Conformal time series decomposition with component-wise exchangeability

Derck W. E. Prinzhorn[*]   DERCK.PRINZHORN@STUDENT.UVA.NL

Thijmen Nijdam[*]   THIJMEN.NIJDAM@STUDENT.UVA.NL

Putri A. van der Linden   P.A.VANDERLINDEN@UVA.NL

Alexander Timans   A.R.TIMANS@UVA.NL

University of Amsterdam, The Netherlands

Abstract

Conformal prediction offers a practical framework for distribution-free uncertainty quantification, providing finite-sample coverage guarantees under relatively mild assumptions on data exchangeability. However, these assumptions cease to hold for time series due to their temporally correlated nature. In this work, we present a novel use of conformal prediction for time series forecasting that incorporates time series decomposition. This approach allows us to model different temporal components individually. By applying specific conformal algorithms to each component and then merging the obtained prediction intervals, we customize our methods to account for the different exchangeability regimes underlying each component. Our decomposition-based approach is thoroughly discussed and empirically evaluated on synthetic and real-world data. We find that the method provides promising results on well-structured time series, but can be limited by factors such as the decomposition step for more complex data.

Keywords: Conformal prediction, time series decomposition, exchangeability regimes

1. Introduction

Time series forecasting is a central task that sees widespread occurrence in various sectors such as finance, business, energy, or weather (Sezer et al., 2020; Deb et al., 2017; Lim and Zohren, 2021), and decision-makers value the ability to quantify the uncertainty associated with a given forecast (Makridakis and Bakas, 2016; Padilla et al., 2021). Conformal prediction (CP) has emerged as a popular method for such uncertainty quantification, providing distribution-free, finite-sample coverage guarantees under assumptions on *data exchangeability* (Shafer and Vovk, 2008; Vovk et al., 2005). However, data modalities such as time series are generally non-exchangeable due to their sequentially correlated nature, rendering applications of CP challenging. Recent CP approaches catered to non-exchangeable data such as Barber et al. (2023), Xu and Xie (2021) or Gibbs and Candès (2021) relax these exchangeability assumptions through weighting strategies emphasizing sample contributions, or rely on alternate assumptions altogether. These methods, however, always operate directly on the original time series.

Alternatively, temporal signals can be represented through a set of components that reflect varying informative temporal patterns, obtained via a *time series decomposition*

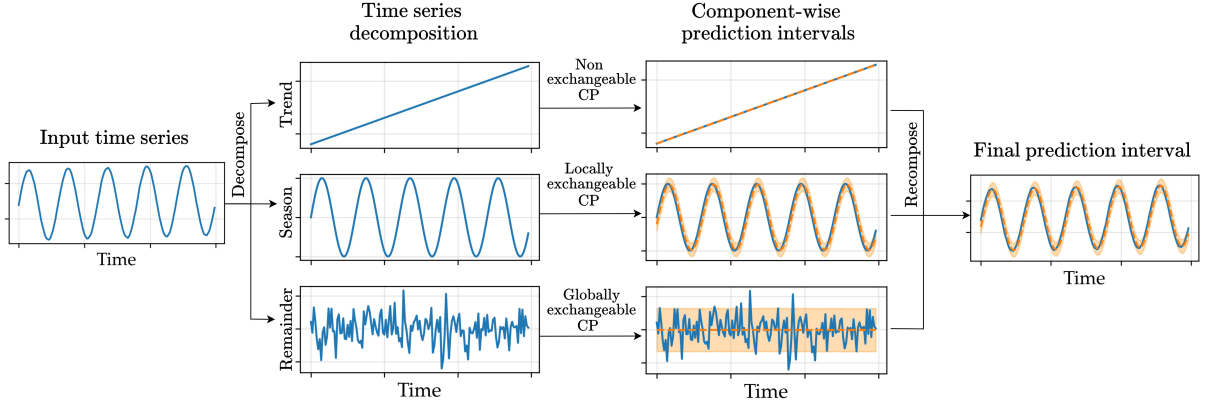[*] Equal contributions. Correspondence to <derck.prinzhorn@student.uva.nl>.

Figure 1: A high-level overview of our conformal time series decomposition approach. A time series signal is *decomposed* into individual components, and each component is associated with a specific regime of exchangeability (none, local or global). Relevant conformal methods are applied to each component, producing intermediate prediction intervals which are then *recomposed* to obtain a prediction interval for the overall time series.

(TSD, Hyndman and Athanasopoulos (2018)). Such a decomposition provides a flexible way of capturing and analysing temporal correlations at different scales, and has been employed to improve forecasting quality or identify outliers (Theodosiou, 2011; Wen et al., 2019). Interestingly, the interplay between possible decomposition models and CP approaches for time series remains unexplored.

In our work, we investigate this connection by leveraging TSD principles for conformal time series forecasting (see Fig. 1). Our approach decomposes the time series and relates each extracted component to a specific underlying *exchangeability regime*, which guides the application of tailored CP methods per component. Obtained component-wise prediction intervals (PIs) are then recomposed to obtain aggregated intervals for the overall time series. We validate our methods on both synthetic and real-world data, showing promising results when exchangeability assumptions seem satisfied. To the best of our knowledge, we are the first to thoroughly explore the intersection of conformal methods and time series decomposition. In summary, our key contributions include:

- a proposed pipeline for conformal time series forecasting via time series decomposition, with several crucial design choices outlined and motivated; and

- conceptually relating temporal dependencies underlying each time series component to different notions of exchangeability, including novel weighting strategies for the seasonal time series component which are framed in terms of local exchangeability.

2. Background: time series decomposition and exchangeability regimes

We begin by introducing notation for our time series and conformal prediction settings. We next provide general background on time series decomposition, and follow with details on selected CP methods for identified regimes of data exchangeability relevant to our approach.

Time series data. We define a time series of length T as a sequence of sampled covariate-response pairs $\mathcal{D} = \{(x_t, y_t)\}_{t=1}^T \subseteq \mathcal{X} \times \mathcal{Y}$ from P_{XY}. We denote feature and target spaces $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$, since we consider a univariate time series (as opposed to multivariate responses). Observed pairs are split into three disjoint sets: a training set $\mathcal{D}_{train}$, calibration set $\mathcal{D}_{cal} = \{(x_i, y_i)\}_{i=1}^n$ and test set $\mathcal{D}_{test} = \{(x_j, y_j)\}_{j=n+1}^{n_t}$. $\mathcal{D}_{train}$ is used to fit a prediction model $\hat{f}_\theta : \mathcal{X} \to \mathcal{Y}$, $\mathcal{D}_{cal}$ is leveraged for any conformal procedure, and we evaluate on $\mathcal{D}_{test}$.

Conformal prediction. In order to build coverage-controlling prediction sets, we first define a scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ applied to $\mathcal{D}_{cal}$, resulting in a set of nonconformity scores $S = \{s(\hat{f}_\theta(X_i), Y_i)\}_{i=1}^n \triangleq \{s_i\}_{i=1}^n$[1]. These encode a desired notion of dissimilarity (*i.e.*, nonconformity) between predictions and responses, such as the plain residual $|\hat{f}_\theta(X_i) - Y_i|$. The scores are leveraged to compute a conformal quantile $\hat{Q}(1 - \alpha; \hat{F}(S))$, where $\hat{F}(S)$ denotes the empirical c.d.f. of the score set $S \cup \{+\infty\}$[2]. The target coverage level $(1 - \alpha)$ is defined for a tolerated miscoverage rate $\alpha \in (0, 1)$. For a new test sample (X_{n+1}, Y_{n+1}), a valid prediction set for X_{n+1} is then constructed as $\hat{C}(X_{n+1}) = \{y \in \mathcal{Y} : s(\hat{f}(X_{n+1}), y) \leq \hat{Q}(1 - \alpha; \hat{F}(S))\}$. Assuming data exchangeability holds, the set $\hat{C}(X_{n+1})$ then guarantees the desired marginal coverage, *i.e.*, we have that $\mathbb{P}_{\mathcal{D}_{cal}, \mathcal{D}_{test} \sim P_{XY}}(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha$. For our regression setting, the sets $\hat{C}(X_{n+1}) \subseteq \mathbb{R}$ take the form of prediction intervals (PIs).

2.1. Time series decomposition

Time series decomposition (TSD) involves the analysis of a time series by decomposing it into several components, each representing distinct time-dependent patterns across varying durations. Methods for TSD can vary in the type of time-dependent patterns they extract (*e.g.*, components in target or frequency domains), and how these are combined (*e.g.*, explicitly or implicitly). Most commonly, one assumes some form of explicit decomposition model with components in target space, *i.e.*, that of the observed responses $\{y_t\}_{t=1}^T$ (see, *e.g.*, Hyndman and Athanasopoulos (2018)). These components typically include a long-term *trend*, reflecting the overarching direction of the series over extended periods; a *seasonal* component, capturing recurrent patterns within a specific cycle, such as yearly or quarterly variations; and a noise term or *remainder* component, accounting for fluctuations, irregularities and randomness not explained by the trend or seasonal influences.

Specifically, let Y_t denote the random variable of the response at time t. Then, a TSD into trend T, seasonality S and remainder component R can be modelled as

$$Y_t = T_t \circ S_t \circ R_t, \tag{1}$$

where $\circ$ is a binary operator, often chosen to be either addition ($+$) or multiplication ($\times$). Generally, an additive model is suitable when the time series is relatively stable and exhibits low volatility, while a multiplicative structure is useful when there is strong proportionality of components to the level of the time series, *i.e.*, a dependency on the magnitude of the data (Hyndman and Athanasopoulos, 2018). Alternatively, data transformations such as into log-space or via Box-Cox power transforms (Osborne, 2010) may alleviate such dependencies and permit the use of additive models nonetheless.

---

1. Random variables are denoted upper-case (X) and their observed realizations as lower-case (x).
2. The inclusion of $\{+\infty\}$ is a notational formality, see, *e.g.*, Barber et al. (2023).

Various methods exist for estimating T, S and R. Classical decomposition approaches following Macaulay (1931) involve isolating the trend through a smoothing mechanism, such as a weighted moving average. The estimated trend can in turn be used to extract seasonal patterns by de-trending the time series and averaging over seasonal instances, where the cycle frequency (*e.g.*, yearly, monthly or daily) is pre-specified. Finally, the remainder is defined as the portion of the signal that is not explained by either component. This general framework has been extended and improved upon by a range of TSD methods which vary in expressivity, assumptions and use-cases (see § 4).

A popular approach is *Seasonal-Trend decomposition based on LOESS* (STL) introduced by Cleveland et al. (1990), which leverages local weighted polynomial regression for the decomposition (see Cleveland and Devlin (1988) for details). STL continues to see use in recent applications due to favourable properties such as its non-parametric formulation, ability to handle varying cycle frequencies of unconstrained form, and outlier robustness due to its weighting mechanism (Hyndman and Athanasopoulos, 2018; Theodosiou, 2011). It is thus also our employed decomposition method throughout the experiments (§ 5).

Regardless of a particular choice, TSD methods remain constrained by the inherent limitations of an attempted decomposition, which assumes clear underlying patterns and is susceptible to real-world data irregularities, adversely affecting any decomposition results. Such irregularities can include data outliers, fluctuating seasonal patterns, complex systematic interactions (*e.g.*, irregular trend behaviour), or inherently low signal-to-noise ratios (Wen et al., 2019). While expectations on a proper decomposition for 'well-behaved' time series may be warranted, *e.g.*, by arguments such as Wold's decomposition (Miamee and Pourahmadi, 1988), retrieving interpretable components remains challenging in practice.

2.2. Conformal methods for different exchangeability regimes

Conformal prediction methods were first proposed assuming global exchangeability, *i.e.*, data exchangeability across all samples (Vovk et al., 2005). We define data exchangeability similar to Shafer and Vovk (2008) as follows:

Definition 1 (Exchangeability) *A sequence of random variables $X_1, \ldots, X_n$ is exchangeable if for any permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ with $n \geq 1$ we have that*

$$P(X_1, \ldots, X_n) = P(X_{\pi(1)}, \ldots, X_{\pi(n)}).$$

That is, the joint distribution of $X_1, \ldots, X_n$ remains invariant to any permutation of indices. Notably, the setting of *i.i.d* random variables is a special case thereof. Relaxing this assumption to weaker notions of exchangeability has proven fruitful and permitted broadening the application of CP methods to settings with (temporally) dependent data (see § 4), for which Definition 1 is violated.

