Transformer Conformal Prediction for Time Series

Junghwan Lee [1] **Chen Xu** [1] **Yao Xie** [1]

Abstract

We present a conformal prediction method for time series using the Transformer architecture to capture long-memory and long-range dependencies. Specifically, we use the Transformer decoder as a conditional quantile estimator to predict the quantiles of prediction residuals, which are used to estimate the prediction interval. We hypothesize that the Transformer decoder benefits the estimation of the prediction interval by learning temporal dependencies across past prediction residuals. Our comprehensive experiments using simulated and real data empirically demonstrate the superiority of the proposed method compared to the existing state-of-the-art conformal prediction methods.

1. Introduction

Uncertainty quantification has become crucial in many scientific domains where black-box machine learning models are often used (Angelopoulos & Bates, 2021). Conformal prediction has emerged as a popular and modern technique for uncertainty quantification by providing valid predictive inference for those black-box models (Shafer & Vovk, 2008; Barber et al., 2023).

Time series prediction aims to forecast future values based on a sequence of observations that are sequentially ordered in time (Box et al., 2015). With recent advances in machine learning, numerous models have been proposed and adopted for various time series prediction tasks. The increased use of black-box machine learning models necessitates uncertainty quantification, particularly in high-stakes time series prediction tasks such as medical event prediction, stock prediction, and weather forecasting.

While conformal prediction can provide valid predictive inference for uncertainty quantification, applying conformal prediction to time series is challenging since time series data often violate the exchangeability assumption. Additionally, real-world time series data typically exhibit significant stochastic variations and strong temporal correlations. Many efforts have been made to develop valid and effective conformal prediction methods for time series (Xu & Xie, 2023a). Sequential Predictive Conformal Inference (SPCI), a recently proposed conformal prediction framework for time series, has shown state-of-the-art performance by using Quantile Random Forest (Meinshausen & Ridgeway, 2006) as a conditional quantile estimator to predict the quantiles of future prediction residuals, which are used to estimate prediction interval (Xu & Xie, 2023b).

In this study, we employed Transformer decoder (Vaswani et al., 2017; Radford et al., 2018) as a conditional quantile estimator in the SPCI framework. Specifically, the Transformer decoder takes a sequence of past residuals and features as input to predict the quantiles of future residuals. Given that Transformer decoder-only architecture has already shown impressive performance in many sequential modeling tasks, we hypothesize that utilizing it in the SPCI framework benefits the estimation of prediction interval by learning temporal dependencies between the residuals.

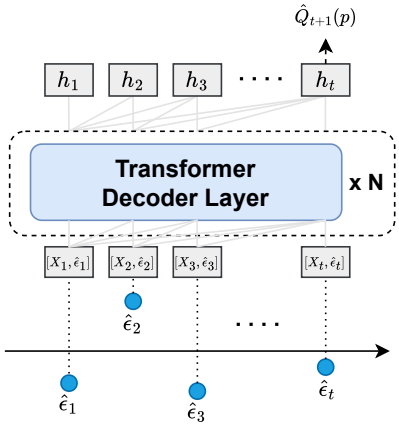
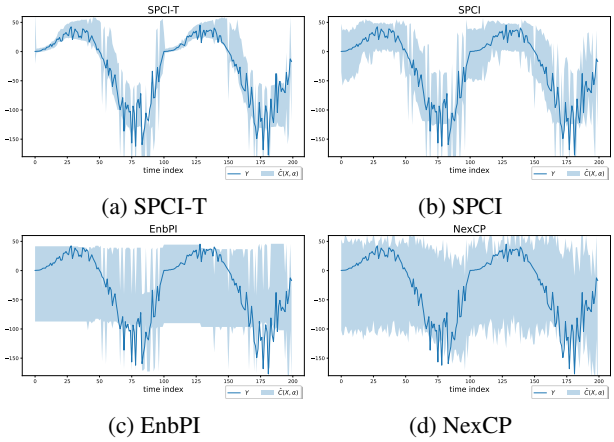


Figure 1. Visual description of our proposed method. Prediction residuals ($\hat{\epsilon}_t$) and corresponding features (X_t) are used as input to a stack of Transformer decoder layers to predict the quantiles of future residual(s), which are used to estimate prediction intervals.

[1]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA. Correspondence to: Yao Xie <yao.xie@isye.gatech.edu>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria. Copyright 2024 by the author(s).



