Neural Conformal Control for Time Series Forecasting

Ruipu Li, Alexander Rodríguez

University of Michigan
liruipu@umich.edu, alrodri@umich.edu

Abstract

We introduce a neural network conformal prediction method for time series that enhances adaptivity in non-stationary environments. Our approach acts as a neural controller designed to achieve desired target coverage, leveraging auxiliary multi-view data with neural network encoders in an end-to-end manner to further enhance adaptivity. Additionally, our model is designed to enhance the consistency of prediction intervals in different quantiles by integrating monotonicity constraints and leverages data from related tasks to boost few-shot learning performance. Using real-world datasets from epidemics, electric demand, weather, and others, we empirically demonstrate significant improvements in coverage and probabilistic accuracy, and find that our method is the only one that combines good calibration with consistency in prediction intervals.

Project page — https://github.com/complex-ai-lab/ncc

Introduction

Quantifying uncertainty in predictions is crucial for various practical applications. Rigorous frameworks are necessary to assess the reliability of these predictions and inform high-stakes decisions, such as hospital resource allocation during epidemics or planning energy supply for cities (Gawlikowski et al. 2023). Among these frameworks, conformal prediction (CP) (Vovk, Gammerman, and Shafer 2005; Shafer and Vovk 2008) stands out for its ability to provide a theoretical coverage guarantee under mild distribution assumptions. However, the guarantee often breaks down in time series forecasting due to temporal dependencies and distribution shifts inherent in time series data. To address these challenges, multiple methods have been proposed, each making specific assumptions about the time series, such as stationarity (Oliveira et al. 2024), dependence structures (Xu and Xie 2023), or exchangeable data and forecasting strategies (Stankeviciute, M Alaa, and van der Schaar 2021). Another line of work (Gibbs and Candes 2021), which we refer to as 'conformal control', circumvents those assumptions by adopting a control perspective in an online setting, where prediction intervals dynamically adapt to coverage errors to achieve a desired target coverage. Based on this, several variations have been introduced to refine the controller logic and enhance adaptivity (Angelopoulos, Candes, and Tibshirani 2024; Zaffran et al. 2022; Bhatnagar et al. 2023).

Despite these advancements, significant gaps remain between current CP methods and the demands of real-world applications. For instance, the epidemic forecasting initiatives hosted by the US Centers for Disease Control and Prevention (CDC) in response to diseases such as Ebola (Viboud et al. 2018; Johansson et al. 2019), influenza (Reich et al. 2019; Mathis et al. 2024), and COVID-19 (Cramer et al. 2022) highlight these challenges. Because these predictions inform critical policymaking decisions, the CDC requires forecasters to provide a comprehensive view of future possibilities. Consequently, submissions must include different confidence levels that adherence to *distributional consistency* (Reich et al. 2019; Cramer et al. 2022), which prohibits crossing prediction intervals. Failure to meet these standards makes forecasts invalid for CDC acceptance. Conformal control methods do not guarantee this because of disjoint calibration for each confidence level and our experiments revealed that some of them fail to provide valid forecasts as low as 0% of cases. In addition, forecasting teams are encouraged to utilize any relevant datasets, and previous work has demonstrated the effectiveness of leveraging *multi-view datasets* from heterogeneous sources (Rodríguez et al. 2021; Qian, Alaa, and van der Schaar 2020) and across multiple modalities (Deng et al. 2020; Kamarthi et al. 2022). For example, distribution shifts driven by human behavior can be anticipated by analyzing digital surveys and mobility data (Rodríguez et al. 2024). However, CP methods often underutilize the auxiliary information, which could better inform adaptivity mechanisms and improve calibration. Moreover, high-stakes applications like epidemic forecasting require readiness to address emerging outbreaks, such as recent cases of H5N1 and measles, where data is scarce. This underscores the need to enhance the *few-shot learning* capabilities of current CP methods.

In response to these challenges, we propose integrating recent ideas from conformal control with advances in deep learning. Our method, Neural Conformal Control for Time Series Forecasting (NCC), incorporates control-inspired loss functions to enable the neural network to function as a predictive controller. This approach offers a flexible, end-to-end learning framework for conformal prediction using neural networks without losing the theoretical guarantee on coverage. We summarize our main contributions as follows.

1. **Neural Control for Conformal Prediction:** We introduce NCC, a conformal control method designed to harness the capabilities of neural networks to capture patterns that inform adaptivity to changes in time series. To achieve this, we develop multiple neural modules to identify these patterns and incorporate loss functions that enhance control and improve the efficiency of prediction intervals.
2. **Enabling Consistent Prediction Intervals and Few-Shot Learning:** We leverage the flexibility of NCC to constrain our neural network to yield consistent prediction intervals through a process involving a monotonicity loss and a test-time adaptation (TTA) procedure. Additionally, our neural architecture is designed to facilitate transfer learning across multiple tasks (e.g., regions) and enhance few-shot learning (FSL) performance.
3. **Extensive Experiments and Theoretical Guarantees:** We evaluate NCC on a variety of real-world datasets and against other state-of-the-art CP methods. We show that NCC achieves superior empirical performance while maintaining theoretical long-term coverage guarantees.

More details in our contributions are outlined in Table 1.

	Conformal control	Dist. assumptions	NCC (this paper)
Control-based adaptivity	✓	✗	✓
Coverage guarantees	✓	✓	✓
Incorporates multi-view data	✗	✗	✓
Consistent prediction intervals	✗	✓	✓
Data efficiency for FSL	✗	✗	✓
End-to-end learning	✗	✗	✓

Table 1: Comparison with other time series CP methods. Conformal control methods include ACI (Gibbs and Candes 2021) and variants (Angelopoulos, Candes, and Tibshirani 2024; Zaffran et al. 2022; Gibbs and Candès 2024). Methods with dist. assumptions include (Stankeviciute, M Alaa, and van der Schaar 2021; Barber et al. 2023; Oliveira et al. 2024).

Preliminaries

Here we provide background on CP methods in time series, highlighting their limitations in real-world applications, which motivate our proposed method.

