Channel Dependence, Limited Lookback Windows, and the Simplicity of Datasets: How Biased is Time Series Forecasting?

Ibram Abdelmalak[*†]
abdelmalak@ismll.de
ISMLL & VWFS-DARC
Hildesheim, Niedersachsen
Germany

Kiran Madhusudhanan[*†]
kiranmadhusud@ismll.de
ISMLL & VWFS-DARC
Hildesheim, Niedersachsen
Germany

Jungmin Choi[*†]
choi@ismll.de
ISMLL & VWFS-DARC
Hildesheim, Niedersachsen
Germany

Christian Klötergens[†]
kloetergens@ismll.de
ISMLL & VWFS-DARC
Hildesheim, Niedersachsen
Germany

Vijaya Krishna Yalavarthi[†]
yalavarthi@ismll.de
ISMLL & VWFS-DARC
Hildesheim, Niedersachsen
Germany

Maximilian Stubbemann[†]
stubbemann@ismll.de
ISMLL & VWFS-DARC
Hildesheim, Niedersachsen
Germany

Lars Schmidt-Thieme[†]
schmidt-thieme@ismll.de
ISMLL & VWFS-DARC
Hildesheim, Niedersachsen
Germany

ABSTRACT

In Time Series Forecasting (TSF), the lookback window (the length of historical data used for prediction) is a critical hyperparameter that is often set arbitrarily, undermining the validity of model evaluations. We argue that the lookback window must be tuned on a per-task basis to ensure fair comparisons. Our empirical results show that failing to do so can invert performance rankings, particularly when comparing univariate and multivariate methods. Experiments on standard benchmarks reposition Channel-Independent (CI) models, such as PatchTST, as state-of-the-art methods. However, we reveal this superior performance is largely an artifact of weak inter-channel correlations and simplicity of patterns within these specific datasets. Using Granger causality analysis and ODE datasets (with implicit channel correlations), we demonstrate that the true strength of multivariate Channel-Dependent (CD) models emerges on datasets with strong, inherent cross-channel dependencies, where they significantly outperform CI models. We conclude with three key recommendations for improving TSF research: (i) treat the lookback window as a critical hyperparameter to be tuned, (ii) use statistical analysis of datasets to inform the choice between CI and CD architectures, and (iii) favor CD models in applications with limited data.

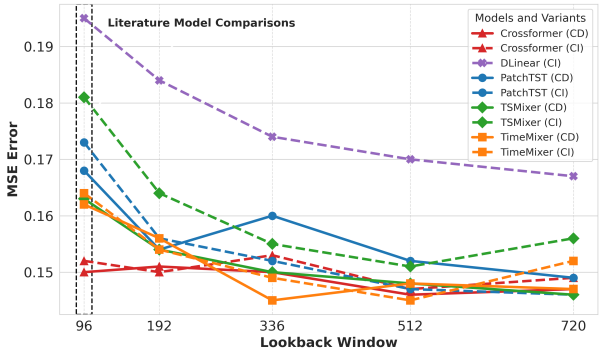


Figure 1: Test-error (MSE) on y-axis for various lookback windows on the x-axis for the weather dataset and a forecast horizon of 96. CI denotes the channel-independent variant and CD denotes the channel-dependent variant of TSF architectures. The black dotted line shows lookback window of 96, where most methods compare performance.

1 INTRODUCTION

Time Series Forecasting (TSF) is a fundamental challenge in machine learning that drives key innovations across domains such as finance, energy, healthcare, and climate science [6]. The core task of TSF is to predict future observations by leveraging patterns within historical time-dependent data. Nevertheless, the literature suffers from two significant experimental shortcomings that compromise the fairness and reliability of model comparisons.

First, in the multivariate TSF learning setting, the historical behavior is modeled using a fixed-length sequence of past observations—known as the *lookback window*—to forecast a set of future values over a given prediction horizon. Despite its fundamental role, the choice of lookback window remains surprisingly underexplored in TSF literature. Most existing Long-term Time Series Forecasting (LTSF) studies adopt a fixed lookback window (mostly 96) for all models and datasets, often without justification or transparency [2, 3, 8, 11–13, 20–22].

[*]Authors contributed equally to this research.
[†]Full Affiliation: Information System and Machine Learning Lab (ISMLL) & VWFS Data Analytics Research Center (VWFS-DARC), University of Hildesheim, Hildesheim, Niedersachsen, Germany.

While this practice simplifies experimental comparisons, it masks a critical reality: *the optimal lookback window can vary significantly across models and datasets, and failing to tune it can lead to misleading conclusions*. As demonstrated in Figure 1, for the weather dataset, the optimal lookback window length differs across various state-of-the-art TSF models. Consequently, performance comparisons at a single, fixed lookback window may misrepresent which models are truly state-of-the-art under realistic, well-optimized conditions. Comparable studies have been conducted by Zeng et al. [21] on the Electricity dataset and by Tong and Yuan [17] on the Weather dataset. However, unlike previous studies, we explicitly tune all the hyperparameters for each lookback window to generate the figure ensuring reliable comparisons.

The optimal lookback window is highly dependent on both the intrinsic properties of the dataset and the specific architecture of the forecasting model. For example, when comparing within the same weather dataset (Figure 1), Channel Independent (CI) or *univariate* models tend to favor longer lookback windows (e.g., 336 or 512), enabling deeper temporal modeling per variable [3, 14, 21]. Whereas the Channel Dependent (CD) or *multivariate* models jointly model all channels and often operate with shorter lookback windows (e.g., 96), leveraging channel dependence to compensate for the reduced temporal depth [13, 19]. Even within CD class of methods (i.e, across different multivariate models) certain architectures benefit more from long histories than others. For example, Crossformer-CD significantly outperforms TimeMixer-CD at a lookback window of 96, yet its performance drops below TimeMixer-CD when the lookback window is extended to 336. These findings motivate our central claim:

> *Tuning the lookback window as a model-specific hyperparameter is essential for fair evaluation and reliable assessment in time series forecasting.*

Beyond the issue of lookback tuning, a second critical limitation undermines the fairness and practical relevance of current TSF benchmarks. Most widely-used multivariate forecasting datasets such as ETTh, ETTm, Electricity, Weather, and Traffic exhibit minimal channel dependence among their variables. This lack of inter-variable relational complexity renders these benchmarks overly simplistic; even simple linear models [21] can achieve performance comparable to sophisticated nonlinear architectures. This raises a key question:

> *Does channel dependence truly matter in multivariate LTSF?*

To answer this, we conduct an in-depth empirical analysis, combining traditional performance metrics with statistical tools such as *Granger causality* [7]. Our analysis reveals that the benefits of multivariate modeling depend strongly on two factors: (1) the amount of historical data available, and (2) the degree of causal influence among channels. When inter-channel interactions are weak or datasets are large enough for univariate models to exploit long-term patterns, the added complexity of modeling channel dependence may offer limited returns. Additionally, through experiments on complex channel-dependent chaotic Ordinary Differential Equation (ODE) datasets [5], we reveal that the standard benchmarks lack realistic cross-channel interactions essential for multivariate forecasting. This simplicity disproportionately favors CI models and reduce the ability to fairly assess models designed to exploit channel dependencies as also shown in [3, 14].

Based on these experimental and statistical findings, we offer the following practical recommendations for evaluating and designing multivariate time series forecasting models:

- Channel dependence is unnecessary for top performance on standard benchmarks; tuning the lookback window enables CI models like PatchTST to match or surpass CD ones (Section 6.1).
- On complex datasets with strong channel correlations (supported by granger causality Section 6.2), CD models outperform CI models (Section 6.3).
- Longer lookback windows generally favor univariate models, highlighting the need for careful lookback tuning to ensure fair comparisons (Section 6.4).

2 RELATED WORKS

LTSF models typically fall into two categories: *Channel-Dependent (CD)*, which model cross-channel interactions, and *Channel-Independent (CI)*, which treat each channel in isolation and focus purely on temporal dynamics. CI models such as DLinear [21] and PatchTST [14] process each feature independently and consistently achieve strong performance on widely-used benchmarks, casting doubt on the necessity of cross-channel modeling. In contrast, MLP-based CD methods like TSMixer [3], TimeMixer [18], and transformer-based architectures including iTransformer [13] and Crossformer [22] aim to capture inter-channel dependencies explicitly. However, the predominantly univariate nature of standard benchmark datasets limits the ability of CD methods to demonstrate their advantages. To mitigate this issue, we incorporate evaluations on ODE-based datasets [5] designed to embed implicit channel correlations and enabling more representative comparisons on channel-dependent datasets.

The effectiveness of CI methods also appears to be amplified by longer lookback windows. For instance, DLinear adopts a lookback of 336 steps, whereas transformer-based baselines were defaulted to a lookback of 96. PatchTST, a successor to DLinear, employs a fixed lookback window of 512 or 336 and addresses the limitations of its predecessor by systematically tuning baseline methods over a broader range of lookback lengths. Nonetheless, the widespread adoption of a fixed 96-step lookback, popularized by iTransformer has constrained CI models and impeded equitable benchmarking. Subsequent studies [2, 8, 11, 12, 17] frequently report improved performance of proposed methods by retaining short fixed windows for strong CI baselines like PatchTST, thereby hindering fair comparisons due to inconsistent tuning.

Several recent studies [14, 17, 18, 21] acknowledge the importance of lookback window selection in forecasting performance. However, lookback tuning is frequently confined to ablation studies, with many works failing to apply consistent tuning across both baselines and proposed methods in their main experiments. For instance, TimeMixer reports results using a unified set of hyperparameters (including a fixed lookback window) in the main results, while providing tuned results only in the appendix. Another work proposed IRPA method [17] that explores various lookback lengths (480, 1920, 3600) for their method, yet retains a fixed lookback of 96

for all baselines, diminishing their comparative performance and inflating IRPA's gains. In contrast, our evaluations perform systematic lookback window tuning for all models, uncovering genuinely optimal configurations.

3 PROBLEM STATEMENT

Multivariate time series is a sequence $(\boldsymbol{x}^l)_{l=1}^T \in (\mathbb{R}^C)^T$, where each $\boldsymbol{x}^l \in \mathbb{R}^C$ represents observations from C channels at time step l. This sequence can be represented as a matrix $\boldsymbol{X}^{\text{full}} \in \mathbb{R}^{C\times T}$. If $C = 1$, this represents the *univariate* forecasting setting.

To train forecasting models effectively, especially when T is large, we partition the long sequence into multiple shorter *subsequences* using a sliding window approach. Each subsequence is split into two parts: an input segment (lookback window) of length L and a target segment (forecast horizon) of length H. Specifically, a time-series forecasting dataset is composed of input-output pairs $(\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{R}^{C\times L} \times \mathbb{R}^{C\times H}$, where:

$\boldsymbol{X} = (\boldsymbol{x}^t, \boldsymbol{x}^{t+1}, \ldots, \boldsymbol{x}^{t+L-1})$ is referred to as the *forecasting query*, and

$\boldsymbol{Y} = (\boldsymbol{x}^{t+L}, \boldsymbol{x}^{t+L+1}, \ldots, \boldsymbol{x}^{t+L+H-1})$ is the corresponding *forecasting answer*.

The goal of the time-series forecasting problem is to train a model $\hat{\boldsymbol{Y}} : \mathbb{R}^{C\times L} \to \mathbb{R}^{C\times H}$ on a dataset $\mathcal{D}_{\text{train}} \subseteq \mathbb{R}^{C\times L} \times \mathbb{R}^{C\times H}$, constructed from many such overlapping or non-overlapping windows, sampled from the full sequence. This dataset is assumed to be drawn i.i.d. from some underlying distribution p. The objective is to minimize the expected loss:

$$\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim p}\, \ell(\boldsymbol{Y}, \hat{\boldsymbol{Y}}(\boldsymbol{X})) \tag{1}$$

where $\ell : \mathbb{R}^{C\times H} \times \mathbb{R}^{C\times H} \to \mathbb{R}$ is a suitable loss function, commonly the *mean squared error* (MSE).

4 BACKGROUND

Fixing Lookback Windows: The problem formulation described earlier places no intrinsic requirement on the forecasting query $\boldsymbol{X}$ to have a fixed lookback length L. Nevertheless, most recent LTSF models fix the lookback window during training, typically to a value of $L = 96$. This constraint is usually imposed for implementation convenience, computational efficiency, or alignment with benchmarking conventions, rather than for a principled reason. As we have argued, the lookback window should instead be treated as an integral property of the model, similar to any other model-specific hyperparameter.

Channel Independence: One of the main findings of PatchTST [14] and later by TSMixer [3] is that, on widely used benchmark datasets, CI models are sufficient for accurate forecasting. This suggests that interactions between different channels are not essential, and that each channel can be modeled in isolation. Formally, a model $\hat{\boldsymbol{Y}} : \mathbb{R}^{C\times L} \to \mathbb{R}^{C\times H}$ is said to be *channel-independent* if for all $\boldsymbol{X} \in \mathbb{R}^{C\times L}$ and any channel input $X_c \in \mathbb{R}^L$,

$$\hat{\boldsymbol{Y}}(\boldsymbol{X})_c = \hat{\boldsymbol{Y}}(X_c) \tag{2}$$

i.e., the prediction for any given channel c is generated using only its own historical data, independent of all other channels.

Granger Causality: The Granger causality test [7] evaluates whether past values of X_{c_2} improve forecasts of Y_{c_1} beyond what is achievable using only past values of X_{c_1}. If including X_{c_2} reduces prediction error in an autoregressive model, then X_{c_2} provides a useful linear signal for forecasting channel c_1. To quantify this directional relationship, we compute the F-statistic (Equation (5)) for each channel pair, measuring the predictive gain from adding X_{c_2}.

The F-statistic compares the sum of squared residuals (SSR) of two models: a univariate model using only lagged values of c_1 (SSR_u, Equation (3)), and a multivariate model that additionally includes lagged values of c_2 (SSR_{mv}, Equation (4)). The univariate model uses coefficients $\hat{a}$ applied to $X_{c_1}^{t-i}$ over a lag window of length L, where $i = 1, \ldots, L$; the multivariate model extends this with coefficients $\hat{b}$ for $X_{c_2}^{t-i}$. Residuals are computed for $t = L+1, \ldots, L+H$.

In Equation (5), k_u is the number of added parameters associated with channel c_2, and k_{mv} is the total number of parameters in the multivariate model. The denominator degrees of freedom are $N - k_{mv}$, where N is the number of observations. A large F-statistic indicates a significant improvement in explained variance when including c_2. We reject the null hypothesis H_0 : "c_2 does not Granger-cause c_1" when this improvement is statistically significant.

$$SSR_u = \sum_{t=L+1}^{L+H} \left(Y_{c_1}^t - \hat{a}_0 - \sum_{i=1}^{L} \hat{a}_i X_{c_1}^{t-i} \right)^2, \tag{3}$$

$$SSR_{mv} = \sum_{t=L+1}^{L+H} \left(Y_{c_1}^t - \hat{a}_0 - \sum_{i=1}^{L} \hat{a}_i X_{c_1}^{t-i} - \sum_{i=1}^{L} \hat{b}_i X_{c_2}^{t-i} \right)^2, \tag{4}$$

$$F = \frac{(SSR_u - SSR_{mv})/k_u}{SSR_{mv}/(N - k_{mv})}, \tag{5}$$

5 GENERAL EXPERIMENTAL SETUP

5.1 Models

We evaluate a diverse set of strong baseline models, varying in architectural complexity and treatment of channel dependencies. These include MLP-based models (TSMixer [3], TimeMixer [19]), attention-based models (PatchTST [14], iTransformer [13], Crossformer [22]), and the simple yet competitive DLinear [21]. We use only the CI version of DLinear, consistent with its design as a lightweight baseline. For iTransformer, we use only the CD version, as removing inter-channel attention would reduce it to a basic MLP, diverging from its core design. Rest of the methods are available in or allow derivation of both CD and CI forms. Additional details on the derivation of both variants are included in the appendix. For additional experimental details, please refer to our code repository[1].

5.2 Datasets

As noted in the introduction, we evaluate two main dataset categories to examine how increasing complexity particularly channel correlation impacts model performance.

[1] https://github.com/anonsub936/How-Biased-is-Time-Series-Forecasting-

Table 1: Average MSE over the 4 forecasting horizons for each variant of the baseline evaluated for the standard datasets. The best variant for each model is in bold and the best overall model is in blue. Note here due to running out of memory (OOM) for some experiments related to electricity and traffic, we reported the results with "*" from TimeMixer paper from the experiment where they tune the lookback window as well as the hyperparameters (please refer to appendix E in [18]).

Dataset	PatchTST		TSMixer		Crossfor.		DLin.	iTrans.	TimeMixer	
	CI	CD	CI	CD	CI	CD	CI	CD	CI	CD
ETTh1	**0.422**	0.437	**0.438**	0.453	0.477	**0.456**	0.423	0.511	**0.434**	0.489
Weather	0.245	**0.233**	**0.230**	0.241	**0.236**	0.296	0.289	0.253	0.247	**0.246**
Electricity	**0.159**	0.170	**0.162**	0.170	**0.166**	0.188	0.162	0.220	**0.173**	0.260
ETTh2	**0.365**	0.382	**0.378**	0.390	0.741	**0.691**	0.507	**0.363**	**0.371**	0.378
ETTm1	**0.356**	0.371	**0.356**	0.370	**0.393**	0.426	0.359	0.374	**0.355**	0.408
ETTm2	**0.258**	0.259	**0.257**	0.273	**0.433**	0.485	0.289	**0.257**	**0.275**	0.342
Traffic	**0.388**	OOM	**0.407**	0.417	OOM	0.542	0.426	OOM	OOM	**0.388**
Wins	**3**	0	2	0	0	0	0	2	1	0
Avg. Rank	**2.14**	5.43	2.71	5.00	7.29	8.14	5.43	5.86	5.29	6.57

Table 2: Average F-scores and percentage of rejected null hypotheses H_0 ($p = 0.05$) across different lag values, comparing the standard datasets with the ODE benchmark. Bold values indicate the highest correlation at each lag.