Relevant to our problem setting, we identify three distinct notions of exchangeability of varying assumption strength: *(i)* the strongest condition of global exchangeability; *(ii)* a total lack of exchangeability, instead relying on alternate assumptions; and *(iii)* in-between forms of local exchangeability based on weighted CP. We next describe selected CP methods for each of these three exchangeability regimes in more detail, which are subsequently leveraged for our TSD approach in § 3.

Global exchangeability. The well-known standard split CP (Papadopoulos et al., 2007) belongs to this case. Herein, the set of nonconformity scores S is obtained from a single-split calibration set to compute the conformal quantile $\hat{Q}(1 - \alpha; \hat{F}(S))$. To alleviate dependency on the single split and improve adaptivity, principles from cross-validation can be employed (Kim et al., 2020; Vovk, 2015). Specifically, Barber et al. (2021) propose the jackknife+ as a per-sample, and CV+ as a per-fold *leave-one-out* approach. The joint set $\mathcal{D}_{train} \cup \mathcal{D}_{cal}$ is split into K distinct folds and predictors $\hat{f}_{\theta, -k}$ are fitted with the k-th fold removed, which in turn acts as an unbiased set to compute out-of-fold nonconformity scores. For any new test sample (X_{n+1}, Y_{n+1}), both split and cross-conformal approaches assume data exchangeability of $\mathcal{D}_{cal} \cup \{(X_{n+1}, Y_{n+1})\}$ for coverage guarantees to hold.

Lack of exchangeability. In the presence of temporal dependency between samples, proposed CP methods employ update strategies to achieve target coverage over time under alternate assumptions (see § 3.2), and provided guarantees do not require data exchangeability. Ensemble batch prediction intervals (EnbPI, Xu and Xie (2021)), a popular method, makes use of repeated updates to the calibration set. An ensemble of bootstrapped estimators is used to obtain a score set S_t of *leave-one-out* nonconformity scores, up to some timestep t. A prediction interval $\hat{C}(X_{t+1})$ for $t + 1$ then uses a time-dependent quantile $\hat{Q}_{t+1}(1 - \alpha; \hat{F}(S_t))$, where S_t is updated in a rolling fashion for the next step as $S_{t+1} := (S_t \backslash \{s_1\}) \cup \{s_{t+1}\}$. Adaptive conformal inference (ACI, Gibbs and Candès (2021)), another popular approach, proposes quantile updating on the basis of previously observed coverage. That is, given the current timestep t, PIs for $t + 1$ utilize a time-dependent quantile $\hat{Q}_{t+1}(1 - \alpha_{t+1}; \hat{F}(S_t))$, where the target coverage level α_{t+1} is affected by (mis-)coverage at the previous step via the update $\alpha_{t+1} = \alpha_t + \gamma (\alpha - \mathbb{1}[Y_t \notin \hat{C}(X_t)])$, with step size $\gamma > 0$.

Local exchangeability. A number of CP approaches operate in-between the two mentioned regimes. These primarily leverage weighting mechanisms to assign varying measures of importance to individual scores, where such importance may relate to sample distances in time, feature space, distribution, and others (see § 4). A general formulation of *weighted CP* is given by computing a conformal quantile $\hat{Q}(1 - \alpha; \hat{F}_w(S))$, where $\hat{F}_w(S)$ is now the empirical c.d.f. over a weighted score set and takes the form

$$\hat{F}_w(S) = \sum_{i=1}^{n} \tilde{w}_i \, \delta_{s_i} + \tilde{w}_{n+1} \, \delta_{+\infty}, \tag{2}$$

with δ_{s_i} denoting the dirac delta centered at score s_i, and $\tilde{w}_i$ its associated normalized weight such that $\sum_{i=1}^{n} \tilde{w}_i = 1$. Barber et al. (2023) suggest fixed weights which aim to minimize the total variational distance, *e.g.*, by upweighting more recent samples in case of distribution shifts; whereas Guan (2023) employ data-dependent weights determined by feature distance via a 'localizer' function, such as the kernel distance $\exp\{-h\,|X_i - X_{n+1}|\}$.

Interestingly, a relationship between particular weighting schemes and exchangeability assumptions can be established. For instance, it is straightforward to see that Eq. 2 with the choice $\tilde{w}_i = 1/(n + 1) \; \forall i = 1, \ldots, n$ recovers standard split CP, and thus subsumes global exchangeability for coverage guarantees to hold. We hypothesize that similar relations can be established for locally restricted (approximate) data exchangeability, as illustrated in § 3.3. We suggest defining such local exchangeability similarly to Definition 1 as below:

Definition 2 (Local exchangeability) *Consider a sequence of random variables $X_1, \ldots, X_n$ and denote its index set by $\mathcal{I} := \{1, \ldots, n\}$. For any subset $\mathcal{I}' := \{1, \ldots, k\} \subset \mathcal{I}$ the associated random variables $X_1, \ldots, X_k$ are locally (approximately) exchangeable if for any permutation $\tilde{\pi} : \mathcal{I}' \to \mathcal{I}'$ with $|\mathcal{I}'| \geq 1$ we have that $P(X_1, \ldots, X_k) \simeq P(X_{\tilde{\pi}(1)}, \ldots, X_{\tilde{\pi}(k)})$.*

3. Leveraging exchangeability regimes for time series components

We now explicitly match CP methods for varying regimes of exchangeability to time series components obtained from its decomposition. Our key insight is the correspondence between obtained components and their potential to fulfill respective exchangeability assumptions, permitting the application of tailored CP algorithms to each component. While the correspondence is straightforward for remainder and trend terms, we explore different weighting mechanisms for the seasonal term, and relate them to notions of local exchangeability by Definition 2. Our exchangeability hypotheses are empirically challenged in § 5.

3.1. Remainder component: globally exchangeable

The remainder component R (Eq. 1) accounts for fluctuations that are left unexplained by other components. Also referred to as error or noise ϵ, this term is frequently paired with a systematic component $f_\theta(X)$ in many statistical models, typically via the additive structure $f_\theta(X) + \epsilon$. Additional assumptions may be placed on ϵ, such as on *i.i.d* sample draws or its distribution origin (*e.g.*, the common assumption $\epsilon \sim N(0, \sigma^2)$). These conditions apply to a broad spectrum of (non-)parametric models, including linear regression, generalized linear models, regression splines, ARIMA, and more (see, *e.g.*, Hastie et al. (2009)). Similarly, STL and its regression mechanism are based on a linear additive structure, but do not impose any distributional form (Cleveland et al., 1990).

Aligned with these common assumptions, we hypothesize R to satisfy a lenient *i.i.d* condition. If the signal is perfectly decomposed according to Eq. 1, the remainder term will ideally reflect uncorrelated functional randomness only – what is referred to as 'white noise' in time series modelling (Parzen, 1966). While practicly unattainable due to approximation errors, one can nonetheless expect an accurate TSD to capture a large part of the non-exchangeable time series signal, thereby justifying a relaxed *i.i.d* assumption on R via global exchangeability. This permits the use of CP methods requiring such a condition, including standard split CP and its cross-validation variants. Specifically, we consider the CV+ algorithm by Barber et al. (2021), which minimizes dependency on a single calibration split and is observed to yield smaller prediction intervals than split CP. On the other hand, its finite-sample, distribution-free coverage guarantee is limited to $(1 - 2\alpha)$, similar to other cross-conformal approaches (Vovk, 2015). In practice, empirical coverage tends to satisfy $(1 - \alpha)$ and above, as noted by Barber et al. (2021) and corroborated in our experiments.

3.2. Trend component: non-exchangeable

The trend reflects long-term temporal patterns in the time series, and captures the strongest proportion of temporally correlated signal in the data. As such, observed trend values $\{T_t\}_{t=1}^T$ exhibit strong serial correlation, and any permutation risks modifying their joint distribution, thus violating Definition 1. In such settings, one may directly leverage any

proposed CP methods designed for non-exchangeable time series, relying on the relationship between a time series and its extracted trend. This strong correspondence is evident, for example, in the common practice of detrending, a preprocessing step used to achieve data stationarity (Wu et al., 2007).

In our experiments, we rely on the two popular CP approaches EnbPI (Xu and Xie, 2021) and ACI (Gibbs and Candès, 2021), which we briefly outlined in § 2.2. Since data exchangeability is considered violated, these methods rely on alternate assumptions to derive nominal coverage guarantees. EnbPI assumes a linear model structure $Y_t = f_\theta(X_t) + \epsilon_t$, and a stationary and strongly mixing error process $\{\epsilon_t\}_{t=1}^T$[3]. Under such conditions, a lower bound on the procedure's under-coverage is derived in finite samples (Xu and Xie (2021), Thm. 1), and for $t \to \infty$ asymptotically attains marginal coverage $(1 - \alpha)$. ACI does not require additional assumptions beyond conditions on the step size γ to provide bounds on the miscoverage rate (Gibbs and Candès (2021), Prop. 4.1), which likewise recovers asymptotic marginal coverage at level $(1 - \alpha)$. In both instances, we generally find that empirical coverage is achieved in finite samples. We refer to their respective works for further details.

3.3. Seasonal component: locally exchangeable

The seasonal component S is designed to capture regular patterns in time series, specifically periodically recurring variations like weekly or daily peaks. Although extracted patterns are time-dependent, their repetitive nature implies that such dependency might be confined to shorter time intervals. Specifically, for a consistent periodic schedule, strong temporal correlation is confined to samples within a single period (*e.g.*, from one trough to the next), leading us to hypothesize that samples *across* periods can be considered locally (approximately) exchangeable. For example, permuting the indices of samples at the period peaks or troughs is likely to have minimal impact on the overall pattern. In practice, we may test such local exchangeability by means of weighted CP according to Eq. 2, with particular weights $\tilde{w}$ reflecting different exchangeability conditions. We follow by detailing three such weighting schemes, qualitatively illustrated in Fig. 2.

Weighting strategies. Consider a seasonal pattern with fixed period length $\tau > 0$. In addition, let us denote by τ_{n+1} the position of a given test sample with respect to its period start. For example, if we consider the sample to be located at the season peak in Fig. 2, then $\tau_{n+1} = 4$ for the given period length $\tau = 7$ (equating the particular time step). Next, let $\mathcal{I}_{cal} := \{1, \ldots, n\}$ denote the index set of $\mathcal{D}_{cal}$. Our first approach **BinaryPoint** subsets $\mathcal{I}_{cal}$ based on the test sample's precise period location to obtain a binary weighting mechanism. That is, the filtered index set $\mathcal{I}_{cal}^{\mathrm{BP}} \subset \mathcal{I}_{cal}$ takes the form
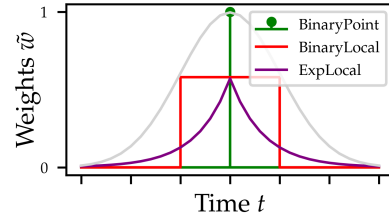

Figure 2: Illustration of weighting schemes for an exemplary season (grey) with period $\tau = 7$. The test sample's location is at the period peak, thus $\tau_{n+1} = 4$.

$$\mathcal{I}_{cal}^{\mathrm{BP}} := \{i \in \mathcal{I}_{cal} : i \bmod \tau = \tau_{n+1}\}. \tag{3}$$

---

3. The *i.i.d* setting can be highlighted as a special mixing case.

The employed weighting mechanism for Eq. 2 is then given by $\tilde{w}_i = 1/(|\mathcal{I}_{cal}^{\mathrm{BP}}| + 1) \cdot \mathbb{1}[i \in \mathcal{I}_{cal}^{\mathrm{BP}}] \; \forall i = 1, \ldots, n$, where equal weights are assigned to every sample whose index is in the filtered set, and zero otherwise[4]. Due to the risk of small calibration set sizes, we further consider an extension mechanism **BinaryLocal**, wherein BinaryPoint is expanded to include indices within local neighbourhoods $k > 0$ of every sample in $\mathcal{I}_{cal}^{\mathrm{BP}}$. In Fig. 2 such a neighbourhood is set to $k = 1$, *i.e.*, we include samples on either side of τ_{n+1} at the peak. As for BinaryPoint, the samples are weighted equally if included and zero otherwise, producing a hard binary weighting scheme. As an alternative, we also consider a softer weighting mechanism **ExpLocal**, which applies a repeated exponential decay to all samples in $\mathcal{D}_{cal}$. Specifically, an unnormalized weight for the i-th calibration sample is given by the decay $\lambda e^{-\lambda \cdot \Delta\tau_i}$, where $\lambda \in (0, 1)$ denotes the decay rate and $\Delta\tau_i = |\tau_i - \tau_{n+1}|$ the sample's distance in regards to the target location τ_{n+1} for it associated period. Thus with increasing distance from τ_{n+1} the weights exponentially shrink, as visualised in Fig. 2.