(a) SPCI-T (b) SPCI

(c) EnbPI (d) NexCP

Figure 2. Prediction intervals estimated by `SPCI-T` and baselines on non-stationary simulated dataset.

We empirically demonstrate the superiority of the proposed method through experiments with simulated and real data, comparing it to state-of-the-art conformal prediction methods.

2. Problem setup

Consider a sequence of observations $\{(X_t, Y_t) : t = 1, 2, \ldots\}$, where $X_t \in \mathbb{R}^d$ denotes d-dimensional feature and $Y_t \in \mathbb{R}$ denotes continuous scalar outcome. Assume that the first T samples $\{(X_t, Y_t)\}_{t=1}^{T}$ be the data for training and validation. We also assume that we have a point prediction method $\hat{f}$ that provides a point prediction $\hat{Y}_t$ for Y as $\hat{Y}_t = \hat{f}(X_t)$.

Our goal is to sequentially construct a prediction interval $\hat{C}_{t-1}(X_t)$, starting from time $T+1$, that desirably contains the true outcome Y_t with the probability at least $1-\alpha$. $\hat{C}_{t-1}(X_t)$ is constructed using the past observations and predictions up to $t-1$ time. The *significance level* α is predefined and $\hat{C}_{t-1}(X_t)$ is a set includes Y_t with probability at least $1-\alpha$.

We use prediction residual (i.e., prediction error) as a nonconformity score, which is defined as:

$$\hat{\epsilon} = Y_t - \hat{Y}_t. \tag{1}$$

There are two types of coverage guarantees that should be satisfied with prediction intervals: *marginal coverage* and *conditional coverage*. *Marginal coverage* is defined as follows:

$$\mathbb{P}\left(Y_t \in \hat{C}_{t-1}(X_t)\right) \geq 1-\alpha, \forall t. \tag{2}$$

Conditional coverage is a stronger guarantee that given each X_t, the true observation Y_t in included in $\hat{C}_{t-1}(X_t)$ with at least $1-\alpha$ probability, which can be defined as follows:

$$\mathbb{P}\left(Y_t \in \hat{C}_{t-1}(X_t) | X_t\right) \geq 1-\alpha, \forall t. \tag{3}$$

If $\hat{C}_{t-1}(X_t)$ satisfies eq (2) or eq (3), it is called *marginally valid* or *conditionally valid*, respectively. The desired aim is to construct $\hat{C}_{t-1}(X_t)$ that satisfies both marginally and conditionally valid.

While infinite-width prediction intervals always satisfy both coverage guarantees, such intervals are pointless since they do not carry any information to quantify uncertainty. Therefore, we aim to minimize the width of prediction intervals while satisfying coverage.

Table 1. Empirical coverage and interval width of `SPCI-T` and baselines on the simulated datasets. The target coverage is set to 0.9, and the past window is set to 100. We report the average value with standard deviation calculated based on three independent trials with different random seeds.

	Non-stationary		Heteroskedasticity	
	Coverage	Width	Coverage	Width
SPCI-T	$0.91_{\pm.008}$	**52.38**$_{\pm.584}$	$0.88_{\pm.002}$	**9.56**$_{\pm.082}$
SPCI	$0.96_{\pm.004}$	$76.35_{\pm.527}$	$0.89_{\pm.006}$	$10.03_{\pm.001}$
EnbPI	$0.86_{\pm.002}$	$134.7_{\pm.743}$	$0.90_{\pm.006}$	$10.72_{\pm.008}$
NexCP	$0.91_{\pm.002}$	$156.4_{\pm2.73}$	$0.92_{\pm.004}$	$11.73_{\pm.198}$

Table 2. Empirical coverage and interval width of multi-step prediction using `SPCI-T` on the simulated datasets. The target coverage is set to 0.9, and the past window is set to 100.