Dataset Type (lag)	Average F-score	H_0 rejected (%, p=0.05)
ODEs (30)	**335.21**	**73%**
Standard (30)	2.46	61%
ODEs (96)	**88.50**	**68%**
Standard (96)	1.40	45%
ODEs (192)	**35.29**	**63%**
Standard (192)	1.23	27%

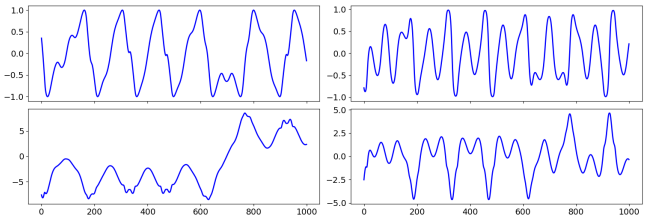


Figure 2: Time series visualization of the Double Pendulum ODE system. Notably, the bottom two channels (angular velocities of the two pendulums) exhibit synchronized behavior in the last 200 time steps, demonstrating time-varying inter-channel correlation.

The Standard Datasets: The first category includes widely used benchmarks from recent state-of-the-art studies and surveys [3, 13, 14, 19, 21, 22]: *Weather*, *Electricity*, *Traffic*, and the *Electricity Transformer Temperature (ETT)* datasets. These benchmarks gained popularity following their adoption in the Informer paper [23]. We refer to this group as *the standard datasets*. Details on data statistics are provided in the appendix.

The Chaotic ODE Benchmark: As a second dataset collection, we consider the benchmark proposed by Gilpin [5]. This benchmark was introduced as a time-series collection that features complex time series dynamics with clear real world data properties, providing a challenging and realistic setting for channel-dependent time-series forecasting. A demonstration for one of these datasets and how it simulates a recurrent patterns across different channels can be seen in Figure 2. This figure particularly emphasizes periods of synchronization in the angular velocities of the two pendulums during the final 200 timesteps, indicative of inter-channel correlation. Additional plots of the selected ODE system are provided in the appendix.

We select a subset of datasets from this benchmark based on two distinctive criteria: *(i)* The datasets follow complex trajectories that challenge most time series models and *(ii)* they are derived directly from the underlying ODEs, ensuring a clear mathematical correlation between variables. For criterion (i), we use the *Largest Lyapunov exponent* [5] to quantify dataset complexity, selecting datasets with high values to ensure forecasting remains non-trivial task that cannot be solved easily based on one channel information. To illustrate criterion (ii), we show below a simplified form of the Double Pendulum system [4], assuming equal masses and rod lengths. The coupling between angular velocities $\frac{d\theta_1(t)}{dt}$, $\frac{d\theta_2(t)}{dt}$ and angles $\theta_1(t), \theta_2(t)$ at time t from Equations (6) and (7) is evident, with parameters g (gravity) and l (rod length) incorporated. Such systems are essential in modeling real-world physics applications, including robotics [9].

Due to space constraints, we defer detailed discussion of another example, *Cell Cycle*, to the appendix. This dataset reflects strong variable coupling and captures real-world measurements of Cyclin-Dependent Kinase (CDK) activity, crucial to cell cycle progression [15]. Our selection yields six datasets with rich temporal dynamics and strong inter-channel dependencies, referred to as *the ODE datasets*. Further reasoning for particularly choosing this benchmark follows through the statistical analysis in Section 6.2.

Table 3: Average MSE across four forecasting horizons for each model variant on the ODE Benchmark. Bold values highlight the best variant within each model; the best overall result per row is shown in blue.

Dataset	PatchTST		TSMixer		CrossFor.		DLin.	iTrans.	TimeMixer	
	CI	CD	CI	CD	CI	CD	CI	CD	CI	CD
Lorenz	**0.839**	0.841	0.880	**0.868**	0.667	**0.643**	0.934	0.675	0.764	**0.673**
BlinkingRotlet	**0.424**	0.426	0.580	**0.487**	0.340	**0.311**	0.522	0.426	**0.433**	0.520
CellCycle	**0.635**	0.667	0.792	**0.771**	0.429	**0.428**	0.935	0.808	0.679	**0.556**
DoublePendulum	**0.653**	0.668	0.768	**0.737**	0.553	**0.541**	0.805	0.656	0.595	**0.529**
Hopfield	0.420	**0.346**	0.507	**0.435**	0.335	**0.316**	0.690	0.300	0.622	**0.245**
LorenzCoupled	0.881	**0.866**	0.950	**0.900**	0.700	**0.666**	0.963	0.857	**0.788**	0.832
Wins	0	0	0	0	0	**4**	0	0	0	2
Avg. Rank	5.17	5.67	8.83	7.50	2.50	**1.50**	9.83	5.17	5.50	3.30

$$\frac{d^2\theta_1(t)}{dt^2} = -\frac{2g}{l}\theta_1(t) + \frac{g}{l}\theta_2(t) \tag{6}$$

$$\frac{d^2\theta_2(t)}{dt^2} = \frac{2g}{l}\theta_1(t) - \frac{2g}{l}\theta_2(t) \tag{7}$$

5.3 Hyperparameter Tuning

To ensure fair evaluation, we apply a consistent tuning procedure across all models. Each hyperparameter is defined either by a discrete set (categorical) or a continuous range (real-valued). Performance for each hyperparameter configuration is validated on the designated validation set through the Mean Squared Error (MSE). The best model is chosen based on the lowest validation error, then it is tested on a the separate test split to get the final evaluation. We use Bayesian optimization via Optuna [1] to efficiently explore promising configurations, running 20 trials per setup (i.e., a specific model, dataset, and forecasting horizon).

Crucially, the lookback window is treated as a tunable hyperparameter, selected from $\{96, 192, 336, 512, 720\}$. Also more hyperparameters that are specific for each model are tuned thoroughly, allowing each model to operate under optimal conditions for fair comparison. Detailed search spaces and best hyperparameters chosen for each model are provided in the appendix.

5.4 Granger Causality

We conduct Granger causality analysis following the standard procedure outlined by Granger [7], with several preprocessing steps to ensure robust and meaningful results:

(i) We first assess the stationarity of each dataset. If non-stationarity is detected, differencing is applied iteratively until stationarity is achieved.

(ii) To remove redundancy, we compute the Pearson correlation coefficient for all channel pairs. Channels with correlation exceeding 0.95 are considered highly similar, and one from each such pair is excluded from further analysis.

After the preprocessing steps, we conduct the F-test described in Section 4. Consequently, we report two important metrics: the average F-score over all valid channel pairs, and the percentage of pairs for which the null hypothesis is rejected (indicating significant Granger causality). These results are summarized in Section 6.2.

Additional implementation details, including parameter settings and the formal hypothesis definition, are provided in the appendix.

6 RESULTS

6.1 Evaluation of Standard Datasets

We reproduce the current time-series forecasting models in our above-mentioned evaluation settings, where we conduct fair and extensive hyperparameter tuning, including tuning the lookback window. Unless stated otherwise, all results are averaged over 5 different random seeds. Table 1 presents the results averaged across forecasting horizons 96, 192, 336, and 720 for the standard datasets. Detailed values for each individual horizon and standard-deviations are provided in the appendix.

PatchTST remains the state-of-the-art for time-series forecasting. In 3 out of 7 datasets, the standard channel-independent version of PatchTST achieves the best performance. For three other datasets, this version ranks second or third. Its performance is not among the top three only on the Weather dataset, which may be attributed to the relatively stronger inter-channel dependencies in this dataset, as indicated by Granger causality. The average rank of PatchTST across all datasets is 2.14 showing a significant gap of 0.57 compared to the second-best model, the channel-independent version of TSMixer.

Linear models, meanwhile, remain competitive with recent forecasting approaches. The performance of DLinear on the standard datasets is comparable to more recent models. Across the 6 datasets, it achieves an average rank of 5.43 out of 10. Notably, on ETTh1, it ranks second, trailing PatchTST by only a small margin.

6.2 Granger Causality for Evaluation of Channel Correlations

Based on the findings from the previous section, which showed that multivariate models do not consistently outperform univariate ones on the *standard datasets*, we turn to a statistical evaluation of channel correlations across a broader set of datasets. Our goal is to identify benchmarks where inter-channel dependencies exist, potentially finding datasets and evaluation setup where multivariate models are needed.

Although the full test results are presented in the appendix, we focus here on a representative comparison between the *ODE*

datasets and the *standard datasets*, to highlight key patterns. We conduct Granger causality tests using three different lag values: 30, 96, and 192. These lags correspond to different lookback window sizes used in the autoregressive models underlying the tests. To further analyze channel dependence across a broad range of time series domains, we computed Granger causality values for 41 real-world datasets from the extensive Monash Time Series Forecasting Archive [6]. These results, calculated for lag values of 7, 30, 96, and 192, are provided in the appendix. Throughout this section, we use the terms *lag* and *lookback window* interchangeably.

Varying the lag allows us to analyze how inter-channel correlations evolve with different temporal resolutions. We can see the F-scores and the percentage of pairs of channels that exhibits granger causality in Table 2. For both metrics, the higher values indicate higher correlation in the datasets. As seen in the table, the *ODE datasets* exhibit significantly higher F-scores and pass rates across all lag values, strongly indicating the presence of meaningful inter-channel dependencies. On contrary, the *standard datasets* show considerably weaker correlations, with average F-scores dropping below 1.0 for ETTm2 dataset at the highest lag.

These findings align with our earlier empirical observations: univariate models often match or outperform multivariate models on the *standard datasets* due to weak inter-channel coupling. Moreover, we observe that increasing the lag generally reduces the measured correlation for both dataset types. This effect is particularly pronounced for the standard datasets, suggesting that longer lookback windows benefit more univariate models than multivariate ones, since the former cannot leverage strong inter-channel signals.

In contrast, the *ODE datasets* show much stronger correlations at longer lags, e.g., average F-scores of 88.5 (lag 96) and 35.29 (lag 192), compared to 1.40 and 1.23 for the *standard datasets*. These results support using the *ODE datasets* as a valuable testbed for evaluating models under strong cross-channel dynamics.

6.3 Evaluation on the ODE Benchmark

As discussed in Section 6.1, our results reaffirm DLinear as a competitive baseline and also suggest the superiority of CI models over their CD counterparts. In this section, we evaluate how the strong correlations between channels in the *ODE datasets* affect model performance. Table 3 presents the results obtained using the same evaluation protocol described in Section 6.1.

The evaluation on the *ODE datasets* reveals a markedly different performance hierarchy among models. First of all, CD variants clearly outperform CI ones. The CD variants of the evaluated models consistently outperform their CI counterparts, highlighting the importance of modeling inter-channel dependencies, especially in datasets governed by coupled differential equations. As shown in Table 4, CD models outperform CI versions in 18 out of 24 total experiments. This strongly supports the claim that model architecture should be tailored to the underlying data structure. Secondly, Linear models perform poorly on the ODE benchmark. Contrary to their performance on standard datasets, linear models such as DLinear are no longer competitive here. Averaged over all forecasting horizons, DLinear ranks worst in 5 out of the 6 *ODE datasets*, with an average rank of 9.83. Compared to Crossformer, DLinear suffers from performance drops ranging from 45% to 118%, with a median degradation of 58.3% and a mean of 74%.

In summary, the commonly held view that linear models are competitive with deep architectures does *not* generalize to datasets with strong inter-channel dynamics. Our findings underline the importance of evaluating forecasting models on datasets with diverse structural properties. In cases where inter-channel correlation is intrinsic to the data—as with ODE systems—model architectures must be explicitly designed to capture these dependencies in order to perform well.

6.4 Lookback Window Analysis

In this section, we extend our evaluation by analyzing the distribution of selected lookback window values, grouped by both datasets and models. This analysis highlights how an optimal lookback window varies significantly depending on the model type and dataset characteristics.

As shown in Figure 3, we report the frequency of best-performing lookback window across CI models, CD models, and datasets. In the Figure 3a, CI models exhibit substantial variation in their preferred lookback windows. For instance, DLinear consistently favors longer windows (336 or 720), while shorter windows are rarely optimal. This finding is soldified by the fact that overall, CI models choose 96 lookback window only 11% times. In contrast, Figure 3b shows that CD Models more often benefit from shorter windows, with 96 or 192 selected in 55% of the cases. These counts are aggregated across the four forecasting horizons (96, 192, 336, and 720).

Figure 3c shows that optimal lookback windows also vary widely by dataset. This aligns with expectations: for hourly datasets like ETTh1 and ETTh2, a window of 96 covers four days, often sufficient for effective forecasting. In contrast, for 15-minute interval datasets like ETTm1 and ETTm2, more timesteps are required to capture similar temporal spans. Accordingly, ETTm2 frequently favors longer windows larger than or equal to 336, while ETTh2 consistently prefers windows shorter than 336.

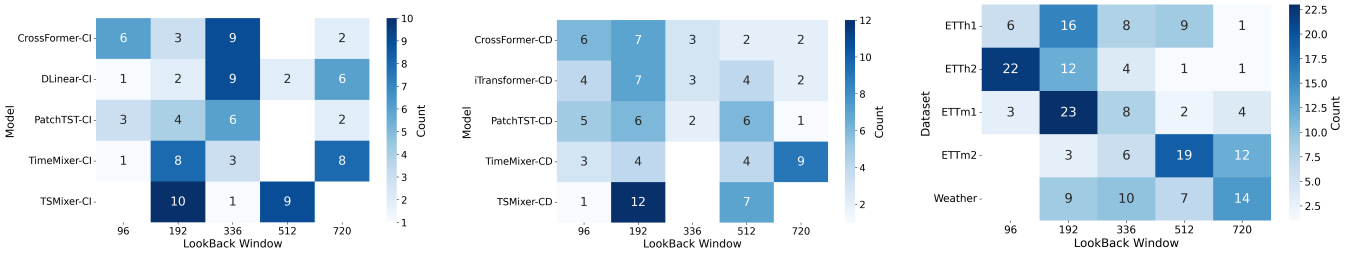
Overall, this analysis reinforces the importance of tuning the lookback window, not just across model types, but also in response to dataset-specific temporal characteristics. For per-horizon results, refer to the bar charts in the appendix.

7 CONCLUSION AND RECOMMENDATIONS

7.1 Conclusion

Model	ODE-6		Standard-7	
	CI	CD	CI	CD
PatchTST	**4**	2	**6**	1
TSMixer	0	**6**	**7**	0
Crossformer	0	**6**	**4**	3
TimeMixer	**2**	**4**	**4**	3
Totals	6	**18**	**21**	7

Table 4: Frequency of CI and CD variants outperforming each other across models and datasets.



(a) CI models **(b) CD models** **(c) Datasets**

Figure 3: Frequencies of best-performing lookback windows grouped by: (a) CI models, (b) CD models, and (c) datasets.

In this paper, as outlined in the introduction, our primary goal is to establish a fair and unbiased evaluation framework for multivariate TSF models. To achieve this, we make several critical adjustments to the experimental setup including: tuning the lookback window, analyzing statistical inter-channel correlations, and conducting empirical evaluations on datasets with varying characteristics. From this comprehensive setup, we draw the following key conclusions:

(1) **Tuning the lookback window is essential for fairness in multivariate TSF.** Fixing the lookback window (commonly at 96) can unfairly disadvantage univariate models. Allowing models to tune this parameter ensures a more balanced comparison and reveals meaningful performance differences. This insight directly enables the second key finding below.

(2) **Modeling channel dependencies is unnecessary for standard datasets.** For all models where a CI variant is available, it consistently outperforms the CD counterpart, by a margin of 21 to 7 on standard benchmarks. This is a significant result, especially considering that models like TSMixer, Crossformer, and TimeMixer are inherently designed to mix information across channels. Our findings indicate that inter-channel mixing is not a major contributor to performance on these datasets and can, in fact, degrade it (Table 4).

(3) **Dataset nature plays a critical role in multivariate TSF.** This aspect has been largely overlooked in prior work. Our experiments (Table 4) reveal that datasets with different underlying dynamics require different modeling strategies. For instance, *ODE Datasets* exhibit significantly higher inter-channel correlation, as confirmed by statistical tests. This justifies the use of multivariate models in such cases and is further supported by the superior performance of channel-dependent models on this dataset family.

7.2 Recommendations for Multivariate Time-Series Forecasting Evaluation

In this paper, we strive to come with a useful insights that can advance the field of multivariate time series forecasting and this was done through a detailed experiments and statistical analysis from which we draw the following recommendations:

- Do not use a fixed lookback window when a vast amount of historical data is available to ensure full potential of all models considered as seen in Section 6.1 and Section 6.3.
- For proper evaluation and before choosing model architecture, consider statistically testing the channel correlation of the datasets as shown in Section 6.2, it can help a lot in making an informed decision between deploying a CD or a CI model.
- As noted in Section 6.4, shorter lookback window favors in general CD models, which means for applications with limited amount of historical data, an architecture capturing channel dependency is generally recommended.

REFERENCES

[1] Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631.

[2] Chen, H.; Luong, V.; Mukherjee, L.; and Singh, V. 2025. SimpleTM: A Simple Baseline for Multivariate Time Series Forecasting. In *The Thirteenth International Conference on Learning Representations*.