Since our suggested weighting mechanisms are based on sample distance across time and assign fixed weights, they primarily relate to schemes discussed in Barber et al. (2023). Such weights leveraging temporal distance are particularly useful for time series, where auto-regressive effects are prevalent. Inspired by Guan (2023), we additionally explore data-dependent weighting based on feature distances in § E.

Relation to local exchangeability. An intuitive connection emerges between the described weighting mechanisms and underlying assumptions about local exchangeability following Definition 2. For BinaryPoint (Eq. 3), a highly selective calibration subset is chosen, and identical weights assume global exchangeability for all included samples. This equates a strong assumption of local exchangeability confined to a precisely defined calibration set. The condition is relaxed for BinaryLocal, wherein a larger set size is traded for a more approximate notion of local exchangeability, especially as the neighborhood expands or when the period length τ is small. ExpLocal further loosens the assumption by not selectively filtering calibration samples but rather weighting them based on their distance within periods. The associated notion of locality depends on factors like period length and decay rate λ, with the assumption approaching global exchangeability as λ tends towards zero.

3.4. Prediction interval recomposition

After applying selected CP methods to each component of the time series, the component-level prediction intervals (PIs) must be *recomposed* to establish final interval boundaries for the appropriate region in target space $\mathcal{Y}$. The recomposition approach is influenced by the chosen decomposition structure detailed in Eq. 1, linking directly to the choice of TSD model. Since we opt for STL, which is based on a linear decomposition, we utilize an additive structure for recomposition as well. Specifically, the lower and upper interval bounds for our conformal PI $\hat{C}(X_{n+1}) = [\hat{L}(X_{n+1}), \hat{U}(X_{n+1})]$ for Y_{n+1} are computed as

$$
\begin{aligned}
\hat{L}(X_{n+1}) &= \hat{L}_T(X_{n+1}) + \hat{L}_S(X_{n+1}) + \hat{L}_R(X_{n+1}) \\
\hat{U}(X_{n+1}) &= \hat{U}_T(X_{n+1}) + \hat{U}_S(X_{n+1}) + \hat{U}_R(X_{n+1}),
\end{aligned}
\tag{4}
$$

---

4. Observe that any split of $\mathcal{D}_{cal}$ can be defined as beginning at a seasonal period, which can take different shapes as long as the period length is consistent, *e.g.*, trough to trough, peak to peak, *etc.*

8

where the right-hand side represents the respective component-level bounds for trend T, season S and remainder R. In our experiments, this linear recomposition sometimes proved overly conservative, for example when interactions between components tend towards a multiplicative nature. To obtain optimally tight PIs, identifying an appropriate recomposition model is as crucial as an accurate decomposition. In addition, one might consider adjusting coverage levels or a re-weighting of individual components based on observed PI widths. For instance, relaxing coverage requirements for components with a small magnitude can reduce the interval size of the recomposed PI without notably affecting its coverage.

Aggregated coverage guarantee. Different CP methods can be employed across time series components, which complicates providing a clear statement about the nominal coverage for the recomposed PI detailed in Eq. 4. Under simplified conditions, a loose lower bound on nominal coverage is obtainable at level $1 - 3\alpha$ (see § A). In practice, traditional CP methods offer finite-sample guarantees, contrasting with the asymptotic guarantees provided by time series approaches. Additionally, any coverage guarantees applicable to settings of local exchangeability are highly dependent on specific data conditions, which limits the feasibility of generalizations. For example, the coverage bounds cited in Barber et al. (2023) are significantly influenced by the total variational distance term, whereas Guan (2023) highlight the need for careful tuning of the miscoverage rate to avoid overcoverage in practice. Consequently, deriving an overall coverage guarantee for the recomposed PI, and its practical applicability, remain open questions. The effectiveness of such guarantees is likely to be substantially affected by modelling factors such as decomposition quality and chosen weighting schemes.

4. Related work

Time series decomposition. Building upon classical TSD concepts (Macaulay, 1931), popular techniques include X-11 (Cleveland and Tiao, 1976), SEATS (Dagum and Bianconcini, 2016), TBATS (De Livera et al., 2011) and STL (Cleveland et al., 1990). We refer to Esling and Agon (2012) for an older survey and Mbuli et al. (2020) for a more recent overview with applications to power systems. In particular, STL sees continued use in recent works due to its favourable properties over, *e.g.*, X-11 or SEATS (Hyndman and Athanasopoulos, 2018). This includes domain-specific applications such as traffic, weather or air quality (Huo et al., 2019; He et al., 2022; Li and Jiang, 2023), as well as extensions tackling multi-seasonality (Bandara et al., 2021; Trull et al., 2022), robustness (Wen et al., 2019, 2020), online settings (He et al., 2023; Mishra et al., 2022) or uncertainty (Krake et al., 2024; Dokumentov and Hyndman, 2022). However, the connection between TSD models such as STL and CP frameworks remains unexplored. TSD models considering alternative decomposition structures (*i.e.*, not into trend, seasonality and remainder terms) include theta-differences (Assimakopoulos and Nikolopoulos, 2000), wavelet (Soltani, 2002) and spectral models (Bonizzi et al., 2014). More recently, implicit decomposition structures in deep learning-based time series models have been explored, *e.g.*, via self-attention mechanisms in transformers (Jiang et al., 2022; Wang et al., 2022; Zhou et al., 2022; Wu et al., 2021; Lin et al., 2021). However, retrieved structures are typically fuzzy and sensitive to model specification; see Wen et al. (2023) for a recent overview.

Conformal prediction for time series. Recent CP approaches for time series typically employ updating strategies to account for time-dependency of observations. Gibbs and Candès (2021) fix the calibration set but dynamically adjust the conformal quantile on the basis of observed coverage. Zaffran et al. (2022) build upon this idea with a more adaptive weighted quantile, while Angelopoulos et al. (2024) design a quantile tracker borrowing ideas from control theory. Auer et al. (2023) obtain quantile weights by leveraging the attention mechanism of a Hopfield Network model. Alternatively, Xu and Xie (2021) select a fixed quantile and instead periodically update the underlying calibration set. Jensen et al. (2022) combine this approach with CQR (Romano et al., 2019) for quantile regressors, while Xu and Xie (2023) replace the fixed quantile with a quantile estimator for more flexibility. While all the above consider a single time series, CP methods have been also proposed for the multivariate case. Assuming exchangeability between the observed time series, these can be used to build individual calibration sets for multiple prediction steps. Stankeviciute et al. (2021) treat these steps as independent, Sun and Yu (2023) account for dependency with a copula-based quantile, and Xu et al. (2024) design ellipsoidal sets. Lin et al. (2022) suggest quantile updates that account for coverage across both time steps and time series.

Local exchangeability. CP methods controlling for global exchangeability violations (and not specifically designed for time series problems) rely on notions of locally restricted or *weighted* exchangeability. In practice, this suggests subsetting or reweighting of calibration samples by some notion of importance. Tibshirani et al. (2019) suggest weights determined by the likelihood ratio of calibration and test samples to counter covariate shifts. Applications include Lei and Candès (2021) for counterfactuals, Fannjiang et al. (2022) for biomolecular design, and Candès et al. (2023) for survival analysis. Podkopaev and Ramdas (2021) extend ideas to account for potential shifts in target space. Guan (2023) propose weights determined by kernel distances in feature space, which is leveraged by Hore and Barber (2023) for improved test-conditional coverage. Györfi and Walk (2019) suggest a nearest-neighbour reordering of scores based on feature distances, while Amoukou and Brunel (2023) and Han et al. (2022) reweigh implicitly via density estimates of the empirical c.d.f. $\hat{F}(S)$. Recently, Barber et al. (2023) suggest fixed weights on the basis of total variational distances quantifying sequence exchangeability. Conceptually related to weighted CP are *mondrian* CP methods, wherein $\mathcal{D}_{cal}$ is partitioned according to different requirements and exchangeability is assumed to hold within subgroups (Cauchois et al., 2021; Jung et al., 2023; Toccaceli and Gammerman, 2019).

5. Experiments

We next empirically verify the effectiveness of our TSD approach across a range of time series datasets. In order to validate our design choices, we first evaluate on a synthetic time series with known decomposition structure. We then test our approach on three real-world time series from different domains: *San Diego energy consumption*[5], *Rossman store sales*[6] and *Beijing air quality*[7] (Zhang et al., 2017). We begin by outlining our experimental setup,

---

5. https://github.com/pratha19/Hourly_Energy_Consumption_Prediction

6. https://github.com/juniorcl/rossman-store-sales

7. https://github.com/Afkerian/Beijing-Multi-Site-Air-Quality-Data-Data-Set

which includes employed regression models, CP baselines and evaluation metrics. Our code is publicly available at https://github.com/dweprinz/CP-TSD.

Model setup. We consider three different regression models $\hat{f}_\theta$ for the underlying predictor: a simple linear model, an MLP regressor, and a gradient boosting model. We train the models in an auto-regressive fashion, *i.e.*, model covariates at time t may include previous observations $y_{t-k}, \ldots, y_{t-1}$ up to lag order k alongside other available features. In order to handle extrapolation for non-stationary data, the gradient boosting regressor is trained on differenced data $\Delta y_t = y_t - y_{t-1}$. Such auto-regressive training procedures are in line with our *sequential* setup, wherein we simulate receiving a single observation per time step in a streaming fashion[8]. Our regressors are implemented via `scikit-learn` (Pedregosa et al., 2011) and minimally tuned (see § G), while employed conformal methods leverage the library `MAPIE` (Taquet et al., 2022) when possible.

Unless stated otherwise, our default target coverage level is fixed at 90% ($\alpha = 0.1$). Throughout, we employ STL (Cleveland et al., 1990) for time series decomposition, as motivated in § 2.1. Our TSD models are compared using the different weighting schemes described in § 3.3 for the seasonal component, while CP algorithms EnbPI and CV+ are employed for the trend and remainder terms, respectively. Results on real-world data in Table 2 are reported for the linear regressor, which proved most robust in terms of hyperparameter selection. All results are averaged across multiple seeds. Further results across different combinations of methods and coverage levels can be found in § B and § C. Predictive model performance as measured via RMSE is reported in § G.

Baselines. We baseline our TSD approaches against CP algorithms EnbPI and ACI in two ways. Firstly, we consider their direct application to the raw time series signal without a decomposition (marked as (✗) in results tables below). This baseline aims to quantify performance differences attributable to the employed TSD model, or an attempted decomposition in the more general sense. Additionally, we consider their application to each time series component *after* applying a TSD model. These decomposition baselines help quantify the efficacy of tailoring CP methods to the exchangeability regimes underlying each component, as opposed to repeatedly applying the same (non-exchangeable) procedures.

