Dynamic Dual-Feedback Conformal Inference for Time Series Forecasting

Songlin Du Ling Luo Uwe Aickelin
Faculty of Engineering and Information Technology
School of Computing and Information Systems
University of Melbourne
sodu@student.unimelb.edu.au ling.luo@unimelb.edu.au
uwe.aickelin@unimelb.edu.au

Abstract

Quantifying uncertainty in time series forecasting is particularly demanding because sequential data exhibit temporal dependence and are prone to distributional changes. Conformal inference has emerged as a powerful uncertainty quantification approach through the construction of prediction sets. Recent advances have introduced online conformal methods that adaptively adjust prediction thresholds through feedback mechanisms. However, the existing feedback mechanism typically relies solely on miscoverage indicators (actual feedback)—whether the true label falls within the interval at each time step—while overlooking the empirical prediction threshold (estimated feedback) that is derived from the oracle conformal method. In this paper, we propose *Dynamic Dual-feedback Conformal Inference* (DDCI), which incorporates a dual-feedback mechanism consisting of *actual feedback* and *estimated feedback*. The former drives the primary adjustment of the intervals based on true observations, while the latter dampens excessive expansions or contractions by leveraging empirical thresholds from conformal inference during updates. By balancing these two signals, DDCI achieves more stable and narrower prediction intervals in sequential settings while maintaining near-nominal coverage in practice. Code is available here: `https://github.com/sdu1998/DDCI`.

1 Introduction

Machine learning models are increasingly being applied in many real-world domains such as health care, energy and transportation, where uncertainty quantification is essential for decision-making [17, 10, 29]. Generally speaking, prediction sets/intervals are a common approach to reflect models' uncertainty. However, traditional machine learning or statistical models themselves are difficult to quantify uncertainty, or require strong assumptions on the data distributions [19, 21, 36].

Conformal inference, firstly introduced by Vovk et al. [43], is a non-parametric framework to construct prediction sets, which enjoys statistically marginal coverage guarantee under a lenient assumption called exchangeability. This framework offers a general methodology for converting the outputs of black-box models into prediction sets and can be applied on top of any underlying algorithm [5]. The simplicity and generality of conformal methods have led to their wide adoption in real-world applications [20, 30]. Fundamentally, the prediction sets generated by conformal methods are guaranteed to contain the true response with a user-defined probability. However, exchangeability often fails to hold in sequential data and the validity of conformal inference may be compromised because of distributional changes arising from external factors such as seasonality or events. Motivated by this challenge, this work studies the problem of conformal inference for time series data.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In recent years, substantial effort has been devoted to developing conformal methods tailored for non-exchangeable time series data to achieve the target coverage guarantee. A prominent line of research is to update the prediction interval via feedback mechanisms. The very first method, Adaptive Conformal Inference (ACI), attains the desired coverage rate by dynamically adjusting target quantiles, treating distributional shifts as a single-parameter learning problem [22]. Following that, multiple methods have been designed to improve the predictive efficiency [50, 13, 4]. These methods update the length of the interval by determining whether the true label falls within the interval (binary feedback from current thresholds). Recently, Wu et al. [47] proposed Error-quantified Conformal Inference (ECI), which incorporates the degree of under- or over-coverage to refine interval lengths. However, their approach does not fully exploit the magnitude of the error, thereby weakening the corrective effect of the feedback. More critically, a major limitation of all existing methods is that all of them rely exclusively on the signal from the current thresholds, which makes the update process unstable and prone to fluctuations. In other words, the interval will gradually shrink until it can no longer cover the true value at a certain point in time, after which it will expand.

Due to the limitations mentioned above, we first propose a dynamic feedback function that more precisely quantifies the *actual feedback*. Secondly, to mitigate potential overreaction to this signal, we introduce an additional component, termed *estimated feedback*, which is from empirical thresholds derived from split conformal methods. This second feedback introduces a counter-signal with opposite sign to the actual feedback through the empirical thresholds, serving as a control mechanism that suppresses excessive adjustments and ensures smoother, more stable updates. Our main contributions are summarised as follows:

- We propose a novel online conformal methods *Dynamic Dual-feedback Conformal Inference* (DDCI) for time series forecasting, which is able to quantify the actual feedback precisely and explores how empirical thresholds from conformal methods influence the stability of the updating process.
- Furthermore, we investigate the impact of different degrees of estimated feedback on the smoothness and efficiency of interval outputs. We also compare different kinds of estimated feedback from the original conformal method and the non-exchangeable conformal method and analyse the differences among them.
- We conduct considerable experiments on 3 synthetic and 5 real datasets to demonstrate that, in the vast majority of cases, our proposed methods can substantially shorten the prediction intervals without affecting the target coverage rate, compared to all existing methods.

2 Related Work

2.1 Conformal Prediction

The task of uncertainty quantification for unobserved data has been widely studied [1, 37]. Conformal inference, pioneered by Vovk et al. [44], is a significant branch of uncertainty quantification methods. The power of conformal methods stems largely from its flexibility, simplicity and theoretical guarantee. Conformal methods are able to quantify distribution-free uncertainty under a lenient assumption of data distribution called exchangeability. Multiple variants of conformal methods and score functions have been developed to improve the adaptivity and efficiency [42, 27, 35, 3, 33, 11, 2] and have been applied to many areas [45, 38, 15, 28, 51, 18]. A detailed introduction of conformal inference and the theory behind it are provided by Angelopoulos et al. [5, 7].

2.2 Conformal Inference under Non-exchangeability

Conformal methods highly rely on the exchangeability, but it is often violated when they are applied to practical problems (e.g., sequential data). Recent works of developing conformal methods beyond exchangeability can be summarised into two categories. The first is to treat the past conformal scores non-equally by assigning weights, giving greater emphasis to those more similar to the present [40, 48, 12]. Recent approaches model temporal dependence and rely on asymptotic guarantees that require assumptions on the underlying data-generating process [49, 9]. The second is to adaptively update the significance level α or the interval directly based on the miscoverage feedback [22, 50, 13, 23]. The following work [4] advanced this line of research by framing PID control theory to prospectively model conformal scores in a predictive and adaptive manner for online scenarios. Angelopoulos

et al. [6] decay the learning rate based on time steps. Following that, Wu et al. [47] introduced an error-quantified term to update the interval by using a smoothed quantile loss. Although our work is highly inspired by the error-quantified term of Wu et al. [47], it differs in two aspects: (i) our feedback mechanism is residual-aware which makes it more applicable to different datasets and underlying algorithms; (ii) ours also considers an estimated feedback that is from conformal inference itself, where all existing methods did not take this empirical threshold into account. Overall, our method is therefore able to produce substantially tighter prediction sets while maintaining its validity. There are some very recent works related to this area [31, 25, 26]. Closely related to our work, the concurrent work Conformal Optimistic Prediction (COP) proposes an optimistic conformal prediction framework [25]. It leverages the distribution information of non-conformity scores as an optimistic hint within the Online Optimistic Gradient Descent framework to refine the prediction intervals.

3 Problem Setup

Conformal Inference for Time Series. Under the background of time series forecasting, we observe a sequence of covariates $x_t \in \mathcal{X}$ and responses $y_t \in \mathcal{Y}$ for $t \in \mathbb{N} = \{1, 2, 3, ..., T\}$. Then, we train an underlying model $\theta_t(\cdot)$ on the data before time t. As in oracle conformal methods, the conformal score function $s_t(\cdot, \cdot)$ can be designed in a flexible manner, which is able to measure the accuracy of model's prediction at time t. A typical choice of score function for regression problems is using the absolute error $s_t(x_t, y_t) = |y_t - \theta_t(x_t)|$. The main goal is to construct a prediction interval:

$$C_t(x_t) = \{y_t \in \mathcal{Y} : s_t(x_t, y_t) \leq q_t^*\}, \tag{1}$$

where q_t^*, the empirical threshold, is the $(1-\alpha)$-th quantile from the distribution of past w scores $S_t = \{s_i(x_i, y_i) : t - w \leq i < t\}$. In principle, if the data is exchangeable, the interval output is expected to be marginally valid: $P(y_t \in C_t(x_t)) \geq 1 - \alpha$.

Nevertheless, the challenge for time series forecasting is that the sequence data can be non-exchangeable due to external factors (e.g., distribution shifts). Therefore, a general goal is to achieve a long-run coverage guarantee that $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} err_t = \alpha$, where $err_t = \mathbb{1}\{y_t \notin C_t(x_t)\}$ [4, 34].

Online Conformal Inference. The online update of prediction interval started from ACI [22] and evolved into the quantile tracking component inside the conformal PID control method [4], called online gradient descent (OGD). It considers an optimisation problem to minimise the quantile loss $\ell_t(q_t)$,

$$q_{t+1} := q_t - \eta_t \nabla \ell_t(q_t) = q_t + \eta_t \big(\text{err}_t - \alpha\big), \tag{2}$$

where η_t is a non-negative learning rate and $\ell_t(q_t) = (s_t - q_t)\big(\mathbf{1}\{s_t > q_t\} - \alpha\big)$ means the $(1-\alpha)$-th quantile loss. Simply put, if the interval $C_t(x_t)$ fails to cover the true label at time t, the interval is expanded to make it conservative in the next step; otherwise, it is shortened.

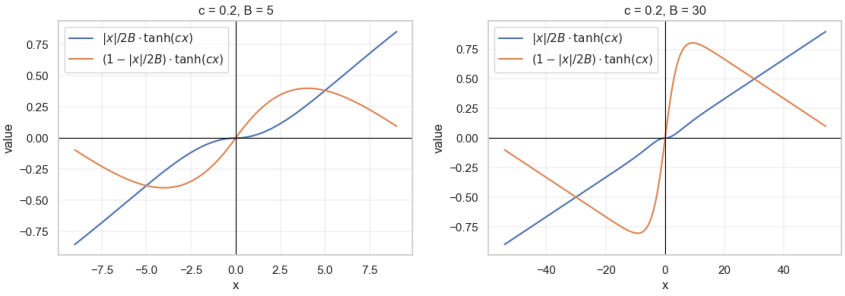
4 Methodology

4.1 Dynamic Dual-feedback Conformal Inference

In the standard OGD update rule (Eq. 2), the threshold adjustment relies solely on the binary coverage feedback ($\text{err}_t - \alpha$). While effective, this signal is too coarse as it only indicates whether the prediction interval covers the true label, without reflecting the magnitude or reliability of the error. Additionally, relying on a single feedback can make the update process unstable and inclined to volatility.

To address this limitation, we augment the update rule with two complementary forms of feedback. To clarify, q_t and q_t^* denote the current threshold and the estimated threshold from conformal inference respectively. The first is the *actual feedback*, expressed as

$$\mathcal{A}(e_t) = \rho_t h(c \cdot e_t), \quad \rho_t = \min\Big(\frac{|e_t|}{\max_{t-w\leq i,j\leq t} |s_i - s_j|}, 1\Big) \in [0, 1], \tag{3}$$



(a) $B = 5$ (b) $B = 30$

Figure 1: Functional shape comparison of the actual feedback $\frac{|x|}{2B}\tanh(cx)$ (blue) and the estimated feedback $(1 - \frac{|x|}{2B})\tanh(cx)$ (orange) with $c = 0.2$ and $B = 5$ or 30. The two functions are plotted on the same axis to highlight their distinct regimes of dominance: the former grows in magnitude and dominates for large |x|, whereas the latter exhibits attenuation and dominates near the origin.

which incorporates the observed error $e_t = s_t - q_t$ and uses a squashing function $h(\cdot)$ to weight the update by the severity of miscoverage. This squashing function is capable of mapping the error to $[-d, d]$ and is origin-symmetric, e.g., $tanh(x)$. This ensures that large deviations from the threshold lead to stronger corrective actions. The parameter c governs the steepness of the curve.

Furthermore, we employ the second feedback, termed *estimated feedback*, given by

$$\mathcal{E}(e_t^*) = (1 - \rho_t^*)\, h(c \cdot e_t^*), \qquad \rho_t^* = \frac{|e_t^*|}{\max_{t-w \le i,j \le t} |s_i - s_j|} \in [0, 1], \tag{4}$$

which provides a stabilising signal based on the estimated error $e_t^* = s_t - q_t^*$ and q_t^* is the empirical $(1 - \alpha)$-th quantile of scores $S_t = \{s_i(x_i, y_i) : t - w \le i < t\}$.