Conformal Prediction (CP) Setup. Let us introduce the basic setup for conformal predictions in time series forecasting. We focus our attention on an online learning setting where feature-target pairs are observed sequentially (Gibbs and Candes 2021). At each time step t, we have a univariate response $y_t \in \mathcal{Y}$, features X_t (X_t may include y_t), and use a forecasting model to generate predictions $\hat{y}_{t+\tau}$ for future time steps, where $\tau \in \mathbb{N}^+$ is the forecasting horizon. A user shall specify one or several target error rate $\alpha \in (0, 1)$ (also known as miscoverage rate). Then, we will refer to a method as well-calibrated in uncertainty if it constructs prediction interval $\hat{C}_t$ such that y_t, the actual future value of the response, falls within $\hat{C}_t$ at least $100(1-\alpha)\%$ of the time.

A CP method typically involves splitting the observed data into a training set, used to fit the forecasting model, and a calibration set used for creating the prediction sets (Papadopoulos, Vovk, and Gammerman 2007). A non-conformity score $s_t(y_t, \hat{y}_t)$ is defined to measure how closely the observed value y_t 'conforms' with the predictions from the fitted model. For example, one of the most widely-used and simple non-conformity score in regression problems is $s_t(y_t, \hat{y}_t) = |y_t - \hat{y}_t|$, which we adopt in this paper. Then, within the calibration set, we calculate the non-conformity scores and determine q_t^α, which is the $(1-\alpha)$ quantile of these scores. In conventional CP settings, where all data points are assumed to be exchangeable, the prediction interval $\hat{C}_t^\alpha = \{y \in \mathcal{Y} : s_t(y, \hat{y}_t) \leq q_t^\alpha\}$ is guaranteed to include y_t with a probability of at least $1-\alpha$ (Vovk, Gammerman, and Shafer 2005).

CP in Time Series. In time series forecasting, the exchangeability assumption does not hold. As a result, a coverage guarantee at time t can only be achieved with trivial methods that produce prediction intervals of infinite size (Angelopoulos, Candes, and Tibshirani 2024). The most feasible outcome is achieving a valid long-term coverage, defined as $\frac{1}{T}\sum_{t=1}^{T}(\text{err}_t^\alpha - \alpha) < o(1)$. Here err_t^α is the coverage error at time step t indicating that the ground truth data is not covered by the prediction interval associated with α. This error is defined as $\text{err}_t^\alpha := \mathbb{1}(s_t > \hat{q}_t^{1-\alpha})$, where $\mathbb{1}$ is the indicator function. $o(1)$ denotes a quantity that approaches zero as $T \to \infty$. Such guarantees can be achieved using an adaptive error rate α_t (Gibbs and Candes 2021), different from the constant target error rate α specified by the user. α_t evolves over time according to the update rule: $\alpha_{t+1} = \alpha_t + \eta(\alpha - \text{err}_t^\alpha)$, where η is a learning rate. This approach allows α_t to adjust dynamically based on the observed error err_t^α, enhancing the adaptivity of the prediction intervals. Recent work has recast the task of adaptivity as a control problem with a feedback loop (Angelopoulos, Candes, and Tibshirani 2024; Zaffran et al. 2022; Bhatnagar et al. 2023) and achieved notable performance improvements maintaining theoretical guarantees on long-term coverage. As mentioned earlier, we refer to these approaches as 'conformal control' methods.

Practical Limitations in Conformal Control. One key limitation of existing conformal control methods is the lack of mechanisms to ensure adaptivity of prediction intervals. As exemplified by the update rule above, the adaptivity of these methods relies solely on adjustments based on observed coverage errors, potentially overlooking other important patterns, such as recurring trends in non-conformity scores. Additionally, these methods have not yet explored how to incorporate multi-view auxiliary data into their adaptive updates. Another issue is the inconsistency among prediction intervals for a given set of error rates. In standard CP methods, the error rate α is fixed, making the non-conformity score–the $(1-\alpha)$ quantile of calibration set scores–naturally monotonic with respect to α (Shafer and Vovk 2008). However, conformal control methods such as ACI (Gibbs and Candes 2021) and variants (Angelopoulos, Candes, and Tibshirani 2024; Gibbs and Candès 2024; Zaffran et al. 2022), and some ML-based CP methods (Xu and Xie 2023), where the quantile of non-conformity scores are constantly adapted, the non-conformity score is not necessarily monotonic. This can lead

to the overlapping and crossing of prediction intervals, a problem known as quantile crossing. We formalize this property with the following definition, which comes naturally from the non-decreasing property of the quantile function.

Definition 1. (Distributional Consistency) Given the n prediction intervals $(\hat{y}^{\alpha_1}_{lower}, \hat{y}^{\alpha_1}_{upper}), \ldots, (\hat{y}^{\alpha_n}_{lower}, \hat{y}^{\alpha_n}_{upper})$ and a set of n error rates $\mathbf{A} := \alpha_1, \ldots, \alpha_n$, we say the prediction intervals are consistent if, for any $i, j \in [1, n]$, where $\alpha_i < \alpha_j$, it holds that $\hat{y}^{\alpha_i}_{lower} \leq \hat{y}^{\alpha_j}_{lower}$ and $\hat{y}^{\alpha_i}_{upper} \geq \hat{y}^{\alpha_j}_{upper}$.

Connection to Related Works. As mentioned earlier, due to temporal distribution shifts, using the quantile of scores in the calibration set no longer gives "conformal" non-conformity scores on test data. Therefore, we can view the efforts made by time series CP methods as proposing better quantile predictors that adapts to the temporal changes. For ACI and variants (Gibbs and Candes 2021; Bhatnagar et al. 2023; Gibbs and Candès 2024), the quantile function evolves across time through changing the error rate. C-PID (Angelopoulos, Candes, and Tibshirani 2024) directly updates the quantile using online gradient descent. SPCI (Xu and Xie 2023) learns a random forest quantile estimator. Our work follows this line of work by using a neural network. It is worth noting that conformalized quantile regression (Romano, Patterson, and Candes 2019) is quite different. CQR and other quantile regression methods aim to learn a conditional quantile function on the target variable Y (Xu and Xie 2023). Thus, they fit the quantile regressor on Y directly, while ours fits the quantile regressor on user-defined non-conformity scores. More on related work can be found in our Appendix.