	Non-stationary		Heteroskedasticity	
	Coverage	Width	Coverage	Width
$s=2$	$0.88_{\pm.015}$	$51.54_{\pm.314}$	$0.84_{\pm.003}$	$9.67_{\pm.134}$
$s=3$	$0.86_{\pm.008}$	$51.46_{\pm.183}$	$0.80_{\pm.003}$	$9.72_{\pm.151}$
$s=4$	$0.84_{\pm.010}$	$51.49_{\pm.256}$	$0.80_{\pm.010}$	$9.76_{\pm.140}$

3. Method

3.1. Sequential Predictive Conformal Inference

Sequential Predictive Conformal Inference (SPCI) is a conformal prediction framework for time series proposed by (Xu & Xie, 2023b). SPCI adopted Quantile Random Forest (Meinshausen & Ridgeway, 2006) as a conditional quantile estimator to estimate $\hat{\epsilon}_t$ sequentially, leveraging the dependencies of the past residuals. Specifically, SPCI estimates prediction intervals as follows:

$$\left[\hat{f}(X_t) + \hat{Q}_t(\hat{\beta}), \hat{f}(X_t) + \hat{Q}_t(1-\alpha+\hat{\beta})\right], \tag{4}$$

where $\hat{Q}$ denotes a conditional quantile estimator, $\hat{Q}_t(p)$ is the estimation for the true p-th quantile of $\hat{\epsilon}_t$. $\hat{\beta}$ denotes the

Table 3. Empirical coverage and interval width of the proposed method and baselines. The target coverage is 0.9, and the past window is set to 50 for all experiments. We report the average value with standard deviation calculated based on three independent trials with different random seeds.

	Wind		Electricity		Solar	
	Coverage	Width	Coverage	Width	Coverage	Width
SPCI-T	$0.93_{\pm.006}$	$\mathbf{2.08}_{\pm.072}$	$0.92_{\pm.009}$	$\mathbf{0.18}_{\pm.005}$	$0.93_{\pm.005}$	$\mathbf{50.70}_{\pm3.84}$
SPCI	$0.96_{\pm.016}$	$2.41_{\pm.016}$	$0.92_{\pm.004}$	$0.22_{\pm.001}$	$0.91_{\pm.006}$	$88.76_{\pm.245}$
EnbPI	$0.48_{\pm.006}$	$4.10_{\pm.009}$	$0.79_{\pm.002}$	$0.22_{\pm.001}$	$0.88_{\pm.002}$	$86.91_{\pm.363}$
NexCP	$0.92_{\pm.016}$	$6.27_{\pm.145}$	$0.89_{\pm.001}$	$0.46_{\pm.001}$	$0.86_{\pm.002}$	$114.98_{\pm.201}$

Table 4. Empirical coverage and interval width of the proposed method and baselines. The target coverage is 0.9, and the past window is set to 100 for all experiments. Note that the performance of NexCP is identical to the performance in Table 3 since it does not use the past window.

	Wind		Electricity		Solar	
	Coverage	Width	Coverage	Width	Coverage	Width
SPCI-T	$0.91_{\pm.011}$	$\mathbf{1.96}_{\pm.094}$	$0.92_{\pm.013}$	$\mathbf{0.17}_{\pm.014}$	$0.90_{\pm.006}$	$\mathbf{45.35}_{\pm1.67}$
SPCI	$0.94_{\pm.006}$	$2.39_{\pm.030}$	$0.93_{\pm.005}$	$0.22_{\pm.002}$	$0.93_{\pm.002}$	$88.21_{\pm.434}$
EnbPI	$0.74_{\pm.038}$	$4.65_{\pm.026}$	$0.85_{\pm.001}$	$0.26_{\pm.001}$	$0.88_{\pm.002}$	$86.64_{\pm.138}$
NexCP	$0.92_{\pm.016}$	$6.27_{\pm.145}$	$0.89_{\pm.001}$	$0.46_{\pm.001}$	$0.86_{\pm.002}$	$114.98_{\pm.201}$

value that minimizes the prediction interval as:

$$\hat{\beta} = \arg\min_{\beta \in [0,\alpha]} \left(\hat{Q}_t(1-\alpha+\beta) - \hat{Q}_t(\beta) \right).$$

3.2. Transformer Conformal Prediction for Time Series

In this study, we employ Transformer (Vaswani et al., 2017) as a conditional quantile estimator $\hat{Q}$ to estimate the true quantiles of $\hat{\epsilon}_t$ within the SPCI framework. Specifically, we use decoder-only architecture (Radford et al., 2018) since it can generalize to variable lengths of sequences without strictly partitioning the sequences for encoding and decoding. Throughout the paper, we refer to the proposed method as Sequential Predictive Conformal Inference with Transformer (`SPCI-T`).