Evaluation metrics. We employ standard metrics to assess the two key desiderata of *validity* and *efficiency* for our conformal approaches. We measure validity via the *prediction interval coverage probability* (PICP) – also referred to as empirical coverage – and efficiency via the *prediction interval average width* (PIAW). These are defined as

$$\text{PICP} = \frac{1}{n_t} \sum_{j=1}^{n_t} \mathbb{1}[Y_j \in \hat{C}(X_j)] \quad \text{and} \quad \text{PIAW} = \frac{1}{n_t} \sum_{j=1}^{n_t} |\hat{C}(X_j)|, \tag{5}$$

and metrics are computed equivalently for the overall time series (both in raw and recomposed forms) or specific components using respective prediction intervals[9].

---

8. Our approaches are extendable to batch and online settings, which we do not consider here.
9. By annotating Eq. 5 with component indices T, S or R respectively, which we omit for clarity.
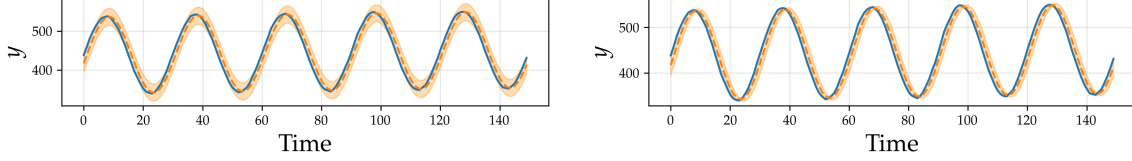
Figure 3: Qualitative results for a segment of the synthetic time series for the EnbPI decomposition baseline (*left*) and our TSD approach with BinaryPoint (*right*). We observe more efficient PIs across all time steps, but in particular at seasonal peaks and troughs.

5.1. Synthetic data: time series with known decomposition structure

We first test our model on a synthetic time series that follows a generative process with known decomposition structure. To this end, we construct a time series with an additive decomposition (in line with Eq. 1) consisting of the following components:

$$T_t = 0.1t, \quad S_t = 100 \cdot \sin\left(2\pi \cdot \frac{t}{30}\right), \quad R_t \sim \mathcal{N}(0,1), \quad t = 1, \ldots, T.$$

In this controlled setting we identify the trend as non-exchangeable, the seasonal term as locally exchangeable, and the remainder as globally exchangeable, thus applying our TSD approaches leveraging component-specific CP methods.

Results for different regressors are displayed in Table 1. Firstly, a comparison of baselines on the raw and decomposed time series suggest that a TSD approach is meaningful, since similarly sized PIs are obtained at much higher coverage. Given the observed overcoverage tendency, a tuning of nominal coverage levels would permit trading said coverage for improved interval widths, as suggested in § 3.4. Our TSD approaches with BinaryLocal and ExpLocal record similar behaviour, but achieve marginally tighter PIs than the decomposition baselines. Finally, BinaryPoint is able to fully exploit the sinusoidal pattern underlying the seasonal term, obtaining much-improved intervals over other TSD approaches and baselines. That is, the underlying assumption on strong local exchangeability is warranted by the data's consistent and accurately retrievable seasonality. This behaviour is highlighted in Fig. 3, where the largest reductions in PI width are observed at the seasonal extrema.

Table 1: Results on synthetic data for a fixed target coverage of 90% ($\alpha = 0.1$). Table entries for our TSD approaches denote different weighting schemes for the seasonal component, while EnbPI and CV+ are used for trend and remainder. Overall best results are underlined, while best decomposition results are **boldened**.

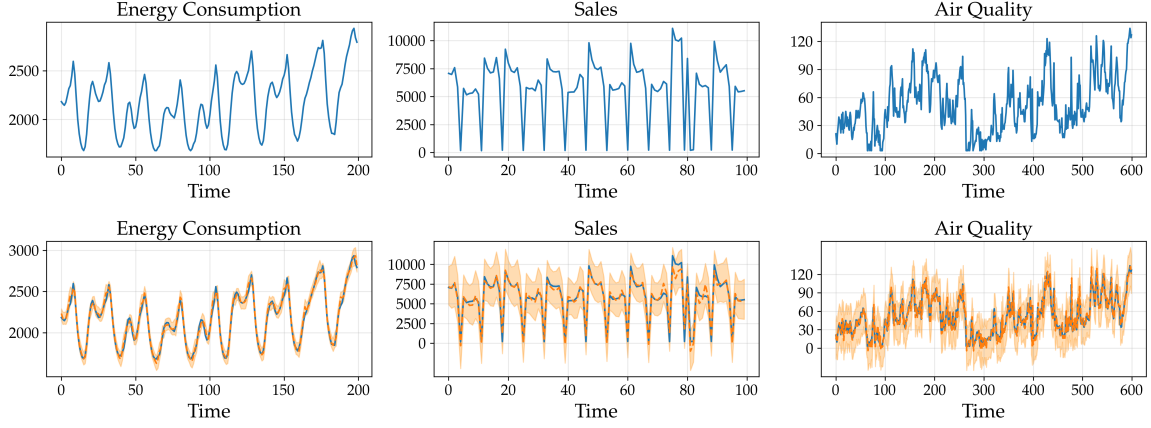| Method	Decomposed	Linear Reg.		MLP Reg.		Gradient Boosting	
		$PICP$	$PIAW$ ($\downarrow$)	$PICP$	$PIAW$ ($\downarrow$)	$PICP$	$PIAW$ ($\downarrow$)
EnbPI	✗	0.889	41.694	0.887	41.626	0.897	11.922
EnbPI	✓	0.985	44.013	0.985	44.197	0.987	12.536
ACI	✗	0.898	42.065	0.898	42.168	0.900	11.991
ACI	✓	0.985	43.995	0.986	44.237	0.994	554.298
BinaryPoint (Ours)	✓	0.946	**29.755**	0.947	**29.771**	0.947	**9.524**
BinaryLocal (Ours)	✓	0.994	41.323	0.994	41.347	0.990	11.942
ExpLocal (Ours)	✓	0.994	44.086	0.993	44.009	0.990	12.510

Figure 4: Segments of the three considered real-world time series from multiple domains. *From left to right*: Different time series complexities are observed as we consider *San Diego energy consumption* with a shifting trend, *Rossman store sales* with varying seasonalities, and *Beijing air quality* exhibiting stronger irregularity and noise. The second row includes prediction intervals leveraging our TSD approaches with the best-performing seasonal weights following Table 2 (*left to right*: BinaryPoint, ExpLocal, ExpLocal).

5.2. Real-world data

We proceed by evaluating our approach on real-world datasets from different domains, of which exemplary segments are shown in Fig. 4. Such real-world time series are generally more volatile and contain less structured regularity. This poses challenges for any TSD model, in particular regarding retrieved seasonal patterns. While a relatively stable seasonality is observed for energy consumption, the sales data exhibits more erratic layered seasons, and the air quality dataset shows no clearly discernible patterns. Thus, the considered time series exhibit varying levels of complexity affecting both prediction and decomposition quality, as corroborated by empirical results. Dataset descriptions and detailed result tables for all three datasets can be found in § B, while qualitative decomposition plots are in § F.

Table 2: Results on real-world data for a fixed target coverage of 90% ($\alpha = 0.1$) and linear regressor. Table entries for our TSD approaches denote different weighting schemes for the seasonal component, while EnbPI and CV+ are used for trend and remainder. Overall best results are <u>underlined</u>, while best decomposition results are **boldened**.

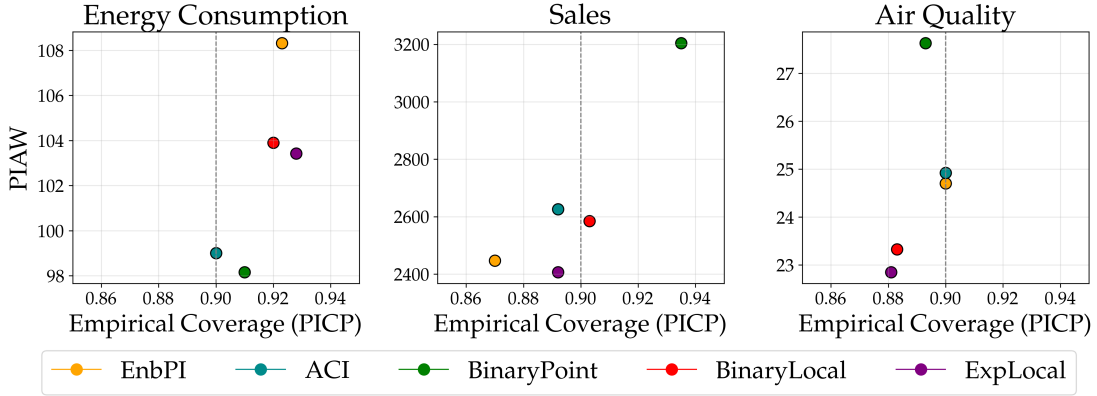| Method	Decomposed	Energy		Sales		Air Quality	
		$PICP$	$PIAW$ ($\downarrow$)	$PICP$	$PIAW$ ($\downarrow$)	$PICP$	$PIAW$ ($\downarrow$)
EnbPI	✗	0.907	222.866	0.892	1572.343	0.902	<u>42.918</u>
EnbPI	✓	0.964	243.005	1.0	5303.825	0.954	62.721
ACI	✗	0.901	<u>218.301</u>	0.895	<u>1539.925</u>	0.901	43.549
ACI	✓	0.960	**229.425**	1.0	5447.390	0.958	63.445
BinaryPoint (Ours)	✓	0.971	234.969	1.0	6045.373	0.958	67.014
BinaryLocal (Ours)	✓	0.967	240.710	1.0	5426.439	0.955	62.716
ExpLocal (Ours)	✓	0.967	240.230	1.0	**5248.193**	0.954	**62.240**

Figure 5: A comparison of decomposition baselines EnbPI and ACI and our weighting mechanisms (§ 3.3) across real-world datasets. Metrics are computed based on obtained intervals for the *seasonal* component only. The dashed line marks target coverage of 90%.

Results for all three datasets are summarized in Table 2. Notably, there is a discrepancy in obtained PI widths between baseline methods applied to the raw and decomposed time series, particularly evident for the sales data. This strongly relates to the practical limitations of an attempted decomposition, in that inaccurately retrieved components propagate down-stream and affect the performance of subsequently applied CP algorithms (see also § 2.1). For example, STL is only capable of retrieving a single seasonal pattern, whereas the sales data might exhibit multiple seasonalities at different frequencies. This limitation also affects the performance of our TSD approaches, preventing them from surpassing time series-specific CP methods used on the raw signals directly. Among decomposition-based approaches, however, leveraging exchangeability regimes does appear to enhance performance for more complex datasets, with ExpLocal achieving best results on sales and air quality data. In contrast to synthetic data, the irregular seasonal patterns in real-world data benefit from softer weighting mechanisms like BinaryLocal and ExpLocal, promoting associated notions of approximate local exchangeability (Definition 2). Overcoverage tendencies caused by our recomposition model in Eq. 4 could be addressed by dataset-specific tuning, *e.g.*, on nominal coverage levels. Particularly for more consistent time series such as energy consumption, such adjustments may help narrow the performance gaps observed.

5.3. Comparison of methods for specific time series components

We now more closely examine and compare parts of our TSD approaches to decomposition baselines for individual components, particularly focusing on the seasonal and remainder terms, which are more susceptible to errors (see § G). This analysis allows us to better understand how assumptions on underlying exchangeability impact the effectiveness of CP methods we employ on real-world data.