Our Approach

NCC consists of a predictive controller for coverage errors based on neural networks, control-inspired losses, and a test-time adaptation procedure. NCC enables: **(1)** *Improved adaptivity through representation learning and multi-view data:* Leveraging a neural network allows us to exploit patterns beyond error rates and seamlessly incorporate multi-view data, which provides valuable information for adapting prediction intervals. Our architecture is designed with encoders that incorporate appropriate inductive biases and control-inspired losses to effectively learn to function as a predictive controller. **(2)** *Improved consistency of prediction intervals without sacrificing performance*: We design our neural network architecture and loss functions to ensure that predicted quantiles are monotonically decreasing (i.e., non-increasing) with respect to α. Although our prediction intervals are constructed sequentially, an adjustment step required for coverage guarantees (known as the 'conformalization step') may disrupt this property. To address this issue, we introduce a monotonicity loss during training and a test-time adaptation (TTA) procedure to preserve this property during inference. **(3)** *Guaranteed long-term coverage through a differentiable conformalization step*: While the prediction intervals produced by our neural network are well-calibrated and generally consistent, they do not inherently provide the theoretical coverage guarantees offered by other CP methods. By incorporating an adjustment term to these prediction intervals, we can conformalize them, thereby preserving the theoretical coverage guarantees of traditional conformal control methods.

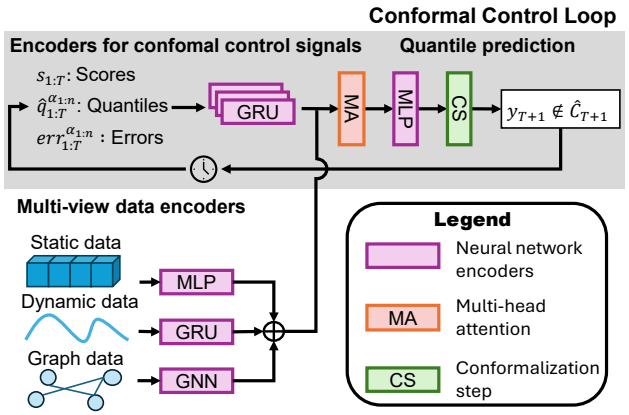

Figure 1: Model architecture of NCC. The shaded area represents the conformal control loop of our method. The neural network encoders used here include multilayer perceptron (MLP), gated recurrent unit (GRU), and graph neural network (GNN). The embeddings of multi-view data and conformal control signals are integrated using multi-head cross-attention mechanisms and subsequently processed through an MLP to predict the quantiles of the scores. Finally, the predicted quantiles are refined through a conformalization step.

Neural Network Architecture

Our model consists of two major modules: a conformal control loop and multi-view data encoders.

Conformal Control Loop. The main part of our model is a conformal control loop as shown in the shaded area in Figure 1. At time T, we will have available the conformal control signals err^{α}_{T} and $\hat{q}^{\alpha}_{T}$ for each $\alpha \in \{\alpha_1, \ldots, \alpha_n\}$, and non-conformity score s_T. Then, considering their past, we have three conformal signals: $\text{err}^{\alpha}_{1:T} := \{\text{err}^{\alpha}_{t} | 1 \leq t \leq T\}$, $\hat{q}^{\alpha}_{1:T} := \{\hat{q}^{\alpha}_{t} | 1 \leq t \leq T\}$ for each α and $s_{1:T} := \{s_t | 1 \leq t \leq T\}$. These conformal control signals decide how we adapt our predicted quantiles of the non-conformity scores. Therefore, we encode the three sequences using GRUs (Chung et al. 2015). The learned embedding is then dynamically combined using a multi-head attention mechanism. Specifically, we have a combined embedding $z_{\text{combined}} = \text{Multi-head}(z_{\text{data}}, z_{\text{err}}, z_q, z_s)$ which is subsequently fed into an MLP with a ReLU layer as the final layer. We denote the output from the quantile predictor as $\{\tilde{q}^{\alpha_1}, \ldots, \Delta\tilde{q}^{\alpha_n}\}$, which holds $\Delta\tilde{q}^{\alpha_i} \geq 0$ because of the final ReLU layer. Therefore, for two positions i and j, we will have $i < j$, $\alpha_i > \alpha_j$ (i.e., α's are sorted in decreasing order). Note that $\tilde{q}$ is not yet our final prediction interval $\hat{q}^{\alpha}_{t}$, but instead an intermediate value. $\tilde{q}$ from the quantile predictor is then conformalized through a conformalization step (we detail this step later in this section).

Multi-view Data Encoders. Following previous work in multi-view learning (Yan et al. 2021), we use different encoders for each view. A natural way to determine views is based on data modalities, which are typically time se-

ries (sequential), graph (relational) and static for forecasting tasks (Kamarthi et al. 2022). Within each modality, we find different views (e.g., one for each data source), which can be particularly useful when each view provides a different perspective or has varying noise levels. Let j be the number of data views; then, the combined representation of the multi-view data is $z_{\text{data}} = \text{Multi-head}\big\{g_{\theta_j}(X^{(j)})\big\}$, where Multi-head is a multi-head attention layer (Vaswani et al. 2017), and g_{θ_j} is a neural network module parameterized by θ_j, chosen based on the modality of view j. For example, we use a GRU for views with sequence data, a GCN (Kipf and Welling 2016) to encode relational data, and feed-forward networks for static fixed-sized features. This flexible setup allows us to leverage the representational power of neural networks to transfer knowledge across multiple tasks. Building on previous successes in transfer learning for CP (Fisch et al. 2021), our framework can facilitate this process, for instance, by incorporating region encoding as one of our static views.

Control-inspired Losses for Improved Adaptivity

To enable a data-driven model to effectively control the coverage of our conformal prediction intervals, we incorporate losses that promote controller-like behavior.

Coverage Adaptivity via Quantile Loss. To learn a conditional quantile distribution of the nonconformity scores, we use the quantile (pinball) loss. Denote the quantile loss (Romano, Patterson, and Candes 2019) at the level $(1-\alpha)$ as $\rho_{1-\alpha}(z)$, which is defined as $(1-\alpha)|z|$ for $z > 0$ and $(\alpha)|z|$ for $z \leq 0$ (Koenker 2005). When the predicted interval covers the ground truth data point (i.e., $s_t \leq \hat{q}_t^{\alpha}$), the quantile loss is $(-\alpha)(s_t - \hat{q}_t^{\alpha})$; otherwise, it is $(1-\alpha)(s_t - \hat{q}_t^{\alpha})$. Thus:

$$\mathcal{L}_Q = \max\big\{(1-\alpha)(s_t - \hat{q}_t^{\alpha}), (-\alpha)(s_t - \hat{q}_t^{\alpha})\big\}, \quad \forall \alpha \in \mathbf{A}. \tag{1}$$

Coverage Stability via Integrative Coverage Loss. Distribution shifts can lead to fluctuating coverage errors (sometimes largely positive, sometimes largely negative). These fluctuations can be misleading for decision makers trying to understand the reliability of prediction intervals. Therefore, an important feature for our prediction intervals is not only to adapt but also maintain a stable coverage error, a property we refer to as coverage stability. To address this, we design an integrative *coverage loss* to ensure coverage stability.