[3] Chen, S.-A.; Li, C.-L.; Arik, S. O.; Yoder, N. C.; and Pfister, T. 2023. TSMixer: An All-MLP Architecture for Time Series Forecast-ing. *Transactions on Machine Learning Research*.

[4] Elbori, A.; and Abdalsmd, L. 2017. Simulation of double pendulum. *J. Softw. Eng. Simul*, 3(7): 1–13.

[5] Gilpin, W. 2021. Chaos as an interpretable benchmark for forecasting and data-driven modelling. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[6] Godahewa, R.; Bergmeir, C.; Webb, G. I.; Hyndman, R. J.; and Montero-Manso, P. 2021. Monash Time Series Forecasting Archive. In *Neural Information Processing Systems Track on Datasets and Benchmarks*.

[7] Granger, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424–438.

[8] Huang, S.; Zhao, Z.; Li, C.; and Bai, L. 2025. Timekan: Kan-based frequency decomposition learning architecture for long-term time series forecasting. In *The Thirteenth International Conference on Learning Representations*.

[9] Jadlovská, S.; and Sarnovskỳ, J. 2012. Classical double inverted pendulum—A complex overview of a system. In *2012 IEEE 10th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 103–108. IEEE.

[10] Kim, T.; Kim, J.; Tae, Y.; Park, C.; Choi, J.-H.; and Choo, J. 2021. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*.

[11] Lin, S.; Chen, H.; Wu, H.; Qiu, C.; and Lin, W. 2025. Temporal Query Network for Efficient Multivariate Time Series Forecasting. In *Forty-second International Conference on Machine Learning*.

[12] Lin, S.; Lin, W.; HU, X.; Wu, W.; Mo, R.; and Zhong, H. 2024. CycleNet: Enhancing Time Series Forecasting through Modeling Periodic Patterns. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[13] Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2024. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*.

[14] Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations*.

[15] Romond, P.-C.; Rustici, M.; Gonze, D.; and Goldbeter, A. 1999. Alternating oscillations and chaos in a model of two coupled biochemical oscillators driving successive phases of the cell cycle. *Annals of the New York Academy of Sciences*,

879(1): 180–193.
[16] Seabold, S.; and Perktold, J. 2010. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, 92–96.
[17] Tong, S.; and Yuan, J. 2025. Efficiently Enhancing Long-term Series Forecasting via Ultra-long Lookback Windows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 20912–20920.
[18] Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J. Y.; and ZHOU, J. 2024. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*.
[19] Wang, Y.; Wu, H.; Dong, J.; Liu, Y.; Long, M.; and Wang, J. 2024. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*.
[20] Yi, K.; Zhang, Q.; Fan, W.; Wang, S.; Wang, P.; He, H.; An, N.; Lian, D.; Cao, L.; and Niu, Z. 2023. Frequency-domain MLPs are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems*, 36: 76656–76679.
[21] Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9): 11121–11128.
[22] Zhang, Y.; and Yan, J. 2023. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In *The Eleventh International Conference on Learning Representations*.
[23] Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. arXiv:2012.07436.

A DESCRIPTION OF ALL DATASETS

In this section, we would like to show in detail the different datasets used in our experiments. Starting with the standard datasets which were explained a lot in literature, so we will just give a brief description of them. First ETT- datasets which representing indication of long term power deployment which are recorded for 2 different granalurities (15 mins for 'm' and 1 hour for 'h') for 2 different counties (hence the post-fix '1' and '2'). Electricity dataset follows the consumption of electricity over an extended amount of time (3 years) recorded hourly for 321 different households. Traffic dataset is representing the flow of traffic recorded through multiple sensors in an hourly frequency spanning over a period of approximately 2 years. All of this information as well as the train/validation/test split is shown in Table 5

Second type of datasets used in our experiments is based on the chaotic ODE Benchmark. Here, we elaborate more on the general common features of the used ODE datasets and we give further description of the underlying processes for generating them. For the generation process, we use a common protocol for the chosen 6 attractors which is generating 20 points per time unit, splitting the data into 7:1:2 train:validation:test split which is the common practice for the standard datasets as well. Finally, we generate a very long time series 60000 timesteps which is divided into time series forecasting samples though a rolling window (same approach that is adopted on TSF papers for the standard datasets). For reproducibility, we provide the code stating the random seed that should be used to produce the same datasets and subsequently the same results we got. For understanding more the chosen processes as well as the dimensionality of each dataset, please refer to the chaotic ODE Benchmark paper [5].

B HYPERPARAMETER TUNING

We evaluate various Transformer-based and MLP-based time series forecasting models (e.g., DLinear, TSMixer, PatchTST, Crossformer, and our proposed FaCT) under a unified set of hyperparameter search ranges as shown in Table Table 6.

For specific architectures, additional parameters that are model specific were tuned on the same range of parameters specified on the respective papers. For all the range [] or set of values {}, refer to Table 7.

To ensure easy and full reproducibility of our experiments, we include all the best performing hyperparameter sets corresponding to all models and datasets. Please see tables from Table 16 to Table 28 for the full list. We hope inclusion of these lists, make reproducibility of this paper's results straight forward.

C IMPLEMENTATION DETAILS

C.1 Libraries and Hardware

All models were implemented using the PyTorch library (version `2.4`) in Python (version `3.12.1`). For hyperparameter optimization, we employed Optuna (version `3.6.1`). Experiments were conducted on machines equipped with NVIDIA RTX 4090 and RTX 3090 GPUs.

C.2 Training Procedure

All models are trained using the Adam optimizer. We use mean squared error (MSE) as the default loss function, and each epoch iterates over mini-batches of size 32 by default. Training proceeds for a maximum of 100 epochs, or until early stopping is triggered if the validation loss fails to improve for 10 consecutive epochs. We use Optuna to conduct 20 trials per model, selecting configurations from the defined set of hyperparameters B that yield the best validation performance.

C.3 Granger Causality

Here, we list in more details the procedure to do the granger causality test which, alongside our code, should guide an easy usage of the framework for other datasets as well as easy reproduction of current results. We pick directly the first **1000 timesteps** from the loaded dataset for this statistical test purposes, which we think is reasonable sample size with regard to the lags considered here. First, we check the pearson correlation between each pair of channels, and if it is too high (greater than 0,95 in our case), then one of the two channels is dropped. This is done, because channels which are very similar can cause results that are misleading when testing pair of channels that are identical (that can easily yield a very high score that will push the average scores way higher than expected). After, the initial filtering, we get dataset with diverse channels that can still have strong correlation, but one more important aspect we address next is stationarity as grangar causality works best on stationary time series data (i.e. mean and variance should not be shifting over time). Here, we apply the augmented Dickey-Fuller test on each channel of the time series to check if it is stationary. If the channel is not stationary, we apply differencing which should help bring stationarity into the time series channel. Now we have our preprocessed data which is ready for the grangar causality to be applied on. We use the packages from `statsmodels` package [16].

D DERIVATION OF CI/CD VERSION OF MODELS

In our paper, 6 strong recent baselines are used throughout evaluation, they vary in complexity and structure. The aim of this

Dataset/Feature	Channels	Granularity	Train/val/test	Total Timesteps
ETTh (1/2)	7	1 hour	12/4/4 (months)	17420
ETTm (1/2)	7	15 mins	12/4/4 (months)	69680
electricity	321	1 hour	7:1:2	26304
traffic	862	1 hour	7:1:2	17544
weather	21	10 mins	7:1:2	52696

Table 5: Basic statistics about the standard datasets widely used in the time series forecasting literature. This table shows the number of features, frequency of recording the data, train/validation/test split as well as the total timesteps recorded in a given dataset.

Hyperparameter	Search Range
Learning Rate	$[10^{-7}, 10^{-2}]$
Hidden Dimension (d_model)	{128, 256, 512, 1024}
Feedforward Dimension (d_ff)	{128, 256, 512, 1024}
# Encoder/Mixer Layers	[1, 10]
Dropout	[0, 0.9]
Sequence Length (seq_len)	{96, 192, 336, 512, 720}

Table 6: Common Hyperparameter Search Ranges with all parameters being integer datatypes except for learning rate and dropout which span the whole range of floats on the respective range of values

Model	Hyperparameter	Range
PatchTST	Patch Size	{8, 16}
	Stride	{4, 8}
TSMixer	Hidden Size	{32, 64, 256, 1024}
Crossformer/FaCT	Segement Length	[3, 12]
	Baseline	{0, 1}
	Cross Factor	[3, 20]

Table 7: Model Specific Hyperparameter Search Ranges/set of possible values with all of them having a datatype of integer in this case.

appendix section is to ease the understanding of making CI and CD versions of the different baselines.

Simpler Linear/MLP-based Models. We start with the simpler models used in our experiments which are DLinear and TSMixer. DLinear as mentioned earlier is based on being the simplest model that applies a trend seasonality deceomposition followed by simpler linear layers applied on both the trend and the seasonal components [21]. For **TSMixer**, the authors had implemented a CD version which is the original version of the model and removed the channel mixing component which is called in their paper **TMix-Only**. We reimplemnted the model using pytorch and made sure of reproducing the paper results before using both versions in our experiments. Throughout our paper, we call both versions as TSMixer (CD) and TSMixer (CI) for the CD and CI versions respectively.

PatchTST. Now we move into transformer-based models, starting with PatchTST as one of the first transformer models to reestablish a boost of performance over the DLinear model. In their paper, the original version of the model is CI where the transformer encoder backbone is applied independently on different channels before the application of a linear projection layer. For the CD version, they briefly discuss it in the appendix of applying the same backbone but on a flattened dimension of both number of patches and channels which means the attention is now applied jointly over patches and channels. We use the original implementation of the paper but we had to add the part related to CD implementation ourselves as it was not readily available on the paper at the time of writing our paper.

iTransformer/Crossformer. Next, transformer model that proved to be a strong baseline is Crossformer which applies a Depth-Segment-Wise embedding before application of a Two Stage Attention (TSA). Finally a decoder layer is applied through cross attention between the positional embedding of output time info and encoder output. The TSA is applied on the segments dimension (temporal) as well as followed by the channel dimension which constitutes the original CD version of the model [22]. For a CI version, we just removed the second stage of the TSA to make the attention only applied on the segments dimension. We used authors implementation of the CD version and made the mentioned edits on the implementation to produce the CI version used in our experiments. For iTransformer [13], we don't introduce additional CI version to the existing CD version as otherwise the model will be simplified to a linear model which is already covered in this paper.

TimeMixer. One more prominent model proposed recently which proves be competitive on the standard datasets is TimeMixer [18]. In this model, the authors provide an MLP-based approach which is applied over multiple resolutions of the time series. This model original setting is CD where the MLP-mixing is done over both temporal and channel dimensions. For a CI version, the channel mixing component can be removed yielding a multiresolution temporal mixing based model. We use the authors implementation for both the CI and CD versions of the model.

E DETAILED RESULTS

In this appendix, we add the full results over all forecasting horizons for both types of introduced datasets. In Table 9, we can see in detail the results for the chaotic ODE datasets, confirming the results from the main paper that Crossformer is clearly the SoTA on this

benchmark with it being the best model in 15 out of possible 24 settings(for the CD version). For the standard datasets, the results are detailed in Table 8 which also confirms the state of the current datasets where no model is clearly the best even emphasized with DLinear being the best model on **5 occurrences**. This also fits the picture that was formulated on the table from the main paper text that when proper tuning is carried out for lookback window as well as all related hyperparameters, PatchTST remains competitive being the best overall **6 times**. To conclude how comprehensive and reproducible our experiments are, these results represents the mean of 5 random seeds {3001, 3002, 3003, 3004, 3005} MSE error on the test split of each dataset.

F DETAILED GRANGER CAUSALITY RESULTS

We report results across three dataset groups (i) standard benchmark datasets, (ii) chaotic ODE datasets, and (iii) the large-scale Monash forecasting collection. For datasets with a large number of columns, we restrict the Granger causality analysis to pairwise comparisons among 20 randomly selected channels.

For the standard benchmark datasets such as electricity and traffic, Granger causality tests reveal that these datasets exhibit moderate inter-channel correlation, with higher causality (F-score) predominantly at shorter lookback windows (lag=30) (Table 11, Table 12, and Table 14). This suggests that, for short lags, information from other channels can slightly aid forecasting, but as the lag increases, most predictive power lies within each channel's own history. This outcome aligns with the observation that univariate models (CI) can often perform well for these datasets at longer lookback settings, as there is diminishing marginal benefit from including other channels when window size grows.

In contrast, chaotic ODE datasets (Table 11, Table 12, and Table 14) consistently display higher Granger causality scores across all lag settings (with the exception of the electricity dataset), indicating robust and persistent dependencies between channels. This is particularly evident even at longer lookback windows. Such behavior is characteristic of deterministic physical systems or simulated multivariate dynamics where the evolution of one variable intrinsically influences the others. The implication is that multivariate models leveraging channel dependencies are especially suitable for these ODE datasets.

We also report results for the Monash repository's [6] large-scale real-world datasets (over 40 time series from diverse domains). We select 40 datasets from a total of 58 datasets after removing datasets with missing values. Granger causality analysis Table 13 reveals that only a minority—six of forty—show an inter-channel F-score above 10, and that F-scores tend to decline with increasing lag (except isolated cases like the Kaggle dataset). This pattern suggests that most Monash datasets lack strong lateral dependencies and instead behave like collections of independent univariate series. For instance, many resemble traffic datasets from the M1 or M3 collections: aggregations rather than integrated multivariate systems. As lags increase (e.g., at lag=196), additional channels yield minimal information gain, hence the F-score drops.

Overall, these findings highlight that Granger causality excels at disentangling the nature of multivariate time series dependencies. Standard benchmarks typically permit effective univariate treatment at larger lags, the ODE datasets consistently demand modeling of channel interactions, and most Monash datasets demonstrate little true multivariate dependence. These results underscore the value of focusing research on genuinely multivariate datasets, such as chaotic ODEs, to advance understanding and modeling of channel dependence in time series forecasting.

G FULL LOOKBACK WINDOW ANALYSIS

We show here the detailed results of best performing lookback windows on seperate figures for each of the forecasting horizons. We can see from Figure 4, Figure 5, Figure 6, and Figure 7 the analysis of the results based on forecasting horizons 96, 192, 336, and 720 respectively. We can see that on average the same insights noticed from the discussed forecasting horizon 96, still applies to the rest of experiments. The patterns discussed with respect to ETTh1, and ETTh2 tending to use smaller lookback windows still applies clearly on horizons 192 and 336. This changes when 720 horizon is deployed which makes sense as the difficulty of forecasting such a large amount of timesteps promotes the usage of a larger lookback windows for all datasets.

Regarding the tendency of CI models to benefit more from a larger lookback window than CD models, that still holds for the different horizons as seen in the different figures. With also more models on the larger forecasting horizon, such as 720, tend to use larger lookback window more often. This pattern follows the same pattern we discussed for the performance on the ETTh datasets. The bottom line from this analysis is to show as well the complexity of forecasting different multivariate datasets across different horizons which makes the tuning of the lookback window clearly the standard approach to go for.

H ODE DATASETS SAMPLE VISUALIZATION

In this section, a visualization is shown for the rest of the 6 ODE datasets used in this research. In Figure 8, a visualization of a 1000 timesteps from cell cycle based ODE is shown. Two different cyclin degradation activation are shown with channel 2 (middle top) activating the degradation of cyclin C1 levels (middle bottom). Same phenomena is shown clearly on the right 2 channels of the figure for cyclin C2 levels. In Figure 9, Figure 10, Figure 11, and Figure 12, we can see a visualization of the first 1000 time steps of the blinking rotlet, hopfield, lorenz and lorenz coupled time series based datasets. We can see clearly on these figures, how the dynamics of these different datasets generate complex yet real-world like time series data, with highs and lows from one channel triggering highs and lows in other channels.