Seasonal component. In Fig. 5, we evaluate our weighting mechanisms – BinaryPoint, BinaryLocal, and ExpLocal – against non-exchangeable baselines ACI and EnbPI, assessing the suitability of local exchangeability concepts in handling irregular real-world seasonal patterns. BinaryPoint performs best on energy consumption data, but yields significantly wider
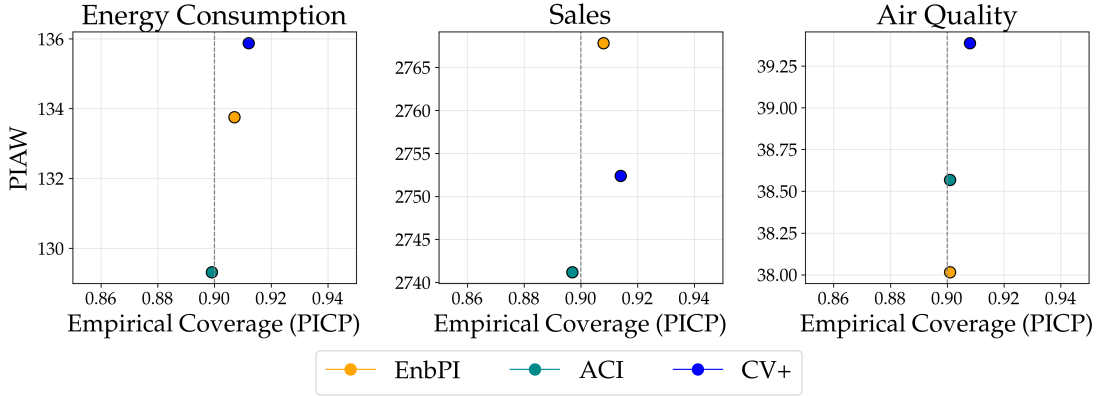
Figure 6: A comparison of decomposition baselines EnbPI and ACI and the CV+ algorithm (Barber et al., 2021) across real-world datasets. Metrics are computed based on obtained intervals for the *remainder* component only. The dashed line marks target coverage of 90%.

PIs for the other datasets, where BinaryLocal and particularly ExpLocal show strong performance. Conversely, EnbPI and ACI generally produce slightly broader PIs but maintain more consistency across datasets. These findings indicate that the effectiveness of employed CP methods is indeed highly dependent on a dataset's ability to meet its exchangeability assumption. Stronger exchangeability assumptions may benefit datasets with consistent seasonal patterns, while more approximate notions better suit complex and variable seasonality. However, it remains unclear whether using locally exchangeable CP methods for seasonal components offers a distinct advantage over generic time series CP methods like EnbPI, which are less dependent on satisfying such specific assumptions.

Remainder component. As discussed in § 3.1, under an effective TSD model the remainder should ideally resemble uncorrelated 'white noise'. Consequently, we investigate whether a method like CV+, which assumes global exchangeability, is practically suitable, or if non-exchangeable approaches that recognize temporal dependencies might be more appropriate. Our results, depicted in Fig. 6, confirm that all CP methods reach the intended coverage levels and provide similar prediction interval quality, suggesting that global exchangeability can indeed be sufficient for modelling the remainder term.

6. Conclusion

In this work, we investigate the use of conformal prediction methods for time series forecasting through the framework of time series decomposition. We examine both conceptually and empirically the relationship between the components obtained from such a decomposition, and the applicability of different exchangeability regimes that these components may adhere to. By leveraging these exchangeability regimes, we apply specific CP algorithms to each component and combine them to construct prediction intervals that empirically meet conformal guarantees for the overall time series. Our experiments on both synthetic and real-world data indicate that while promising, the effectiveness of obtained uncertainty estimates largely depends on the complexity of the time series, the quality of the decomposi-

tion, and other methodological choices. Consequently, our approaches appear most effective for 'well-behaved' time series that exhibit high regularity, or when prior knowledge about the data is available to guide modelling decisions. In such settings, closely aligning the employed CP methods with specific exchangeability regimes can be highly beneficial.

Currently, our exploratory investigation focused on an additive decomposition model using the STL algorithm. Future work should address the robustness of our results by *(i)* evaluating more advanced time series decomposition algorithms such as MSTL (Bandara et al., 2021), RobustSTL (Wen et al., 2019), or OnlineSTL (Mishra et al., 2022), which may offer more flexible modelling capabilities; *(ii)* considering different de- and recomposition models, such as for multiplicative structures; and *(iii)* exploring more advanced seasonality weighting schemes, such as adaptive weights that respond to component behaviour, or draw inspiration from local exchangeability conditions seen in *mondrian* CP methods (Toccaceli and Gammerman, 2019). Additionally, open questions remain related to nominal coverage guarantees, and quantifying the impact of decomposition quality on such guarantees.

By illustrating the potential synergies between conformal prediction and classical time series decomposition, we hope to encourage further research at their intersection, highlighting the critical role of thoughtfully assessing underlying exchangeability assumptions.

References

Salim Amoukou and Nicolas Brunel. Adaptive conformal prediction by reweighting non-conformity score. *arXiv preprint arXiv:2303.12695*, 2023.

Anastasios Angelopoulos, Emmanuel Candès, and Ryan J Tibshirani. Conformal PID control for time series prediction. *Advances in Neural Information Processing Systems*, 2024.

Vassilis Assimakopoulos and Konstantinos Nikolopoulos. The Theta model: A decomposition approach to forecasting. *International Journal of Forecasting*, 2000.

Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with Modern Hopfield Networks. *Advances in Neural Information Processing Systems*, 2023.

Kasun Bandara, Rob J Hyndman, and Christoph Bergmeir. MSTL: A seasonal-trend decomposition algorithm for time series with multiple seasonal patterns. *International Journal of Operational Research*, 2021.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 2021.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 2023.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for Outliers with Conformal p-values. *The Annals of Statistics*, 2023.

Pietro Bonizzi, Joël MH Karel, Olivier Meste, and Ralf LM Peeters. Singular spectrum decomposition: A new method for time series decomposition. *Advances in Adaptive Data Analysis*, 2014.

Emmanuel Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2023.

Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 2021.

Robert B Cleveland, William S Cleveland, Jean E McRae, and Irma Terpenning. STL: A seasonal-trend decomposition. *Journal of Official Statistics*, 1990.

William P Cleveland and George C Tiao. Decomposition of seasonal time series: A model for the census X-11 program. *Journal of the American Statistical Association*, 1976.

William S Cleveland and Susan J Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 1988.

Estela Bee Dagum and Silvia Bianconcini. *Seasonal adjustment methods and real time trend-cycle estimation*. Springer, 2016.

Alysha M De Livera, Rob J Hyndman, and Ralph D Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 2011.

Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 2017.

Alexander Dokumentov and Rob J Hyndman. STR: Seasonal-trend decomposition using regression. *INFORMS Journal on Data Science*, 2022.

Philippe Esling and Carlos Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 2012.

Clara Fannjiang, Stephen Bates, Anastasios N Angelopoulos, Jennifer Listgarten, and Michael I Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 2022.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 2021.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 2023.

Laszlo Gyôrfi and Harro Walk. Nearest neighbor based conformal prediction. *Annales de l'ISUP*, 2019.

Xing Han, Ziyang Tang, Joydeep Ghosh, and Qiang Liu. Split localized conformal prediction. *arXiv preprint arXiv:2206.13092*, 2022.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: Data mining, inference, and prediction.* Springer, 2009.

Renfei He, Limao Zhang, and Alvin Wei Ze Chew. Modeling and predicting rainfall time series using seasonal-trend decomposition and machine learning. *Knowledge-Based Systems*, 2022.

Xiao He, Ye Li, Jian Tan, Bin Wu, and Feifei Li. OneShotSTL: One-Shot seasonal-trend decomposition for online time series anomaly detection and forecasting. *Proceedings of the VLDB Endowment*, 2023.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: Randomization enables local guarantees. *arXiv preprint arXiv:2310.07850*, 2023.

Yonghua Huo, Yu Yan, Dan Du, Zhihao Wang, Yixin Zhang, and Yang Yang. Long-term span traffic prediction model based on STL decomposition and LSTM. *Asia-Pacific Network Operations and Management Symposium (APNOMS)*, 2019.

Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice.* OTexts, 2018.

Vilde Jensen, Filippo Maria Bianchi, and Stian Normann Anfinsen. Ensemble conformalized quantile regression for probabilistic time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Song Jiang, Tahin Syed, Xuan Zhu, Joshua Levy, Boris Aronchik, and Yizhou Sun. Bridging self-attention and time series decomposition for periodic forecasting. *ACM International Conference on Information & Knowledge Management*, 2022.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *International Conference on Learning Representations*, 2023.

Byol Kim, Chen Xu, and Rina Barber. Predictive inference is free with the Jackknife+-after-bootstrap. *Advances in Neural Information Processing Systems*, 2020.

Tim Krake, Daniel Klötzl, David Hägele, and Daniel Weiskopf. Uncertainty-aware seasonal-trend decomposition based on Loess. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2021.

Wenlin Li and Xuchu Jiang. Prediction of air pollutant concentrations based on TCN-BiLSTM-DMAttention with STL decomposition. *Scientific Reports*, 2023.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A*, 2021.

Yang Lin, Irena Koprinska, and Mashud Rana. SSDNet: State space decomposition neural network for time series forecasting. *IEEE International Conference on Data Mining*, 2021.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Conformal prediction with temporal quantile adjustments. *Advances in Neural Information Processing Systems*, 2022.

Frederick R Macaulay. *The smoothing of time series*. National Bureau of Economic Research New York, 1931.

Spyros Makridakis and Nikolas Bakas. Forecasting and uncertainty: A survey. *Risk and Decision Analysis*, 2016.

Nhlanhla Mbuli, Malusi Mathonsi, Modisane Seitshiro, and Jan-Harm C Pretorius. Decomposition forecasting methods: A review of applications in power systems. *Energy Reports*, 2020.

AG Miamee and M Pourahmadi. Wold decomposition, prediction and parameterization of stationary processes with infinite variance. *Probability Theory and Related Fields*, 1988.

Abhinav Mishra, Ram Sriharsha, and Sichen Zhong. OnlineSTL: Scaling time series decomposition by 100x. *Proceedings of the VLDB Endowment*, 2022.

Jason Osborne. Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation*, 2010.

Lace MK Padilla, Maia Powell, Matthew Kay, and Jessica Hullman. Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. *Frontiers in Psychology*, 2021.

Harris Papadopoulos, Volodya Vovk, and Alex Gammerman. Conformal prediction with Neural Networks. *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2007.

Emanuel Parzen. *Time series analysis for models of signal plus white noise*. Department of Statistics, Stanford University, 1966.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *the Journal of Machine Learning Research*, 2011.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. *Uncertainty in Artificial Intelligence*, 2021.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 2019.

Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review 2005–2019. *Applied Soft Computing*, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 2008.

Skander Soltani. On the use of the wavelet decomposition for time series prediction. *Neurocomputing*, 2002.

Kamile Stankeviciute, Ahmed M Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. *Advances in Neural Information Processing Systems*, 2021.

Sophia Huiwen Sun and Rose Yu. Copula conformal prediction for multi-step time series prediction. *International Conference on Learning Representations*, 2023.

Vianney Taquet, Vincent Blot, Thomas Morzadec, Louis Lacombe, and Nicolas Brunel. MAPIE: An open-source library for distribution-free uncertainty quantification. *arXiv preprint arXiv:2207.12274*, 2022.

Marina Theodosiou. Forecasting monthly and quarterly time series using STL decomposition. *International Journal of Forecasting*, 2011.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 2019.

Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, and Eric Nalisnick. A powerful rank-based correction to multiple testing under positive dependency. *arXiv Preprint (arXiv:2311.10900)*, 2023.

Paolo Toccaceli and Alexander Gammerman. Combination of inductive mondrian conformal predictors. *Machine Learning*, 2019.

Oscar Trull, J Carlos García-Díaz, and Angel Peiró-Signes. Multiple seasonal STL decomposition with discrete-interval moving seasonalities. *Applied Mathematics and Computation*, 2022.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 2015.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world.* Springer, 2005.

Zhiyuan Wang, Xovee Xu, Weifeng Zhang, Goce Trajcevski, Ting Zhong, and Fan Zhou. Learning latent seasonal-trend representations for time series forecasting. *Advances in Neural Information Processing Systems*, 2022.

Qingsong Wen, Jingkun Gao, Xiaomin Song, Liang Sun, Huan Xu, and Shenghuo Zhu. RobustSTL: A robust seasonal-trend decomposition algorithm for long time series. *AAAI Conference on Artificial Intelligence*, 2019.

Qingsong Wen, Zhe Zhang, Yan Li, and Liang Sun. Fast RobustSTL: Efficient and robust seasonal-trend decomposition for time series with complex patterns. *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *International Joint Conference on Artificial Intelligence*, 2023.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 2021.