First, let us define the running coverage error ($\bar{\text{err}}$), which integrates errors over the most recent w time steps. At time $T+1$, this error is given by: $\bar{\text{err}}_w^{\alpha}(T) := \frac{1}{w}\sum_{t=T-w+1}^{T} \text{err}_t^{\alpha}$. Using this error, we can determine whether our future prediction intervals need to cover the ground truth data to achieve the target error rate. To facilitate this, we define the indicator function $\text{cov}^{\alpha}(T)$, which indicates whether coverage is necessary for a prediction interval made at T (remember that at time T we predict for $T+\tau$). Our intuition is the following: if $\bar{\text{err}}_w^{\alpha}(T) > \alpha$, we want the next prediction interval to be large enough to cover the ground truth, i.e., $\text{err}_{T+1}^{\alpha}$ should be 0. In this case, we set $\text{cov}^{\alpha}(T) = 1$. Otherwise, if $\bar{\text{err}}_w^{\alpha}(T) \leq \alpha$, the prediction interval should be smaller, i.e., $\text{err}_{T+1}^{\alpha}$ should be 1. In this case, we set $\text{cov}^{\alpha}(T) = 0$. This is:

$$\text{cov}^{\alpha}(T) = \begin{cases} 0, & \bar{\text{err}}_w^{\alpha}(T) \leq \alpha \\ 1, & \text{otherwise.} \end{cases}$$

We can think of $\text{cov}^{\alpha}(T)$ as a binary label, which should align with the current coverage condition indicated by $1 - \text{err}_{T+1}^{\alpha}$. Therefore, we choose to define our coverage loss based on the log loss. For simplicity, we denote $\text{err}_{T+1}^{\alpha}$ as err^{α} and denote $\text{cov}^{\alpha}(T)$ as cov^{α} in the loss function.

$$\mathcal{L}_C = -(1-\text{cov}^{\alpha})\log(\text{err}^{\alpha}) - \text{cov}^{\alpha}\log(1-\text{err}^{\alpha}), \quad \forall \alpha \in \mathbf{A}. \tag{2}$$

To make err_t^{α} differentiable, we approximate it as:

$$\text{err}_t^{\alpha} \approx \text{sigmoid}\Big(\frac{s_t - \hat{q}_t^{\alpha}}{K}\Big),$$

where $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, and K is a constant. With an appropriate choice of K, err_t^{α} is close to 1 when $s_t > \hat{q}_t^{\alpha}$ and close to 0 otherwise.

Designing Efficient Prediction Intervals. Previous work has shown that quantile regression methods trained with quantile loss sometimes produce unnecessarily wide prediction intervals (Romano, Patterson, and Candes 2019). To prevent the prediction intervals from becoming excessively wide, we incorporate a regularization term, the efficiency loss $\mathcal{L}_E$. This term penalizes large prediction interval sizes when the ground truth is covered, thus promoting more efficient intervals.

$$\mathcal{L}_E = (y_t - \hat{q}_t^{\alpha})^2 \times (1 - p_{\text{cov}}), \quad \forall \alpha \in \mathbf{A}. \tag{3}$$

Long-term Coverage Guarantees Through Conformalization

The quantiles of nonconformity scores predicted by the neural network do not inherently ensure valid long-run coverage (Romano, Patterson, and Candes 2019). To achieve theoretical guarantees, we further conformalize the outputs of our model as per Algorithm 1, line 6. Note that our algorithm only considers the case where we calibrate for one step ahead predictions for simplicity, but it works for any prediction horizon τ. Refer to the Appendix for more details. For some α, we will consider the prediction interval constructed as $\hat{C}_t = \{y \in \mathcal{Y} : s_t(y_t, \hat{y}_t) \leq (\hat{q}_t)\}$, where $s_t(y_t, \hat{y}_t) = |y_t - \hat{y}_t|$. Unlike quantile tracking from C-PID, we do not use the update rule: $q_{t+1} = q_t + \eta(\text{err}_t - \alpha)$. Instead, we replace err_t with the running average error rate $\frac{1}{w}\sum_{t=T-w+1}^{T} \text{err}_t$. In the setting of C-PID, since err_t is either 1 or 0, $q_{t+1} \neq q_t$. This means that the prediction intervals predicted by the model are always adjusted. Our objective is to make the least adjustments to the prediction intervals. Therefore, instead of using error at t, we use the average of errors in a window of size w. Then when the prediction intervals are well calibrated, the average is close to α and the prediction interval would not be changed significantly.

Theorem 1. (Long-term coverage (Angelopoulos, Candes, and Tibshirani 2024)) For sufficiently large T, assume that $\Delta q \leq b$ for some $b \in \mathbb{R}$. Then the prediction interval $\hat{C}_t$ constructed by our method NCC satisfies $\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}(y_t \notin \hat{C}_t) = \alpha + o(1)$.

See proof of this theorem in our Appendix.

Remark: The predicted quantiles from the neural network lack a theoretical guarantee of long-term coverage. However, as demonstrated in the proof of **Theorem 1**, applying the

Algorithm 1: Forward pass over NCC at time step T.

Parameter: User-specified window size w, learning rate η, and a set of n decreasing target error rates $\mathbf{A} := \alpha_1, \cdots, \alpha_n$.
Input: Historical data, along with past non-conformative scores, coverage errors and quantiles, which we group and denote as $\mathcal{D} := \{X_t, y_t, s_t, \{\text{err}_t^\alpha, \hat{q}_t^\alpha\}_{\alpha \in \mathbf{A}}\}_{t=1}^T$. Quantile predictor $\mathcal{M}$.
Output: Conformalized prediction intervals $\hat{C}_{T+1}^\alpha, \forall \alpha \in \mathbf{A}$.