The past w residuals and features are used as input for the model to predict the quantiles of the future unobserved residuals. Specifically, $\{Z_t\}_{t-(w+1)}^{t-1}$ is used to predict $\hat{Q}_t(p)$, where $Z_t := [X_t, \hat{\epsilon}_t]$. A fully connected layer without activation converts Z_t to the input representation of the model dimension. Note that Z_t can include other features that are useful for prediction besides X_t. Figure 1 visually describes the model architecture.

A fully connected layer without activation transforms output representation into the prediction of the quantile of $\hat{\epsilon}_t$. Training is done by sequentially minimizing the quantile loss as follows:

$$\mathcal{L}(\hat{\epsilon}, \hat{\epsilon}', p) = \begin{cases} p(\hat{\epsilon} - \hat{\epsilon}') & \text{if } \hat{\epsilon} - \hat{\epsilon}' \geq 0, \\ (1-p)(\hat{\epsilon}' - \hat{\epsilon}) & \text{if } \hat{\epsilon}' - \hat{\epsilon} \geq 0, \end{cases} \quad (5)$$

where p is the target quantile and $\hat{\epsilon}'$ is the predicted value of $\hat{\epsilon}$ corresponding to the target quantile.

We hypothesize that using the Transformer decoder as a conditional quantile estimator can offer the following advantages:

- Transformer decoder can effectively learn temporal dependencies, including long-term dependencies, across residuals.
- We can incorporate additional features, such as X_t, for conditional quantile estimation, allowing the Transformer to learn potential dependencies between these additional features and the residuals.
- Transformer decoder can perform multi-step predictions using known features through generative inference without needing explicit residuals as input for prediction.

4. Experiments

We evaluate `SPCI-T` and baselines on simulated and real data. The code for all experiments is available at `https://github.com/Jayaos/TCPTS`. Hyperparameters and implementation details are available in Appendix A.

4.1. Setup

We first obtain point predictions for all Y_t and corresponding residuals $\hat{\epsilon}$ by using leave-one-out (LOO) point predictors in

Table 5. Empirical coverage and interval width of multi-step prediction using SPCI-T on real datasets. The target coverage is 0.9, and the past window is set to 100.

	Electricity		Solar	
	Coverage	Width	Coverage	Width
$s = 2$	$0.87_{\pm.004}$	$0.18_{\pm.007}$	$0.88_{\pm.005}$	$42.70_{\pm1.20}$
$s = 3$	$0.82_{\pm.004}$	$0.18_{\pm.007}$	$0.86_{\pm.010}$	$44.56_{\pm1.82}$
$s = 4$	$0.78_{\pm.007}$	$0.18_{\pm.007}$	$0.84_{\pm.020}$	$46.83_{\pm2.08}$

all experiments. We use the ensemble of 25 random forests as the LOO point predictor.

We use state-of-the-art conformal prediction methods as baselines, which include SPCI (Xu & Xie, 2023b), EnbPI (Xu & Xie, 2021), and NexCP (Barber et al., 2023). We evaluate `SPCI-T` and the baselines regarding interval coverage and width. We also evaluate the multi-step prediction of `SPCI-T`. In a multi-step prediction setup, we aim to estimate the prediction intervals at s-step ahead, assuming that we only have known features (X_t) for multi-step prediction.

4.2. Simulation

Dataset We generate two simulated time series datasets. The data generating process follows $Y_t = f(X_t) + \epsilon_t$. The first dataset is a non-stationary time series, which contains periodicity and autoregressive ϵ_t. The second dataset is time series with heteroscedastic errors where the variance of ϵ_t depends on X_t. Details on simulated data are provided in Appendix A.

Results Table 1 shows empirical coverage and interval width of `SPCI-T` and baselines on the two simulated datasets. We observe that `SPCI-T` achieves the narrowest interval width without losing coverage. Figure 2 displays the prediction interval estimated by `SPCI-T` and baselines on the non-stationary dataset, confirming that `SPCI-T` obtains significantly narrower interval width compared to the baselines. Table 2 presents multi-step prediction results of `SPCI-T`. While `SPCI-T` maintains its prediction interval width, it loses coverage for multi-step prediction with larger s.