For theoretical simplicity, we set $\max_{t-w \le i,j \le t} |s_i - s_j| = 2B$ based on the Assumption 4.1. In practice, we use the maximum and minimum from the past window. This empirical quantile q_t^* acts as a stabilising anchor. By referencing the robust empirical threshold q_t^*, it provides a local distributional context that prevents the interval from overfitting to instantaneous noise. Notice the term $(1 - \rho_t^*)$ acts as a natural adaptive gate. When e_t^* is small ($|e_t^*| \ll 2B$), the term allows the estimated feedback to effectively dampen noise and stabilise the interval. When e_t^* are large ($|e_t^*| \to 2B$), the term vanishes, automatically diminishing the estimated feedback. This allows the algorithm to prioritise the actual feedback for rapid adaptation without interference. Thus, the mechanism intelligently balances stability and responsiveness based on the magnitude of the prediction error relative to the global bound $2B$. A visual comparison of these two feedback is given in Fig. 1, with $tanh(x)$ as an example squashing function.

Therefore, we propose *Dynamic Dual-feedback Conformal Inference* (DDCI), based on the OGD method (Eq. 2) and the dual-feedback mechanism, the update rule is given below:

$$q_{t+1} := q_t + \eta_t(\text{err}_t - \alpha) + \eta_t(\underbrace{\mathcal{A}(s_t - q_t)}_{\text{Actual feedback}} - \underbrace{\epsilon\mathcal{E}(s_t - q_t^*)}_{\text{Estimated feedback}}), \tag{5}$$

where $\epsilon > 0$ is a parameter to regulate the influence of the estimated feedback term. The estimated feedback enters the update rule with an opposite sign to the actual feedback term, forming a push-pull mechanism that prevents over-reaction. Specifically, when both $\mathcal{A}(e_t)$ and $\mathcal{E}(e_t^*)$ are positive, the actual feedback drives q_t upward while the estimated feedback applies a counteracting downward pull. This mechanism is particularly effective when $|q_t - q_t^*|$ is small: $\mathcal{A}$ and $\mathcal{E}$ have similar magnitudes, allowing the estimated feedback to dampen excessive expansion. Conversely, when q_t deviates substantially from q_t^* (indicating a genuine distribution shift), the actual feedback dominates and drives the threshold to adapt quickly. As shown in Fig. 1, the actual feedback term $\mathcal{A}(e_t)$ is highly sensitive to sudden shifts or anomalous values of s_t, which can exert a strong influence on the update. By balancing these signals, the update adapts flexibly to distribution changes while maintaining robustness, thereby constructing prediction intervals that are both responsive and stable.

To further understand this dynamic feedback mechanism, a mathematical decomposition of the additional feedback mechanism within Appendix A.2 (Stabilised Control term) reveals that the

mechanism acts as an adaptive gating function: it damps high-frequency noise when $|e_t|$ is small while restoring corrective force during significant distribution shifts.

4.2 Theoretical Analysis

The main theoretical analysis of regret bound and long-run coverage of the proposed method DDCI is presented in this section. More details and the analysis on the parameter ϵ are provided in the Appendix A.

At first, we start from a general assumption for conformal scores, which sets a boundary [4, 47].

Assumption 4.1. *For any $t \in \mathbb{N}_+$, there exists $0 < B < \infty$ such that $s_t \in [-B, B]$.*

Assumption 4.2. *The origin-symmetric squashing function $h(\cdot)$ is chosen to be a bounded ($|h(\cdot)| \leq d$ for $x \in R$), and L_h-Lipschitz continuous function. Consequently, for any $c > 0$ and $x, y \in \mathbb{R}$: $|h(cx) - h(cy)| \leq L_h|cx - cy| = cL_h|x - y|$.*

Based on these assumptions, there is a boundary for q_t, which are given in the following Proposition.

Proposition 4.3. *Fix an initial threshold $q_1 \in [-B, B]$. Assume that $\eta_t \leq \frac{2B}{1-\alpha+d(1+\epsilon)}$. The update rule in Eq. 5 with arbitrary non-negative learning rate η_t satisfies that,*

$$-B - (\alpha + d(1+\epsilon))M_{t-1} \leq q_t \leq B + (1 - \alpha + d(1+\epsilon))M_{t-1} \quad \forall t \geq 1, \tag{6}$$

where $M_0 = 0$, $M_t = \max_{1 \leq r \leq t} \eta_r$ for $t \geq 1$. The upper bound $\eta_t \leq \frac{2B}{1-\alpha+d(1+\epsilon)}$ ensures that a single update step cannot move q_t across the entire range $2B$.

4.2.1 Regret Bound

At first, we analyse the theoretical stability of DDCI using the framework of Dynamic Regret [16, 32, 52]. Dynamic regret generalises the standard regret framework to non-stationary environments by bench-marking the learner against a sequence of time-varying optimal comparators u_t, rather than a single fixed parameter [53].

Theorem 4.4. *DDCI achieves the following dynamic regret bound for arbitrary u_t with a fixed learning rate η and a small ϵ ($\epsilon \leq 1$):*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(\ell_t(q_t) - \ell_t(u_t)) \leq & \frac{2B^2}{\eta T} + \frac{2B + \eta K_+}{\eta T}\sum_{t=1}^{T}||u_{t+1} - u_t|| + 2\epsilon^2\gamma^2\frac{\eta}{T}\sum_{t=1}^{T}||q_t - q_t^*||^2 \\ & + \frac{(2B + \eta K_+)\epsilon\gamma}{T}\sum_{t=1}^{T}||q_t - q_t^*|| + (2B + \eta K_+)d + \eta(1 + 2d^2),\end{aligned} \tag{7}$$

where $\gamma = cL_h + \frac{d}{2B}$ and $K_+ = 1 - \alpha + d(1+\epsilon)$. The term $\sum_{t=1}^{T}\|u_{t+1} - u_t\|$ represents the path length of the optimal comparator sequence and captures the non-stationarity of the environment. The proof is provided in the Appendix A.2.

A critical distinction between DDCI and single feedback methods (e.g., ECI) lies in their sensitivity to the stochastic nature of conformal scores. Typically, when the underlying algorithm is well-trained, the conformal score s_t tends to be highly stochastic, exhibiting significant variance. Existing methods often suffer from updates that are explicitly dependent on the magnitude of observed error $||s_t - q_t||$. Their structure couples the algorithmic performance with the inherent uncertainty of the data. Consequently, even if the algorithm converges to the optimal parameter, high-variance noise in s_t inevitably limits the algorithm's stability and makes it sensitive to outliers. In contrast, our analysis reveals that the stability of DDCI can be controlled by the empirical quantile q_t^*. This formulation mitigates the irreducible data noise from the learning process but its effectiveness highly relies on the quality of q_t^*. Theorem 4.4 presents a worst-case adversarial bound containing a non-vanishing penalty $\mathcal{O}(d)$, which is the limitation of introducing the additional feedback. Paying this $\mathcal{O}(d)$ in the worst-case bound is the cost for a strong insurance policy against extreme scenarios. However, this cost would be small since it is linearly related to the steepness parameter c.

Remark 4.5. Specifically, during time intervals where the score distribution remains invariant (i.e., locally exchangeable, which means that the distribution of the current score s_t is consistent with the

scores in its preceding window of size w in our case), the empirical q_t^* acts as a theoretically consistent surrogate for the true local measure, and our analysis of the additional feedback mechanism implies that using q_t^* can weaken the influence from the s_t. This capability to adapt to local stationarity is formally substantiated by our Localised Static Regret analysis in the Appendix A.3. Furthermore, under locally exchangeable periods, because q_t^* is an aggregate order statistic computed over a trailing window of size w, its asymptotic variance is approximated by: $\text{Var}(q_t^*) \approx \frac{\alpha(1-\alpha)}{w \cdot [f_t(q_t^*)]^2}$ where $f_t(q_t^*)$ is the probability density. This formulation demonstrates that the variance of q_t^* shrinks on the order of $O(1/w)$, making it more stable than using s_t merely (which has an $O(1)$ variance). By considering q_t^*, the feedback mechanism can reduce the influence from the irreducible data noise (e.g., outliers), thereby promoting greater stability during the learning process. As q_t converges to q_t^*, the difference term in our bound vanishes asymptotically.

4.2.2 Long-run Coverage Analysis

Given the proposition above, the following theorem shows the worst-case upper bound of *long-run average miscoverage rate*.

Theorem 4.6. *Let $\{\eta_t\}_{t\geq 1}$ be an arbitrary positive sequence. Under Assumption 4.1, 4.2 and $\eta_t \leq \frac{2B}{1-\alpha+d(1+\epsilon)}$, the prediction interval generated by Eq. 5 with learning rate η_t satisfies:*

$$|\frac{1}{T}\sum_{t=1}^{T}(err_t - \alpha)| \leq \frac{(2B + (1 + 2d(1+\epsilon))M_T)\,\|\Delta_{1:T}\|_1}{T} + \underbrace{(cL_h + \epsilon\gamma)(2B + K_+ M_T)}_{Bias}, \quad (8)$$

where $\|\Delta_{1:T}\|_1 = |\eta_1^{-1}| + \sum_{t=2}^{T} |\eta_t^{-1} - \eta_{t-1}^{-1}|$, $M_T = \max_{1\leq r\leq T} \eta_r$, $K_+ = 1 - \alpha + d(1+\epsilon)$, $\gamma = cL_h + \frac{d}{2B}$, c and d are the steepness parameter and the range of the squashing function $h(\cdot)$.

The first term follows the analysis from Angelopoulos et al. [6] and this term vanishes asymptotically as $T \to \infty$ as long as the decay of the learning rate is not too fast (e.g., sublinear decay such as $\eta_t \propto t^{-1/2}$). The rest part reveals that the empirical miscoverage is bounded by a constant bias, which is caused by introducing additional feedback terms. A detailed proof is given in the Appendix A.4. While a non-zero asymptotic bias appears as a theoretical limitation, it is a deliberate and controllable bias-variance trade-off. Specifically, this bias can be decoupled into two distinct components governed by our dual-feedback mechanism, as detailed in the following remark.

Remark 4.7. The bias in Theorem 4.6 is a worst-case quantity derived under adversarial conditions, obtained by bounding all error terms by the domain diameter. In practice, this bound is far from tight. The bias is from two sources and governed by the hyperparameters c and ϵ: (1) $cL_h(2B + K_+M_T)$: The actual feedback introduces an empirical sum $\frac{1}{T}\sum_{t=1}^{T} |h(c \cdot e_t)|$, which is bounded via the Lipschitz continuity of $h(\cdot)$. This worst-case bound scales linearly with c, so selecting a small c directly suppresses this bias term without affecting the qualitative behaviour of the feedback mechanism, as corroborated by the sensitivity analysis in the Appendix B.2.6. (2) $\epsilon\gamma\frac{1}{T}\sum_{t=1}^{T} ||q_t - q_t^*||$ (Structural Lag Bias): This term stems from the estimated quantile. During non-stationary distribution shifts, the empirical threshold q_t^* naturally lags behind the instantaneous optimal quantile. However, under locally exchangeable regimes, q_t^* stabilises towards the true optimal quantile. In such periods, this lag penalty collapses into a statistical variance penalty bounded by $\mathcal{O}(\epsilon/\sqrt{(}w))$.

5 Experiments

5.1 Settings

Datasets. Three synthetic datasets (described in the Section 5.2.1) and five public datasets are used to evaluate methods (Google/Amazon/Microsoft stock price, Delhi temperature and Electricity, described in the Appendix B.1). Due to the limit of pages, the results and analysis for Delhi temperature and Electricity datasets are given in the Appendix B.2.1.

Underlying Time Series Forecasting Algorithms. Four algorithms are used: Prophet [39], AutoRegressive (AR) [14], Theta [8] and Transformer [41]. The chosen models span different paradigms of time series forecasting.

Online Conformal Baselines. (1) ACI [22]: updates the intervals via adjusting the quantile threshold to select from calibration set; (2) OGD family [13]: uses an iterative optimization algorithm that

updates model parameters incrementally; (3) Conformal PID Control [4]: uses control theory and is capable to model scores to adjust intervals; (4) ECI [47]: uses a smoothed quantile loss to update intervals; (5) COP [25]: uses the distribution information to optimise the update.

Common Setup. We set the target miscoverage rate to 10% and construct asymmetric prediction intervals [35]. For baselines, the parameter settings (*T_burnin*, window size, learning rate) are consistent with those described in the open-source implementations [4, 47]. The results and analysis of varying significance levels and learning rates are in Appendix B.2.2 and B.2.3.

