A Proof of Theorem 1

Theorem 1. *Let $\mathcal{D} = \{(y^{(i)}_{1:t}, y^{(i)}_{t+1:t+H})\}_{i=1}^{l}$ be the dataset of exchangeable time-series observations and their H-step forecasts obtained from the same underlying probability distribution. Let M be the recurrent neural network predicting H-step forecasts using the direct strategy. For a target coverage level $\alpha \in (0, 1)$, the intervals obtained by the ICP-based conformal forecasting algorithm satisfy*

$$\mathbb{P}\left(\forall h \in \{1, \ldots, H\}.\, y_{t+h} \in [\hat{y}_{t+h} - \hat{\varepsilon}_h, \hat{y}_{t+h} + \hat{\varepsilon}_h]\right) \geq \alpha$$

Proof. Due to the direct forecasting strategy, every step in the horizon can be treated as a separate inductive conformal predictor that uses the same underlying model M (with the final predictions derived from the internal state being independent) and the same dataset $\mathcal{D}$. The independent validity of each of the H ICPs follows from Proposition 1 in Vovk [33]. Setting the error rate of each of the H ICPs to $(1 - \alpha)/H$ and applying Boole's inequality we obtain that the combined error rate of the H-step forecaster is $1 - \alpha$, as required. □

B Experiments

B.1 Synthetic data

The hyperparameters are provided in Table 7 and the prediction interval widths for different realisations of randomly generated datasets are given in Table 8.

Table 7: Training hyperparameters.

Parameter	**Value**
Training samples	1000
Calibration samples	1000
Test samples	500
Sequence length L	10
Prediction horizon S	5
Autoregressive mean μ_x	1
Autoregressive variance σ_x^2	2
Periodicity s	None
Amplitude u	5
Epochs	1000
Batch size	100
Embedding size	20
Learning rate	0.01
Underlying RNN type	LSTM
Target coverage α	90%

Table 8: CoRNN prediction interval lengths for different random seeds.

Noise mode		**CoRNN PI lengths**				
Static $\sigma_t^2 = 0.1n$	$n = 1$	17.33 ± 4.04	18.01 ± 4.15	19.02 ± 5.20	18.33 ± 4.63	18.32 ± 4.72
	$n = 2$	18.73 ± 5.09	19.85 ± 4.92	18.30 ± 4.46	18.33 ± 4.66	17.49 ± 3.87
	$n = 3$	16.99 ± 3.81	18.42 ± 4.91	16.86 ± 4.21	17.83 ± 4.35	18.75 ± 4.55
	$n = 4$	18.77 ± 5.10	18.51 ± 4.01	17.59 ± 4.73	18.01 ± 4.28	17.71 ± 4.64
	$n = 5$	18.47 ± 4.14	18.00 ± 3.82	19.08 ± 4.94	19.31 ± 4.80	19.09 ± 5.01
Time-dependent $\sigma_t^2 = 0.1tn$	$n = 1$	20.07 ± 4.89	18.91 ± 3.66	18.37 ± 3.82	21.40 ± 5.40	19.58 ± 3.69
	$n = 2$	22.98 ± 3.88	23.13 ± 4.43	22.55 ± 4.37	22.97 ± 4.59	23.31 ± 4.00
	$n = 3$	27.90 ± 4.74	27.75 ± 4.10	27.97 ± 4.81	27.10 ± 4.38	26.78 ± 4.53
	$n = 4$	32.56 ± 6.68	35.05 ± 6.89	32.73 ± 5.30	33.70 ± 6.00	33.00 ± 6.33
	$n = 5$	37.97 ± 5.96	38.99 ± 7.71	38.68 ± 7.07	39.21 ± 7.43	38.79 ± 6.85

B.2 Real-world datasets

MIMIC-III We collect the data of patients on antibiotics from the MIMIC-III dataset [35], and filter out the sequences of total length at least 5, resulting in 4323 sequences. From these sequences

we pick out the white blood cell (high) count as the feature for the univariate time-series. We split the sequences randomly into training, calibration and test datasets. We pick the constant time horizon of 2, which is to account for the shortest sequences being of length 5, and use the rest of the sequence as model input.

EEG The EEG dataset, available at https://archive.ics.uci.edu/ml/datasets/EEG+Database, was used as the source for the EEG signal time-series. The dataset contains the data for control and alcoholic subjects responding to a visual stimulus of three types. We used the medium version of the dataset, involving 10 control and 10 alcoholic subjects, though for the experiments we only used the control subjects—from the summaries provided, control subject EEG responses seemed to be more difficult to predict. Each subject had repeated trials for every type of stimulus, and each trial had a time-series for the 64 channels obtained from their corresponding sensor. We treated every individual trial and each of the 64 channels as a separate time-series example, resulting in 19200 sequences in the training set. To keep training efficient, we downsampled the sequences (normally of length 255) to a total length 50, which we further split into the training sequence of 40 and prediction horizon 10. The training and test dataset splits are readily provided in the UCI repository, and for repeated trials we used different subsets for calibration and different model training seeds.

COVID-19 The data is available at https://coronavirus.data.gov.uk/. We picked the data of different regions of the same country in order to follow the exchangeability assumption as closely as possible, while the data from different countries risks having much larger distribution shifts due to a large variation of factors like government lockdown policies. Given the setup of the conformal prediction framework, we looked for the data that would have a sufficiently large number of independent sequences—the lower tier local authority split gives a total number of 380 sequences, which over repeated trials we would randomly split into the test set of 80 sequences and the the rest between the training and calibration sets. We picked the data of daily new cases over the course of 150 days starting mid-September 2020 and ending mid-February 2021, which we further split into the input sequence of 100 examples (ending Christmas 2020) and using the remaining 50 days as the testing sequence. We chose these dates to capture interesting properties of changing government lockdown policies and so that the two waves are separated between the observed and the to-be-predicted sequence.

Hyperparameters The training hyperparameters (Table 9) mostly follow those provided in previous work [21] and are kept the same for new experiments in order to ensure fair comparison between the baselines. For the CoRNN model, the total training data available was split between the training and calibration sets, and the other baselines used all available training data to train the underlying RNN model.

Table 9: Training hyperparameters for the real-world datasets.

Parameter	**MIMIC-III**	**EEG**	**COVID-19**
Training samples	3823 (2000)	19200 (15360)	300 (200)
Calibration samples	1823	3840	100
Test samples	500	19200	80
Sequence length L	[3, 47]	40	100
Prediction horizon S	2	10	50
Epochs		1000	
Batch size		150	
Embedding size		20	
Learning rate		0.01	
Dropout probability (for DP-RNN)		0.5	
Underlying RNN type		LSTM	
Target coverage α		90%	