4.3. Real data Experiments

Dataset We use three time-series datasets from the real world: solar, wind, and electricity. The solar dataset (Zhang et al., 2021) contains solar radiation information in Atalanta downtown measured in terms of diffuse horizontal irradiance, provided by the United States National Solar Radiation Database. The wind dataset consists of wind speed records measured by the Midcontinent Independent System Operator every 15 minutes over a week period in September 2020 (Zhu et al., 2021). The electricity dataset contains electricity usage and pricing in the states of New South Wales and Victoria in Australia, observed between 1996 and 1999 (Harries et al., 1999). All three datasets were widely adopted benchmark datasets in conformal prediction literature. Details of the three datasets are provided in Appendix A.

Table 6. Empirical coverage and interval width of `SPCI-T` and baselines on a solar dataset with additional features. The target coverage is 0.9, and the past window is set to 50 or 100. Note that NexCP showed the identical performance regardless of w since NexCP does not use past windows.

	Solar with additional features			
	$w = 50$		$w = 100$	
	Coverage	Width	Coverage	Width
SPCI-T	$0.93_{\pm.013}$	$\mathbf{33.51}_{\pm1.99}$	$0.91_{\pm.006}$	$\mathbf{28.14}_{\pm1.60}$
SPCI	$0.92_{\pm.002}$	$89.63_{\pm.651}$	$0.92_{\pm.005}$	$87.28_{\pm.782}$
EnbPI	$0.88_{\pm.005}$	$86.94_{\pm.181}$	$0.89_{\pm.000}$	$85.73_{\pm.459}$
NexCP	$0.90_{\pm.002}$	$100.92_{\pm6.64}$	$0.90_{\pm.002}$	$100.92_{\pm6.64}$

Results Table 3 and Table 4 shows empirical coverage and interval width of `SPCI-T` and baselines on three real datasets with $w = 50$ and $w = 100$, respectively. We observe that `SPCI-T` consistently outperforms all other baselines. We also observe that the interval width of `SPCI-T` was narrow with a longer window, which shows that `SPCI-T` utilizes the long history of residuals by learning dependencies. Table 5 shows multi-step prediction results of `SPCI-T` on electricity and solar datasets. We exclude the results on the wind dataset since the coverage deteriorated significantly with $s > 2$. Similarly, in simulation, we observed that `SPCI-T` lost the coverage with the increasing s.

We add additional time features to each X_t in the solar dataset to see how these additional features can influence the performance of `SPCI-T`. Table 6 presents empirical coverage and interval width of `SPCI-T` and baselines on a solar dataset with additional features. `SPCI-T` again consistently outperforms baselines and shows significantly improved performance compared to the experiment on the solar dataset without the additional features. These results empirically demonstrate the advantage of `SPCI-T` in utilizing features with the residuals for conditional quantile estimation.

5. Conclusion

In this study, we propose `SPCI-T`, which incorporates a Transformer decoder into the recently developed SPCI framework. `SPCI-T` uses a Transformer decoder to learn temporal dependencies between the residuals and features to predict the conditional quantile of future residuals, which are then used to estimate the prediction interval. Our simulated and real data experiments empirically demonstrate the superiority of `SPCI-T`. Future directions include tailoring the model architecture for the time series conformal prediction task and conducting more comprehensive evaluations with state-of-the-art methods.

Software and Data

Datasets used in simulation and real data experiments are available at https://github.com/Jayaos/TCPTS.

Acknowledgements

This work is partially supported by an NSF CAREER CCF-1650913, NSF DMS-2134037, CMMI2015787, CMMI-2112533, DMS-1938106, DMS-1830210, and the Coca-Cola Foundation.

References

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

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

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

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

Meinshausen, N. and Ridgeway, G. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. 2018.

Sengupta, M., Xie, Y., Lopez, A., Habte, A., Maclaurin, G., and Shelby, J. The national solar radiation data base (nsrdb). *Renewable and sustainable energy reviews*, 89: 51–60, 2018.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

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

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

Xu, C. and Xie, Y. Conformal prediction for time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023a.

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

Zhang, M., Xu, C., Sun, A., Qiu, F., and Xie, Y. Solar radiation ramping events modeling using spatio-temporal point processes. *arXiv preprint arXiv:2101.11179*, 2021. Accepted at the 2021 INFORMS Conference on Service Science (ICSS 2021).

Zhu, S., Zhang, H., Xie, Y., and Van Hentenryck, P. Multi-resolution spatio-temporal prediction with application to wind power generation. *arXiv preprint arXiv:2108.13285*, 2021.