I CELL CYCLE ODE DISCUSSION

We discussed already the interesting properties of Double Pendulum ODEs. In this part of the appendix, we would like to extend more the discussion on the coupling between the ODEs and how variables interact across time. Equations 1-6 presents the system of ODEs, we used to generate the cell cycle dataset [15]. These 6 equations show two different oscillators based on the concentration levels of cyclin (C_1 and C_2) and how they motivate the progression of the cell cycle. The cyclin concentration reflects directly on the activated Cyclin

Dataset / Model	Horizon	PatchTST		TSMixer		CrossFormer		DLin.	iTrans.	TimeMixer	
		CI	CD	CI	CD	CI	CD	CI	CD	CI	CD
ETTh1	96	**0.375**	0.381	**0.374**	0.404	0.396	**0.400**	**0.372**	0.413	**0.391**	0.426
	192	0.419	**0.407**	**0.432**	0.438	**0.452**	0.446	**0.406**	0.457	**0.434**	0.580
	336	**0.459**	0.460	**0.447**	0.471	**0.459**	0.480	**0.435**	0.500	**0.444**	0.500
	720	**0.437**	0.499	0.500	**0.499**	**0.601**	0.497	0.481	0.541	**0.498**	0.473
Weather	96	0.151	**0.150**	**0.152**	0.155	**0.150**	0.153	0.374	0.156	**0.158**	0.199
	192	**0.193**	0.197	**0.196**	0.202	0.199	**0.193**	0.210	0.201	**0.198**	0.289
	336	0.278	**0.247**	**0.247**	0.255	0.261	0.495*	0.255	0.263	**0.266**	0.343
	720	0.359	**0.336**	**0.330**	0.354	0.335	0.342	**0.316**	0.319	**0.342**	0.442
Electricity	96	**0.129***	0.141	**0.131**	0.137	0.139	0.150	0.135	0.135	**0.130**	0.143
	192	**0.147***	0.156	**0.149**	0.156	0.177	0.168	0.149	0.151	**0.149**	0.191
	336	**0.163***	0.173	**0.165**	0.176	0.194	0.182*	**0.164**	0.175	**0.171**	0.174
	720	**0.197***	0.210	**0.203**	0.210	0.261	0.251*	0.199	**0.196**	0.212	**0.197**
ETTh2	96	0.286	**0.285**	**0.291**	0.304	0.397	0.537	0.303	0.324	**0.285**	0.377
	192	**0.352**	0.381	**0.376**	0.392	**0.713**	0.794	0.397	0.396	**0.360**	0.459
	336	**0.392**	0.422	**0.402**	0.421	0.727	0.553	0.518	0.447	**0.410**	0.476
	720	**0.431**	0.438	0.444	**0.441**	1.126	0.880	0.811	0.441	**0.431**	0.603
ETTm1	96	**0.293**	0.306	**0.292**	0.294	0.309	0.364	0.299	0.314	**0.299**	0.366
	192	**0.334**	0.339	**0.333**	0.344	**0.398**	0.411	0.334	0.359	**0.330**	0.340
	336	**0.373**	0.400	**0.374**	0.389	0.399	0.463	0.368	0.405	**0.363**	0.468
	720	**0.425**	0.441	**0.426**	0.452	0.466	0.465	0.434	0.437	**0.418**	0.607
ETTm2	96	**0.161**	0.166	**0.163**	0.176	0.199	0.188	0.166	0.180	**0.166**	0.167
	192	0.221	**0.220**	**0.234**	0.235	**0.270**	0.305	0.236	0.230	**0.220**	0.242
	336	0.287	**0.273**	**0.271**	0.275	0.820	0.660	0.296	0.287	**0.272**	0.292
	720	**0.363**	0.375	**0.360**	0.414	0.442	0.787	0.458	0.369	**0.361**	0.384
Traffic	96	0.360*	OOM	0.386	**0.384**	OOM	0.514*	0.395	OOM	OOM	0.360*
	192	0.375*	OOM	**0.392**	0.405	OOM	0.549*	0.406	OOM	OOM	0.375*
	336	0.385*	OOM	**0.407**	0.423	OOM	0.530*	0.436	OOM	OOM	0.385*
	720	0.43*	OOM	**0.443**	0.457	OOM	0.573*	0.466	OOM	OOM	0.43*

Table 8: In this table, the full results over different horizons are shown over all the 7 standard datasets. These results were average of 5 different random seeds {3001,3002,3003,3004,3005}. The best result for each variant per model is in bold, while the best results overall for each horizon per dataset is highlighted in blue. we reported the results with "*" from TimeMixer paper from the experiment where they tune the lookback window as well as the hyperparameters (please refer to appendix E in [18]).

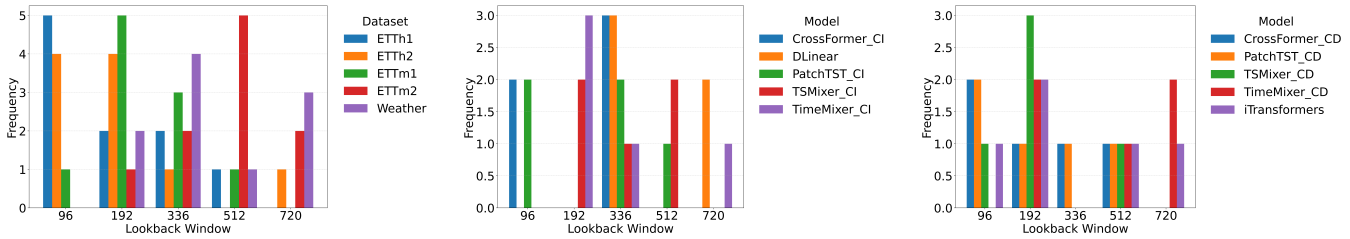


Figure 4: Analysis of Frequencies of Best Performing Lookback Windows Grouped by : (a) datasets, (b) CI models, and (c) CD models. All recorded for the forecasting horizon 96.

Dependent Kinases (cdk1, cdk2) which represented in the equations by M_1 and M_2 respectively. It also affect directly enzymes X_1 and X_2. v_{ij} and v_{dj} are constants that are multiplied in equation 1 for $j = 1$ and equation 4 for $j = 2$ as a rate of cyclin synthesis. In equations 1, we can see clearly the coupling between the ODE variables C_1 and M_2. Whereas, in equation 4, we can see the coupling between C_2 and M_1. All of this shows how information from one channel can cause spikes or change of trend in other channels, showing

Dataset / Model	Horizon	PatchTST		TSMixer		CrossFormer		DLin.	iTrans.	TimeMixer	
		CI	CD	CI	CD	CI	CD	CI	CD	CI	CD
Lorenz	96	0.658	**0.658**	0.747	**0.729**	0.331	**0.266**	0.884	0.433	0.398	**0.338**
	192	**0.825**	0.832	0.862	**0.847**	**0.623**	0.631	0.922	0.712	**0.662**	0.884
	336	**0.906**	**0.906**	0.927	**0.926**	0.792	**0.762**	0.950	0.852	**0.878**	0.895
	720	**0.965**	0.970	0.983	**0.971**	0.920	**0.914**	0.978	0.954	0.962	**0.929**
BlinkingRotlet	96	0.162	**0.156**	0.322	**0.228**	0.064	**0.045**	0.374	0.112	0.116	**0.076**
	192	0.322	**0.321**	0.510	**0.429**	0.230	**0.210**	0.504	0.299	**0.272**	0.315
	336	**0.511**	0.540	0.727	**0.587**	0.491	**0.466**	0.576	0.481	0.599	**0.500**
	720	0.703	**0.688**	0.759	**0.703**	0.576	**0.524**	0.635	0.623	0.710	**0.622**
CellCycle	96	**0.263**	0.311	0.513	**0.505**	0.036	**0.035**	0.866	0.227	0.110	**0.071**
	192	**0.580**	0.624	0.791	**0.740**	**0.244**	0.275	0.932	0.515	0.475	**0.370**
	336	**0.768**	0.795	0.894	**0.876**	0.612	**0.556**	0.957	0.688	0.742	**0.719**
	720	**0.931**	0.939	0.970	**0.963**	**0.825**	0.844	0.984	0.894	**0.867**	0.892
DoublePendulum	96	0.322	**0.278**	0.541	**0.461**	0.090	**0.083**	0.667	0.314	0.250	**0.110**
	192	**0.551**	0.594	**0.713**	0.753	**0.418**	0.466	0.762	0.534	0.440	**0.380**
	336	**0.806**	0.825	0.847	**0.808**	0.802	**0.712**	0.856	0.807	0.702	**0.650**
	720	**0.933**	0.975	0.973	**0.926**	0.902	**0.904**	0.933	0.952	**0.901**	1.018
Hopfield	96	0.156	**0.073**	0.268	**0.185**	0.059	**0.049**	0.472	0.046	0.054	**0.036**
	192	0.311	**0.216**	0.410	**0.321**	0.162	**0.155**	0.641	0.128	0.118	**0.102**
	336	0.486	**0.439**	0.571	**0.494**	**0.399**	0.404	0.759	0.280	0.567	**0.199**
	720	0.727	**0.654**	0.778	**0.742**	0.721	**0.655**	0.886	0.663	**0.699**	0.717
LorenzCoupled	96	0.696	**0.587**	0.840	**0.745**	0.372	**0.276**	0.919	0.610	**0.458**	0.818
	192	**0.883**	0.901	0.949	**0.895**	0.687	**0.635**	0.959	0.840	**0.795**	0.947
	336	**0.957**	0.971	0.997	**0.966**	0.813	**0.810**	0.978	0.924	**0.893**	0.984
	720	**0.990**	1.007	1.013	**0.994**	**0.926**	0.942	0.996	0.984	**0.969**	1.003

Table 9: In this table, the full results over different horizons are shown over all the 6 chaotic ODE datasets. These results were average of 5 different random seeds {3001,3002,3003,3004,3005}. The best result for each variant per model is in bold, while the best results overall for each horizon per dataset is highlighted in blue. we reported the results with "*" from TimeMixer paper from the experiment where they tune the lookback window as well as the hyperparameters (please refer to appendix E in [18]).

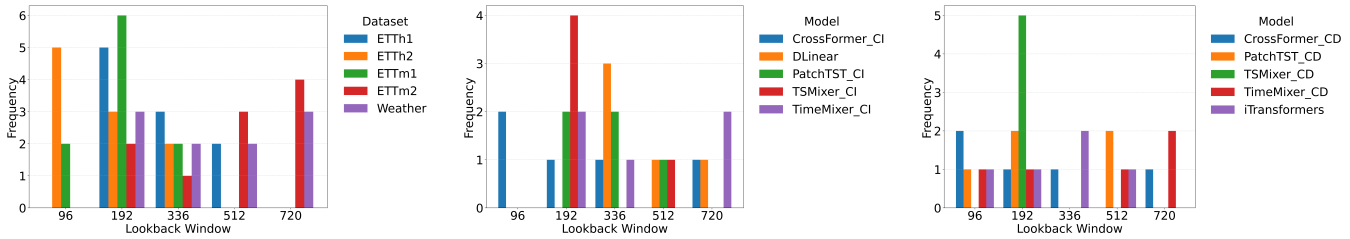


Figure 5: Analysis of Frequencies of Best Performing Lookback Windows Grouped by : (a) datasets, (b) CI models, and (c) CD models. All recorded for the forecasting horizon 192.

clear channel dependence in the generated time series data. Also, this section is valuable to show the very important application of such equation on a real life biomedical process which elevates the importance of our study more. Please note that these equations were taken from [15] and for more details, we encourage the reader to go through their work as well.

$$\frac{dC_1}{dt} = \frac{v_{i1}K_{im1}}{K_{im1} + M_2} - \frac{v_{d1}X_1C_1}{K_{d1} + C_1} - k_{d1}C_1 \tag{8}$$

$$\frac{dM_1}{dt} = \frac{V_1\,(1 - M_1)}{K_1 + (1 - M_1)} - \frac{V_2M_1}{K_2 + M_1} \tag{9}$$

$$\frac{dX_1}{dt} = \frac{V_3\,(1 - X_1)}{K_3 + (1 - X_1)} - \frac{V_4X_1}{K_4 + X_1} \tag{10}$$

$$\frac{dC_2}{dt} = \frac{v_{i2}K_{im2}}{K_{im2} + M_1} - \frac{v_{d2}X_2C_2}{K_{d2} + C_2} - k_{d2}C_2 \tag{11}$$

Dataset / Model	PatchTST		TSMixer		CrossFormer		DLin.	iTrans.	TimeMixer	
	CI	CD	CI	CD	CI	CD	CI	CD	CI	CD
Lorenz	0.0033	0.0025	0.0020	0.0115	0.0002	0.0095	0.0065	0.0303	0.0267	0.0066
BlinkingRotlet	0.0063	0.0072	0.0063	0.0083	0.0023	0.0171	0.0085	0.0580	0.0146	0.0294
CellCycle	0.0030	0.0040	0.0030	0.0047	0.0000	0.0093	0.0163	0.1127	0.0126	0.0818
DoublePendulum	0.0070	0.0067	0.0070	0.0305	0.0005	0.0110	0.0088	0.0136	0.0026	0.0052
Hopfield	0.0057	0.0070	0.0057	0.0075	0.0007	0.0049	0.0129	0.0262	0.3732	0.0076
LorenzCoupled	0.0030	0.0045	0.0030	0.0075	0.0000	0.0127	0.0056	0.0565	0.0088	0.0088
ETTh1	0.0030	0.0092	0.0030	0.0015	0.0002	0.0425	0.0222	0.0156	0.0012	0.0175
Weather	0.0058	0.0010	0.0033	0.0053	0.0004	0.0061	0.0049	0.0172	0.0037	0.0039
Electricity	OOM	OOM	0.0003	0.0020	0.0000	0.0041	0.0124	0.0053	0.0017	0.0037
ETTh2	0.0040	0.0047	0.0040	0.0088	0.0011	0.1265	0.1307	0.0153	0.0036	0.0389
ETTm1	0.0023	0.0047	0.0023	0.0033	0.0000	0.0131	0.0367	0.0059	0.0023	0.0161
ETTm2	0.0010	0.0020	0.0010	0.0040	0.0004	0.0928	0.0364	0.0007	0.0022	0.1017
Traffic	OOM	OOM	0.0007	0.0025	0.0000	OOM	OOM	OOM	OOM	OOM

Table 10: In this table, all standard deviation values are averaged over the 4 used horizons for all datasets. For each horizon, the standard deviation results were acquired by running the model for 5 different seeds with the following values {3001,3002,3003,3004,3005}

. Note for some heavier models and datasets, we got Out of Memory error, specifically speaking some of the runs for electricity and traffic datasets.

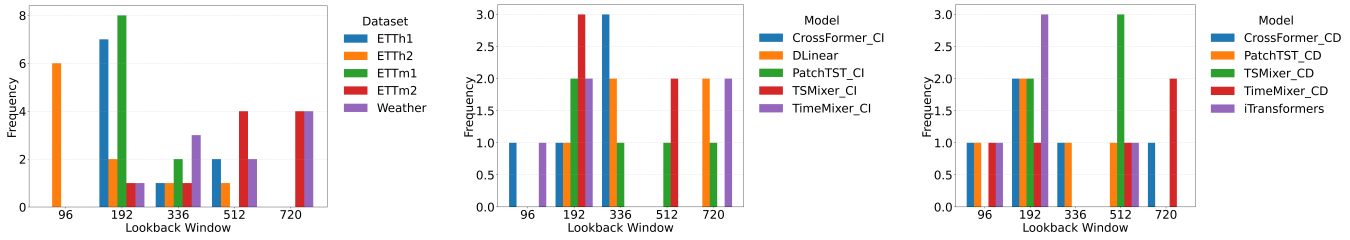


Figure 6: Analysis of Frequencies of Best Performing Lookback Windows Grouped by : (a) datasets, (b) CI models, and (c) CD models. All recorded for the forecasting horizon 336.

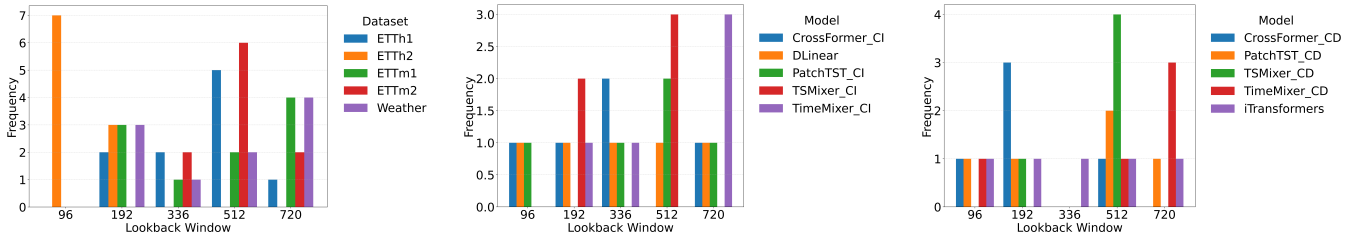


Figure 7: Analysis of Frequencies of Best Performing Lookback Windows Grouped by : (a) datasets, (b) CI models, and (c) CD models. All recorded for the forecasting horizon 720.

$$\frac{dM_2}{dt} = \frac{U_1\,(1-M_2)}{H_1+(1-M_2)} - \frac{U_2 M_2}{H_2+M_2} \tag{12}$$

$$\frac{dX_2}{dt} = \frac{U_3\,(1-X_2)}{H_3+(1-X_2)} - \frac{U_4 X_2}{H_4+X_2} \tag{13}$$

$$V_1 = \frac{C_1}{K_{c1}+C_1} V_{M_1}, \qquad V_3 = M_1 \cdot V_{M_3} \tag{14}$$

$$U_1 = \frac{C_2}{K_{c2}+C_2} U_{M_1}, \qquad U_3 = M_2 \cdot UM_3 \tag{15}$$

J ABLATION STUDY ON REVIN

In this section, we include an important ablation study where we test the importance of using Reversible Instance Normalization (RevIN) [10] on 2 standard datasets and 2 datasets from the chaotic benchmark. Crossformer was originally proposed without RevIN,

Dataset / Metric	Average F-score (30)	H_0 rejected (%, p=0.05)
Double Pendulum	39.60	100%
Lorenz	1378.54	83%
Lorenz Coupled	579.23	97%
Cell Cycle	6.78	50%
Blinking Rotlet	3.98	33%
HopField	2.86	80%
ETTm1	1.48	50%
ETTm2	1.31	30%
ETTh1	1.87	60%
ETTh2	1.94	70%
Weather	1.55	33%
Electricity* (Ch=20)	4.34	93%
Traffic* (Ch=20)	4.71	94%

Table 11: Average F-scores and percentage of H_0 rejected (p=0.05) across datasets (lag=30).

Dataset / Metric	Average F-score (96)	H_0 rejected (%, p=0.05)
Double Pendulum	10.21	100%
Lorenz	353.25	83%
Lorenz Coupled	160.95	66%
Cell Cycle	2.61	53%
Blinking Rotlet	2.59	33%
HopField	1.92	70%
ETTm1	1.13	33%
ETTm2	1.11	15%
ETTh1	1.12	20%
ETTh2	1.19	36%
Weather	1.12	30%
Electricity* (Ch=20)	1.92	84%
Traffic* (Ch=20)	2.22	94%

Table 12: Average F-scores and percentage of H_0 rejected (p=0.05) across datasets (lag=96).