DDCI Setup. The squashing function employed here is $tanh(\frac{1}{2}x)$ and the hyperparamter ϵ is set to 0.2. For DDCI, the estimated threshold q_t^* is from the original split conformal method which is under the risk of under-coverage in time series settings and the size of the calibration set is equal to *T_burnin* which is used as a warm-up data in baselines. For DDCI-Nex, the q_t^* follows the design of $NexCP$ [12], where the weights for conformal scores are $w_i = \frac{0.99^{\,t-i}}{\sum_{j=t-w}^{t-1} 0.99^{\,t-j}}, \quad t-w \le i \le t-1$. We also test another squashing function $softsign(\cdot)$ and the results are given in the Appendix B.2.2. The results and analysis about varying window length w and the squashing function range d and the steepness parameter c are given in the Appendix B.2.4 and B.2.5 and B.2.6.

Table 1: Results on the three synthetic datasets generated under the Prophet model with nominal level 10%. All experiments are repeated 5 times, and the table reports the mean and standard deviation. The best result (shortest width) is marked in bold, and the second-best result is marked with an underline.

Method	Synthetic Data 1			Synthetic Data 2			Synthetic Data 3		
	Coverage (%)	Avg. width	Winkler Score	Coverage (%)	Avg. width	Winkler Score	Coverage (%)	Avg. width	Winkler Score
OGD	90.0 ± 0.12	11.01 ± 0.50	12.82 ± 0.54	89.9 ± 0.17	17.28 ± 7.59	28.13 ± 18.54	85.0 ± 4.78	371.43 ± 350.57	506.07 ± 541.8
SF-OGD	90.7 ± 0.16	75.75 ± 0.11	84.93 ± 0.14	90.3 ± 0.23	76.83 ± 1.50	86.57 ± 2.44	89.4 ± 0.98	183.13 ± 138.25	213.06 ± 162.6
decay-OGD	89.7 ± 0.82	21.34 ± 2.10	23.32 ± 2.44	84.3 ± 4.90	44.05 ± 34.24	56.04 ± 47.25	59.6 ± 22.53	419.60 ± 307.13	2202.6 ± 2148.2
PID	89.9 ± 0.11	13.25 ± 1.18	15.19 ± 1.33	89.8 ± 0.08	18.04 ± 6.53	25.74 ± 11.62	89.8 ± 0.13	124.17 ± 100.70	144.22 ± 115.73
ECI	90.0 ± 0.10	13.07 ± 1.39	15.07 ± 1.53	90.0 ± 0.12	19.95 ± 7.15	30.58 ± 11.92	81.9 ± 17.84	88.85 ± 54.96	119.52 ± 61.22
COP	89.8 ± 0.02	10.07 ± 1.14	11.86 ± 1.33	90.0 ± 0.16	16.69 ± 7.17	26.53 ± 11.52	89.7 ± 0.15	91.25 ± 51.53	109.43 ± 60.39
DDCI	89.9 ± 0.02	**8.08 ± 0.82**	**9.82 ± 1.04**	90.0 ± 0.19	**13.15 ± 5.35**	**22.93 ± 9.78**	89.8 ± 0.10	**73.33 ± 41.58**	**90.00 ± 49.68**
DDCI-Nex	89.9 ± 0.03	$\underline{8.35 \pm 0.84}$	$\underline{10.07 \pm 1.00}$	89.9 ± 0.14	$\underline{15.36 \pm 6.75}$	$\underline{25.20 \pm 11.18}$	89.8 ± 0.10	$\underline{87.10 \pm 51.51}$	$\underline{104.78 \pm 59.85}$

Evaluation Metrics. *Validity*: Average coverage rate on the test set; *Efficiency*: Average width and the Winkler score [46] of prediction intervals on the test set; *Stability*: The mean absolute successive difference (MASD) of interval width. Let W_t denote the interval width at time t. MASD is defined as $\mathrm{MASD}_W = \frac{1}{n-1}\sum_{t=2}^{n} |W_t - W_{t-1}|$. A smaller MASD_W indicates less step-to-step oscillation and hence smoother interval updates.

5.2 Experimental Results

5.2.1 Synthetic Datasets

We constructed three synthetic datasets exhibiting different forms of non-stationarity: (1) **Random Walk Trend:** a smoothed stochastic trend with time-varying volatility, generated as $X_t = X_{t-1} + \sigma_t\eta_t, \eta_t \sim \mathcal{N}(0,1)$, where the volatility evolves according to $\sigma_t = \sigma_0\big(1+\kappa\,|\sin(2\pi t/P)|\big)$. The time-varying term σ_t introduces slow periodic volatility modulation, capturing smoother drift and cyclical heteroskedasticity; (2) **Exponential Trend:** a multiplicative-growth process $X_t = X_{t-1}\exp(\mu + \sigma\varepsilon_t),\ \varepsilon_t \sim \mathcal{N}(0,1)$; (3) **Changepoint:** a piecewise log-linear process with two regime shifts. For changepoints $c_1 = 600$ and $c_2 = 1200$, $\log X_t = \log X_{t-1} + \mu_{(t)} + \sigma_{(t)}\varepsilon_t, \varepsilon_t \sim \mathcal{N}(0,1)$, where $(\mu_{(t)}, \sigma_{(t)})$ vary across segments, producing moderate structural breaks.

The results are averaged over 5 runs under Prophet model and given in the Table 1. Across all synthetic settings, DDCI consistently delivers the most efficient and stable intervals while maintaining the target coverage. In Setting 1 and 2, DDCI achieves the narrowest intervals and lowest Winkler scores, with DDCI-Nex providing the second-best performance. Under the more challenging changepoint setting (Setting 3), where all methods experience increased uncertainty, DDCI still produces the most informative intervals, substantially outperforming the baselines. Notably, ECI (actual feedback only) loses the validity. Overall, the results show that DDCI is robust across different forms of non-stationarity and yields clear efficiency gains over existing methods.

5.2.2 Public Datasets

The results of baselines can also be found within Wu et al. [47] and Hu et al. [25]. Overall, all methods achieve coverage rates close to the target level of 90%.

Table 2 reports the performance across three stock datasets with a miscoverage rate of $\alpha = 10\%$. The baseline comparisons reveal distinct limitations in existing methods. While ACI satisfies the theoretical coverage requirement, it results in infinite interval widths. Similarly, traditional online methods, such as the OGD family and PID, tend to be overly conservative, yielding unnecessarily wide intervals. Although the ECI generally produces narrower intervals than these earlier approaches, its performance exhibits instability. In contrast, the proposed DDCI and DDCI-Nex consistently deliver the most competitive results across all datasets and underlying forecasting algorithms. The robustness of DDCI is further highlighted by its ability to maintain tight intervals even under complex architectures like Prophet and Transformer, drastically outperforming the best baselines.

Table 2: The experimental results on three datasets (Google, Amazon, Microsoft).

Dataset	Method	Prophet Model			AR Model			Theta Model			Transformer		
		Coverage (%)	Avg. width	Winkler Score	Coverage (%)	Avg. width	Winkler Score	Coverage (%)	Avg. width	Winkler Score	Coverage (%)	Avg. width	Winkler Score
Google	ACI	90.0	∞	∞	89.8	∞	∞	90.5	∞	∞	90.2	∞	∞
	OGD	89.7	57.60	74.74	90.7	33.76	70.96	89.9	31.49	47.88	90.1	109.27	148.76
	SF-OGD	89.6	58.92	77.52	89.9	28.31	**41.93**	90.0	34.04	51.01	90.1	88.30	114.34
	decay-OGD	89.9	77.23	92.36	90.2	46.54	58.27	90.2	55.33	67.89	89.9	120.25	144.38
	PID	89.9	56.13	71.18	89.8	64.19	83.58	89.3	75.71	97.37	89.8	77.36	106.78
	ECI	89.9	56.07	75.46	89.7	19.96	69.11	89.6	30.93	47.64	89.9	70.93	110.29
	COP	89.7	49.72	63.75	89.6	**19.87**	69.10	89.3	30.25	47.09	89.7	61.93	99.47
	DDCI	89.8	**39.20**	**52.80**	89.6	19.92	68.97	89.4	29.24	**46.38**	90.0	51.18	86.68
	DDCI-Nex	89.8	39.69	53.30	89.6	19.93	68.97	89.3	**29.15**	46.57	89.9	**51.03**	**86.51**
Amazon	ACI	90.2	∞	∞	89.8	∞	∞	89.7	∞	∞	90.1	∞	∞
	OGD	89.6	55.15	67.57	89.9	19.10	**30.08**	89.8	18.07	28.23	89.4	52.68	69.27
	SF-OGD	89.5	61.47	73.94	89.9	24.44	34.31	90.0	23.88	34.46	89.3	56.56	72.48
	decay-OGD	89.9	97.22	107.79	89.7	20.23	30.74	89.2	17.49	28.56	89.8	93.16	109.19
	PID	89.9	45.90	56.02	89.7	49.40	62.72	88.6	71.11	87.88	89.8	56.05	71.10
	ECI	90.1	47.00	59.74	89.5	17.12	34.23	89.7	17.46	27.86	89.9	49.07	68.77
	COP	89.6	39.86	49.40	89.5	17.09	34.14	89.6	17.21	**27.66**	89.5	42.73	61.32
	DDCI	89.7	**30.87**	41.06	89.3	**17.04**	34.30	89.5	**17.05**	27.90	89.7	**34.46**	**51.61**
	DDCI-Nex	89.7	30.92	**40.95**	89.3	17.06	34.29	89.6	17.06	27.87	89.7	35.16	52.25
Microsoft	ACI	90.0	∞	∞	90.7	∞	∞	89.9	∞	∞	89.9	∞	∞
	OGD	89.9	3.77	5.24	90.7	4.37	7.54	89.9	2.49	**3.78**	89.9	3.80	6.98
	SF-OGD	90.2	6.91	8.07	90.0	12.62	14.48	90.2	6.98	8.23	89.9	7.08	8.61
	decay-OGD	90.1	6.02	7.30	90.2	7.97	8.94	90.2	4.92	6.00	89.7	5.41	7.05
	PID	90.0	6.30	7.66	89.6	5.60	**6.96**	89.8	4.89	6.41	89.9	6.16	8.13
	ECI	90.4	4.85	6.14	90.2	3.76	8.50	90.2	2.48	3.94	89.9	4.35	7.45
	COP	89.7	3.78	5.16	89.7	**2.66**	7.38	89.4	**2.38**	3.81	89.6	3.77	6.89
	DDCI	89.8	**3.26**	**4.55**	89.8	2.91	7.72	89.8	2.39	3.82	89.9	**3.34**	**6.43**
	DDCI-Nex	89.8	3.40	4.71	89.8	2.90	7.67	89.8	2.39	3.84	89.8	3.47	6.51

Fig. 2 presents a visualisation of comparison result between ECI (single feedback only) and DDCI (ours) on the Google stock dataset using the Prophet model. Generally, ECI tends to produce wider ranges, which often appears overly conservative, with noticeable fluctuations in the interval boundaries. By contrast, DDCI yields narrower and more stable intervals, reducing redundancy while maintaining strong alignment with the true series.

In conclusion, the proposed methods generally achieve tighter prediction intervals while maintaining validity. However, their improvements are less pronounced under the AR model. These observations can be attributed to the behaviour of the conformal estimated threshold q_t^*. When the conformal prediction yields relatively conservative intervals, characterised by achieving the target coverage with a large average width, the estimated threshold tends to overestimate local uncertainty.

5.2.3 Stability Evaluation and Ablation Study

We show *stability evaluation* on all five public datasets under Prophet in Tab. 3. We compare PID and ECI with three configurations of DDCI. DDCI with $c = 0$ disables both additional feedback terms because $h(0) = 0$, and is therefore equivalent to OGD under the same proportional learning rate.

Table 3: Stability evaluation (MASD, ↓) and ablation study on the five public datasets under Prophet.

Method	GOOGL	AMZN	MSFT	Delhi	Electricity
PID	8.55	6.94	1.02	1.05	0.0645
ECI	9.66	7.78	0.81	0.43	0.0628
DDCI ($c = 0$; no additional feedback)	<u>8.32</u>	<u>6.64</u>	<u>0.63</u>	0.32	0.0622
DDCI ($\epsilon = 0$; actual feedback only)	9.93	8.05	0.68	<u>0.31</u>	<u>0.0617</u>
DDCI ($\epsilon = 0.2$; dual feedback)	**6.89**	**5.35**	**0.54**	**0.28**	**0.0603**

DDCI with $\epsilon = 0$ retains only the actual feedback $\mathcal{A}(e_t)$, whereas the full DDCI combines the actual and estimated feedbacks.