Zhaohua Wu, Norden E Huang, Steven R Long, and Chung-Kang Peng. On the trend, detrending, and variability of nonlinear and nonstationary time series. *Proceedings of the National Academy of Sciences*, 2007.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. *International Conference on Machine Learning*, 2021.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. *International Conference on Machine Learning*, 2023.

Chen Xu, Hanyang Jiang, and Yao Xie. Conformal prediction for multi-dimensional time series by ellipsoidal sets. *International Conference on Machine Learning*, 2024.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. *International Conference on Machine Learning*, 2022.

S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, and S.X. Chen. Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A*, 2017.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. *International Conference on Machine Learning*, 2022.

Appendix

The appendix contains additional empirical results and is organized as follows:

Appendix A. Coverage guarantee for the recomposed prediction interval.

Consider the time series decomposition from Eq. 1 and the recomposed prediction interval obtained by applying Eq. 4. Assume that the conformal methods applied to each component (trend T, seasonality S and remainder R) provide finite-sample, nominal coverage at target level $1 - \alpha$, which we select to be identical for all components ($\alpha_T = \alpha_S = \alpha_R = \alpha$). Then a lower bound on nominal coverage for the recomposed prediction interval $\hat{C}(X_{n+1})$ can be obtained by the union bound as

$$
\begin{aligned}
\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})) \\
&= \mathbb{P}(T_{n+1} \in \hat{C}_T(X_{n+1}) \wedge S_{n+1} \in \hat{C}_S(X_{n+1}) \wedge R_{n+1} \in \hat{C}_R(X_{n+1})) \\
&= 1 - \mathbb{P}(T_{n+1} \notin \hat{C}_T(X_{n+1}) \vee S_{n+1} \notin \hat{C}_S(X_{n+1}) \vee R_{n+1} \notin \hat{C}_R(X_{n+1})) \\
&\geq 1 - \Big[\underbrace{\mathbb{P}(T_{n+1} \notin \hat{C}_T(X_{n+1}))}_{\leq \alpha_T} + \underbrace{\mathbb{P}(S_{n+1} \notin \hat{C}_S(X_{n+1}))}_{\leq \alpha_S} + \underbrace{\mathbb{P}(R_{n+1} \notin \hat{C}_R(X_{n+1}))}_{\leq \alpha_R} \Big] \\
&\geq 1 - 3\,\alpha.
\end{aligned}
\tag{6}
$$

The given bound is loose and does not guarantee target coverage of $1 - \alpha$ for $\hat{C}(X_{n+1})$, an effect that can also be interpreted from a *multiple hypothesis testing* perspective on the conformal procedures applied to each component (see, *e.g.*, Bates et al. (2023)). A naive Bonferroni correction can be achieved by selecting $\alpha/3$ as target level per component, but other corrections for the conformal setting have been proposed (Sun and Yu, 2023; Timans et al., 2023). Since our decomposition approach already produces empirically conservative coverage levels and consists of a combination of different guarantees, we do not evaluate such corrections here.

Appendix B. Detailed results tables.

We present detailed results tables for our considered methods both on synthetic and real-world experiments. In particular, we report results for a wide range of possible combinations of considered CP methods across decomposition components. That is, we consider combinations of the options {EnbPI, ACI} for the trend T, {EnbPI, ACI, BinaryPoint, BinaryLocal, ExpLocal} for the season S, and {EnbPI, ACI, CV+} for the remainder R. Results are structured by dataset in individual tables. We additionally provide brief dataset descriptions. Overall best results are underlined, while best decomposition results are **boldened**.

B.1. Synthetic data.

Conformal algorithm			Linear Reg.		MLP		Gradient Boost.	
Trend	Season	Remainder	PICP	PIAW	PICP	PIAW	PICP	PIAW
	Original (EnbPI)		0.889	41.694	0.887	41.626	0.897	11.922
	Original (ACI)		0.898	42.065	0.898	42.168	0.9	11.991
	Original (CV+)		0.879	41.645	0.917	42.41	0.904	12.031
EnbPI	EnbPI	EnbPI	0.985	44.013	0.985	44.197	0.987	12.536
EnbPI	BinaryPoint	EnbPI	0.944	29.764	**0.946**	**29.768**	**0.948**	**9.494**
EnbPI	BinaryLocal	EnbPI	0.994	41.332	0.994	41.344	0.99	11.913
EnbPI	ExpLocal	EnbPI	0.994	44.096	0.993	44.005	0.99	12.481
EnbPI	EnbPI	CV+	0.985	44.003	0.985	44.201	0.988	12.566
EnbPI	BinaryPoint	CV+	**0.946**	**29.755**	0.947	29.771	0.947	9.524
EnbPI	BinaryLocal	CV+	0.994	41.323	0.994	41.347	0.99	11.942
EnbPI	ExpLocal	CV+	0.994	44.086	0.993	44.009	0.99	12.510
ACI	ACI	ACI	0.985	43.995	0.986	44.237	0.994	554.298
ACI	BinaryPoint	ACI	0.946	29.793	0.946	29.822	0.969	551.304
ACI	BinaryLocal	ACI	0.995	41.361	0.995	41.398	0.996	553.722
ACI	ExpLocal	ACI	0.994	44.125	0.994	44.06	0.996	554.29
ACI	ACI	CV+	0.984	43.956	0.985	44.186	0.993	554.321
ACI	BinaryPoint	CV+	**0.946**	**29.755**	0.947	29.771	0.968	551.326
ACI	BinaryLocal	CV+	0.994	41.323	0.994	41.347	0.995	553.744
ACI	ExpLocal	CV+	0.994	44.086	0.993	44.009	0.995	554.312

Table 3: Detailed results for synthetic data.

B.2. Real-world data: San Diego energy consumption.

This dataset records energy usage in the city of San Diego, collected over a period of five years. The response consists of hourly energy consumption measurements in Watt hours. Available time-dependent features include seasonal indicators, public holidays, temperature, and air conditioning use, among others.

Conformal algorithm			Linear Reg.		MLP		Gradient Boost.	
Trend	Season	Remainder	PICP	PIAW	PICP	PIAW	PICP	PIAW
	Original (EnbPI)		0.907	222.866	0.865	161.238	0.899	199.334
	Original (ACI)		<u>0.9</u>	<u>218.301</u>	0.9	187.367	0.9	204.007
	Original (CV+)		0.912	227.846	0.898	168.878	<u>0.9</u>	<u>198.872</u>
EnbPI	EnbPI	EnbPI	0.964	243.005	**0.944**	**161.773**	0.966	271.898
EnbPI	BinaryPoint	EnbPI	0.97	232.852	0.961	172.657	0.973	272.751
EnbPI	BinaryLocal	EnbPI	0.968	238.593	0.963	177.651	0.977	281.453
EnbPI	ExpLocal	EnbPI	0.968	238.113	0.963	176.555	0.977	281.194
EnbPI	EnbPI	CV+	0.963	245.122	0.956	170.989	0.967	273.554
EnbPI	BinaryPoint	CV+	0.971	234.969	0.972	181.872	0.975	274.407
EnbPI	BinaryLocal	CV+	0.967	240.71	0.974	186.867	0.977	283.108
EnbPI	ExpLocal	CV+	0.967	240.23	0.974	185.77	0.977	282.849
ACI	ACI	ACI	0.96	229.425	0.966	190.995	0.968	278.927
ACI	BinaryPoint	ACI	**0.969**	**228.552**	0.969	181.361	**0.97**	**270.637**
ACI	BinaryLocal	ACI	0.967	234.293	0.97	186.355	0.974	279.338
ACI	ExpLocal	ACI	0.968	233.814	0.97	185.259	0.973	279.07
ACI	ACI	CV+	0.962	235.855	0.971	191.174	0.972	286.686
ACI	BinaryPoint	CV+	0.971	234.983	0.971	181.54	0.974	278.396
ACI	BinaryLocal	CV+	0.967	240.724	0.973	186.534	0.977	287.097
ACI	ExpLocal	CV+	0.967	240.245	0.973	185.438	0.977	286.838

Table 4: Detailed results for San Diego energy consumption data.

B.3. Real-world data: Rossman store sales.

This dataset contains aggregated daily sales from the chain of Rossmann drug stores. The response consists of average daily sales numbers. Available time-dependent features include weekday, average number of customers, and whether holiday indicators.

Conformal algorithm			Linear Reg.		MLP		Gradient Boost.	
Trend	Season	Remainder	PICP	PIAW	PICP	PIAW	PICP	PIAW
	Original (EnbPI)		0.892	1572.343	0.923	1231.569	0.932	5476.084
	Original (ACI)		0.895	1539.925	0.906	1172.582	0.919	4867.123
	Original (CV+)		0.892	1512.231	0.928	1276.386	0.924	5470.683
EnbPI	EnbPI	EnbPI	1.0	5303.825	0.975	4870.155	**0.932**	**6075.189**
EnbPI	BinaryPoint	EnbPI	1.0	6060.792	0.989	6404.697	0.978	8051.558
EnbPI	BinaryLocal	EnbPI	1.0	5441.858	0.984	5553.258	0.973	7547.193
EnbPI	ExpLocal	EnbPI	1.0	5263.612	0.984	5381.6	0.658	7373.728
EnbPI	EnbPI	CV+	1.0	5288.406	0.986	5297.233	0.935	6284.836
EnbPI	BinaryPoint	CV+	1.0	6045.373	1.0	6831.775	0.98	8261.205
EnbPI	BinaryLocal	CV+	1.0	5426.439	0.996	5980.336	0.968	7756.839
EnbPI	ExpLocal	CV+	1.0	5248.193	0.996	5808.679	0.976	7583.375
ACI	ACI	ACI	1.0	5447.39	**0.984**	**4615.531**	0.944	6373.45
ACI	BinaryPoint	ACI	1.0	6015.128	0.993	6313.72	0.977	8003.03
ACI	BinaryLocal	ACI	1.0	5396.194	0.987	5462.281	0.973	7498.665
ACI	ExpLocal	ACI	**1.0**	**5217.948**	0.974	5290.623	0.966	7325.201
ACI	ACI	CV+	1.0	5477.696	0.989	5201.327	0.944	6623.43
ACI	BinaryPoint	CV+	1.0	6045.434	1.0	6899.516	0.98	8253.01
ACI	BinaryLocal	CV+	1.0	5426.5	0.996	6048.076	0.968	7748.645
ACI	ExpLocal	CV+	1.0	5248.254	0.996	5876.419	0.968	7575.181

Table 5: Detailed results for Rossman store sales data.

B.4. Real-world data: Beijing air quality.

This dataset consists of hourly readings of air pollutants and meteorological data from the city of Beijing. The primary response is the concentration of PM2.5 pollutants in the air. Available time-dependent features include various pollutant concentrations (*e.g.*, SO2, NO2, CO, O3); air temperature and pressure; or rainfall, wind direction and speed.

Conformal algorithm			Linear Reg.		MLP		Gradient Boost.	
Trend	Season	Remainder	PICP	PIAW	PICP	PIAW	PICP	PIAW
	Original (EnbPI)		0.902	42.918	0.892	40.098	0.905	44.889
	Original (ACI)		0.901	43.549	0.901	44.733	0.901	46.435
	Original (CV+)		0.908	44.499	0.915	44.398	0.913	46.909
EnbPI	EnbPI	EnbPI	0.954	62.721	0.955	64.756	0.951	64.74
EnbPI	BinaryPoint	EnbPI	0.956	65.643	0.959	68.696	0.952	68.127
EnbPI	BinaryLocal	EnbPI	0.953	61.345	0.956	64.376	0.948	63.489
EnbPI	ExpLocal	EnbPI	**0.953**	**60.869**	**0.956**	**63.69**	**0.964**	**62.882**
EnbPI	EnbPI	CV+	0.956	64.092	0.962	67.586	0.954	66.537
EnbPI	BinaryPoint	CV+	0.958	67.014	0.963	71.526	0.954	69.924
EnbPI	BinaryLocal	CV+	0.955	62.716	0.96	67.206	0.951	65.286
EnbPI	ExpLocal	CV+	0.954	62.24	0.96	66.52	0.950	64.68
ACI	ACI	ACI	0.958	63.445	0.961	67.267	0.957	74.507
ACI	BinaryPoint	ACI	0.962	66.195	0.965	70.108	0.96	76.922
ACI	BinaryLocal	ACI	0.957	61.897	0.961	65.789	0.958	72.284
ACI	ExpLocal	ACI	0.957	61.421	0.961	65.103	0.976	71.678
ACI	ACI	CV+	0.958	64.264	0.965	68.233	0.961	74.603
ACI	BinaryPoint	CV+	0.958	67.014	0.963	71.074	0.956	77.019
ACI	BinaryLocal	CV+	0.955	62.716	0.96	66.755	0.953	72.381
ACI	ExpLocal	CV+	0.954	62.24	0.959	66.068	0.953	71.774

Table 6: Detailed results for Beijing air quality data.