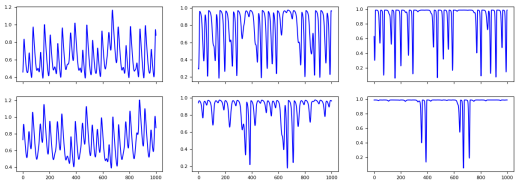


Figure 8: Time series visualization of the cell cycle ODE-based time series sample.

so we added the RevIN component to both the CI and CD versions to measure how much of an effect window size this can have on different types of datasets. Moreover we included one more strong baseline in PatchTST in the study measuring the effect of RevIN on its main CI version.

We got 2 interesting patterns that can be seen also from table Table 15 which are the following:

- Introducing revin helps to improve the performance on the both variants of Crossformer on the standard datasets.

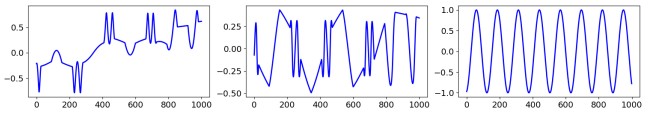


Figure 9: Time series visualization of the blinking rotlet ODE-based time series sample.

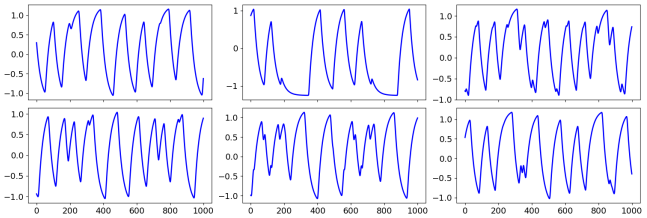


Figure 10: Time series visualization of the hopfield ODE-based time series sample.

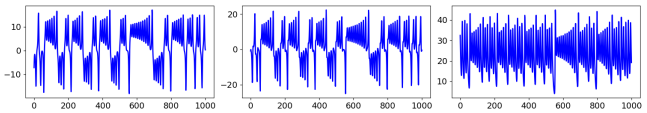


Figure 11: Time series visualization of the lorenz ODE-based time series sample.

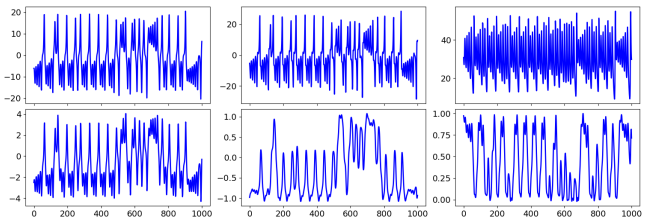


Figure 12: Time series visualization of the Lorenz Coupled ODE-based time series sample.

- On the contrary on the chaotic ODEs the model without revin is performing better than its counterpart with revin.

These findings shows that in general as confirmed in the literature RevIN can help in the performance of different model on standard datasets. While this being not totally true for PatchTST, it has been true for most models in the literature [3]. Important as emphasized throughout this paper, that is not easy task to generalize to different types of datasets, and that is no different for RevIN as it deteriorates performance if introduced on Crossformer on the chaotic ODE datasets. This shows that the different nature of Chaotic ODEs with the complex patterns can not benefit from the normalization+denormalization technique introduced in RevIN.

Monash Dataset	7	30	96	192
australian electricity demand dataset	29.38	18.06	2.71	1.62
bitcoin without missing values	1.88	1.44	1.66	2.47
car parts without missing values	1.25			
cif 2016	1.90			
dominick	1.26			
electricity hourly	32.55	4.66	1.96	1.21
electricity weekly	1.18	1.11		
fred md	1.92	1.49	1.32	1.14
hospital	1.38			
kaggle web traffic without missing values	3.02	3.25	12.96	45.63
kaggle web traffic weekly	1.14	1.56		
kdd cup 2018 without missing values	2.48	1.53	1.18	1.08
london smart meters without missing values	1.69	1.26	1.07	1.03
m1 monthly	2.02			
m1 quarterly				
m1 yearly				
m3 monthly	1.59			
m3 other	1.46			
m3 quarterly				
m3 yearly				
m4 daily	1.10	1.05		
m4 hourly	18.06	3.48	1.96	1.20
m4 monthly	1.76			
m4 quarterly	1.06			
m4 weekly	1.14			
m4 yearly				
nn5 daily without missing values	9.45	2.10	1.17	1.02
nn5 weekly	1.17	1.03		
pedestrian counts	10.85	3.14	2.65	
rideshare without missing values	3.73	2.15	1.31	
solar 10 minutes	8.63	3.87	1.52	1.56
solar weekly	0.78			
temperature rain without missing values	2.85	1.54	1.20	1.06
tourism monthly	12.12	1.58		
tourism quarterly	1.98			
tourism yearly				
traffic hourly	16.00	4.44	2.08	1.25
traffic weekly	1.12	1.21		
vehicle trips without missing values	1.38			
weather	1.22	1.17	1.12	1.03

Table 13: Average F-scores across 40 Monash datasets for lags 7, 30, 90, 192 .

Dataset / Metric	Average F-score (192)	H_0 rejected (%, p=0.05)
Double Pendulum	6.35	100%
Lorenz	126.38	83%
Lorenz Coupled	72.09	57%
Cell Cycle	1.58	53%
Blinking Rotlet	3.76	33%
HopField	1.55	50%
ETTm1	1.04	23%
ETTm2	0.99	0%
ETTh1	1.07	7%
ETTh2	1.06	17%
Weather	1.82	32%
Electricity* (Ch=20)	1.21	43%
Traffic* (Ch=20)	1.39	69%

Table 14: Average F-scores and percentage of rejected H_0 (p=0.05) across datasets (lag=192). Missing values indicate that the Granger Causality failed due to statistically limited number of available time points for a proper analysis.

Dataset / Model	PatchTST_CI		CrossFormer_CD		CrossFormer_CI	
	with RevIN	w/o RevIN	with RevIN	w/o RevIN	with RevIN	w/o RevIN
Hopfield	**0.151**	0.156	0.058	**0.049**	0.073	**0.059**
LorenzCoupled	**0.689**	0.696	0.453	**0.276**	0.515	**0.372**
ETTh2	0.298	**0.287**	**0.322**	0.537	**0.318**	0.397
ETTm1	0.314	**0.313**	**0.339**	0.364	**0.297**	0.309

Table 15: This table shows an ablation study on the reversible instance normalization technique [10] which showed to improve performance for most model on the standard datasets. The best variant of model is in bold and best overall for each dataset is highlighted in blue. These results are based on the forecasting horizon 96.

Model	Setting	Best Hyper. Params
PatchTST	CI-96	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
PatchTST	CD-96	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
TSMixer	CI-96	lr:0.0001,num_blocks:5,hidden_size:32,dropout:0.336,activation:relu,seq_len:512
TSMixer	CD-96	lr:0.0010,num_blocks:1,hidden_size:64,dropout:0.033,activation:relu,seq_len:192
CrossFormer	CI-96	lr: 0.0002, e_layers: 7, d_ff: 256, d_model: 512, dropout: 0.248, seg_len: 7, baseline: 0, cross_factor: 20, seq_len: 336
CrossFormer	CD-96	lr: 0.0003, e_layers: 7, d_ff: 256, d_model: 512, dropout: 0.285, seg_len: 7, baseline: 0, cross_factor: 5, seq_len: 336
DLinear	CI-96	lr: 0.0026, seq_len: 720
iTransformers	CD-96	lr:1.60e-05, e_layers:9, d_ff:2048, d_model:1024, dropout:0.012, seq_len:192
TimeMixer	CI-96	lr:0.0028, d_ff:256, d_model:128, e_layers:4, seq_len:192
TimeMixer	CD-96	lr:0.0023, d_ff:256, d_model:128, e_layers:5, seq_len:192
PatchTST	CI-192	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
PatchTST	CD-192	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
TSMixer	CI-192	lr:0.0005,num_blocks:8,hidden_size:1.02e+03,dropout:0.705,activation:relu,seq_len:512
TSMixer	CD-192	lr:0.0003,num_blocks:1,hidden_size:64,dropout:0.205,activation:gelu,seq_len:512
CrossFormer	CI-192	lr: 0.0003, e_layers: 7, d_ff: 256, d_model: 512, dropout: 0.312, seg_len: 7, baseline: 0, cross_factor: 18, seq_len: 336
CrossFormer	CD-192	lr: 0.0003, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.266, seg_len: 7, baseline: 0, cross_factor: 7, seq_len: 192
DLinear	CI-192	lr: 0.0055, seq_len: 512
iTransformers	CD-192	lr:0.0006, e_layers:9, d_ff:2048, d_model:128, dropout:0.286, seq_len:720
TimeMixer	CI-192	lr:0.0023, d_ff:256, d_model:128, e_layers:5, seq_len:192
TimeMixer	CD-192	lr:0.0028, d_ff:256, d_model:128, e_layers:4, seq_len:192
PatchTST	CI-336	lr':1.96e-07, 'e_layers':5, 'd_ff': 256, 'd_model':1024, 'dropout':0.295,'fc_dropout':0.622, 'patch_size':8,'stride':8,'seq_len':720
PatchTST	CD-336	lr': 9.1666e-05, 'e_layers': 6, 'd_ff': 1024, 'd_model': 1024, 'dropout': 0.086, 'fc_dropout': 0.5099, 'patch_size': 16, 'stride': 8, 'seq_len': 192
TSMixer	CI-336	lr:0.0012,num_blocks:6,hidden_size:32,dropout:0.264,activation:relu,seq_len:512
TSMixer	CD-336	lr:0.0002,num_blocks:5,hidden_size:64,dropout:0.287,activation:relu,seq_len:512
CrossFormer	CI-336	lr: 0.0002, e_layers: 7, d_ff: 256, d_model: 256, dropout: 0.309, seg_len: 8, baseline: 0, cross_factor: 5, seq_len: 336
CrossFormer	CD-336	lr: 0.0002, e_layers: 7, d_ff: 256, d_model: 256, dropout: 0.309, seg_len: 8, baseline: 0, cross_factor: 18, seq_len: 336
DLinear	CI-336	lr: 0.0004, seq_len: 720
iTransformers	CD-336	lr:0.0006, e_layers:9, d_ff:2048, d_model:128, dropout:0.286, seq_len:720
TimeMixer	CI-336	lr:0.0028, d_ff:256, d_model:128, e_layers:4, seq_len:192
TimeMixer	CD-336	lr:0.0010, d_ff:1024, d_model:1024, e_layers:10, seq_len:192
PatchTST	CI-720	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
PatchTST	CD-720	lr:0.0003,e_layers:9,d_ff:256,d_model:128,dropout:0.228,fc_dropout:0.072,patch_size:8,stride:8,seq_len:336
TSMixer	CI-720	lr:0.0004,num_blocks:8,hidden_size:1.02e+03,dropout:0.754,activation:relu,seq_len:512
TSMixer	CD-720	lr:0.0001,num_blocks:2,hidden_size:32,dropout:0.297,activation:relu,seq_len:512
CrossFormer	CI-720	lr: 0.0002, e_layers: 7, d_ff: 512, d_model: 256, dropout: 0.317, seg_len: 7, baseline: 1, cross_factor: 17, seq_len: 96
CrossFormer	CD-720	lr: 2.20e-05, e_layers: 6, d_ff: 256, d_model: 512, dropout: 0.085, seg_len: 12, baseline: 1, cross_factor: 11, seq_len: 96
DLinear	CI-720	lr: 0.0042, seq_len: 512
iTransformers	CD-720	lr:0.0001, e_layers:8, d_ff:2048, d_model:1024, dropout:0.356, seq_len:192
TimeMixer	CI-720	lr:2.30e-05, d_ff:512, d_model:128, e_layers:1, seq_len:512
TimeMixer	CD-720	lr:0.0021, d_ff:256, d_model:128, e_layers:4, seq_len:192

Table 16: Hyperparameter settings for the BlinkingRotlet dataset. The best variant for each model is shown.

Model	Setting	Best Hyper. Params
PatchTST	CI-96	lr:0.0003,e_layers:10,d_ff:256,d_model:128,dropout:0.081,fc_dropout:0.390,patch_size:8,stride:8,seq_len:336
PatchTST	CD-96	lr:0.0003,e_layers:9,d_ff:256,d_model:128,dropout:0.228,fc_dropout:0.072,patch_size:8,stride:8,seq_len:336
TSMixer	CI-96	lr:0.0055,num_blocks:10,hidden_size:256,dropout:0.197,activation:gelu,seq_len:512
TSMixer	CD-96	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
CrossFormer	CI-96	lr: 0.0002, e_layers: 7, d_ff: 256, d_model: 512, dropout: 0.248, seg_len: 7, baseline: 0, cross_factor: 5, seq_len: 336
CrossFormer	CD-96	lr: 0.0002, e_layers: 7, d_ff: 256, d_model: 512, dropout: 0.248, seg_len: 7, baseline: 0, cross_factor: 5, seq_len: 336
DLinear	CI-96	lr: 0.0012, seq_len: 336
iTransformers	CD-96	lr:0.0001, e_layers:8, d_ff:2048, d_model:1024, dropout:0.356, seq_len:192
TimeMixer	CI-96	lr:0.0027, d_ff:256, d_model:256, e_layers:8, seq_len:96
TimeMixer	CD-96	lr:0.0017, d_ff:256, d_model:512, e_layers:4, seq_len:96
PatchTST	CI-192	lr:0.0003,e_layers:10,d_ff:256,d_model:128,dropout:0.081,fc_dropout:0.390,patch_size:8,stride:8,seq_len:336
PatchTST	CD-192	lr:0.0003,e_layers:9,d_ff:256,d_model:128,dropout:0.228,fc_dropout:0.072,patch_size:8,stride:8,seq_len:336
TSMixer	CI-192	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
TSMixer	CD-192	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
CrossFormer	CI-192	lr: 0.0002, e_layers: 7, d_ff: 256, d_model: 512, dropout: 0.248, seg_len: 7, baseline: 0, cross_factor: 5, seq_len: 336
CrossFormer	CD-192	lr: 0.0015, e_layers: 5, d_ff: 512, d_model: 128, dropout: 0.228, seg_len: 12, baseline: 1, cross_factor: 17, seq_len: 336
DLinear	CI-192	lr: 0.0066, seq_len: 720
iTransformers	CD-192	lr:1.30e-05, e_layers:4, d_ff:512, d_model:1024, dropout:0.440, seq_len:336
TimeMixer	CI-192	lr:0.0009, d_ff:1024, d_model:1024, e_layers:10, seq_len:192
TimeMixer	CD-192	lr:0.0021, d_ff:256, d_model:128, e_layers:4, seq_len:192
PatchTST	CI-336	lr': 0.00027, 'e_layers': 10, 'd_ff': 256, 'd_model': 128, 'dropout': 0.081, 'fc_dropout': 0.390, 'patch_size': 8, 'stride': 8, 'seq_len': 336
PatchTST	CD-336	lr': 1.441e-05, 'e_layers': 10, 'd_ff': 256, 'd_model': 1024, 'dropout': 0.0137, 'fc_dropout': 0.53896, 'patch_size': 16, 'stride': 8, 'seq_len': 192
TSMixer	CI-336	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
TSMixer	CD-336	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
CrossFormer	CI-336	lr: 0.0006, e_layers: 8, d_ff: 512, d_model: 128, dropout: 0.361, seg_len: 6, baseline: 1, cross_factor: 8, seq_len: 512
CrossFormer	CD-336	lr: 0.0015, e_layers: 5, d_ff: 512, d_model: 128, dropout: 0.228, seg_len: 12, baseline: 1, cross_factor: 17, seq_len: 336
DLinear	CI-336	lr: 0.0049, seq_len: 512
iTransformers	CD-336	lr:0.0004, e_layers:1, d_ff:2048, d_model:1024, dropout:0.388, seq_len:192
TimeMixer	CI-336	lr:0.0020, d_ff:512, d_model:256, e_layers:7, seq_len:192
TimeMixer	CD-336	lr:0.0021, d_ff:512, d_model:1024, e_layers:8, seq_len:96
PatchTST	CI-720	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
PatchTST	CD-720	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
TSMixer	CI-720	lr:0.0006,num_blocks:8,hidden_size:64,dropout:0.181,activation:relu,seq_len:192
TSMixer	CD-720	lr:0.0010,num_blocks:1,hidden_size:64,dropout:0.033,activation:relu,seq_len:192
CrossFormer	CI-720	lr: 0.0006, e_layers: 7, d_ff: 128, d_model: 256, dropout: 0.309, seg_len: 7, baseline: 1, cross_factor: 18, seq_len: 336
CrossFormer	CD-720	lr: 0.0002, e_layers: 6, d_ff: 128, d_model: 256, dropout: 0.302, seg_len: 7, baseline: 1, cross_factor: 5, seq_len: 512
DLinear	CI-720	lr: 0.0033, seq_len: 336
iTransformers	CD-720	lr:0.0001, e_layers:8, d_ff:2048, d_model:1024, dropout:0.356, seq_len:192
TimeMixer	CI-720	lr:0.0019, d_ff:512, d_model:512, e_layers:8, seq_len:96
TimeMixer	CD-720	lr:0.0022, d_ff:1024, d_model:512, e_layers:8, seq_len:96

Table 17: Hyperparameter settings for the CellCycle dataset. The best variant for each model is shown.