For all $\alpha \in \mathbf{A}$, initialize $\Delta\tilde{q}_1^\alpha = 0$.
Forward pass over quantile predictor: $\tilde{q}_{T+1}^{\alpha_1}, \ldots, \tilde{q}_{T+1}^{\alpha_n} = \mathcal{M}(\mathcal{D})$
for all $i \in 1, \ldots, n$ **do**
  Calculate the average error rate for window $\bar{\text{err}}_w^{\alpha_i}(T) = \frac{1}{w}\sum_{t=T-w+1}^{T} \text{err}_t^{\alpha_i}$
  Update $\Delta\tilde{q}^{\alpha_i}$ as $\Delta\tilde{q}_{T+1}^{\alpha_i} = \Delta\tilde{q}_T^{\alpha_i} + \eta(\bar{\text{err}}_w^{\alpha}(T) - \alpha_i)$
  Conformalization step: $\hat{q}_{T+1}^{\alpha_i} = \tilde{q}_{T+1}^{\alpha_i} + \Delta\tilde{q}_{T+1}^{\alpha_i}$
  if forward pass is used at test time (inference) **then**
    TTA: Train learnable vector $h \in \mathbb{R}^n$ (until desired DCS) so that $\hat{q}_{T+1}^{\alpha_i} + h_i$ monotonically increases as α_i decreases.
  end if
  $\hat{C}_{T+1}^{\alpha_i} = \{y | s_t(\hat{y}_{T+1}, y) \leq \hat{q}_{T+1}^{\alpha_i}\}$
end for

aforementioned algorithm conformalizes the predicted quantiles. This ensures that NCC possesses the same long-term coverage guarantee as ACI and C-PID.

Improved Distributional Validity Through Monotonic Quantiles

Quantile crossing leads to invalid prediction intervals. To address this issue, we first construct the prediction intervals cumulatively. We use a ReLU layer as the final layer of the quantile predictor. As noted before, we denote the output of the quantile predictor as $\tilde{q}^{\alpha_1}, \ldots, \Delta\tilde{q}^{\alpha_n}$, where $\Delta\tilde{q}^{\alpha_i} \geq 0$ and for $i < j$, $\alpha_i > \alpha_j$. Note that $\tilde{q}^{\alpha_1}$ is inherently non-negative, as our chosen absolute residue score function constrains the nonconformity scores to be non-negative. If a different score function is employed, other requirements on the nonconformity scores can be similarly implemented. Then we cumulatively sum the output: $\tilde{q}^{\alpha_k} = \tilde{q}^{\alpha_1} + \sum_{i=2}^{k} \Delta\tilde{q}^{\alpha_i}, k \in [2, n]$. Although the cumulative design ensures that $\tilde{q}$ is monotonic with respect to α, the conformalization step disrupts this consistency by adding an adjustment term. Thus, similar to (Rodríguez et al. 2023), we apply a monotonicity loss to $\hat{q}$:

$$L_M(\hat{q}_t^\alpha, \alpha) = \max\left(0, \frac{\partial \hat{q}_t^\alpha}{\partial \alpha}\right)$$

We approximate the derivatives using a finite difference scheme, which makes our monotonicity loss take the form:

$$L_M = \sum_{i=1}^{n-1} \max\left(0, \frac{\hat{q}_t^{\alpha_{i+1}} - \hat{q}_t^{\alpha_i}}{\alpha_{i+1} - \alpha_i}\right)$$

Such loss encourages the model to produce monotonic output, i.e, when $\alpha_1 > \alpha_2$, $\hat{q}^{\alpha_1} < \hat{q}^{\alpha_2}$.

At test time, while our cumulative design and monotonicity loss enhance overall monotonic behavior, there are instances where distribution validity is compromised due to few non-monotonic terms. To address this, we apply a TTA procedure right before conformalization. Specifically, we pass the combined embedding from the model through a MLP layer, which outputs a term h_i for each $\alpha_i, i \in [1, n]$. For each i, we adjust $\tilde{q}^{\alpha_i}$ by adding h_i and then apply the monotonicity loss to the conformalized $\hat{q}$s. If the DCS (i.e., the number of instances where the model provides consistent quantiles) meets the user-specified threshold, we return the prediction and proceed to the next time step. Otherwise, we repeat the adaptation process until the desired DCS is achieved.

Experiments

We comprehensively compare NCC with popular CP time series methods across multiple datasets and base forecasters. To align with real-world applications, all experiments emulate a real-time forecasting setup. Additionally, we conduct experiments on few-shot learning and model ablation. Detailed information about our experimental setup and results is provided in the Appendix.

Experimental Setting

Datasets and Base Forecasters. We evaluate on a diverse set of publicly available real-world datasets, including commonly-used datasets in time series forecasting (`weather`, `electric`) and datasets that demonstrate significant amount of temporal distribution shifts (`flu`, `covid-19`, `smd`). To illustrate that our method does not depend on a specific base forecaster, we get the point predictions using three base forecasters: seq2seq GRU model, Theta model (Assimakopoulos and Nikolopoulos 2000), and Informer model (Zhou et al. 2021).

Baselines. We compare the calibration power of NCC to four popular time series conformal prediction methods. Conformal forecasting RNNs (CF-RNN) (Stankeviciute, M Alaa, and van der Schaar 2021) generate prediction intervals for multi-horizon forecasts while assuming data exchangeability, similar to conventional conformal prediction methods. Nonexchangeable conformal prediction (NEXCP) (Barber et al. 2023) introduces weighted quantiles to deal with distribution shifts. ACI (Gibbs and Candes 2021), adaptive conformal inference, which adapts the user-specified error rate using online gradient descent. C-PID (Angelopoulos, Candes, and Tibshirani 2024), conformal PID control, which introduces PID control to time series CP. Since using a scorecaster (D-part) in C-PID breaks the theoretical coverage guarantee, we only use the quantile tracker and error integrator, which are the P and I part in PID. As is shown in Table 1, CF-RNN and NEXCP are methods under distribution shift assumptions. ACI and C-PID are conformal control methods. Note that these baselines cannot incorporate auxiliary data. Therefore, to provide a fair comparison, NCC uses multi-view data only when studying the effect of incorporating multi-view data.