Appendix C. Results across varying coverage levels.

We validate our results from Table 1 and Table 2 across varying nominal coverage levels $(1 - \alpha) \in \{0.75, 0.8, 0.85, 0.9, 0.95\}$ in Fig. 7. Once again we consider the linear regressor with CP methods EnbPI and CV+ for trend and remainder terms, and vary our approaches for the seasonal term. We continue to observe that overcoverage persists across all coverage levels and datasets. For synthetic data, our weighting schemes (in particular BinaryPoint) consistently outperform other methods. For real-world data, methods show similar performance across coverage levels, with relatively small deviations. The general trends in terms of empirical coverage and interval width across nominal coverage levels are in line with expectations, *i.e.*, higher coverage requirements result in both larger PICP and PIAW.
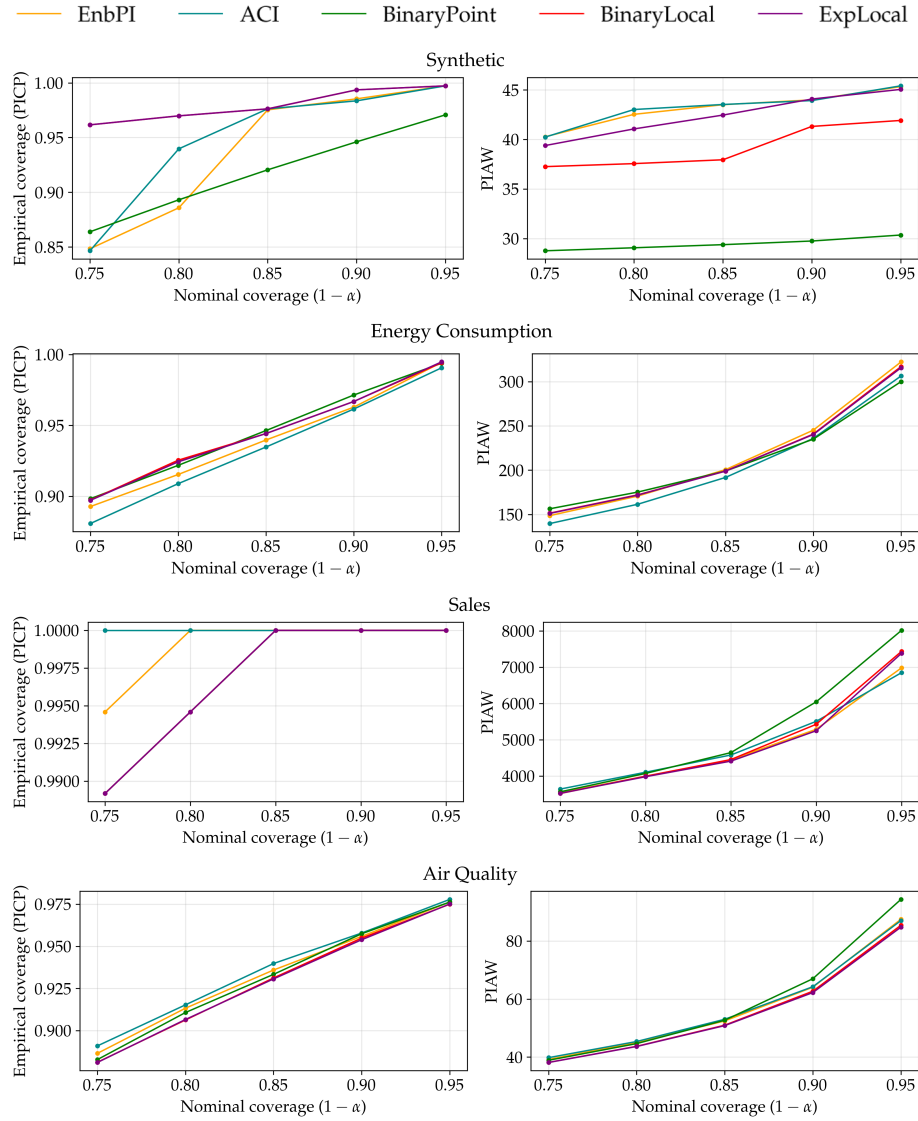

Figure 7: Results for varying seasonal CP methods across nominal coverage levels.

Appendix D. Varying the number of selected periods based on recency.

We conduct an ablation study to assess how results are affected by calibration set sizes in terms of temporal proximity in Fig. 8. That is, for a given seasonal weighting mechanism (BinaryPoint, BinaryLocal or ExpLocal) we consider only leveraging a subset of the most recent periods in time, inspired by the recency-based weighting schemes employed in Barber et al. (2023). We observe that while BinaryPoint experiences generally narrow PIs when fewer periods are used, its empirical coverage also decreases below the target level of 90%. In contrast, both BinaryLocal and ExpLocal maintain coverage closer to the desired level throughout, whilst occasionally displaying more efficient PIs. Overall, the results across datasets are mixed, but indicate that temporal proximity can indeed play a role in further reducing interval sizes whilst guaranteeing target coverage. Since slow-moving trend or seasonality shifts might occur over long periods of time, discarding time-distant observations can be sensible.
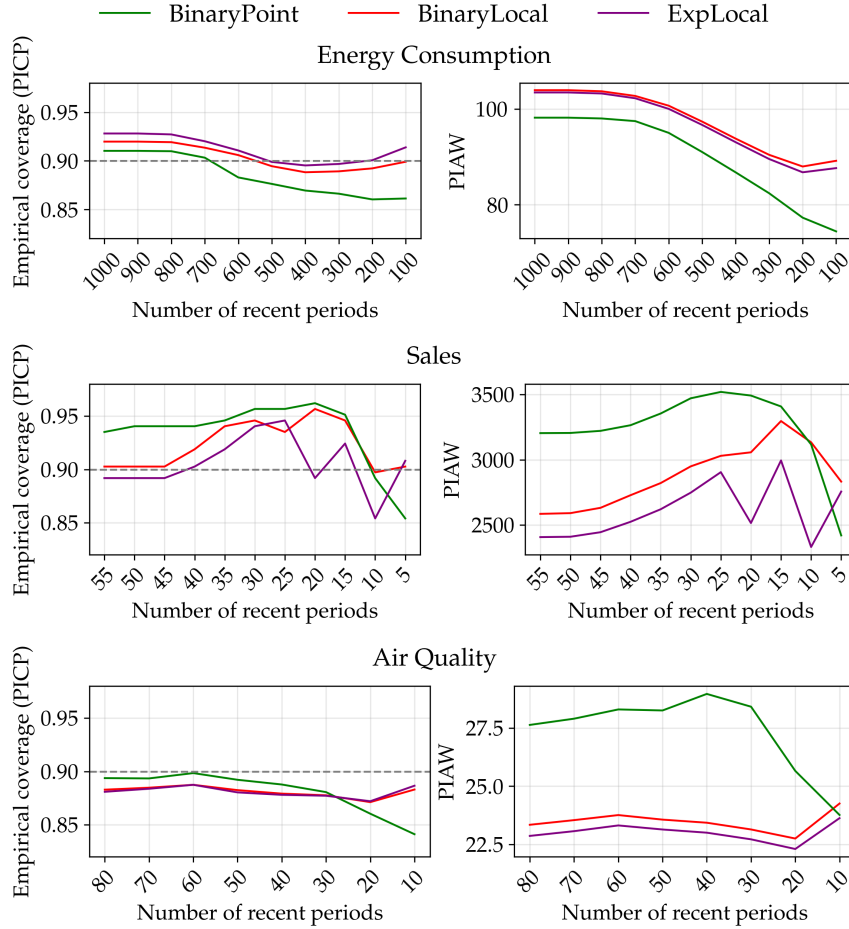

Figure 8: Ablation on temporal proximity of periods on real-world data.

Appendix E. Seasonal weighting mechanisms based on feature distance.

Related to the approach by Guan (2023), we also consider weighting schemes for the seasonal component which are data-dependent and based on feature distances. We consider two approaches which weigh calibration samples proportional to the euclidean feature distance from the test sample, one with clustering (**KNN**) and one without (**FeatDistPoint**). Our ablation is performed on the air quality dataset, which is well-suited for feature-based methods due to its large set of natural features (beyond lags).

KNN. For the first approach we employ the k-nearest neighbours algorithm (KNN), computing the feature distances $d_i := \|X_{n+1} - X_i\|_2 \; \forall i \in \mathcal{D}_{cal}$ and assigning the k-th nearest neighbours an unnormalized weight of one, while any other calibration samples are discarded. This provides a feature-based binary weighting scheme, similar to BinaryPoint. We investigate the effect of choosing the neighbourhood size k in Fig. 9, observing a mixed impact on coverage but a reduction in interval size when only about 50% of available samples ($k = 4000$) are selected.

FeatDistPoint. Our second approach provides a softer weighting scheme similar to ExpLocal, wherein each calibration sample is assigned a weight inversely proportional to its distance (*i.e.*, $\tilde{w}_i = 1/d_i$), ensuring that samples closer to X_{n+1} have larger weights in Eq. 2.

We compare our feature-based weighting schemes against EnbPI, ACI, and approaches based on temporal distance (§ 3.3) in Fig. 10 (we select $k = 7000$ for KNN for maximal coverage). Feature-based weights perform similarly well to other approaches, suggesting that leveraging the feature domain is a viable option. However, this may come at the cost of potentially more involved computations in its generally higher dimensionality.
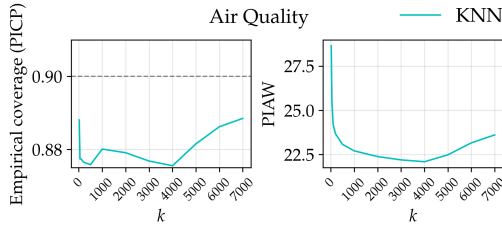

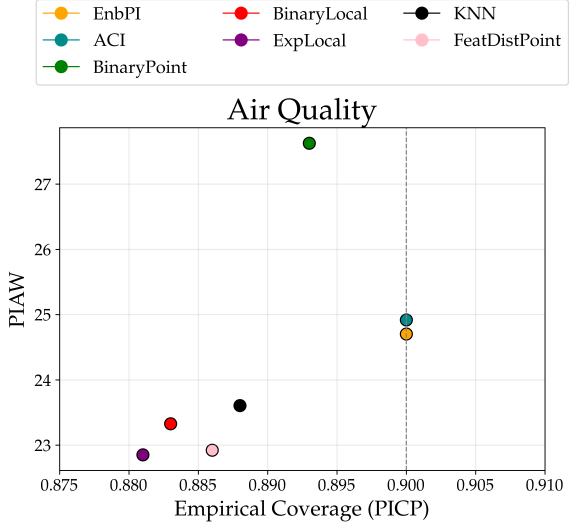Figure 9: The effect of varying the neighbourhood size k for KNN.

Figure 10: Comparison of CP methods for the *seasonal component* on Air Quality data.

Appendix F. Qualitative decomposition plots across datasets.

Here, we display qualitative decomposition plots for every dataset: synthetic (Fig. 11), San Diego energy consumption (Fig. 12), Rossman store sales (Fig. 13) and Beijing air quality (Fig. 14). Specifically, we overlay an exemplary section of the original time series (*blue*) with the conformal prediction intervals produced via our TSD approaches for every component (trend, season and remainder) as well as for the recomposition (Eq. 4). We display these plots three times, once for every seasonal weighting scheme (BinaryPoint, BinaryLocal, ExpLocal; see § 3.3). EnbPI and CV+ are used for the trend and remainder terms, respectively. We add some comments and observations below.