Model	Setting	Best Hyper. Params
PatchTST	CI-96	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
PatchTST	CD-96	lr:0.0003,e_layers:9,d_ff:256,d_model:128,dropout:0.228,fc_dropout:0.072,patch_size:8,stride:8,seq_len:336
TSMixer	CI-96	lr:0.0012,num_blocks:6,hidden_size:64,dropout:0.302,activation:relu,seq_len:192
TSMixer	CD-96	lr:0.0010,num_blocks:1,hidden_size:64,dropout:0.033,activation:relu,seq_len:192
CrossFormer	CI-96	lr: 0.0015, e_layers: 9, d_ff: 512, d_model: 128, dropout: 0.228, seg_len: 6, baseline: 0, cross_factor: 20, seq_len: 336
CrossFormer	CD-96	lr: 0.0003, e_layers: 7, d_ff: 256, d_model: 512, dropout: 0.285, seg_len: 7, baseline: 0, cross_factor: 5, seq_len: 336
DLinear	CI-96	lr: 0.0080, seq_len: 192
iTransformers	CD-96	lr:0.0001, e_layers:8, d_ff:2048, d_model:1024, dropout:0.356, seq_len:192
TimeMixer	CI-96	lr:0.0026, d_ff:256, d_model:256, e_layers:8, seq_len:192
TimeMixer	CD-96	lr:0.0018, d_ff:1024, d_model:128, e_layers:6, seq_len:192
PatchTST	CI-192	lr:1.53e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.112,fc_dropout:0.457,patch_size:16,stride:4,seq_len:192
PatchTST	CD-192	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
TSMixer	CI-192	lr:0.0012,num_blocks:4,hidden_size:64,dropout:0.323,activation:relu,seq_len:192
TSMixer	CD-192	lr:0.0032,num_blocks:7,hidden_size:1.02e+03,dropout:0.006,activation:relu,seq_len:192
CrossFormer	CI-192	lr: 0.0006, e_layers: 7, d_ff: 128, d_model: 256, dropout: 0.309, seg_len: 7, baseline: 1, cross_factor: 18, seq_len: 336
CrossFormer	CD-192	lr: 0.0009, e_layers: 8, d_ff: 512, d_model: 128, dropout: 0.193, seg_len: 9, baseline: 0, cross_factor: 7, seq_len: 336
DLinear	CI-192	lr: 0.0097, seq_len: 192
iTransformers	CD-192	lr:0.0002, e_layers:8, d_ff:2048, d_model:512, dropout:0.114, seq_len:192
TimeMixer	CI-192	lr:0.0027, d_ff:256, d_model:256, e_layers:8, seq_len:96
TimeMixer	CD-192	lr:0.0006, d_ff:256, d_model:256, e_layers:7, seq_len:96
PatchTST	CI-336	lr': 1.441e-05, 'e_layers': 10, 'd_ff': 256, 'd_model': 1024, 'dropout': 0.0137, 'fc_dropout': 0.539, 'patch_size': 16, 'stride': 8, 'seq_len': 192
PatchTST	CD-336	lr': 1.441e-05, 'e_layers': 10, 'd_ff': 256, 'd_model': 1024, 'dropout': 0.0137, 'fc_dropout': 0.5389, 'patch_size': 16, 'stride': 8, 'seq_len': 192
TSMixer	CI-336	lr:0.0012,num_blocks:8,hidden_size:1.02e+03,dropout:0.307,activation:relu,seq_len:192
TSMixer	CD-336	lr:0.0010,num_blocks:1,hidden_size:64,dropout:0.033,activation:relu,seq_len:192
CrossFormer	CI-336	lr: 0.0005, e_layers: 6, d_ff: 512, d_model: 256, dropout: 0.326, seg_len: 7, baseline: 1, cross_factor: 8, seq_len: 512
CrossFormer	CD-336	lr: 0.0002, e_layers: 7, d_ff: 512, d_model: 128, dropout: 0.309, seg_len: 11, baseline: 0, cross_factor: 8, seq_len: 336
DLinear	CI-336	lr: 0.0011, seq_len: 192
iTransformers	CD-336	lr:0.0001, e_layers:8, d_ff:2048, d_model:1024, dropout:0.356, seq_len:192
TimeMixer	CI-336	lr:0.0022, d_ff:1024, d_model:512, e_layers:8, seq_len:96
TimeMixer	CD-336	lr:0.0019, d_ff:256, d_model:256, e_layers:8, seq_len:96
PatchTST	CI-720	lr:0.0001,e_layers:9,d_ff:512,d_model:512,dropout:0.065,fc_dropout:0.539,patch_size:8,stride:4,seq_len:96
PatchTST	CD-720	lr:0.0002,e_layers:1,d_ff:1.02e+03,d_model:512,dropout:0.471,fc_dropout:0.710,patch_size:8,stride:8,seq_len:336
TSMixer	CI-720	lr:0.0005,num_blocks:7,hidden_size:1.02e+03,dropout:0.169,activation:relu,seq_len:192
TSMixer	CD-720	lr:0.0010,num_blocks:1,hidden_size:64,dropout:0.033,activation:relu,seq_len:192
CrossFormer	CI-720	lr: 0.0002, e_layers: 6, d_ff: 512, d_model: 128, dropout: 0.266, seg_len: 8, baseline: 1, cross_factor: 5, seq_len: 336
CrossFormer	CD-720	lr: 0.0001, e_layers: 3, d_ff: 256, d_model: 512, dropout: 0.123, seg_len: 8, baseline: 0, cross_factor: 7, seq_len: 336
DLinear	CI-720	lr: 0.0032, seq_len: 512
iTransformers	CD-720	lr:8.00e-06, e_layers:2, d_ff:2048, d_model:1024, dropout:0.012, seq_len:192
TimeMixer	CI-720	lr:0.0018, d_ff:512, d_model:512, e_layers:7, seq_len:96
TimeMixer	CD-720	lr:0.0020, d_ff:256, d_model:256, e_layers:8, seq_len:96

Table 18: Hyperparameter settings for the DoublePendulum dataset. The best variant for each model is shown.

Model	Setting	Best Hyper. Params
PatchTST	CI-96	lr:0.0001,e_layers:1,d_ff:512,d_model:256,dropout:0.572,fc_dropout:0.631,patch_size:16,stride:4,seq_len:96
PatchTST	CD-96	lr:0.0001,e_layers:3,d_ff:512,d_model:512,dropout:0.518,fc_dropout:0.802,patch_size:16,stride:4,seq_len:96
TSMixer	CI-96	lr:0.0097,num_blocks:9,hidden_size:32,dropout:0.748,activation:gelu,seq_len:336
TSMixer	CD-96	lr:0.0002,num_blocks:6,hidden_size:256,dropout:0.139,activation:gelu,seq_len:96
CrossFormer	CI-96	lr: 0.0002, e_layers: 5, d_ff: 512, d_model: 256, dropout: 0.212, seg_len: 8, baseline: 0, cross_factor: 10, seq_len: 96
CrossFormer	CD-96	lr: 0.0080, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.189, seg_len: 8, baseline: 0, cross_factor: 9, seq_len: 192
DLinear	CI-96	lr: 0.0094, seq_len: 336
iTransformers	CD-96	lr:0.0007, e_layers:3, d_ff:1024, d_model:512, dropout:0.539, seq_len:96
TimeMixer	CI-96	lr:0.0004, d_ff:256, d_model:256, e_layers:6, seq_len:192
TimeMixer	CD-96	lr:0.0002, d_ff:1024, d_model:512, e_layers:4, seq_len:512
PatchTST	CI-192	lr:7.27e-06,e_layers:4,d_ff:1.02e+03,d_model:512,dropout:0.053,fc_dropout:0.183,patch_size:16,stride:4,seq_len:192
PatchTST	CD-192	lr:0.0001,e_layers:3,d_ff:1.02e+03,d_model:128,dropout:0.705,fc_dropout:0.640,patch_size:16,stride:8,seq_len:512
TSMixer	CI-192	lr:0.0005,num_blocks:1,hidden_size:64,dropout:0.173,activation:relu,seq_len:192
TSMixer	CD-192	lr:0.0003,num_blocks:1,hidden_size:64,dropout:0.166,activation:relu,seq_len:192
CrossFormer	CI-192	lr: 0.0080, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.189, seg_len: 8, baseline: 0, cross_factor: 9, seq_len: 192
CrossFormer	CD-192	lr: 0.0080, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.189, seg_len: 8, baseline: 0, cross_factor: 9, seq_len: 192
DLinear	CI-192	lr: 0.0010, seq_len: 336
iTransformers	CD-192	lr:1.60e-05, e_layers:9, d_ff:2048, d_model:1024, dropout:0.012, seq_len:336
TimeMixer	CI-192	lr:1.60e-05, d_ff:512, d_model:256, e_layers:3, seq_len:336
TimeMixer	CD-192	lr:0.0001, d_ff:1024, d_model:256, e_layers:9, seq_len:512
PatchTST	CI-336	lr': 5.207e-05, 'e_layers': 1, 'd_ff': 512, 'd_model': 256, 'dropout': 0.572, 'fc_dropout': 0.6308, 'patch_size': 16, 'stride': 4, 'seq_len': 96
PatchTST	CD-336	lr': 2.920e-06, 'e_layers': 6, 'd_ff': 1024, 'd_model': 512, 'dropout': 0.342, 'fc_dropout': 0.404, 'patch_size': 16, 'stride': 4, 'seq_len': 192
TSMixer	CI-336	lr:0.0002,num_blocks:2,hidden_size:64,dropout:0.205,activation:relu,seq_len:512
TSMixer	CD-336	lr:0.0002,num_blocks:1,hidden_size:64,dropout:0.207,activation:relu,seq_len:192
CrossFormer	CI-336	lr: 0.0080, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.189, seg_len: 8, baseline: 0, cross_factor: 9, seq_len: 192
CrossFormer	CD-336	lr: 0.0080, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.189, seg_len: 8, baseline: 0, cross_factor: 9, seq_len: 192
DLinear	CI-336	lr: 0.0011, seq_len: 336
iTransformers	CD-336	lr:1.60e-05, e_layers:9, d_ff:2048, d_model:1024, dropout:0.012, seq_len:192
TimeMixer	CI-336	lr:1.50e-05, d_ff:256, d_model:256, e_layers:2, seq_len:512
TimeMixer	CD-336	lr:0.0009, d_ff:1024, d_model:512, e_layers:8, seq_len:512
PatchTST	CI-720	lr:1.78e-06,e_layers:2,d_ff:512,d_model:512,dropout:0.438,fc_dropout:0.405,patch_size:16,stride:4,seq_len:336
PatchTST	CD-720	lr:0.0024,e_layers:2,d_ff:1.02e+03,d_model:512,dropout:0.450,fc_dropout:0.374,patch_size:16,stride:4,seq_len:512
TSMixer	CI-720	lr:0.0002,num_blocks:2,hidden_size:64,dropout:0.205,activation:relu,seq_len:512
TSMixer	CD-720	lr:0.0003,num_blocks:1,hidden_size:64,dropout:0.169,activation:relu,seq_len:512
CrossFormer	CI-720	lr: 0.0080, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.189, seg_len: 8, baseline: 0, cross_factor: 9, seq_len: 192
CrossFormer	CD-720	lr: 0.0080, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.189, seg_len: 8, baseline: 0, cross_factor: 9, seq_len: 192
DLinear	CI-720	lr: 0.0078, seq_len: 512
iTransformers	CD-720	lr:1.30e-05, e_layers:8, d_ff:512, d_model:512, dropout:0.017, seq_len:336
TimeMixer	CI-720	lr:1.00e-05, d_ff:512, d_model:512, e_layers:1, seq_len:720
TimeMixer	CD-720	lr:0.0005, d_ff:512, d_model:256, e_layers:8, seq_len:512

Table 19: Hyperparameter settings for the ETTh1 dataset. The best variant for each model is shown.

Model	Setting	Best Hyper. Params
PatchTST	CI-96	lr:3.50e-05,e_layers:1,d_ff:512,d_model:256,dropout:0.571,fc_dropout:0.486,patch_size:16,stride:4,seq_len:96
PatchTST	CD-96	lr:0.0001,e_layers:3,d_ff:512,d_model:512,dropout:0.688,fc_dropout:0.393,patch_size:16,stride:4,seq_len:96
TSMixer	CI-96	lr:0.0005,num_blocks:1,hidden_size:64,dropout:0.166,activation:relu,seq_len:192
TSMixer	CD-96	lr:0.0011,num_blocks:6,hidden_size:1.02e+03,dropout:0.746,activation:gelu,seq_len:192
CrossFormer	CI-96	lr: 0.0003, e_layers: 5, d_ff: 512, d_model: 128, dropout: 0.707, seg_len: 3, baseline: 1, cross_factor: 10, seq_len: 96
CrossFormer	CD-96	lr: 0.0076, e_layers: 2, d_ff: 512, d_model: 128, dropout: 0.494, seg_len: 5, baseline: 1, cross_factor: 13, seq_len: 96
DLinear	CI-96	lr: 0.0010, seq_len: 336
iTransformers	CD-96	lr:0.0001, e_layers:3, d_ff:1024, d_model:512, dropout:0.518, seq_len:96
TimeMixer	CI-96	lr:0.0001, d_ff:256, d_model:256, e_layers:3, seq_len:192
TimeMixer	CD-96	lr:0.0002, d_ff:256, d_model:256, e_layers:6, seq_len:192
PatchTST	CI-192	lr:0.0003,e_layers:1,d_ff:256,d_model:128,dropout:0.553,fc_dropout:0.631,patch_size:8,stride:4,seq_len:336
PatchTST	CD-192	lr:0.0002,e_layers:3,d_ff:512,d_model:512,dropout:0.688,fc_dropout:0.594,patch_size:16,stride:4,seq_len:96
TSMixer	CI-192	lr:0.0097,num_blocks:5,hidden_size:32,dropout:0.114,activation:relu,seq_len:192
TSMixer	CD-192	lr:0.0001,num_blocks:2,hidden_size:64,dropout:0.252,activation:relu,seq_len:192
CrossFormer	CI-192	lr: 0.0076, e_layers: 2, d_ff: 512, d_model: 128, dropout: 0.494, seg_len: 5, baseline: 1, cross_factor: 13, seq_len: 96
CrossFormer	CD-192	lr: 0.0076, e_layers: 2, d_ff: 512, d_model: 128, dropout: 0.494, seg_len: 5, baseline: 1, cross_factor: 13, seq_len: 96
DLinear	CI-192	lr: 0.0034, seq_len: 336
iTransformers	CD-192	lr:0.0001, e_layers:3, d_ff:1024, d_model:512, dropout:0.518, seq_len:96
TimeMixer	CI-192	lr:0.0001, d_ff:256, d_model:256, e_layers:3, seq_len:192
TimeMixer	CD-192	lr:0.0004, d_ff:256, d_model:256, e_layers:4, seq_len:96
PatchTST	CI-336	lr': 3.495e-05, 'e_layers': 1, 'd_ff': 512, 'd_model': 256, 'dropout': 0.5706, 'fc_dropout': 0.486, 'patch_size': 16, 'stride': 4, 'seq_len': 96
PatchTST	CD-336	lr': 0.0002, 'e_layers': 3, 'd_ff': 512, 'd_model': 512, 'dropout': 0.6877, 'fc_dropout': 0.5936, 'patch_size': 16, 'stride': 4, 'seq_len': 96
TSMixer	CI-336	lr:0.0097,num_blocks:5,hidden_size:32,dropout:0.331,activation:relu,seq_len:192
TSMixer	CD-336	lr:0.0082,num_blocks:1,hidden_size:256,dropout:0.155,activation:relu,seq_len:512
CrossFormer	CI-336	lr: 0.0017, e_layers: 2, d_ff: 256, d_model: 512, dropout: 0.512, seg_len: 4, baseline: 1, cross_factor: 8, seq_len: 96
CrossFormer	CD-336	lr: 0.0076, e_layers: 2, d_ff: 512, d_model: 128, dropout: 0.494, seg_len: 5, baseline: 1, cross_factor: 13, seq_len: 96
DLinear	CI-336	lr: 0.0036, seq_len: 192
iTransformers	CD-336	lr:0.0001, e_layers:3, d_ff:1024, d_model:512, dropout:0.518, seq_len:96
TimeMixer	CI-336	lr:0.0072, d_ff:256, d_model:256, e_layers:8, seq_len:96
TimeMixer	CD-336	lr:0.0011, d_ff:256, d_model:256, e_layers:4, seq_len:96
PatchTST	CI-720	lr:0.0002,e_layers:3,d_ff:512,d_model:512,dropout:0.714,fc_dropout:0.650,patch_size:16,stride:4,seq_len:96
PatchTST	CD-720	lr:0.0001,e_layers:3,d_ff:512,d_model:512,dropout:0.518,fc_dropout:0.802,patch_size:16,stride:4,seq_len:96
TSMixer	CI-720	lr:0.0021,num_blocks:6,hidden_size:1.02e+03,dropout:0.141,activation:relu,seq_len:192
TSMixer	CD-720	lr:0.0010,num_blocks:1,hidden_size:64,dropout:0.033,activation:relu,seq_len:192
CrossFormer	CI-720	lr: 0.0083, e_layers: 2, d_ff: 512, d_model: 128, dropout: 0.646, seg_len: 6, baseline: 1, cross_factor: 17, seq_len: 96
CrossFormer	CD-720	lr: 0.0076, e_layers: 2, d_ff: 512, d_model: 128, dropout: 0.494, seg_len: 5, baseline: 1, cross_factor: 13, seq_len: 96
DLinear	CI-720	lr: 0.0093, seq_len: 96
iTransformers	CD-720	lr:0.0001, e_layers:3, d_ff:1024, d_model:512, dropout:0.518, seq_len:96
TimeMixer	CI-720	lr:0.0001, d_ff:512, d_model:128, e_layers:3, seq_len:192
TimeMixer	CD-720	lr:0.0033, d_ff:512, d_model:128, e_layers:8, seq_len:96

Table 20: Hyperparameter settings for the ETTh2 dataset. The best variant for each model is shown.

