median forecast unaffected.

Together, these quantitative improvements verify that conformal calibration effectively restores nominal coverage and yields statistically valid, more reliable uncertainty estimates with only a modest increase in interval width.

## 4.4 Summary of Findings

The application of *Conformalized Quantile Regression* (CQR) successfully corrected the systematic under-coverage observed in Chronos-T5 Small's uncalibrated forecasts. Before calibration, the model's nominal 90% prediction intervals achieved only 74.4% empirical coverage, indicating that uncertainty was underestimated. After calibration, the empirical coverage rose to 0.863 (+11.9 percentage points), substantially reducing the Interval Calibration Error from 0.155 to 0.036.

This improvement came with a modest increase in interval width (Sharpness = 449.41 $\rightarrow$ 555.46), confirming the expected trade-off of conformal methods: achieving statistically valid coverage by slightly widening intervals. The combined Winkler Interval Score decreased by 8.4% (from 1052.92 $\rightarrow$ 964.48), indicating that the gain in empirical coverage outweighs the modest increase in interval width." Likewise, the median Pinball Loss stayed at 78.49, demonstrating that calibration did not degrade the accuracy the model's central (median) forecasts.

Visual analyses reinforced these quantitative results. The calibration curve after CQR aligned closely with the diagonal reference, and coverage–horizon plots showed that each forecast step moved toward the nominal 90% target. Moreover, the distribution of empirical coverage across individual time series became sharply centered near 0.9, indicating that calibration improved not only average reliability but also consistency across heterogeneous series.

Mathematically, the conformal adjustment value $q_{\alpha,h}$ is subtracted from the lower quantile $\hat{q}_{\mathrm{lo}}(x,h)$ and added to the upper quantile $\hat{q}_{\mathrm{hi}}(x,h)$, yielding the calibrated interval:

$$[\ \hat{q}_{\mathrm{lo}}(x,h) - q_{\alpha,h},\ \hat{q}_{\mathrm{hi}}(x,h) + q_{\alpha,h}\ ].$$

Because the nonconformity scores

$$s_{i,h} = \max(\hat{q}_{\text{lo}}(x_{i,h}) - y_{i,h},\ y_{i,h} - \hat{q}_{\text{hi}}(x_{i,h}))$$

are non-negative by construction [4], the derived adjustment values $q_{\alpha,h}$ are also non-negative. Consequently, calibration can only widen the original prediction intervals or, in degenerate cases ($q_{\alpha,h} = 0$), leave them unchanged. This property ensures that conformal calibration always increases or preserves empirical coverage, typically at the cost of a small reduction in sharpness [7].

Overall, these findings confirm that Chronos' native quantile forecasts (while accurate in their central tendency) are miscalibrated in their uncertainty estimates. Conformal calibration provides a simple, distribution-free correction that restores nominal coverage without retraining or compromising accuracy. This demonstrates that pretrained time-series foundation models can produce reliable probabilistic forecasts when complemented with lightweight, statistically grounded calibration.

# Chapter 5

# Conclusion and Future Work

## 5.1 Future Work

While Conformalized Quantile Regression (CQR) effectively improved the reliability of Chronos-T5 Small's uncertainty estimates, several extensions could strengthen this research.

First, future studies should evaluate the calibration behaviour across additional datasets and forecasting horizons to confirm the generalizability of the results beyond the M4 Daily benchmark. Second, since conformal prediction assumes exchangeability between calibration and test samples (a condition only approximated in temporal data) further work could explore adaptive or online conformal methods [2] that maintain valid coverage under time dependence or distributional drift.

Another promising direction involves integrating the Chronos pipeline with established calibration libraries such as `MAPIE` or `TorchCP` once dependency conflicts are resolved, enabling systematic benchmarking of computational efficiency and reproducibility. Moreover, extending Chronos to support denser quantile outputs or full probabilistic forecasts would allow the use of richer evaluation metrics such as CRPS or Expected Calibration Error (ECE).

Finally, applying calibrated Chronos forecasts to domain-specific problems, such as energy demand, retail planning, or healthcare, would demonstrate how improved uncertainty calibration can enhance real-world decision-making. These directions aim to transform the present study from a model-specific analysis into a broader framework for reliable, distribution-free uncertainty quantification in time-series foundation models.

# Bibliography


[1] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Syndar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

[2] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference on Learning Theory (COLT)*. PMLR, 2018.

[3] Cathal Murray, Priyanka Chaurasia, L. E. Hollywood, and Damien Coyle. A comparative analysis of state-of-the-art time series forecasting algorithms. In *Proceedings of the 2022 International Conference on Computational Science and Computational Intelligence (CSCI 2022)*, pages 89–95. IEEE, 2023. Accessed: 2025-09-22.

[4] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] Amazon Science. Chronos: Pretrained Time Series Foundation Models. `https://github.com/amazon-science/chronos-forecasting`, 2024. Accessed: 2025-10-12.

[6] Valeman. Awesome Conformal Prediction. `https://github.com/valeman/awesome-conformal-prediction`, 2024. Accessed: 2025-09-22.

[7] Léo Zaffran. Conformal prediction tutorial. `https://github.com/leo-zaffran/conformal-prediction-tutorial`, 2023. EPFL & Google Research.

[8] Margaux Zaffran, Olivier Feron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25834–25866. PMLR, 17–23 Jul 2022.