The full DDCI achieves the lowest MASD on all five datasets. Relative to the actual-feedback-only configuration, adding the estimated feedback reduces MASD by 2.3%–33.5%. The actual feedback alone is not uniformly smoother than the $c = 0$ configuration, whereas combining the actual and estimated feedbacks produces the lowest successive interval variation in every setting. Across these ablations, adding the estimated feedback changes coverage by at most 0.1 percentage points and mean width by at most 5.5%. Together with the coverage and efficiency results in Tables 2 and 4, this shows that the stability improvement is not obtained through a material loss of validity or wider intervals. All results under all models are provided in Tab. 6 (Appendix B.2.7), where DDCI achieves the lowest MASD in 19 of the 20 dataset–forecaster settings.

5.2.4 Sensitivity Analysis

We examine the role of the parameter ϵ in the update rule Eq. 5, taking Google stock data as an illustrative case. Fig. 3 shows that increasing the parameter ϵ consistently enlarges the average interval size across all four models, confirming its role in controlling the balance between efficiency and smoothness. Noticeably, the influence of ϵ becomes remarkable when it is greater than 0.5. The observed robustness to small ϵ aligns perfectly with our theoretical regret bound. Consequently, for small values, the theoretical cost on the regret bound is minimal due to this quadratic scaling. This provides a strong justification for our parameter choice: keeping ϵ small allows the model to benefit from the stability mechanism while incurring a negligible penalty on the dynamic regret.

DDCI-Nex yields less sensitivity to ϵ by employing NexCP for threshold estimation. This shows the influence of estimated threshold q_t^* on the updating process. This highlights that a more reliable estimated threshold is crucial for enabling the dual-feedback mechanism to function effectively. The apparent sensitivity shown in Fig. 3 is not primarily driven by ϵ itself. The different trends across panels arise because each base model (AR, Prophet, Theta, Transformer) employs a distinct learning

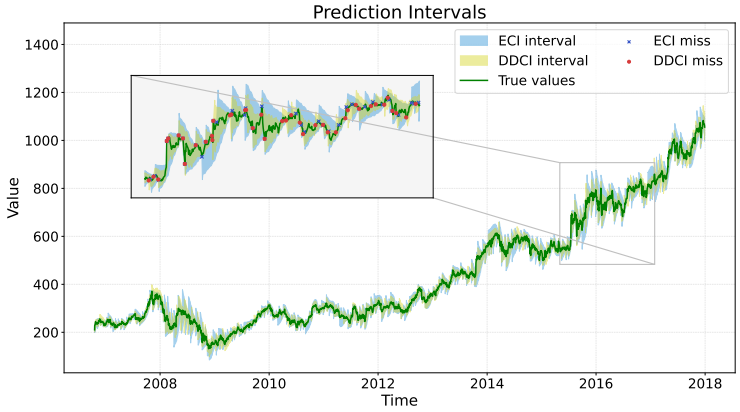


Figure 2: Visualisation of comparison results between ECI (single feedback only) and DDCI (proposed, $\epsilon = 0.2$) on Google stock dataset under Prophet model. The blue crosses and red points represent the miscovered points of each method.

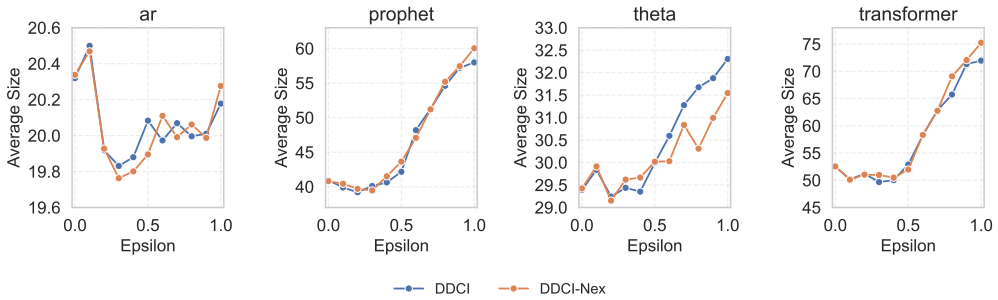


Figure 3: The influence of different smoothing parameter ϵ on the size of prediction interval.

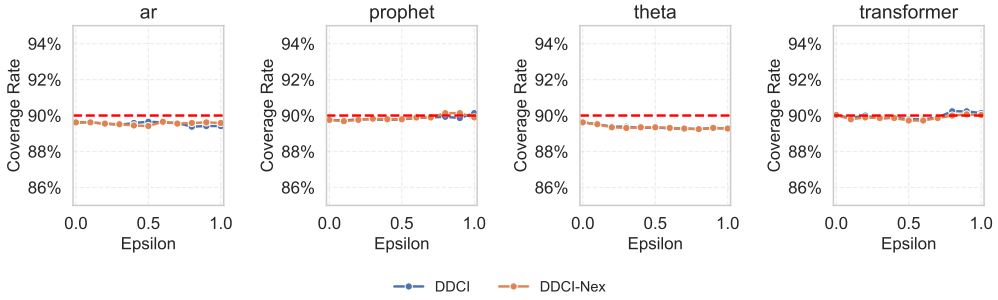


Figure 4: The influence of different smoothing parameter ϵ on the coverage rate.

rate (0.05, 0.5, 0.1, 0.5, respectively). A visualisation of how ϵ can influence the predictive efficiency are displayed in the Appendix B.2.7. In this example, we set ϵ to 0 and 0.8 respectively, which can be regarded as a visualisation for ablation study. Fig. 4 displays that the effect of the parameter ϵ is relatively minor compared to its influence on interval size. This indicates that varying ϵ mainly adjusts efficiency (interval width) without substantially undermining validity.

6 Conclusion

This work introduces a novel framework that incorporates a dual-feedback mechanism. The mechanism combines *actual feedback* and *estimated feedback*, whose signs are intentionally set in opposition to counterbalance increases and decreases when updating the intervals. While the use of *estimated feedback* may slightly widen the intervals when the true label lies inside them, it effectively prevents unnecessary expansions, leading to more stable interval adjustments. Further work could improve the DDCI method by adaptively selecting the parameter ϵ in response to distributional changes, thereby enhancing both stability and efficiency.

References

[1] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

[2] N. Akhavan Attar, H. Asadollahzadeh, L. Luo, and U. Aickelin. Softmax is not enough (for adaptive conformal classification). In *International Conference on Learning Representations*, volume 2026, pages 28009–28058, 2026.

[3] A. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

[4] A. Angelopoulos, E. Candes, and R. J. Tibshirani. Conformal pid control for time series prediction. *Advances in neural information processing systems*, 36:23047–23074, 2023.

[5] A. N. Angelopoulos, S. Bates, et al. Conformal prediction: A gentle introduction. *Foundations and trends® in machine learning*, 16(4):494–591, 2023.

[6] A. N. Angelopoulos, R. F. Barber, and S. Bates. Online conformal prediction with decaying step sizes. *arXiv preprint arXiv:2402.01139*, 2024.

[7] A. N. Angelopoulos, R. F. Barber, and S. Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

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

[9] A. Auer, M. Gauch, D. Klotz, and S. Hochreiter. Conformal prediction for time series with modern hopfield networks. *Advances in neural information processing systems*, 36:56027–56074, 2023.

[10] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixao, F. Mutz, et al. Self-driving cars: A survey. *Expert systems with applications*, 165:113816, 2021.

[11] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

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

[13] A. Bhatnagar, H. Wang, C. Xiong, and Y. Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pages 2337–2363. PMLR, 2023.

[14] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[15] J. C. Cresswell, Y. Sui, B. Kumar, and N. Vouitsis. Conformal prediction sets improve human decision making. *arXiv preprint arXiv:2401.13744*, 2024.

[16] A. Daniely, A. Gonen, and S. Shalev-Shwartz. Strongly adaptive online learning. In *International Conference on Machine Learning*, pages 1405–1411. PMLR, 2015.

[17] F. Díaz-González, A. Sumper, O. Gomis-Bellmunt, and R. Villafáfila-Robles. A review of energy storage technologies for wind power applications. *Renewable and sustainable energy reviews*, 16(4):2154–2171, 2012.

[18] S. Du, L. Luo, I. Nouretdinov, and U. Aickelin. Duen: An ensemble framework enhanced by distribution-free uncertainty for regression. *Proceedings of Machine Learning Research*, 266: 1–23, 2025.

[19] J. Durbin and S. J. Koopman. *Time series analysis by state space methods*. Oxford university press, 2012.

[20] M. Eklund, U. Norinder, S. Boyer, and L. Carlsson. The application of conformal prediction to the drug discovery process. *Annals of Mathematics and Artificial Intelligence*, 74(1):117–132, 2015.

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

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

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

[24] M. Harries, N. S. Wales, et al. Splice-2 comparative evaluation: Electricity pricing. 1999.

[25] D. Hu, J. Wu, S.-T. Xia, and C. Zou. Distribution-informed online conformal prediction. *arXiv preprint arXiv:2512.07770*, 2025.

[26] J. Lee, C. Xu, and Y. Xie. Flow-based conformal prediction for multi-dimensional time series. *arXiv preprint arXiv:2502.05709*, 2025.

[27] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[28] J. Lekeufack, A. N. Angelopoulos, A. Bajcsy, M. I. Jordan, and J. Malik. Conformal decision theory: Safe autonomous decisions from imperfect predictions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11668–11675. IEEE, 2024.

[29] Y. Liu, S. Wijewickrema, D. Hu, C. Bester, S. O'Leary, and J. Bailey. Stochastic diffusion: a diffusion based model for stochastic time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 1939–1950, 2025.

[30] R. Luo, S. Zhao, J. Kuck, B. Ivanovic, S. Savarese, E. Schmerling, and M. Pavone. Sample-efficient safety assurances using conformal prediction. *The International Journal of Robotics Research*, 43(9):1409–1424, 2024.

[31] R. Neglia, A. Cini, M. M. Bronstein, and F. M. Bianchi. Rescp: Reservoir conformal prediction for time series forecasting. *arXiv preprint arXiv:2510.05060*, 2025.

[32] F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

[33] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, pages 345–356. Springer, 2002.

[34] A. Podkopaev, D. Xu, and K.-c. Lee. Adaptive conformal inference by betting. *arXiv preprint arXiv:2412.19318*, 2024.

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

[36] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191, 2020.

[37] R. C. Smith. *Uncertainty quantification: theory, implementation, and applications*. SIAM, 2024.

[38] S. Stanton, W. Maddox, and A. G. Wilson. Bayesian optimization with conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*, pages 959–986. PMLR, 2023.

[39] S. J. Taylor and B. Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

[40] R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[42] V. Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1): 9–28, 2015.

[43] V. Vovk, A. Gammerman, and C. Saunders. Machine-learning applications of algorithmic randomness. 1999.

[44] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[45] V. Vovk, I. Petej, I. Nouretdinov, E. Ahlberg, L. Carlsson, and A. Gammerman. Retrain or not retrain: Conformal test martingales for change-point detection. In *Conformal and Probabilistic Prediction and Applications*, pages 191–210. PMLR, 2021.

[46] R. L. Winkler. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191, 1972.

[47] J. Wu, D. Hu, Y. Bao, S.-T. Xia, and C. Zou. Error-quantified conformal inference for time series. *arXiv preprint arXiv:2502.00818*, 2025.

[48] C. Xu and Y. Xie. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pages 11559–11569. PMLR, 2021.

[49] C. Xu and Y. Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pages 38707–38727. PMLR, 2023.

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

[51] Z. Zhang, Z. Lu, and H. Yang. The benefit of being bayesian in online conformal prediction. *arXiv preprint arXiv:2410.02561*, 2024.

[52] P. Zhao, Y.-J. Zhang, L. Zhang, and Z.-H. Zhou. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *Journal of Machine Learning Research*, 25(98):1–52, 2024.

[53] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

A Theoretical Analysis

Here, we provide the proofs for theoretical analysis in Section 4.2. The proofs for Proposition are given in the A.1. The proofs for regret bound analysis are in the A.2 and A.3. The proof for long-run coverage analysis is provided in the A.4. The analysis for the additional parameter ϵ is in the A.5.