Evaluation Metrics. We evaluate our model and baselines using carefully chosen metrics that are widely used in the literature to measure probabilistic accuracy and calibration (Ka-

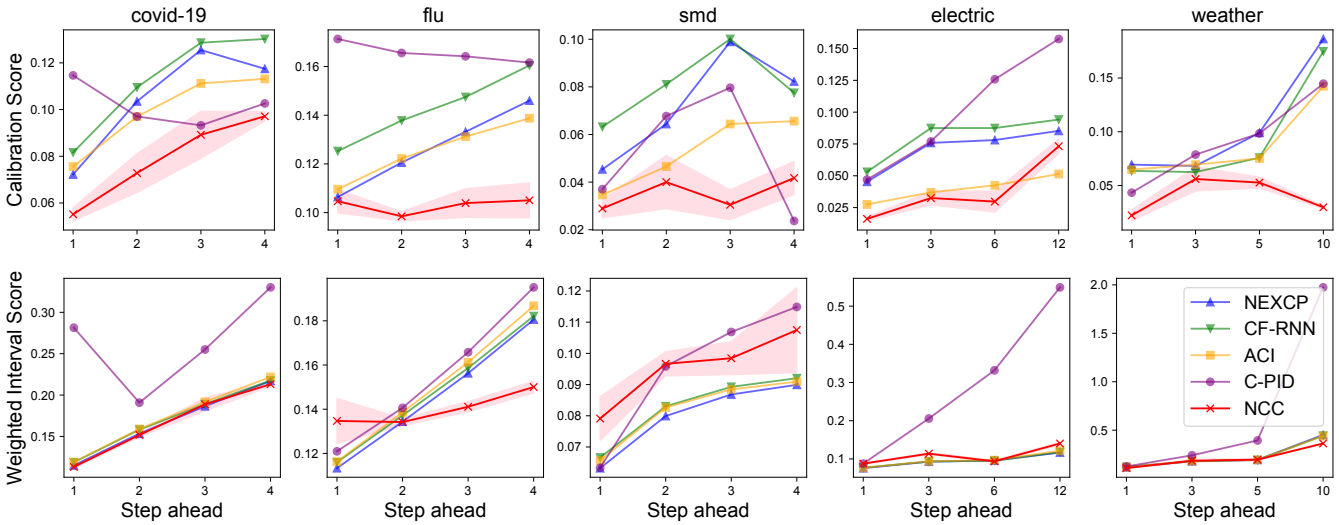

Figure 2: Comparison of CS and WIS across five datasets for four different steps ahead predictions. Our method (NCC) is highlighted with a red line, with the pink-shaded area around it indicating the error margin. NCC has the best CS across all datasets. Except for the `smd` dataset, NCC consistently achieves the best or near-best performance in WIS.

marthi et al. 2022; Xu and Xie 2021). These metrics include Calibration Score (CS), Continuous Ranked Probability Score (CRPS) and Weighted Interval Score (WIS). The CS is used to measure how much the empirical coverage rate differs from the ideal coverage rate. CRPS evaluates the efficiency and coverage at multiple miscoverage levels. WIS is often used as an approximation of CRPS (Bracher et al. 2021). In our experiments, the results for CRPS and WIS are similar. Thus, we present the CRPS in the appendix. For CS, WIS and CRPS, smaller values indicate better preformance. To evaluate distribution consistency, we propose the Distribution Consistency Score (DCS), which quantifies the ratio of consistent predictions to the total number of predictions.

Comparison with Other CP Methods

Here we focus on the seq2seq model as our base forecaster. Additional results with other base forecasters are provided in our Appendix. As discussed before, most time series CP methods exhibit some degree of distribution inconsistency, which makes intervals invalid. For example, C-PID has 0% valid intervals for `flu`. To enable comparisons, we first make the intervals valid by sorting them, which is a commonly used approach (Chernozhukov, Fernández-Val, and Galichon 2010; Chernozhukov, Fernandez-Val, and Galichon 2009).

Figure 2 provides a summary of our main findings, with detailed numerical values available in Table 6 in the Appendix. The plot illustrates the CS and WIS across five datasets, evaluated at four different forecasting horizons. For CS, NCC consistently achieves the smallest values, indicating superior calibration performance, except for the four-step-ahead cases on `smd` and `electric`. Notably, CS tends to increase as the step becomes larger, yet NCC demonstrates stability, maintaining a nearly flat trend even as prediction accuracy decreases at longer horizons. For WIS, NCC outperforms the baselines on the `covid-19`, `flu`, and `weather` datasets. On `electric`, NCC performs comparably to the best results, showcasing its overall effectiveness. It might be surprising that ACI outperforms C-PID in many cases in CS. This occurs because C-PID, though it adapts quickly under distribution shifts, produces less consistent intervals. Then the sorting greatly reduces the performance of the C-PID. In contrast, NCC, which also exhibits strong adaptivity, maintains good performance even when sorting is applied. Figure 3(a) illustrates the reason. The figure shows DCS against CS when the prediction intervals are unsorted. It reveals that, except for our method, the more adaptive a method is (indicated by a lower CS), the more inconsistent its predictions tend to be (indicated by a lower DCS). For example, while C-PID achieves the best CS on most datasets, it shows nearly zero DCS for `smd`, `electric`, and `weather`. NCC, however, successfully finds a balance between adaptivity and distribution consistency. Across all five datasets, NCC's markers are consistently near the top left corner of the plot, which represents the optimal performance region.

Few-shot Learning Ability

In this section, we demonstrate that our method achieves superior performance in a few-shot learning setup. For testing, we used COVID-19 hospitalization data from the 2021-2022 respiratory virus season within the `covid-19` dataset, training on the data from the 10 weeks before the respiratory virus season. The test period spans 28 weeks. Figure 3(c) shows the cumulative average calibration score starting from the sixth week. During the respiratory virus season, COVID-19 hospitalization exhibits a rapid increase followed by a decrease, making calibration particularly challenging at the start and end of the season. As shown by the red line in the plot, despite the aforementioned challenges, the CS of NCC rapidly drops below 0.1, demonstrating its strong adaptability. In contrast, the CS of other methods remains high by week 10.