Model	Setting	Best Hyper. Params
PatchTST	CI-96	lr:0.0003,e_layers:9,d_ff:256,d_model:128,dropout:0.228,fc_dropout:0.072,patch_size:8,stride:8,seq_len:336
PatchTST	CD-96	lr:6.07e-06,e_layers:3,d_ff:256,d_model:1.02e+03,dropout:0.006,fc_dropout:0.642,patch_size:16,stride:8,seq_len:192
TSMixer	CI-96	lr:0.0097,num_blocks:2,hidden_size:1.02e+03,dropout:0.115,activation:relu,seq_len:192
TSMixer	CD-96	lr:0.0084,num_blocks:2,hidden_size:32,dropout:0.264,activation:relu,seq_len:512
CrossFormer	CI-96	lr: 0.0002, e_layers: 6, d_ff: 512, d_model: 128, dropout: 0.266, seg_len: 8, baseline: 1, cross_factor: 5, seq_len: 336
CrossFormer	CD-96	lr: 0.0001, e_layers: 6, d_ff: 512, d_model: 256, dropout: 0.272, seg_len: 7, baseline: 1, cross_factor: 17, seq_len: 96
DLinear	CI-96	lr: 0.0006, seq_len: 336
iTransformers	CD-96	lr:0.0005, e_layers:5, d_ff:512, d_model:128, dropout:0.462, seq_len:192
TimeMixer	CI-96	lr:0.0004, d_ff:256, d_model:128, e_layers:2, seq_len:192
TimeMixer	CD-96	lr:0.0009, d_ff:512, d_model:512, e_layers:8, seq_len:192
PatchTST	CI-192	lr:2.97e-06,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.081,fc_dropout:0.435,patch_size:16,stride:8,seq_len:192
PatchTST	CD-192	lr:2.97e-06,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.081,fc_dropout:0.435,patch_size:16,stride:8,seq_len:192
TSMixer	CI-192	lr:0.0097,num_blocks:2,hidden_size:1.02e+03,dropout:0.115,activation:relu,seq_len:192
TSMixer	CD-192	lr:0.0005,num_blocks:6,hidden_size:64,dropout:0.280,activation:relu,seq_len:192
CrossFormer	CI-192	lr: 0.0002, e_layers: 7, d_ff: 512, d_model: 256, dropout: 0.317, seg_len: 7, baseline: 1, cross_factor: 17, seq_len: 96
CrossFormer	CD-192	lr: 0.0002, e_layers: 7, d_ff: 512, d_model: 256, dropout: 0.317, seg_len: 7, baseline: 1, cross_factor: 17, seq_len: 96
DLinear	CI-192	lr: 0.0010, seq_len: 336
iTransformers	CD-192	lr:1.30e-05, e_layers:4, d_ff:512, d_model:1024, dropout:0.440, seq_len:336
TimeMixer	CI-192	lr:0.0002, d_ff:256, d_model:256, e_layers:4, seq_len:192
TimeMixer	CD-192	lr:0.0002, d_ff:512, d_model:1024, e_layers:7, seq_len:192
PatchTST	CI-336	lr': 0.0003, 'e_layers': 9, 'd_ff': 256, 'd_model': 128, 'dropout': 0.228, 'fc_dropout': 0.0716, 'patch_size': 8, 'stride': 8, 'seq_len': 336
PatchTST	CD-336	lr': 1.441e-05, 'e_layers': 10, 'd_ff': 256, 'd_model': 1024, 'dropout': 0.0137, 'fc_dropout': 0.5389, 'patch_size': 16, 'stride': 8, 'seq_len': 192
TSMixer	CI-336	lr:0.0006,num_blocks:8,hidden_size:1.02e+03,dropout:0.166,activation:relu,seq_len:192
TSMixer	CD-336	lr:0.0004,num_blocks:8,hidden_size:64,dropout:0.215,activation:relu,seq_len:192
CrossFormer	CI-336	lr: 0.0003, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.265, seg_len: 8, baseline: 0, cross_factor: 9, seq_len: 336
CrossFormer	CD-336	lr: 0.0080, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.189, seg_len: 8, baseline: 0, cross_factor: 9, seq_len: 192
DLinear	CI-336	lr: 0.0007, seq_len: 336
iTransformers	CD-336	lr:0.0004, e_layers:1, d_ff:2048, d_model:1024, dropout:0.388, seq_len:192
TimeMixer	CI-336	lr:0.0002, d_ff:256, d_model:256, e_layers:4, seq_len:192
TimeMixer	CD-336	lr:0.0021, d_ff:256, d_model:128, e_layers:4, seq_len:192
PatchTST	CI-720	lr:2.27e-06,e_layers:4,d_ff:1.02e+03,d_model:128,dropout:0.198,fc_dropout:0.356,patch_size:16,stride:8,seq_len:512
PatchTST	CD-720	lr:0.0019,e_layers:7,d_ff:1.02e+03,d_model:512,dropout:0.471,fc_dropout:0.375,patch_size:8,stride:8,seq_len:720
TSMixer	CI-720	lr:0.0097,num_blocks:2,hidden_size:1.02e+03,dropout:0.115,activation:relu,seq_len:192
TSMixer	CD-720	lr:0.0011,num_blocks:6,hidden_size:1.02e+03,dropout:0.339,activation:relu,seq_len:192
CrossFormer	CI-720	lr: 0.0001, e_layers: 6, d_ff: 128, d_model: 256, dropout: 0.272, seg_len: 7, baseline: 1, cross_factor: 20, seq_len: 336
CrossFormer	CD-720	lr: 0.0080, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.189, seg_len: 8, baseline: 0, cross_factor: 9, seq_len: 192
DLinear	CI-720	lr: 0.0097, seq_len: 192
iTransformers	CD-720	lr:0.0001, e_layers:8, d_ff:2048, d_model:128, dropout:0.319, seq_len:720
TimeMixer	CI-720	lr:1.00e-05, d_ff:512, d_model:512, e_layers:1, seq_len:720
TimeMixer	CD-720	lr:0.0001, d_ff:1024, d_model:1024, e_layers:2, seq_len:192

Table 21: Hyperparameter settings for the ETTm1 dataset. The best variant for each model is shown.

Model	Setting	Best Hyper. Params
PatchTST	CI-96	lr:0.0002,e_layers:7,d_ff:1.02e+03,d_model:1.02e+03,dropout:0.648,fc_dropout:0.175,patch_size:8,stride:4,seq_len:512
PatchTST	CD-96	lr:1.60e-06,e_layers:8,d_ff:256,d_model:1.02e+03,dropout:0.307,fc_dropout:0.623,patch_size:8,stride:4,seq_len:512
TSMixer	CI-96	lr:0.0097,num_blocks:5,hidden_size:32,dropout:0.289,activation:relu,seq_len:512
TSMixer	CD-96	lr:0.0003,num_blocks:1,hidden_size:64,dropout:0.077,activation:relu,seq_len:192
CrossFormer	CI-96	lr: 0.0003, e_layers: 6, d_ff: 512, d_model: 128, dropout: 0.298, seg_len: 7, baseline: 1, cross_factor: 5, seq_len: 336
CrossFormer	CD-96	lr: 0.0003, e_layers: 4, d_ff: 512, d_model: 128, dropout: 0.382, seg_len: 11, baseline: 1, cross_factor: 11, seq_len: 512
DLinear	CI-96	lr: 0.0020, seq_len: 720
iTransformers	CD-96	lr:1.20e-05, e_layers:2, d_ff:1024, d_model:512, dropout:0.083, seq_len:512
TimeMixer	CI-96	lr:1.10e-05, d_ff:512, d_model:256, e_layers:4, seq_len:336
TimeMixer	CD-96	lr:5.00e-06, d_ff:512, d_model:512, e_layers:3, seq_len:720
PatchTST	CI-192	lr:7.02e-07,e_layers:2,d_ff:1.02e+03,d_model:512,dropout:0.292,fc_dropout:0.673,patch_size:8,stride:4,seq_len:336
PatchTST	CD-192	lr:1.82e-06,e_layers:4,d_ff:1.02e+03,d_model:1.02e+03,dropout:0.471,fc_dropout:0.405,patch_size:16,stride:8,seq_len:512
TSMixer	CI-192	lr:0.0010,num_blocks:1,hidden_size:64,dropout:0.033,activation:relu,seq_len:192
TSMixer	CD-192	lr:0.0004,num_blocks:1,hidden_size:64,dropout:0.077,activation:relu,seq_len:192
CrossFormer	CI-192	lr: 0.0001, e_layers: 3, d_ff: 512, d_model: 128, dropout: 0.180, seg_len: 9, baseline: 1, cross_factor: 14, seq_len: 720
CrossFormer	CD-192	lr: 0.0001, e_layers: 3, d_ff: 512, d_model: 128, dropout: 0.180, seg_len: 9, baseline: 1, cross_factor: 14, seq_len: 720
DLinear	CI-192	lr: 0.0002, seq_len: 512
iTransformers	CD-192	lr:1.40e-05, e_layers:1, d_ff:1024, d_model:512, dropout:0.083, seq_len:512
TimeMixer	CI-192	lr:1.00e-05, d_ff:512, d_model:512, e_layers:1, seq_len:720
TimeMixer	CD-192	lr:0.0003, d_ff:512, d_model:128, e_layers:4, seq_len:720
PatchTST	CI-336	lr': 0.0002, 'e_layers': 7, 'd_ff': 1024, 'd_model': 1024, 'dropout': 0.6482, 'fc_dropout': 0.1748, 'patch_size': 8, 'stride': 4, 'seq_len': 512
PatchTST	CD-336	lr': 1.602e-06, 'e_layers': 8, 'd_ff': 256, 'd_model': 1024, 'dropout': 0.307, 'fc_dropout': 0.6229, 'patch_size': 8, 'stride': 4, 'seq_len': 512
TSMixer	CI-336	lr:0.0097,num_blocks:5,hidden_size:32,dropout:0.328,activation:relu,seq_len:512
TSMixer	CD-336	lr:0.0087,num_blocks:6,hidden_size:64,dropout:0.759,activation:gelu,seq_len:512
CrossFormer	CI-336	lr: 0.0026, e_layers: 1, d_ff: 512, d_model: 128, dropout: 0.897, seg_len: 6, baseline: 1, cross_factor: 5, seq_len: 336
CrossFormer	CD-336	lr: 0.0001, e_layers: 6, d_ff: 512, d_model: 256, dropout: 0.679, seg_len: 10, baseline: 1, cross_factor: 17, seq_len: 720
DLinear	CI-336	lr: 0.0038, seq_len: 720
iTransformers	CD-336	lr:8.00e-06, e_layers:6, d_ff:2048, d_model:128, dropout:0.184, seq_len:512
TimeMixer	CI-336	lr:1.00e-05, d_ff:512, d_model:512, e_layers:1, seq_len:720
TimeMixer	CD-336	lr:1.00e-05, d_ff:512, d_model:128, e_layers:1, seq_len:720
PatchTST	CI-720	lr:2.27e-06,e_layers:4,d_ff:1.02e+03,d_model:128,dropout:0.198,fc_dropout:0.356,patch_size:16,stride:8,seq_len:512
PatchTST	CD-720	lr:0.0032,e_layers:3,d_ff:1.02e+03,d_model:128,dropout:0.318,fc_dropout:0.402,patch_size:16,stride:8,seq_len:512
TSMixer	CI-720	lr:0.0097,num_blocks:5,hidden_size:32,dropout:0.289,activation:relu,seq_len:512
TSMixer	CD-720	lr:0.0086,num_blocks:9,hidden_size:32,dropout:0.759,activation:relu,seq_len:512
CrossFormer	CI-720	lr: 2.75e-05, e_layers: 8, d_ff: 512, d_model: 256, dropout: 0.895, seg_len: 10, baseline: 1, cross_factor: 12, seq_len: 720
CrossFormer	CD-720	lr: 0.0028, e_layers: 8, d_ff: 512, d_model: 256, dropout: 0.664, seg_len: 10, baseline: 0, cross_factor: 16, seq_len: 512
DLinear	CI-720	lr: 0.0022, seq_len: 336
iTransformers	CD-720	lr:9.00e-06, e_layers:2, d_ff:2048, d_model:128, dropout:0.280, seq_len:512
TimeMixer	CI-720	lr:1.40e-05, d_ff:512, d_model:256, e_layers:4, seq_len:336
TimeMixer	CD-720	lr:1.00e-05, d_ff:512, d_model:128, e_layers:1, seq_len:720

Table 22: Hyperparameter settings for the ETTm2 dataset. The best variant for each model is shown.

Model	Setting	Best Hyper. Params
PatchTST	CI-96	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
PatchTST	CD-96	lr:0.0003,e_layers:10,d_ff:256,d_model:128,dropout:0.081,fc_dropout:0.390,patch_size:8,stride:8,seq_len:336
TSMixer	CI-96	lr:0.0002,num_blocks:6,hidden_size:256,dropout:0.334,activation:relu,seq_len:512
TSMixer	CD-96	lr:0.0097,num_blocks:5,hidden_size:32,dropout:0.311,activation:relu,seq_len:512
CrossFormer	CI-96	lr: 0.0003, e_layers: 7, d_ff: 256, d_model: 512, dropout: 0.336, seg_len: 6, baseline: 0, cross_factor: 17, seq_len: 512
CrossFormer	CD-96	lr: 0.0001, e_layers: 3, d_ff: 256, d_model: 512, dropout: 0.123, seg_len: 8, baseline: 0, cross_factor: 7, seq_len: 336
DLinear	CI-96	lr: 0.0078, seq_len: 720
iTransformers	CD-96	lr:0.0002, e_layers:8, d_ff:2048, d_model:512, dropout:0.114, seq_len:192
TimeMixer	CI-96	lr:0.0015, d_ff:256, d_model:256, e_layers:7, seq_len:192
TimeMixer	CD-96	lr:0.0023, d_ff:256, d_model:128, e_layers:5, seq_len:192
PatchTST	CI-192	lr:0.0003,e_layers:10,d_ff:256,d_model:128,dropout:0.081,fc_dropout:0.390,patch_size:8,stride:8,seq_len:336
PatchTST	CD-192	lr:0.0003,e_layers:10,d_ff:256,d_model:128,dropout:0.081,fc_dropout:0.390,patch_size:8,stride:8,seq_len:336
TSMixer	CI-192	lr:0.0002,num_blocks:8,hidden_size:64,dropout:0.295,activation:relu,seq_len:512
TSMixer	CD-192	lr:0.0010,num_blocks:1,hidden_size:64,dropout:0.033,activation:relu,seq_len:192
CrossFormer	CI-192	lr: 0.0002, e_layers: 7, d_ff: 256, d_model: 512, dropout: 0.248, seg_len: 7, baseline: 0, cross_factor: 20, seq_len: 336
CrossFormer	CD-192	lr: 0.0002, e_layers: 7, d_ff: 256, d_model: 512, dropout: 0.248, seg_len: 7, baseline: 0, cross_factor: 20, seq_len: 512
DLinear	CI-192	lr: 0.0082, seq_len: 512
iTransformers	CD-192	lr:0.0002, e_layers:8, d_ff:2048, d_model:512, dropout:0.114, seq_len:192
TimeMixer	CI-192	lr:0.0023, d_ff:256, d_model:128, e_layers:5, seq_len:192
TimeMixer	CD-192	lr:0.0010, d_ff:1024, d_model:1024, e_layers:10, seq_len:192
PatchTST	CI-336	lr': 1.441e-05, 'e_layers': 10, 'd_ff': 256, 'd_model': 1024, 'dropout': 0.014, 'fc_dropout': 0.539, 'patch_size': 16, 'stride': 8, 'seq_len': 192
PatchTST	CD-336	lr:0.0004,num_blocks:8,hidden_size:1.02e+03,dropout:0.159,activation:relu,seq_len:512
TSMixer	CI-336	lr:0.0011,num_blocks:1,hidden_size:256,dropout:0.322,activation:relu,seq_len:512
TSMixer	CD-336	lr:0.0011,num_blocks:1,hidden_size:256,dropout:0.322,activation:relu,seq_len:512
CrossFormer	CI-336	lr: 0.0006, e_layers: 7, d_ff: 128, d_model: 256, dropout: 0.309, seg_len: 7, baseline: 1, cross_factor: 18, seq_len: 336
CrossFormer	CD-336	lr: 0.0011, e_layers: 8, d_ff: 512, d_model: 128, dropout: 0.357, seg_len: 7, baseline: 1, cross_factor: 8, seq_len: 512
DLinear	CI-336	lr: 0.0002, seq_len: 512
iTransformers	CD-336	lr:0.0001, e_layers:8, d_ff:2048, d_model:1024, dropout:0.356, seq_len:192
TimeMixer	CI-336	lr:0.0041, d_ff:256, d_model:1024, e_layers:6, seq_len:96
TimeMixer	CD-336	lr:0.0018, d_ff:1024, d_model:128, e_layers:6, seq_len:192
PatchTST	CI-720	lr:0.0003,e_layers:10,d_ff:256,d_model:128,dropout:0.081,fc_dropout:0.390,patch_size:8,stride:8,seq_len:336
PatchTST	CD-720	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
TSMixer	CI-720	lr:0.0002,num_blocks:8,hidden_size:1.02e+03,dropout:0.205,activation:relu,seq_len:512
TSMixer	CD-720	lr:0.0010,num_blocks:1,hidden_size:64,dropout:0.033,activation:relu,seq_len:192
CrossFormer	CI-720	lr: 0.0001, e_layers: 3, d_ff: 256, d_model: 512, dropout: 0.123, seg_len: 8, baseline: 0, cross_factor: 7, seq_len: 336
CrossFormer	CD-720	lr: 0.0002, e_layers: 6, d_ff: 512, d_model: 128, dropout: 0.266, seg_len: 8, baseline: 1, cross_factor: 5, seq_len: 336
DLinear	CI-720	lr: 0.0034, seq_len: 512
iTransformers	CD-720	lr:0.0001, e_layers:8, d_ff:2048, d_model:1024, dropout:0.356, seq_len:192
TimeMixer	CI-720	lr:0.0021, d_ff:256, d_model:256, e_layers:8, seq_len:512
TimeMixer	CD-720	lr:0.0010, d_ff:1024, d_model:1024, e_layers:10, seq_len:192

Table 23: Hyperparameter settings for the Hopfield dataset. The best variant for each model is shown.