Recall the update rule:

$$q_{t+1} := q_t + \eta_t(\text{err}_t - \alpha) + \eta_t(\underbrace{\rho_t h(c \cdot e_t)}_{\text{Actual feedback}} - \underbrace{\epsilon\,(1-\rho_t^*)\,h(c \cdot e_t^*)}_{\text{Estimated feedback}}), \quad \rho_t, \rho_t^* \in [0,1]$$

where $\epsilon > 0$ is a parameter to regulate the influence of the estimated feedback term. The actual and estimated feedbacks operate in opposing directions to establish a push-pull mechanism, ensuring dynamic equilibrium. In other words, the actual feedback acts as a driving force to correct immediately, while the estimated feedback serves as a stabilising force to prevent over-adjustment. For the theoretical analysis, we replace $\max_{t-w\le i,j\le t} |s_i - s_j|$ in ρ_t and ρ_t^* with its upper bound 2B according to the Assumption 4.1. This yields a conservative bound; in practice $\max_{t-w\le i,j\le t} |s_i - s_j|$ may be substantially smaller, which explains the empirical efficiency of DDCI.

A.1 Propositions 4.3

Proof of Proposition 4.3. We prove this by induction, following the structure of similar arguments in prior work [6, 47]. First, $q_1 \in [-B, B]$ by assumption. Assume that $\eta_t \le \frac{2B}{1-\alpha+d(1+\epsilon)}$ and $\epsilon \ge 0$. Next fix any $t \ge 1$. Meanwhile, the empirical quantile satisfies $|q_t^*| \le B$. Then we analyse the update in four cases based on the relative positions of s_t, q_t, and q_t^*.

(1): $s_t \ge q_t$ (Interval expands, error = $1-\alpha$)

- $q_t^* \ge s_t \ge q_t$:

$$\begin{aligned} q_{t+1} &= q_t + \eta_t(1-\alpha) + \eta_t\left(\rho_t h(c(s_t - q_t)) - \epsilon\,(1-\rho_t^*)\,h(c(s_t - q_t^*))\right) \\ &\le q_t + \eta_t(1-\alpha) + d\eta_t(1+\epsilon) \\ &\le B + (1-\alpha+d(1+\epsilon))M_t \end{aligned}$$

- $s_t \ge q_t \ge q_t^*$ or $s_t \ge q_t^* \ge q_t$:

$$\begin{aligned} q_{t+1} &= q_t + \eta_t(1-\alpha) + \eta_t\left(\rho_t h(c(s_t - q_t)) - \epsilon\,(1-\rho_t^*)\,h(c(s_t - q_t^*))\right) \\ &\le q_t + \eta_t(1-\alpha) + d(1+\epsilon)\eta_t \\ &\le B + (1-\alpha+d(1+\epsilon))M_t \end{aligned}$$

(2): $s_t \le q_t$ (Interval shrinks, error = $-\alpha$)

- $q_t^* \le s_t \le q_t$:

$$\begin{aligned} q_{t+1} &= q_t + \eta_t(-\alpha) + \eta_t\left(\rho_t h(c(s_t - q_t)) - \epsilon\,(1-\rho_t^*)\,h(c(s_t - q_t^*))\right) \\ &\ge q_t + \eta_t(-\alpha) - d\eta_t(1+\epsilon) \\ &\ge -B - (\alpha+d(1+\epsilon))M_t \end{aligned}$$

- $s_t \le q_t \le q_t^*$ or $s_t \le q_t^* \le q_t$:

$$\begin{aligned} q_{t+1} &= q_t + \eta_t(-\alpha) + \eta_t\left(\rho_t h(c(s_t - q_t)) - \epsilon\,(1-\rho_t^*)\,h(c(s_t - q_t^*))\right) \\ &\ge q_t + \eta_t(-\alpha) - d\eta_t(1+\epsilon) \\ &\ge -B - (\alpha+d(1+\epsilon))M_t \end{aligned}$$

Then combining all cases gives:

$$-B - K_- M_{t-1} \le q_t \le B + K_+ M_{t-1} \quad \forall t \ge 1,$$

where $M_0 = 0$, $M_t = \max_{1\le r\le t} \eta_r$ for $t \ge 1$, $K_- = \alpha + d(1+\epsilon)$ and $K_+ = 1-\alpha+d(1+\epsilon)$.

Remark on the inductive bound. The bound $|q_t| \leq B + K_+ M_{t-1}$ (due to α is usually small, we use K_+ as the upper bound of the absolute value) leverages the structural property that the maximum learning rate M_{t-1} acts as a buffer: once q_t approaches $\pm B$, the self-correcting nature of the dual-feedback signals (when q_t exceeds B, the actual feedback $\mathcal{A}(\cdot)$ becomes negative; symmetric for $q_t < -B$) prevents unbounded drift. The constraint $\eta_t \leq \frac{2B}{K_+}$ ensures that no single update can move q_t across the entire range $2B$.

A.2 Dynamic Regret for Theorem 4.4

To rigorously analyse the dynamic regret of DDCI, we cast the algorithm within the framework of Online Mirror Descent (OMD) [32]. Let $Br_\psi(\cdot,\cdot)$ be the Bregman divergence with respect to $\psi(x) : x \in \mathcal{R}$ and assume ψ to be closed and λ-strongly convex. Specifically, we define: $g_t = g_t^{OGD} + g_t^{AF} - \epsilon g_t^{EF}$. Here $g_t^{OGD} \in \nabla \ell_t(q_t)$ is a subgradient of the pinball loss. The algorithm thus performs inexact subgradient descent: instead of using g_t^{OGD} directly, it uses the perturbed direction $g_t = g_t^{OGD} + g_t^{AF} - \epsilon g_t^{EF}$, where $g_t^{AF} - \epsilon g_t^{EF}$ is the bias. We use Euclidean distance $Br_\psi(q, q_t) = \frac{1}{2}\|q - q_t\|^2$, which is the Bregman divergence induced by $\psi(q) = \frac{1}{2}||q||^2$.

Proof of Theorem 4.4. From the one-step inequality for OMD with competitor u_t and fixed learning rate η, we have

$$\begin{aligned}\eta(\ell_t(q_t) - \ell_t(u_t)) &\leq \eta g_t^{OGD}(q_t - u_t) = \eta(g_t - (g_t^{AF} - \epsilon g_t^{EF}))(q_t - u_t) \\ &= \eta g_t(q_t - u_t) - \eta(g_t^{AF} - \epsilon g_t^{EF})(q_t - u_t)\end{aligned}$$

Applying the standard OMD analysis to the first term, we bound it using the Bregman divergence Br_ψ:

$$\eta g_t(q_t - u_t) \leq Br_\psi(u_t, q_t) - Br_\psi(u_t, q_{t+1}) + \frac{\eta^2}{2}||g_t||^2$$

For the second term, by Cauchy-Schwarz and the bounded domain Assumption 4.1 ($||q_t - u_t|| \leq 2B + K_+ M_{t-1}$), we have:

$$-\eta(g_t^{AF} - \epsilon g_t^{EF})(q_t - u_t) \leq \eta||g_t^{AF} - \epsilon g_t^{EF}|| \cdot ||q_t - u_t|| \leq \eta(2B + K_+ M_{t-1})||g_t^{AF} - \epsilon g_t^{EF}||$$

Combining these, we obtain:

$$\begin{aligned}\eta(\ell_t(q_t) - \ell_t(u_t)) \leq & Br_\psi(u_t, q_t) - Br_\psi(u_{t+1}, q_{t+1}) + Br_\psi(u_{t+1}, q_{t+1}) - Br_\psi(u_t, q_{t+1}) \\ &+ \frac{\eta^2}{2}||g_t||^2 + \eta(2B + \eta K_+)||g_t^{AF} - \epsilon g_t^{EF}||\end{aligned}$$

Dividing by η and averaging over time T, we have:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(\ell_t(q_t) - \ell_t(u_t)) &\leq \frac{2B^2}{\eta T} + \frac{1}{\eta T}\sum_{t=1}^{T}\langle \nabla\psi(q_{t+1}) - \nabla\psi(q_1), u_t - u_{t+1}\rangle \\ &\quad + \frac{1}{T}\sum_{t=1}^{T}\frac{\eta}{2}||g_t||^2 + \frac{2B + \eta K_+}{T}\sum_{t=1}^{T}||g_t^{AF} - \epsilon g_t^{EF}|| \\ &\leq \frac{2B^2}{\eta T} + \frac{2B + \eta K_+}{\eta T}\sum_{t=1}^{T}||u_{t+1} - u_t|| \\ &\quad + \frac{1}{T}\sum_{t=1}^{T}\frac{\eta}{2}||g_t||^2 + \frac{2B + \eta K_+}{T}\sum_{t=1}^{T}||g_t^{AF} - \epsilon g_t^{EF}||\end{aligned}$$

Now we analyse the term g_t:

(1) **The Gradient Descent (g_t^{OGD}):** The standard descent direction derived from the pinball loss, representing the fundamental learning signal: $||g_t^{OGD}|| = ||\nabla \ell_t(q_t)|| \leq 1$;

(2) **Stabilised Control ($g_t^{AF} - \epsilon g_t^{EF}$):** A correction term governed by the dual feedback mechanism. It dynamically "re-calibrates" the update magnitude by referencing the estimated threshold, ensuring stability in stationary regimes and responsiveness during drifts.

$$\begin{aligned} g_t^{AF} - \epsilon g_t^{EF} &= \frac{|e_t|}{2B} h(c \cdot e_t) - \epsilon(1 - \frac{|e_t^*|}{2B}) h(c \cdot e_t^*) \\ &= \left[(1+\epsilon)\frac{|e_t|}{2B} - \epsilon\right] h(c \cdot e_t) - \epsilon \left(\frac{|e_t| - |e_t^*|}{2B}\right) h(c \cdot e_t) + \epsilon \left(1 - \frac{|e_t^*|}{2B}\right) \left[h(c \cdot e_t) - h(c \cdot e_t^*)\right] \end{aligned}$$

The identity decomposition above serves not only to facilitate mathematical bounding but also to elucidate how the dual-feedback mechanism achieves an adaptive toggle between smoothness and responsiveness at a fundamental level.

Specifically, the first term in the decomposition acts as the primary control term for the system update. A closer inspection of its coefficient, $\left[(1+\epsilon)\frac{|e_t|}{2B} - \epsilon\right]$, reveals that $\epsilon \leq 1$ and it essentially functions as an adaptive gating switch:

- Locally exchangeable Regimes: When the instantaneous prediction error $|e_t|$ is small (i.e., $|e_t| < \frac{\epsilon}{1+\epsilon} 2B$), this coefficient becomes negative. This implies that the overall update direction actively opposes the native actual feedback $h(c \cdot e_t)$, allowing the estimated feedback (anchored by q_t^*) to dominate. This mechanism forcefully suppresses overreactions to single-step data noise, providing a dynamic dampening effect that ensures interval stability;
- Drift Regimes: When encountering a severe distribution shift that significantly inflates the instantaneous error $|e_t|$, the coefficient turns positive, and the dampening effect $\epsilon(1 - \frac{|e_t|}{2B})$ decays rapidly. Consequently, the actual feedback regains dominance, driving the threshold q_t to rapidly expand or contract to capture the new underlying data distribution.

Meanwhile, the remaining two terms in the decomposition perfectly isolate the structural bias introduced by the empirical quantile. By leveraging the triangle inequality and the Lipschitz continuity of $h(\cdot)$, these terms can be cleanly bounded proportionally to $||q_t - q_t^*||$. This naturally lays the groundwork for the γ-dependent worst-case bound derived below.

Then combining boundedness and Lipschitz continuity, we have the worst-case bound $||g_t^{AF} - \epsilon g_t^{EF}|| \leq d + \epsilon\gamma||q_t - q_t^*||$, where $\gamma = cL_h + \frac{d}{2B}$. The constant d here comes from $\left[(1+\epsilon)\frac{|e_t|}{2B} - \epsilon\right] h(c \cdot e_t)$ when e_t always reaches the maximum in the worst-case. However, this term is linearly correlated to the steepness parameter c (Lipschitz continuity), which can be a small value without influence the efficiency based on the sensitivity analysis on c (Appendix B.2.6).

Recalling the inequality $||g_t||^2 = ||g_t^{OGD} + (g_t^{AF} - \epsilon g_t^{EF})||^2 \leq 2||g_t^{OGD}||^2 + 2||g_t^{AF} - \epsilon g_t^{EF}||^2$, we substitute $||g_t^{AF} - \epsilon g_t^{EF}||$ and $||g_t||^2$ back into the regret bound derived in the previous step. Then we obtain the final dynamic regret:

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}(\ell_t(q_t) - \ell_t(u_t)) \leq & \frac{2B^2}{\eta T} + \frac{2B + \eta K_+}{\eta T}\sum_{t=1}^{T}||u_{t+1} - u_t|| + 2\epsilon^2\gamma^2\frac{\eta}{T}\sum_{t=1}^{T}||q_t - q_t^*||^2 \\ & + \frac{(2B + \eta K_+)\epsilon\gamma}{T}\sum_{t=1}^{T}||q_t - q_t^*|| + (2B + \eta K_+)d + \eta(1 + 2d^2). \end{aligned}$$

The term $\sum_{t=1}^{T}\|u_{t+1} - u_t\|$ represents the path length of the optimal comparator sequence. This term captures the non-stationarity of the environment.