A. Additional Details of Experiments

A.1. Details of Simulated Data

Non-stationary time series We set $T = 2000$, which means we have $\{(X_t, Y_t)\}_{t=1}^{2000}$. We first generate each dimension of $X_t \in \mathbb{R}^{10}$ from $\text{Unif}(0, e^{0.01\text{mod}(t,100)})$ for all t. Then, $f(X_t)$ is obtained as follows:

$$f(X_t) = g(X_t)h(X_t),$$

where

$$g(t) = \log(t')\sin(2\pi t'/100), \quad t' = \text{mod}(t, 100),$$

and

$$h(X_t) = (|\beta^\top X_t| + (\beta^\top X_t)^2 + |\beta^\top X_t|^3)^{1/4}.$$

We sample ϵ_t from AR(1) as $\epsilon_t = \rho\epsilon_{t-1} + e_t$, where $\rho = 0.6$ and e_t is *i.i.d.* generated from normal distribution with zero mean and unit variance. $\beta \in \mathbb{R}^{10}$ is a sparse vector having only 20% non-zero elements where non-zero elements are generated from $\text{Unif}(0, 1)$.

Time series with heteroskedastic error We set $T = 2000$, then generate $X_t \in \mathbb{R}^{10}$ similarly to non-stationary time series for all t. Then, $f(X_t)$ is obtained as follows:

$$f(X_t) = (|\beta^\top X_t| + (\beta^\top X_t)^2 + |\beta^\top X_t|^3)^{1/4},$$

where β is generated in a similar way as in the non-stationary time series. We also sample ϵ_t from AR(1) similar to non-stationary time series, but with the variance of ϵ_t dependent on X_t as follows:

$$\text{Var}(\epsilon_t) = \sigma(X_t)^2,$$

$$\sigma(X_t) = \mathbf{1}^\top X_t.$$

A.2. Details of Real Data

Solar data The solar dataset contains solar radiation information recorded in every 30 minutes in 2018 in Atalanta downtown (Zhang et al., 2021) provided by the United States National Solar Radiation Database (Sengupta et al., 2018). We used seven covariates: Direct Normal Irradiance (DNI), dew point, surface albedo, wind speed, relative humidity, temperature, and pressure. The outcome variable is Diffuse Horizontal Irradiance (DHI), which reflects radiation levels. For additional time features, we converted the 24-hour period into 24 hourly one-hot encoded features, adding 24 additional features.

Wind data The wind dataset contains wind speed data (measured in m/s) from wind farms operated by the Midcontinent Independent System Operator (MISO) in the United States (Zhu et al., 2021). This 10-dimensional dataset recorded wind speed every 15 minutes over a one-week period in September 2020.

Electricity data The electricity dataset recorded electricity usage and pricing in the states of New South Wales and Victoria in Australia, every 30 minutes over a 2.5 year period between 1996 and 1999 (Harries et al., 1999). We used four covariates: *nswprice* and *vicprice*, the price of electricity in each of the two states; and *nswdemand* and *vicdemand*, the usage demand in each of the two states. The outcome variable is *transfer*, which is the quantity of electricity transferred between the two states.

A.3. Hyperparameters and Implementation Details

Table 7 shows hyperparameters chosen for `SPCI-T`. We conducted a grid search using training and validation set to find the optimal hyperparameters. The model dimension, number of heads, and number of layers were chosen when performance plateaued. Since `SPCI-T` requires validation set to select the best model during training, we split the datasets into training, validation, and test set with 8:1:1 ratio for `SPCI-T`. For all other baselines, the datasets were split into training and test set with a 9:1 ratio. For a fair comparison in terms of data usage, we additionally trained `SPCI-T` on the validation set after it was initially trained on the training set, which was conducted for 10% of the number of epochs used for the initial training.

Table 7. Hyperparameters chosen for `SPCI-T` in experiments using simulated and real data.

	Simulation		Real data			
	Non-stationary	Heteroskedastic	Wind	Electricity	Solar	Solar w/ add. features
batch size	4	4	4	4	4	4
learning rate	0.0001	0.0001	0.0001	0.0001	0.0005	0.0005
model dimension	16	16	32	16	16	32
number of heads	4	4	4	4	4	4
number of layers	4	4	4	4	4	4
dropout	0.2	0.2	0.1	0.2	0.2	0.2
additional training	N	N	Y	Y	Y	Y