Synthetic data. Similarly to Fig. 3 we observe tight PIs for the seasonal component using BinaryPoint. Conversely, wider intervals are observed for BinaryLocal and ExpLocal. The linear regressor correctly predicts zero mean for the $\mathcal{N}(0, 1)$ remainder component, producing a constant interval. The seasonal component clearly dominates the pattern of the time series, as also visible in the recomposition.

San Diego energy consumption. A somewhat periodic trend is revealed, hinting at potentially multiple seasonalities of different time lengths, which could be investigated for example with MSTL (Bandara et al., 2021) (but not the employed STL). For the seasonal patterns, wider intervals are observed at the extrema, which aim to compensate for the larger irregularities.

Rossmann store sales. We observe complex trend and seasonal patterns which are associated with larger errors. The remainder component exhibits significant drops on certain days due to holidays without sales. As such, the remainder displays substantial volatility. Since the regressor tends to underpredict in these instances, it results in the generation of notably wide intervals for this component, contributing to the wide intervals and empirical overcoverage observed in the recomposition results shown in Table 2.

Beijing air quality. Here, we observe a notable difference between BinaryPoint and the other methods, BinaryLocal and ExpLocal, where the PIs of the former are characterized by spikes. It is hard to identify a consistent seasonal pattern, and season and remainder terms exhibit large similarity, suggesting that improvements on the decomposition quality can be made.
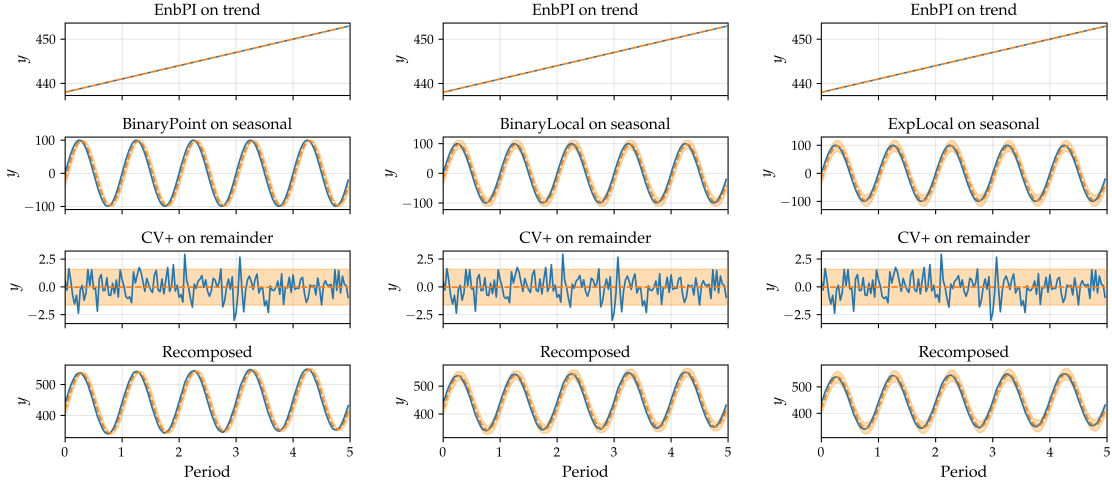
Figure 11: Time series and conformal prediction intervals per component and for the re-composition on **synthetic** data. *From left to right*: The seasonal component leverages BinaryPoint, BinaryLocal or ExpLocal.
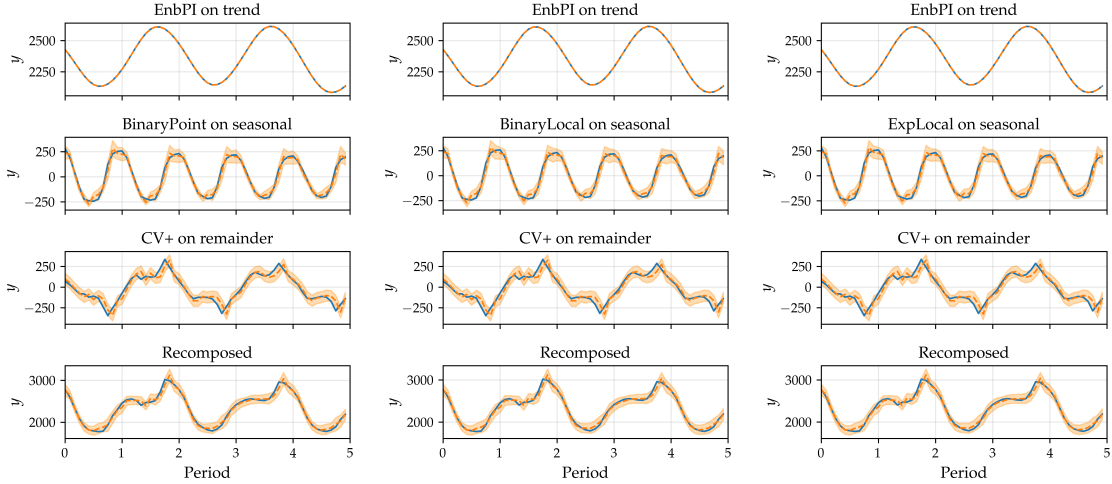

Figure 12: Time series and conformal prediction intervals per component and for the re-composition on **San Diego energy consumption** data. *From left to right*: The seasonal component leverages BinaryPoint, BinaryLocal or ExpLocal.
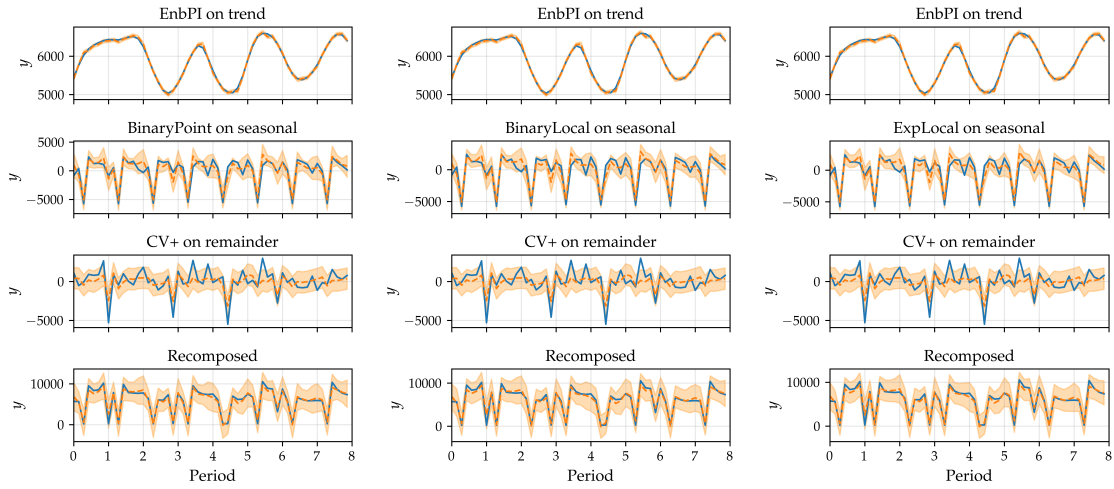
Figure 13: Time series and conformal prediction intervals per component and for the recomposition on **Rossmann store sales** data. *From left to right*: The seasonal component leverages BinaryPoint, BinaryLocal or ExpLocal.
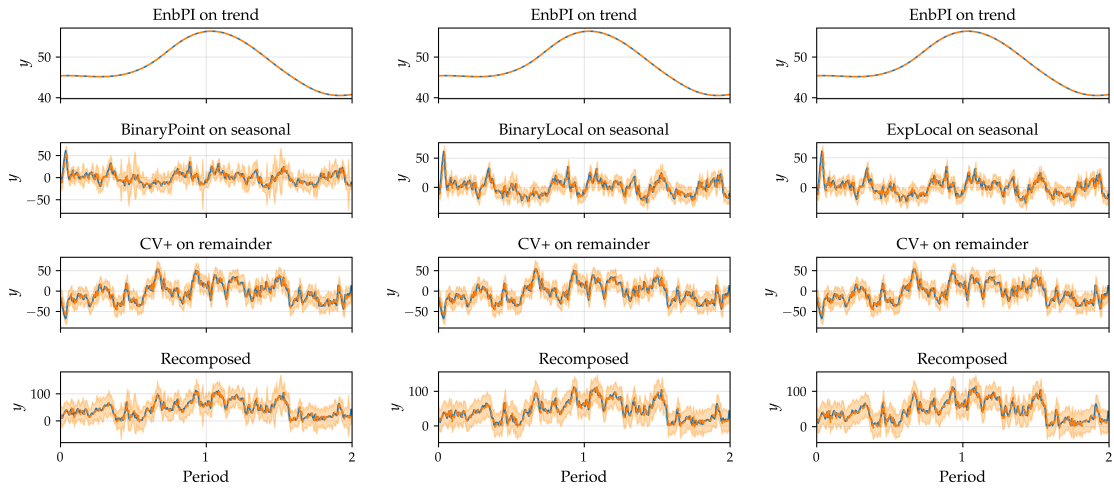

Figure 14: Time series and conformal prediction intervals per component and for the recomposition on **Beijing air quality** data. *From left to right*: The seasonal component leverages BinaryPoint, BinaryLocal or ExpLocal.

Appendix G. Details on regression models.

Predictive performance. We evaluate the three underlying regressors (Linear, MLP and Gradient Boosting) via the *root mean squared error* (RMSE), and report values in Table 7. Overall the trend is generally easy to predict, while the more irregular seasonal and remainder terms record higher errors. The linear and MLP regressors display similar results across datasets except on the sales data, where the MLP outperforms the linear model. Interestingly, this performance gap does not translate into lower PIAW scores when evaluating CP methods (see § B.1). Gradient boosting outperforms the other regressors on synthetic data due to its small sequential differences, but performs notably worse on the sales data, where abrupt changes in sales lead to large differences between time steps.

Parameter settings. We also display hyperparameter settings for employed conformal algorithms and regression models in Table 8. We do not perform extensive tuning and employ many default settings. The decay rate parameter $\lambda \in (0, 1)$ for ExpLocal depends on each dataset since the period length varies across datasets. It is selected such that weights approach zero for data points that are separated by a full period length.

Component	Linear Reg.	MLP	Gradient Boost.
Synthetic data			
Original	14.903	14.898	3.855
Trend	0.0	0.002	0.003
Seasonal	14.702	14.703	3.129
Noise	1.001	1.003	1.242
San Diego energy consumption			
Original	63.207	51.818	60.746
Trend	1.336	3.443	2.552
Seasonal	28.252	23.098	34.983
Noise	38.903	32.755	45.179
Rossmann store sales			
Original	539.921	359.879	1512.717
Trend	28.003	66.599	41.372
Seasonal	782.387	465.386	983.724
Noise	897.016	758.889	1204.878
Beijing air quality			
Original	16.86	17.226	18.296
Trend	0.001	0.877	0.046
Seasonal	8.735	8.753	8.994
Noise	14.204	14.154	15.001

Table 7: RMSE for various datasets across different components and models.

CP algo.	Param.	Value
EnbPI	# bootstrap samples	20
	window length	1
	window overlap	False
	seed	42
ACI	γ	0.01
CV+	k	20
BinaryPoint	use region	False
	λ	1
BinaryLocal	use region	True
	λ	1
ExpLocal	use region	True
	λ	$\in (0, 1)$
Regressor	**Param.**	**Value**
Linear Reg.	–	–
MLP	hidden layers	$(100, 100)$
	activation	ReLU
	solver	Adam
	LR init	0.001
	LR	constant
	max iter	100
	early stop	True
Gradient Boost.	Mean loss	MSE
	Quantile loss	Pinball
	# estimators	100

Table 8: Hyperparameter settings for different conformal algorithms and regressors.