A.3 Localised Static Regret for Remark 4.5

Then, we start showing the Locally Static Regret (LSReg) analysis of DDCI. During the time interval τ where the score distribution remains invariant (i.e., locally exchangeable), the q_t^* from conformal

prediction reduces to a fixed optimal solution u^*. This supports the stabilising role of q_t^* as an empirical anchor.

Corollary A.1. *Let τ be a time interval of length k and assume the score distribution remains invariant within this interval. For a fixed optimal decision u^* within τ, based on the inexact gradient analysis in Theorem 4.4, DDCI achieves the following local regret bound:*

$$\frac{1}{k}\sum_{t=\tau}^{\tau+k-1}(\ell_t(q_t)-\ell_t(u^*)) \leq \frac{(2B+\eta K_+)^2}{2\eta k} + \frac{\eta}{k}\sum_{t=\tau}^{\tau+k-1}\left(2\epsilon^2\gamma^2||q_t-q_t^*||^2\right) + (2B+\eta K_+)d + \frac{(2B+\eta K_+)\epsilon\gamma}{k}\sum_{t=\tau}^{\tau+k-1}||q_t-q_t^*|| + \eta(1+2d^2).$$

Proof of Corollary A.1. For a fixed u^* within the time interval τ, we follow the inexact gradient decomposition $g_t = g_t^{OGD} + g_t^{AF} - \epsilon g_t^{EF}$ as established in the proof of Theorem 4.4.

Starting from the one-step inequality and summing over $t = \tau$ to $\tau + k - 1$, we have:

$$\frac{1}{k}\sum_{t=\tau}^{\tau+k-1}(\ell_t(q_t)-\ell_t(u^*)) \leq \frac{1}{\eta k}\sum_{t=\tau}^{\tau+k-1}\frac{||q_t-u^*||^2-||q_{t+1}-u^*||^2}{2} + \frac{1}{k}\sum_{t=\tau}^{\tau+k-1}\frac{\eta}{2}||g_t||^2 + \frac{1}{k}\sum_{t=\tau}^{\tau+k-1}(2B+\eta K_+)||g_t^{AF}-\epsilon g_t^{EF}||$$

Because the competitor u^* is fixed, the first term forms a perfect telescoping sum. Using the bounded domain assumption $||q_\tau - u^*|| \leq 2B + K_+ M_{t-1}$, we bound the telescoping term as:

$$\frac{||q_\tau-u^*||^2-||q_{\tau+k}-u^*||^2}{2\eta k} \leq \frac{||q_\tau-u^*||^2}{2\eta k} \leq \frac{(2B+\eta K_+)^2}{2\eta k}$$

Notice that the dynamic path length term $\sum ||u_{t+1} - u_t||$ vanishes completely since $u_{t+1} = u_t = u^*$.

Finally, substituting the bound $||g_t^{AF} - \epsilon g_t^{EF}|| \leq d + \epsilon\gamma||q_t - q_t^*||$ into the inequality, we can get the desired local regret bound:

$$\frac{1}{k}\sum_{t=\tau}^{\tau+k-1}(\ell_t(q_t)-\ell_t(u^*)) \leq \frac{(2B+\eta K_+)^2}{2\eta k} + \frac{\eta}{k}\sum_{t=\tau}^{\tau+k-1}\left(2\epsilon^2\gamma^2||q_t-q_t^*||^2\right) + \eta(1+2d^2) + \frac{(2B+\eta K_+)\epsilon\gamma}{k}\sum_{t=\tau}^{\tau+k-1}||q_t-q_t^*|| + (2B+\eta K_+)d.$$

A.4 Coverage Analysis for Theorem 4.6

Proof of Theorem 4.6. Define $\Delta_r := \eta_r^{-1} - \eta_{r-1}^{-1}$ *for* $r \geq 1$*, with the convention* $\eta_0^{-1} := 0$*. Then we can get the long-run coverage bound:*

$$\begin{aligned}
\left|\frac{1}{T}\sum_{t=1}^{T}(\text{err}_t - \alpha)\right| &= \left|\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{r=1}^{t}\Delta_r\right)\cdot\eta_t(\text{err}_t - \alpha)\right| \\
&= \left|\frac{1}{T}\sum_{r=1}^{T}\Delta_r\left(\sum_{t=r}^{T}\eta_t(\text{err}_t - \alpha)\right)\right| \\
&= \left|\frac{1}{T}\sum_{r=1}^{T}\Delta_r\left(\sum_{t=r}^{T}[q_{t+1} - q_t - \eta_t\rho_t h(c(s_t - q_t)) + \eta_t\epsilon(1-\rho_t^*)h(c(s_t - q_t^*)))]\right)\right| \\
&\leq \left|\frac{1}{T}\sum_{r=1}^{T}\Delta_r(q_{T+1} - q_r)\right| + \left|\frac{1}{T}\sum_{r=1}^{T}\Delta_r\sum_{t=r}^{T}\eta_t(\rho_t h(c(s_t - q_t)) - \epsilon(1-\rho_t^*)h(c(s_t - q_t^*)))\right| \\
&\leq \left|\frac{1}{T}\sum_{r=1}^{T}\Delta_r(q_{T+1} - q_r)\right| + \left|\frac{1}{T}\sum_{t=1}^{T}(\rho_t h(c(s_t - q_t)) - \epsilon(1-\rho_t^*)h(c(s_t - q_t^*))\right| \\
&\leq \frac{(2B + (1 + 2d(1+\epsilon))M_T)\,\|\Delta_{1:T}\|_1}{T} + \frac{1}{T}\sum_{t=1}^{T}|h(c\cdot e_t)| + \epsilon\gamma\frac{1}{T}\sum_{t=1}^{T}||q_t - q_t^*|| \\
&\leq \frac{(2B + (1 + 2d(1+\epsilon))M_T)\,\|\Delta_{1:T}\|_1}{T} + (cL_h + \epsilon\gamma)(2B + K_+M_T),
\end{aligned}$$

where $\|\Delta_{1:T}\|_1 = |\eta_1^{-1}| + \sum_{t=2}^{T}|\eta_t^{-1} - \eta_{t-1}^{-1}|$, $K_+ = 1 - \alpha + d(1+\epsilon)$, $\gamma = cL_h + \frac{d}{2B}$ and c is a steepness parameter inside the squashing function, $M_T = \max_{1\leq r\leq T}\eta_r$.

For two bias terms $|h(c\cdot e_t)|$ and $||q_t - q_t^*||$, $2B + K_+M_T$ is the worst-case bound according to the Proposition 4.3.

A.5 The Heuristic Analysis of ϵ

To ensure the update direction is consistent with the primary learning signal, we analyse the condition where the magnitude of the actual feedback exceeds that of the estimated feedback. Incorporating the weighting coefficients from Eq. 5, the condition is:

$$\frac{|e_t|}{2B}|h(ce_t)| > \epsilon\left(1 - \frac{|e_t^*|}{2B}\right)|h(ce_t^*)|.$$

Leveraging the relationship $s_t - q_t = (s_t - q_t^*) - q_t + q_t^*$, this simplifies to a dynamic bound on ϵ:

$$\epsilon < \frac{|e_t^* + q_t^* - q_t|}{2B - |e_t^*|}\frac{|h(c(e_t^* + q_t^* - q_t))|}{|h(ce_t^*)|}.$$

To make this bound tractable, we consider the algorithm in the stationary regime where the parameter q_t fluctuates within a small convergence radius ξ around the estimated threshold, i.e., $|q_t - q_t^*| \leq \xi$. We analyse the worst-case scenario where the parameter deviation weakens the learning signal.

Using the first-order Taylor expansion for the squashing function $h(c\cdot x)$ around e_t^*, we approximate the numerator term:

$$h(c(e_t^* + q_t^* - q_t)) \approx h(ce_t^*) - c\cdot h'(ce_t^*)\cdot\xi.$$

The condition can be simplified to:

$$\epsilon < \frac{|e_t^* - \xi|}{2B - |e_t^*|}\left(\frac{h(ce_t^*) - c\xi h'(ce_t^*)}{h(ce_t^*)}\right) = (1 - \underbrace{\frac{c\xi h'(ce_t^*)}{h(ce_t^*)}}_{\text{Safety Margin}})\frac{|e_t^* - \xi|}{2B - |e_t^*|}$$

This derived bound reveals two critical constraints on ϵ. First, the term involving ξ acts as a safety margin. This margin accounts for inherent parameter oscillation near equilibrium. Second, the bound dynamically tightens as the prediction error $|e_t^*|$ diminishes. Specifically, the scaling factor

$\frac{|e_t^*-\xi|}{2B-|e_t^*|}$ decreases as the error magnitude diminishes. Consequently, the feasible range for ϵ shrinks significantly in regimes when e_t^* is small. To guarantee stability across all scenarios, it is theoretically preferable to select a small ϵ. This justifies our empirical choice (e.g., $\epsilon = 0.2$) to ensure the actual feedback remains dominant, even in the presence of convergence noise.

B Additional Experimental results

B.1 Description of Datasets

- **Amazon and Google**: contain thirty blue-chip stock prices, with daily opening prices forecasted on a log scale, spanning the period from January 1, 2006 to December 31, 2014. `https://www.kaggle.com/datasets/camnugent/sandp500`.
- **Microsoft**: includes daily stock opening prices from April 1, 2015 to May 31, 2021. `https://www.kaggle.com/datasets/camnugent/sandp500`.
- **Delhi temperature**: provides daily measurements of temperature (averaged over eight readings taken at three-hour intervals), humidity, wind speed, and atmospheric pressure in Delhi, from January 1, 2003 to April 24, 2017, collected via the Weather Underground API. `https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data`.
- **Electricity demand** [24]: records electricity demand in New South Wales at half-hour intervals from May 7, 1996 to December 5, 1998.

B.2 Additional Experimental Results

B.2.1 Delhi temperature and Electricity Demand

On the Delhi temperature dataset, DDCI variants consistently achieve the narrowest prediction intervals (except AR model) while preserving nominal coverage. The improvements over the strongest baselines are smaller in magnitude compared with stock datasets but remain consistent, demonstrating the robustness and adaptability of DDCI in settings with relatively stable time series such as temperature.

For the electricity demand dataset, the performance of the DDCI variants is broadly comparable to that of the ECI baselines rather than superior. Across all forecasting models, DDCI and DDCI-Nex achieve interval widths that are similar to the best ECI methods while maintaining the target coverage level. Although the improvements observed in the stock datasets are not replicated here, the results indicate that DDCI remains competitive without loss of validity. This suggests that the performance of DDCI in the electricity demand setting remains stable and consistent with strong baseline approaches.

Table 4: The experimental results on Delhi Temperature and Electricity Demand datasets.