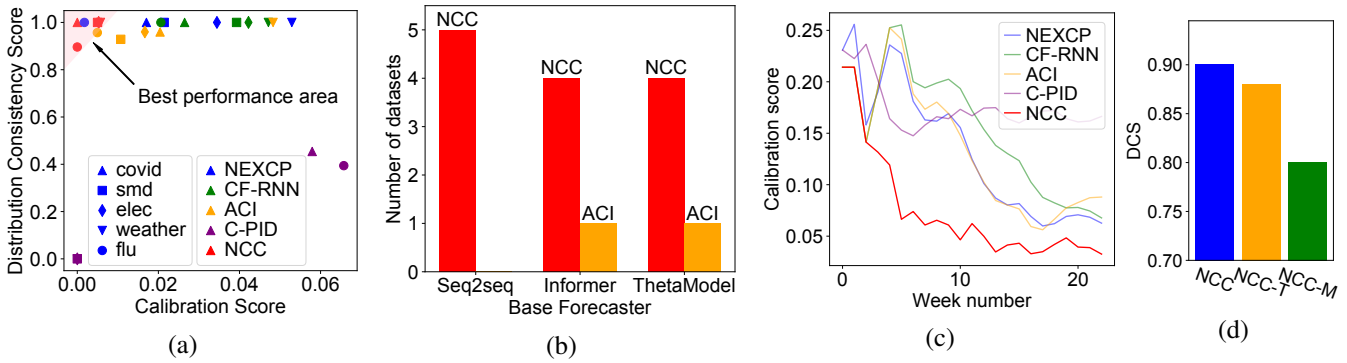

Figure 3: (a) DCS versus CS, using the **unsorted** results. The top left corner represents the optimal performance region. (b) The number of datasets where our method achieves the best CS. The methods that do not achieve the best CS are not shown in this plot. (c) Cumulative average CS on `covid-19` in a few-shot learning setup. (d) Ablation study on TTA and monotonicity loss. Comparison of NCC, NCC without TTA (NCC-T) and NCC without TTA and monotonicity loss (NCC-M).

Furthermore, NCC consistently achieves the lowest calibration scores. Notably, the baselines either require a number of observations of non-conformity scores before they can adapt, or rely on online gradient descent for adjustments, resulting in significantly high initial CS. Since NCC does not rely on online gradient descent and effectively uses additional data from other regions through transfer learning, it achieves optimal calibration performance from the beginning.

Robust Calibration across Various Base Forecasters

In this section, we demonstrate that NCC consistently achieves the best performance, irrespective of the base forecaster used. In addition to the seq2seq model, we further use an Informer and a Theta model as the base forecaster. Detailed results are provided in the Appendix in Table 4. Figure 3(b) presents the number of datasets in which each method excels, based on CS. Across all base forecasters, our method outperforms others by achieving the best CS on at least four datasets, which highlights its superior calibration capabilities.

Ablation Study

Effect of Quantile Loss and Coverage Loss. We evaluate across all five datasets on one step ahead predictions when either quantile loss or coverage loss is excluded. Note that when coverage loss is removed, we also exclude efficiency loss. The results, presented in Table 2 in the Appendix, show a significant decrease in CS when coverage loss is excluded, which suggests that quantile loss alone is insufficient to achieve a valid empirical coverage rate. However, quantile loss remains essential to the model as it enables learning the conditional distribution, which serves as a foundation for coverage loss to further improve performance. When quantile loss is removed, the CS degrades only slightly, but the WIS increases dramatically, indicating that the resulting prediction intervals become much larger (less efficient).

Effect of Incorporating Monotonicity Loss and TTA. In this section, we explore the methods used to enhance the distribution consistency of prediction intervals. In Figure 3(d) we compare the Distribution Consistency Score (DCS) for NCC, NCC without TTA (NCC-T), and NCC without both TTA and monotonicity loss (NCC-M). The results demonstrate that both the monotonicity loss and TTA contribute to improving distribution consistency.

Effect of Incorporating Multi-view Data. We compare the performance of NCC with and without the use of multi-view data. Table 3 in Appendix summarizes the results. Overall, incorporating multi-view data reduces CS by up to 18%.

Conclusion

We introduce NCC, one of the first end-to-end deep learning frameworks for time series CP. By leveraging the powerful data ingestion capabilities and data-driven flexibility of deep learning models, our method provides more adaptive prediction intervals showcasing better calibration. An advantage of our approach is its ability to incorporate domain expert requirements, such as avoiding quantile crossing, which is a common challenge in real-world applications. Through extensive evaluations on diverse datasets and comparisons against competitive baselines, we demonstrate the superior performance and practicality of NCC. We believe that our work paves the way for a tighter integration of conformal prediction techniques with deep learning, enabling more reliable and trustworthy decision-making processes in various domains that rely on accurate time series forecasting and UQ.

Limitations and Future Work: We acknowledge some limitations of our method. In a few datasets, our method constructs wider prediction intervals than others. The reason can be that our model is trained using multiple losses for different objectives, in which case the efficiency loss can be overlooked due to its scale. We also found our model's performance can be sensitive to hyperparameters. Future work could explore advanced methods for balancing losses and making architectural modifications to enhance robustness. Additionally, integrating recent advancements in time series prediction, such as decomposition techniques (Zeng et al. 2023) and generative models (Yuan and Qiao 2024), may offer further improvements.

Acknowledgements

This work was supported in part by the Centers for Disease Control and Prevention Award NU38FT000002 and start-up funds from the University of Michigan.

References

Angelopoulos, A.; Candes, E.; and Tibshirani, R. J. 2024. Conformal PID control for time series prediction. *Advances in neural information processing systems*, 36.

Assimakopoulos, V.; and Nikolopoulos, K. 2000. The theta model: a decomposition approach to forecasting. *International journal of forecasting*, 16(4): 521–530.

Barber, R. F.; Candes, E. J.; Ramdas, A.; and Tibshirani, R. J. 2023. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2): 816–845.

Bhatnagar, A.; Wang, H.; Xiong, C.; and Bai, Y. 2023. Improved online conformal prediction via strongly adaptive online learning. *arXiv preprint arXiv:2302.07869*.

Bracher, J.; Ray, E. L.; Gneiting, T.; and Reich, N. G. 2021. Evaluating epidemic forecasts in an interval format. *PLoS computational biology*, 17(2): e1008618.

Chernozhukov, V.; Fernandez-Val, I.; and Galichon, A. 2009. Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3): 559–575.

Chernozhukov, V.; Fernández-Val, I.; and Galichon, A. 2010. Quantile and probability curves without crossing. *Econometrica*, 78(3): 1093–1125.

Chernozhukov, V.; Wüthrich, K.; and Zhu, Y. 2021. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48): e2107794118.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2015. Gated feedback recurrent neural networks. In *International conference on machine learning*, 2067–2075. PMLR.

Cramer, E. Y.; Ray, E. L.; Lopez, V. K.; Bracher, J.; Brennen, A.; Castro Rivadeneira, A. J.; Gerding, A.; Gneiting, T.; House, K. H.; Huang, Y.; et al. 2022. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, 119(15): e2113561119.