Dataset	Method	Prophet Model			AR Model			Theta Model			Transformer		
		Coverage (%)	Avg. width	Winkler Score	Coverage (%)	Avg. width	Winkler Score	Coverage (%)	Avg. width	Winkler Score	Coverage (%)	Avg. width	Winkler Score
Delhi Temp.	ACI	91.0	∞	∞	90.0	∞	∞	90.2	∞	∞	90.3	∞	∞
	OGD	90.4	7.54	9.78	90.1	6.82	**9.11**	90.0	6.36	**8.48**	89.9	11.08	13.15
	SF-OGD	90.0	10.63	13.24	90.1	10.92	13.72	90.1	11.33	13.99	90.0	10.93	13.64
	decay-OGD	90.1	11.24	13.03	90.2	10.52	12.25	90.2	10.95	12.85	89.7	14.43	16.18
	PID	89.9	10.87	13.36	89.7	12.28	15.11	89.7	12.49	15.13	89.9	11.57	14.15
	ECI	90.0	7.20	9.34	90.1	6.39	9.99	90.0	6.41	8.76	89.9	9.15	12.33
	COP	90.1	7.05	9.35	90.1	**6.35**	9.99	90.0	**6.27**	8.69	89.9	9.12	12.36
	DDCI	90.0	**6.95**	**9.10**	89.9	6.44	10.04	90.0	6.36	8.69	89.9	8.57	**11.80**
	DDCI-Nex	90.0	7.00	9.16	89.9	6.39	10.10	90.0	6.31	8.67	89.9	**8.56**	11.87
Electricity	ACI	90.1	∞	∞	90.1	∞	∞	90.2	∞	∞	90.2	∞	∞
	OGD	89.8	0.433	0.504	90.0	0.133	0.188	90.1	0.081	0.117	90.1	0.139	0.181
	SF-OGD	89.9	0.419	0.479	90.0	0.129	0.180	90.3	0.106	0.136	90.3	0.141	0.179
	decay-OGD	90.1	0.532	0.616	90.1	0.122	**0.154**	90.0	0.100	0.132	90.3	0.147	**0.176**
	PID	90.1	**0.373**	**0.424**	90.0	0.434	0.518	89.9	0.413	0.482	89.9	0.428	0.504
	ECI	90.1	0.395	0.458	90.0	**0.118**	0.173	89.8	**0.069**	0.111	90.2	0.135	0.177
	COP	90.1	0.385	0.448	90.0	**0.118**	0.173	89.8	**0.069**	0.112	90.2	0.136	0.178
	DDCI	90.1	0.392	0.454	90.0	**0.118**	0.174	89.8	**0.069**	**0.111**	90.2	0.135	0.178
	DDCI-Nex	90.1	0.400	0.463	90.0	**0.118**	0.174	89.8	**0.069**	0.111	90.2	**0.134**	0.178

B.2.2 Varying Significance Levels α and Squashing Functions $h(\cdot)$

We conducted experiments on Amazon stock dataset under Prophet model. To comprehensively evaluate the calibration behaviour of the online conformal methods, we conducted a multi-level significance experiment by sweeping α from 0.05 to 0.9. For each level, we measured the empirical coverage, the mean interval width, and the Winkler score across four methods (DDCI, ECI, PID and OGD).

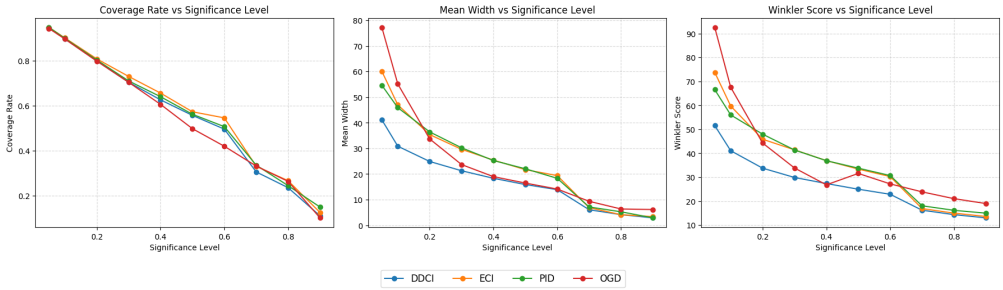


Figure 5: Comparison of empirical coverage rate, mean interval width, and Winkler Score across a wide range of significance levels (α = 0.05–0.9) for four online conformal inference methods: DDCI, ECI, PID, and OGD.

Across the full range of significance levels from 0.05 to 0.9, DDCI consistently demonstrates superior calibration stability and interval efficiency compared with existing online conformal methods. In terms of efficiency, DDCI achieves the narrowest mean interval widths across almost all α, reflecting its ability to maintain reliable calibration without resorting to excessively conservative intervals. This advantage is further reinforced by the Winkler Score results: DDCI attains the lowest Winkler Score throughout the entire spectrum, indicating the most favourable balance between coverage accuracy and interval compactness. Overall, the empirical curves of coverage, width, and Winkler Score collectively show that DDCI offers a more robust and efficient update rule, yielding both tighter and better-calibrated prediction intervals under varying levels of uncertainty.

We also tested different squashing functions with different significance level (0.2, 0.1 and 0.05). Table 5 shows that changing squashing function does not influence the results significantly. Details are given below.

Table 5: Results under different significance levels α on Amazon stock dataset.

Method	$\alpha = 0.20$			$\alpha = 0.10$			$\alpha = 0.05$		
	Coverage (%)	Avg. width	Winkler Score	Coverage (%)	Avg. width	Winkler Score	Coverage (%)	Avg. width	Winkler Score
OGD	79.7	33.70	44.27	89.6	55.15	67.57	94.3	77.24	92.63
SF-OGD	79.7	44.51	58.31	89.5	61.47	73.94	94.5	79.11	92.74
decay-OGD	80.1	46.92	57.98	89.9	97.22	107.79	90.9	107.32	159.64
PID	80.3	36.45	47.97	89.8	52.56	56.02	94.8	54.65	66.58
ECI	80.8	35.39	45.85	90.1	47.00	59.74	94.8	60.10	73.81
ECI-cutoff	80.0	31.81	40.93	89.7	43.46	53.21	94.8	55.97	66.65
ECI-integral	80.0	32.21	41.26	89.8	42.01	51.85	94.8	56.00	67.19
DDCI	79.9	<u>24.91</u>	<u>33.76</u>	89.7	**30.87**	41.06	94.8	<u>41.17</u>	<u>51.56</u>
DDCI-Nex	79.8	25.25	34.27	89.7	<u>30.92</u>	<u>40.95</u>	94.7	**41.09**	**51.32**
DDCI-Softsign	80.1	**24.56**	**33.34**	89.7	31.11	**40.90**	94.8	41.23	51.77

B.2.3 Learning Rate η Analysis

In this section, we analyse the impact of different learning rates on the performance of the proposed method, using the Google stock dataset as an example. The learning rate range follows Wu et al. [47]. For PID, ECI, and DDCI, the base learning rates η considered are [1, 0.5, 0.1, 0.05] and these work used the proportional learning rates $\eta_t = \max_{t-w \le i,j \le t} |s_i - s_j| \cdot \eta$. The results for coverage and prediction set of four underlying algorithms are given from Fig. 6 to Fig. 13. In our experimental

setting, the range d for the squashing function is set to 1. Therefore, the upper bound for η should be $\frac{1}{2-\alpha+\epsilon}$. Since α and ϵ are typically small, the learning rate η should not exceed 0.5 (it is obvious that our proposed method loses validity and generates extremely wide prediction intervals when the learning rate is set to 1), which means that the empirical observations are consistent with the theoretical analysis. In general, such learning rate is suitable for all datasets without influencing the validity of intervals. However, such large learning rate always yields conservative prediction intervals for all online conformal methods.

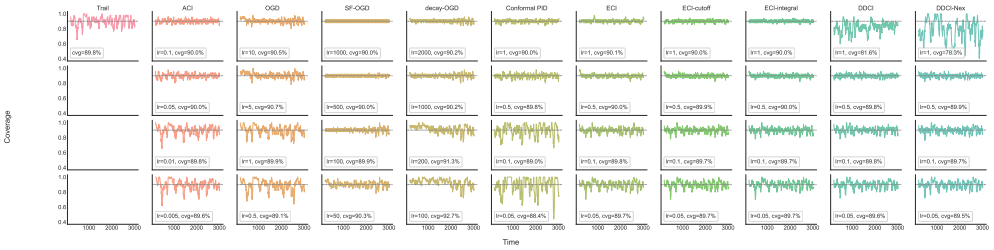


Figure 6: Coverage result for AR model.

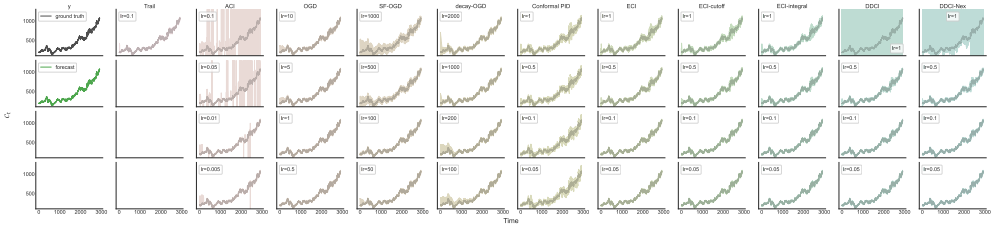


Figure 7: Prediction sets result for AR model.

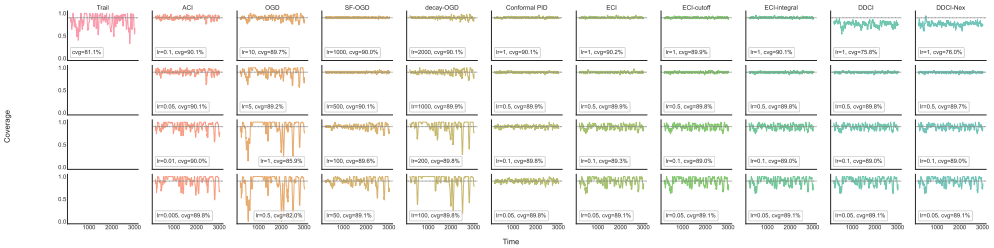


Figure 8: Coverage result for Prophet model.

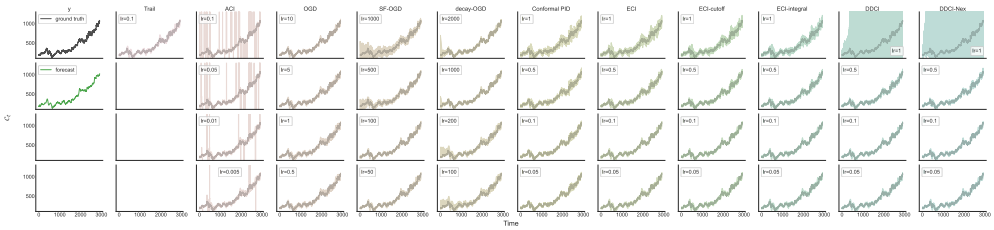


Figure 9: Prediction sets result for Prophet model.

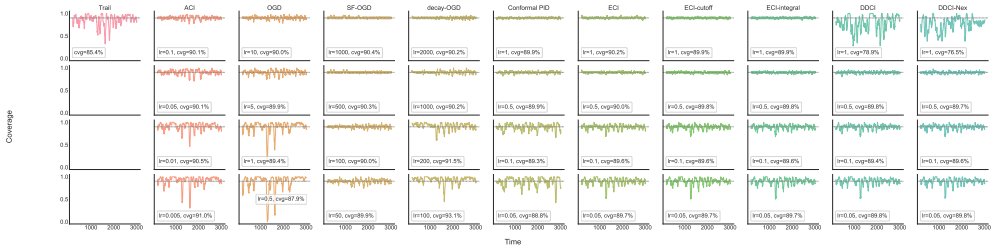


Figure 10: Coverage result for Theta model.

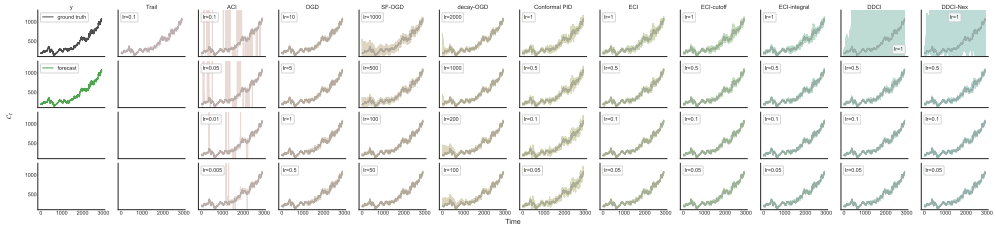


Figure 11: Prediction sets result for Theta model.

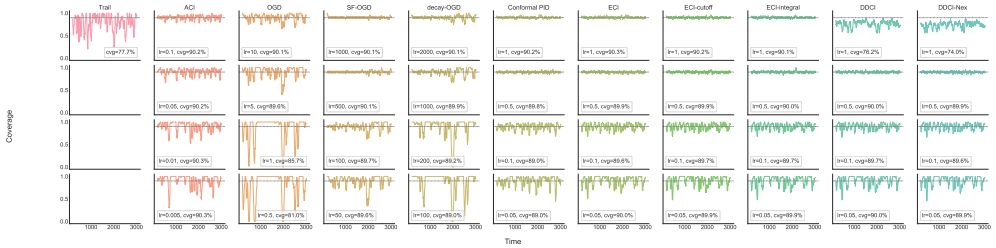


Figure 12: Coverage result for Transformer model.

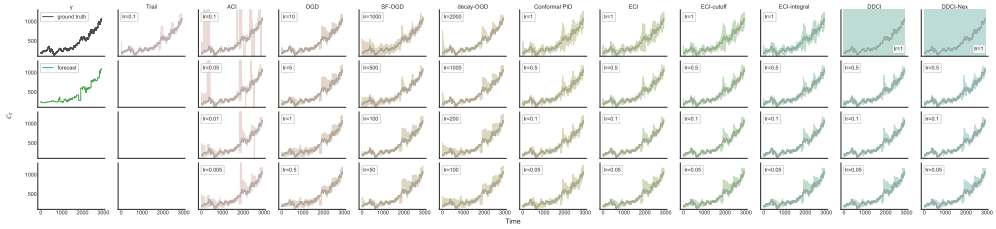


Figure 13: Prediction sets result for Transformer model.

B.2.4 Window Length w Analysis

We conducted experiments on Google and Amazon datasets, varying the window length w across [50, 100, 200, 300, 500] using Prophet. The results are given in the Fig. 14 below. Based on the results, a smaller w (e.g., 50) restricts the "memory" of the estimated feedback. It allows the empirical threshold q_t^* to closely track the current volatility, which causes high miscoverage. A larger w uses a long history to compute q_t^*. This introduces unnecessary lag, making the estimated feedback sluggish, thus inflating the interval width.

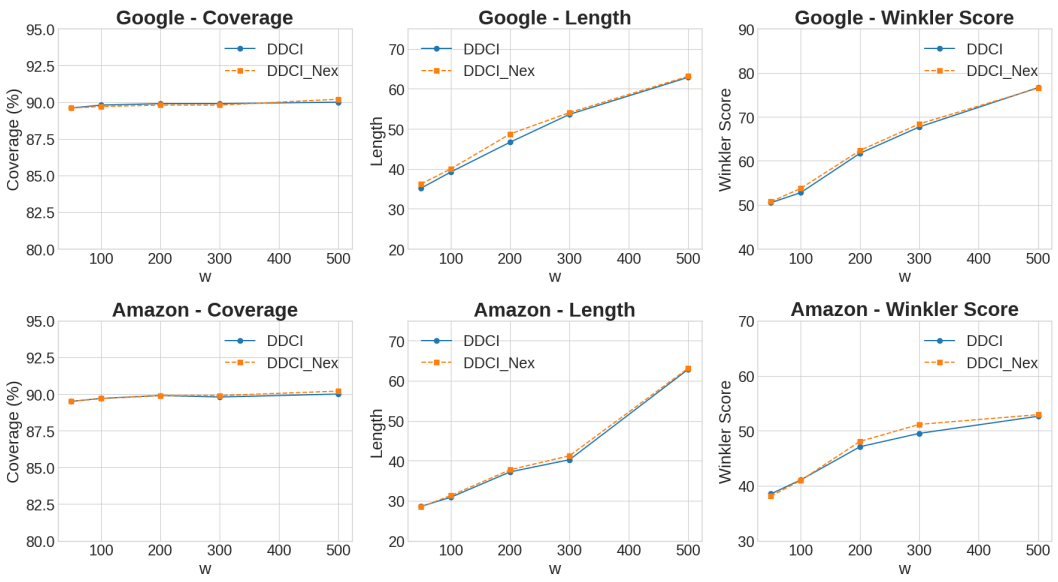


Figure 14: The influence of different window length on Google and Amazon datasets.

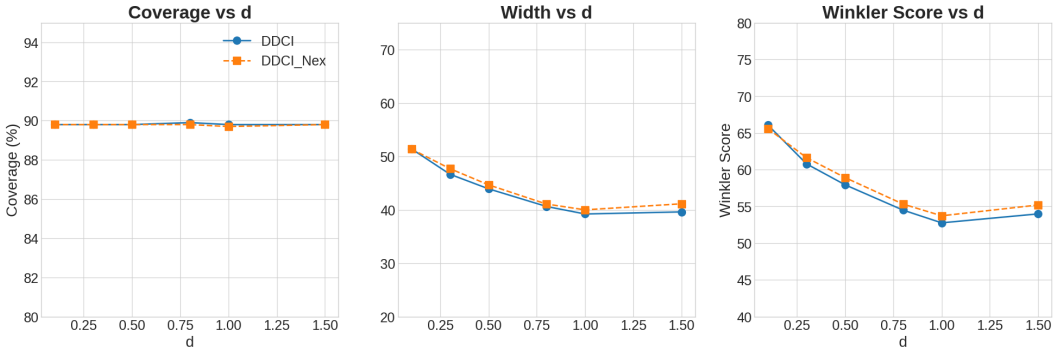


Figure 15: The influence of different d on the Google dataset under the Prophet model.

B.2.5 Squashing Function Range d Analysis

We did an analysis on d (ranging from 0.1 to 1.5) on Google dataset under Prophet model. The results show that the value of d would not influence the validity but it is meaningless when it is small. As shown in the Fig. 15, across a wide and practically meaningful range of values, the empirical coverage consistently concentrates tightly around the 90% target level. However, when d is set to small values, the algorithm effectively degenerates to standard OGD.

B.2.6 Steepness Parameter c Analysis

As illustrated in Fig. 16, we evaluate the impact of the steepness parameter c (ranging from 0.1 to 1.5) on the Prophet model's forecasting performance using the Google stock dataset. The leftmost subplot clearly demonstrates that the empirical coverage rates for both DDCI and DDCI-Nex remain exceptionally stable, forming a nearly flat line that closely adheres to the 90% target baseline across the entire range of c. Similarly, the middle and right subplots reveal that the Mean Interval Width and Mean Winkler Score exhibit strong robustness to variations in c, showing no significant degradation or structural fluctuation. These empirical observations corroborate our worst-case bound in Theorem 4.6, confirming that c controls the theoretical bias without substantially affecting empirical performance.

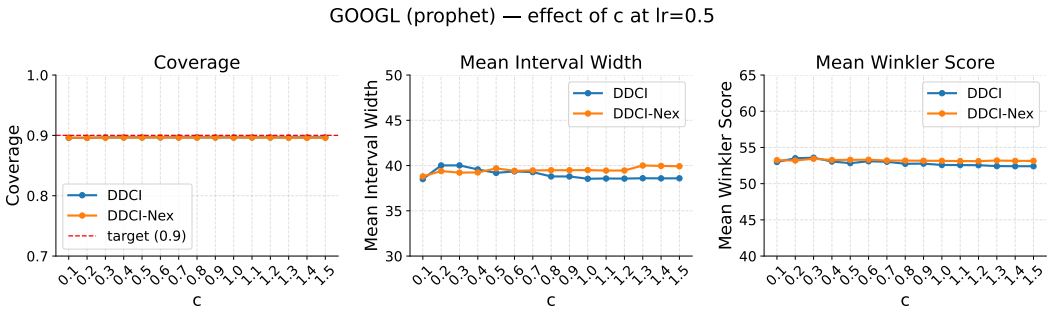


Figure 16: The influence of different c on the Google dataset under the Prophet model.

B.2.7 Ablation Study

Visualisation. Fig. 17 illustrates the effect of different values of ϵ on the proposed DDCI method through AR model. In this example, we set ϵ to 0 and 0.8 respectively, which can be regarded as a visualisation for ablation study. When ϵ is small, the update process becomes more volatile, leading to fluctuations in the prediction intervals. In contrast, larger ϵ values produce smoother and more stable updates, though at the cost of wider intervals when the learning rate is large. Thus, the choice of ϵ entails a trade-off between efficiency and stability.

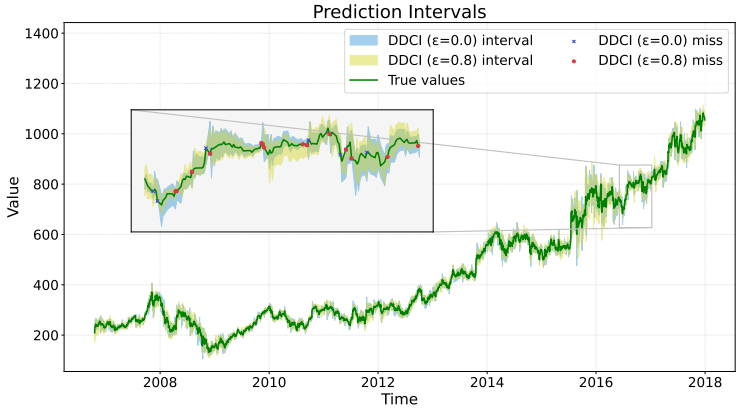


Figure 17: Visualisation of the proposed DDCI method with different ϵ (0.0 and 0.8 respectively) under AR model.

Numerical Results. We conducted the experiments on all datasets under all models for PID, ECI, and dual/actual/no-feedback methods. The results are given in the Tab. 6. The ablation results clearly validate our *core contribution*: introducing the estimated feedback successfully stabilises the update process. While all three methods maintain the target coverage, our dual-feedback method achieves significantly less oscillation, demonstrating superior stability.

Across these 15 additional dataset–forecaster settings, the full DDCI achieves the lowest MASD in 14 settings. The only exception is MSFT under Theta, where the actual-feedback-only configuration obtains a slightly lower MASD (0.0825 versus 0.0842). Nevertheless, the full DDCI outperforms PID, ECI, and the $c = 0$ configuration in all 15 settings. Adding estimated feedback reduces MASD relative to actual feedback alone in 14 of the 15 settings. Combined with the five Prophet results in Table 3, DDCI achieves the lowest MASD in 19 of the 20 dataset–forecaster settings.

C Reproducibility

All experiments are run with Apple M2 Pro. Here is the anonymous link of the code to reproduce the experimental results and figures: `https://anonymous.4open.science/r/DDCI-47A7`.

Table 6: Additional stability results under AR, Theta, and Transformer. Electricity values are displayed after multiplication by 10^3. Bold and underline denote the lowest and second-lowest unrounded values, respectively.

Forecaster	Dataset	PID	ECI	DDCI ($c = 0$)	DDCI ($\epsilon = 0$)	DDCI
AR	GOOGL	1.3204	0.5630	<u>0.4707</u>	0.4710	**0.4294**
	AMZN	1.1762	0.8711	0.7260	<u>0.7227</u>	**0.6330**
	MSFT	0.8166	0.5351	0.4100	<u>0.4034</u>	**0.3768**
	Delhi	1.0344	0.4004	0.2860	<u>0.2757</u>	**0.2654**
	Electricity	10.2395	4.0433	4.0455	<u>4.0011</u>	**3.9913**
Theta	GOOGL	1.6048	1.3182	1.1391	<u>1.1356</u>	**1.0094**
	AMZN	1.1951	0.8819	0.7296	<u>0.7274</u>	**0.6369**
	MSFT	0.5412	0.1214	0.0919	**0.0825**	<u>0.0842</u>
	Delhi	1.0057	0.3627	0.2546	<u>0.2441</u>	**0.2341**
	Electricity	8.9418	2.4094	2.4099	<u>2.4002</u>	**2.3952**
Transformer	GOOGL	10.9010	11.8597	<u>10.4455</u>	12.4885	**8.6863**
	AMZN	7.0344	7.8958	<u>6.7830</u>	8.3479	**5.5897**
	MSFT	0.9441	0.7432	<u>0.5823</u>	0.6178	**0.5038**
	Delhi	1.2411	0.7240	0.5308	<u>0.5229</u>	**0.4725**
	Electricity	11.0325	4.9771	4.9832	<u>4.9430</u>	**4.9241**

D Usage of Language Models

Large Language Models were used to polish and improve the grammar, spelling, and readability of this manuscript, but all research design, analysis, and interpretation were conducted by the authors.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contributions are summarised at the end of the Section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The theoretical limitation of the proposed method is discussed within Section 4.2 and the analysis of the additional parameter ϵ is given within Section A.5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The theoretical analysis is given in the Section 4.2 and the related proof is given in the Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The detailed methodology is described in Section **??** and all experimental settings are provided in Section 5.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The anonymous Github code link is provided in the Appendix C.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: All experimental settings are provided in the Section 5.1

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Most online CP methods (including ACI[22], PID[4], ECI[47]) do not report variance because the update rules are deterministic and our experiments follows their settings strictly. For our synthetic datasets, we used 5 random seeds to generate three settings and reported the variance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments are conducted with Apple M2 Pro.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The baselines and datasets used in this paper are drawn from the open-source implementations and the original works [4, 47] are cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The model introduced in the paper is well documented and the documentation is provided in the anonymous Github link in Appendix C.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is used only for writing and editing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.