Deng, S.; Wang, S.; Rangwala, H.; Wang, L.; and Ning, Y. 2020. Cola-GNN: Cross-location Attention based Graph Neural Networks for Long-term ILI Prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 245–254.

Fisch, A.; Schuster, T.; Jaakkola, T.; and Barzilay, R. 2021. Few-shot conformal prediction with auxiliary tasks. In *International Conference on Machine Learning*, 3329–3339. PMLR.

Fortunato, M.; Blundell, C.; and Vinyals, O. 2017. Bayesian recurrent neural networks. *arXiv preprint arXiv:1704.02798*.

Frazier, P. I. 2018. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Gasthaus, J.; Benidis, K.; Wang, Y.; Rangapuram, S. S.; Salinas, D.; Flunkert, V.; and Januschowski, T. 2019. Probabilistic forecasting with spline quantile function RNNs. In *The 22nd international conference on artificial intelligence and statistics*, 1901–1910. PMLR.

Gawlikowski, J.; Tassi, C. R. N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1): 1513–1589.

Gibbs, I.; and Candes, E. 2021. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34: 1660–1672.

Gibbs, I.; and Candès, E. J. 2024. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162): 1–36.

Gneiting, T.; and Katzfuss, M. 2014. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1: 125–151.

Jain, S.; Liu, G.; Mueller, J.; and Gifford, D. 2020. Maximizing overall diversity for improved uncertainty estimates in deep ensembles. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 4264–4271.

Johansson, M. A.; Apfeldorf, K. M.; Dobson, S.; Devita, J.; et al. 2019. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48): 24268–24274.

Kamarthi, H.; Kong, L.; Rodriguez, A.; Zhang, C.; and Prakash, B. A. 2021. When in doubt: Neural non-parametric uncertainty quantification for epidemic forecasting. *Advances in Neural Information Processing Systems*, 34: 19796–19807.

Kamarthi, H.; Kong, L.; Rodríguez, A.; Zhang, C.; and Prakash, B. A. 2022. CAMul: Calibrated and Accurate Multi-view Time-Series Forecasting. In *Proceedings of the ACM Web Conference 2022*, WWW '22, 3174–3185. New York, NY, USA: Association for Computing Machinery. ISBN 9781450390965.

Kipf, T. N.; and Welling, M. 2016. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

Koenker, R. 2005. *Quantile regression*, volume 38. Cambridge university press.

Kuleshov, V.; Fenner, N.; and Ermon, S. 2018. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, 2796–2804. PMLR.

Lai, G.; Chang, W.-C.; Yang, Y.; and Liu, H. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 95–104.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Mathis, S. M.; Webber, A. E.; León, T. M.; Murray, E. L.; Sun, M.; White, L. A.; Brooks, L. C.; Green, A.; Hu, A. J.; Rosenfeld, R.; et al. 2024. Evaluation of FluSight influenza forecasting in the 2021–22 and 2022–23 seasons with a new target laboratory-confirmed influenza hospitalizations. *Nature communications*, 15(1): 6289.

McDermott, P. L.; and Wikle, C. K. 2019. Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data. *Entropy*, 21(2): 184.

Oliveira, R. I.; Orenstein, P.; Ramos, T.; and Romano, J. V. 2024. Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25(225): 1–38.

Papadopoulos, H.; Vovk, V.; and Gammerman, A. 2007. Conformal prediction with neural networks. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, 388–395. IEEE.

Pearce, T.; Brintrup, A.; Zaki, M.; and Neely, A. 2018. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International conference on machine learning*, 4075–4084. PMLR.

Qian, Z.; Alaa, A. M.; and van der Schaar, M. 2020. When and How to Lift the Lockdown? Global COVID-19 Scenario Analysis and Policy Assessment using Compartmental Gaussian Processes. *Advances in Neural Information Processing Systems*, 33.

Reich, N. G.; Brooks, L. C.; Fox, S. J.; Kandula, S.; McGowan, C. J.; Moore, E.; Osthus, D.; Ray, E. L.; Tushar, A.; Yamana, T. K.; et al. 2019. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences*, 116(8): 3146–3154.

Rodríguez, A.; Cui, J.; Ramakrishnan, N.; Adhikari, B.; and Prakash, B. A. 2023. EINNs: Epidemiologically-informed neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 14453–14460.

Rodríguez, A.; Kamarthi, H.; Agarwal, P.; Ho, J.; Patel, M.; Sapre, S.; and Prakash, B. A. 2024. Machine learning for data-centric epidemic forecasting. *Nature Machine Intelligence*, 1–10.

Rodríguez, A.; Tabassum, A.; Cui, J.; Xie, J.; Ho, J.; Agarwal, P.; Adhikari, B.; and Prakash, B. A. 2021. Deepcovid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15393–15400.

Romano, Y.; Patterson, E.; and Candes, E. 2019. Conformalized quantile regression. *Advances in neural information processing systems*, 32.

Shafer, G.; and Vovk, V. 2008. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(3).

Stankeviciute, K.; M Alaa, A.; and van der Schaar, M. 2021. Conformal time-series forecasting. *Advances in neural information processing systems*, 34: 6216–6228.

Su, Y.; Zhao, Y.; Niu, C.; Liu, R.; Sun, W.; and Pei, D. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2828–2837.

Vaswani, A.; et al. 2017. Attention is all you need. *NIPS*.

Viboud, C.; Sun, K.; Gaffey, R.; Ajelli, M.; Fumanelli, L.; Merler, S.; Zhang, Q.; Chowell, G.; Simonsen, L.; Vespignani, A.; et al. 2018. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, 22: 13–21.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Xu, C.; and Xie, Y. 2021. Conformal prediction interval for dynamic time-series. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 11559–11569. PMLR.

Xu, C.; and Xie, Y. 2023. Sequential Predictive Conformal Inference for Time Series. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 38707–38727. PMLR.

Yan, X.; Hu, S.; Mao, Y.; Ye, Y.; and Yu, H. 2021. Deep multi-view learning methods: A review. *Neurocomputing*, 448: 106–129.

Yuan, X.; and Qiao, Y. 2024. Diffusion-TS: Interpretable Diffusion for General Time Series Generation. In *The Twelfth International Conference on Learning Representations*.

Zaffran, M.; Féron, O.; Goude, Y.; Josse, J.; and Dieuleveut, A. 2022. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, 25834–25866. PMLR.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11121–11128.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.