Model	Setting	Best Hyper. Params
PatchTST	CI-96	lr:0.0003,e_layers:9,d_ff:256,d_model:128,dropout:0.228,fc_dropout:0.072,patch_size:8,stride:8,seq_len:336
PatchTST	CD-96	lr:0.0003,e_layers:9,d_ff:256,d_model:128,dropout:0.228,fc_dropout:0.072,patch_size:8,stride:8,seq_len:336
TSMixer	CI-96	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
TSMixer	CD-96	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
CrossFormer	CI-96	lr: 0.0002, e_layers: 7, d_ff: 256, d_model: 512, dropout: 0.309, seg_len: 3, baseline: 0, cross_factor: 18, seq_len: 336
CrossFormer	CD-96	lr: 0.0004, e_layers: 6, d_ff: 512, d_model: 256, dropout: 0.228, seg_len: 7, baseline: 0, cross_factor: 8, seq_len: 336
DLinear	CI-96	lr: 0.0093, seq_len: 96
iTransformers	CD-96	lr:0.0001, e_layers:8, d_ff:2048, d_model:1024, dropout:0.356, seq_len:192
TimeMixer	CI-96	lr:0.0028, d_ff:256, d_model:128, e_layers:4, seq_len:192
TimeMixer	CD-96	lr:0.0028, d_ff:256, d_model:128, e_layers:4, seq_len:192
PatchTST	CI-192	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
PatchTST	CD-192	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
TSMixer	CI-192	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
TSMixer	CD-192	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
CrossFormer	CI-192	lr: 0.0004, e_layers: 8, d_ff: 256, d_model: 256, dropout: 0.323, seg_len: 10, baseline: 0, cross_factor: 10, seq_len: 336
CrossFormer	CD-192	lr: 0.0002, e_layers: 6, d_ff: 128, d_model: 256, dropout: 0.302, seg_len: 7, baseline: 1, cross_factor: 5, seq_len: 336
DLinear	CI-192	lr: 0.0093, seq_len: 96
iTransformers	CD-192	lr:0.0011, e_layers:4, d_ff:2048, d_model:128, dropout:0.255, seq_len:720
TimeMixer	CI-192	lr:0.0013, d_ff:512, d_model:1024, e_layers:8, seq_len:96
TimeMixer	CD-192	lr:0.0016, d_ff:256, d_model:128, e_layers:5, seq_len:192
PatchTST	CI-336	lr': 0.0003, 'e_layers': 9, 'd_ff': 256, 'd_model': 128, 'dropout': 0.2277, 'fc_dropout': 0.0716, 'patch_size': 8, 'stride': 8, 'seq_len': 336
PatchTST	CD-336	lr': 1.441e-05, 'e_layers': 10, 'd_ff': 256, 'd_model': 1024, 'dropout': 0.0137, 'fc_dropout': 0.5389, 'patch_size': 16, 'stride': 8, 'seq_len': 192
TSMixer	CI-336	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
TSMixer	CD-336	lr:0.0080,num_blocks:7,hidden_size:64,dropout:0.204,activation:relu,seq_len:192
CrossFormer	CI-336	lr: 0.0003, e_layers: 7, d_ff: 256, d_model: 512, dropout: 0.285, seg_len: 7, baseline: 0, cross_factor: 5, seq_len: 336
CrossFormer	CD-336	lr: 0.0003, e_layers: 5, d_ff: 512, d_model: 256, dropout: 0.228, seg_len: 6, baseline: 0, cross_factor: 17, seq_len: 336
DLinear	CI-336	lr: 0.0093, seq_len: 96
iTransformers	CD-336	lr:0.0011, e_layers:4, d_ff:2048, d_model:128, dropout:0.255, seq_len:720
TimeMixer	CI-336	lr:0.0010, d_ff:1024, d_model:1024, e_layers:10, seq_len:192
TimeMixer	CD-336	lr:0.0009, d_ff:1024, d_model:256, e_layers:8, seq_len:192
PatchTST	CI-720	lr:0.0001,e_layers:3,d_ff:512,d_model:512,dropout:0.518,fc_dropout:0.802,patch_size:16,stride:4,seq_len:96
PatchTST	CD-720	lr:0.0001,e_layers:3,d_ff:512,d_model:512,dropout:0.518,fc_dropout:0.802,patch_size:16,stride:4,seq_len:96
TSMixer	CI-720	lr:0.0026,num_blocks:6,hidden_size:256,dropout:0.126,activation:gelu,seq_len:96
TSMixer	CD-720	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
CrossFormer	CI-720	lr: 0.0006, e_layers: 7, d_ff: 128, d_model: 256, dropout: 0.309, seg_len: 7, baseline: 1, cross_factor: 18, seq_len: 336
CrossFormer	CD-720	lr: 0.0006, e_layers: 8, d_ff: 512, d_model: 128, dropout: 0.361, seg_len: 6, baseline: 1, cross_factor: 10, seq_len: 336
DLinear	CI-720	enc_in:3, lr:0.0027, seq_len:96
iTransformers	CD-720	lr:0.0002, e_layers:7, d_ff:1024, d_model:128, dropout:0.480, seq_len:720
TimeMixer	CI-720	lr:0.0010, d_ff:1024, d_model:1024, e_layers:10, seq_len:192
TimeMixer	CD-720	lr:0.0041, d_ff:256, d_model:1024, e_layers:6, seq_len:96

Table 24: Hyperparameter settings for the Lorenz dataset. The best variant for each model is shown.

Model	Setting	Best Hyper. Params
PatchTST	CI-96	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
PatchTST	CD-96	lr:0.0003,e_layers:9,d_ff:256,d_model:128,dropout:0.228,fc_dropout:0.072,patch_size:8,stride:8,seq_len:336
TSMixer	CI-96	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
TSMixer	CD-96	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
CrossFormer	CI-96	lr: 0.0003, e_layers: 7, d_ff: 256, d_model: 512, dropout: 0.336, seg_len: 10, baseline: 0, cross_factor: 5, seq_len: 336
CrossFormer	CD-96	lr: 0.0004, e_layers: 7, d_ff: 512, d_model: 256, dropout: 0.323, seg_len: 6, baseline: 0, cross_factor: 8, seq_len: 336
DLinear	CI-96	lr: 0.0097, seq_len: 192
iTransformers	CD-96	lr:1.30e-05, e_layers:4, d_ff:2048, d_model:128, dropout:0.008, seq_len:720
TimeMixer	CI-96	lr:0.0014, d_ff:1024, d_model:256, e_layers:8, seq_len:96
TimeMixer	CD-96	lr:0.0006, d_ff:1024, d_model:128, e_layers:5, seq_len:192
PatchTST	CI-192	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
PatchTST	CD-192	lr:1.44e-05,e_layers:10,d_ff:256,d_model:1.02e+03,dropout:0.014,fc_dropout:0.539,patch_size:16,stride:8,seq_len:192
TSMixer	CI-192	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
TSMixer	CD-192	lr:0.0097,num_blocks:6,hidden_size:256,dropout:0.254,activation:relu,seq_len:512
CrossFormer	CI-192	lr: 0.0003, e_layers: 6, d_ff: 256, d_model: 512, dropout: 0.309, seg_len: 8, baseline: 0, cross_factor: 5, seq_len: 336
CrossFormer	CD-192	lr: 0.0003, e_layers: 6, d_ff: 512, d_model: 128, dropout: 0.298, seg_len: 7, baseline: 1, cross_factor: 5, seq_len: 336
DLinear	CI-192	lr: 0.0080, seq_len: 192
iTransformers	CD-192	lr:1.30e-05, e_layers:4, d_ff:2048, d_model:128, dropout:0.008, seq_len:720
TimeMixer	CI-192	lr:0.0011, d_ff:256, d_model:256, e_layers:7, seq_len:96
TimeMixer	CD-192	lr:0.0024, d_ff:256, d_model:512, e_layers:6, seq_len:96
PatchTST	CI-336	lr': 1.441e-05, 'e_layers': 10, 'd_ff': 256, 'd_model': 1024, 'dropout': 0.0137, 'fc_dropout': 0.539, 'patch_size': 16, 'stride': 8, 'seq_len': 192
PatchTST	CD-336	lr': 1.441e-05, 'e_layers': 10, 'd_ff': 256, 'd_model': 1024, 'dropout': 0.0137, 'fc_dropout': 0.5389, 'patch_size': 16, 'stride': 8, 'seq_len': 192
TSMixer	CI-336	lr:0.0041,num_blocks:10,hidden_size:256,dropout:0.415,activation:gelu,seq_len:96
TSMixer	CD-336	lr:0.0010,num_blocks:1,hidden_size:64,dropout:0.033,activation:relu,seq_len:192
CrossFormer	CI-336	lr: 0.0001, e_layers: 8, d_ff: 256, d_model: 256, dropout: 0.112, seg_len: 10, baseline: 0, cross_factor: 7, seq_len: 336
CrossFormer	CD-336	lr: 0.0002, e_layers: 6, d_ff: 512, d_model: 128, dropout: 0.266, seg_len: 8, baseline: 1, cross_factor: 5, seq_len: 336
DLinear	CI-336	lr: 0.0093, seq_len: 96
iTransformers	CD-336	lr:0.0003, e_layers:1, d_ff:512, d_model:512, dropout:0.270, seq_len:720
TimeMixer	CI-336	lr:0.0015, d_ff:512, d_model:256, e_layers:8, seq_len:96
TimeMixer	CD-336	lr:0.0011, d_ff:256, d_model:256, e_layers:7, seq_len:96
PatchTST	CI-720	lr:0.0001,e_layers:2,d_ff:1.02e+03,d_model:1.02e+03,dropout:0.463,fc_dropout:0.708,patch_size:16,stride:4,seq_len:192
PatchTST	CD-720	lr:0.0018,e_layers:5,d_ff:1.02e+03,d_model:128,dropout:0.097,fc_dropout:0.623,patch_size:16,stride:8,seq_len:512
TSMixer	CI-720	lr:0.0005,num_blocks:6,hidden_size:64,dropout:0.280,activation:relu,seq_len:192
TSMixer	CD-720	lr:0.0097,num_blocks:6,hidden_size:256,dropout:0.254,activation:relu,seq_len:512
CrossFormer	CI-720	lr: 0.0002, e_layers: 6, d_ff: 128, d_model: 256, dropout: 0.302, seg_len: 7, baseline: 1, cross_factor: 5, seq_len: 336
CrossFormer	CD-720	lr: 0.0001, e_layers: 3, d_ff: 256, d_model: 512, dropout: 0.123, seg_len: 8, baseline: 0, cross_factor: 7, seq_len: 336
DLinear	CI-720	lr: 0.0010, seq_len: 96
iTransformers	CD-720	lr:0.0003, e_layers:5, d_ff:512, d_model:128, dropout:0.305, seq_len:720
TimeMixer	CI-720	lr:0.0005, d_ff:256, d_model:256, e_layers:8, seq_len:96
TimeMixer	CD-720	lr:0.0041, d_ff:256, d_model:1024, e_layers:6, seq_len:96

Table 25: Hyperparameter settings for the LorenzCoupled dataset. The best variant for each model is shown.

Model	Setting	Best Hyper. Params
PatchTST	CI-96	
PatchTST	CD-96	
TSMixer	CI-96	lr:0.0067,num_blocks:9,hidden_size:1.02e+03,dropout:0.318,activation:relu,seq_len:512
TSMixer	CD-96	lr:0.0003,num_blocks:1,hidden_size:64,dropout:0.166,activation:relu,seq_len:512
CrossFormer	CI-96	lr: 0.0024, e_layers: 1, d_ff: 128, d_model: 256, dropout: 0.305, seg_len: 6, baseline: 0, cross_factor: 7, seq_len: 192
CrossFormer	CD-96	lr: 0.0032, e_layers: 1, d_ff: 128, d_model: 256, dropout: 0.294, seg_len: 7, baseline: 0, cross_factor: 8, seq_len: 192
DLinear	CI-96	lr: 0.0079, seq_len: 512
iTransformers	CD-96	lr:0.0032, e_layers:4, d_ff:1024, d_model:128, dropout:0.139, seq_len:720
TimeMixer	CI-96	lr:0.0010, d_ff:512, d_model:256, e_layers:1, seq_len:720
TimeMixer	CD-96	lr:0.0005, d_ff:512, d_model:256, e_layers:8, seq_len:512
PatchTST	CI-192	
PatchTST	CD-192	
TSMixer	CI-192	lr:0.0067,num_blocks:9,hidden_size:1.02e+03,dropout:0.318,activation:relu,seq_len:512
TSMixer	CD-192	lr:0.0097,num_blocks:6,hidden_size:32,dropout:0.315,activation:relu,seq_len:512
CrossFormer	CI-192	OOM
CrossFormer	CD-192	lr: 0.0080, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.189, seg_len: 8, baseline: 0, cross_factor: 9, seq_len: 192
DLinear	CI-192	lr: 0.0058, seq_len: 512
iTransformers	CD-192	lr:0.0004, e_layers:1, d_ff:2048, d_model:1024, dropout:0.080, seq_len:192
TimeMixer	CI-192	lr 0.00008, d_ff 512, d_model 512, e_layers 4, seq_len 192
TimeMixer	CD-192	lr:0.0015, d_ff:256, d_model:256, e_layers:7, seq_len:512
PatchTST	CI-336	lr:0.0067,num_blocks:9,hidden_size:1.02e+03,dropout:0.318,activation:relu,seq_len:512
PatchTST	CD-336	
TSMixer	CI-336	
TSMixer	CD-336	lr:0.0097,num_blocks:5,hidden_size:32,dropout:0.336,activation:relu,seq_len:512
CrossFormer	CI-336	lr: 0.0080, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.189, seg_len: 8, baseline: 0, cross_factor: 9, seq_len: 192
CrossFormer	CD-336	lr': 0.00017626081318364395, 'e_layers': 7, 'd_ff': 128, 'd_model': 512, 'dropout': 0.3870070203430017, 'seg_len': 10, 'baseline': 0, 'cross_factor': 16, 'seq_len': 192, 'affine': 1
DLinear	CI-336	lr: 0.0098, seq_len: 512
iTransformers	CD-336	lr:0.0003, e_layers:8, d_ff:2048, d_model:1024, dropout:0.277, seq_len:192
TimeMixer	CI-336	lr:0.0001, d_ff:256, d_model:128, e_layers:2, seq_len:720
TimeMixer	CD-336	lr:0.0021, d_ff:256, d_model:128, e_layers:4, seq_len:192
PatchTST	CI-720	
PatchTST	CD-720	
TSMixer	CI-720	
TSMixer	CD-720	lr:0.0090,num_blocks:5,hidden_size:64,dropout:0.133,activation:relu,seq_len:192
CrossFormer	CI-720	OOM
CrossFormer	CD-720	OOM
DLinear	CI-720	lr: 0.0098, seq_len: 512
iTransformers	CD-720	lr:0.0012, e_layers:4, d_ff:2048, d_model:512, dropout:0.360, seq_len:192
TimeMixer	CI-720	lr:0.0001, d_ff:256, d_model:512, e_layers:2, seq_len:720
TimeMixer	CD-720	lr:0.0021, d_ff:256, d_model:128, e_layers:4, seq_len:192

Table 26: Hyperparameter settings for the electricity dataset. The best variant for each model is shown.

Model	Setting	Best Hyper. Params
PatchTST	CI-96	lr:0.0003,e_layers:9,d_ff:256,d_model:128,dropout:0.228,fc_dropout:0.072,patch_size:8,stride:8,seq_len:336
PatchTST	CD-96	lr:0.0001,e_layers:1,d_ff:1.02e+03,d_model:128,dropout:0.098,fc_dropout:0.465,patch_size:8,stride:8,seq_len:336
TSMixer	CI-96	lr:0.0002,num_blocks:5,hidden_size:1.02e+03,dropout:0.205,activation:relu,seq_len:512
TSMixer	CD-96	lr:0.0010,num_blocks:1,hidden_size:64,dropout:0.033,activation:relu,seq_len:192
CrossFormer	CI-96	lr: 1.43e-05, e_layers: 5, d_ff: 512, d_model: 256, dropout: 0.098, seg_len: 9, baseline: 0, cross_factor: 11, seq_len: 336
CrossFormer	CD-96	lr: 0.0001, e_layers: 3, d_ff: 256, d_model: 512, dropout: 0.123, seg_len: 8, baseline: 0, cross_factor: 7, seq_len: 336
DLinear	CI-96	lr: 0.0020, seq_len: 720
iTransformers	CD-96	lr:0.0001, e_layers:8, d_ff:2048, d_model:1024, dropout:0.356, seq_len:192
TimeMixer	CI-96	lr:6.00e-06, d_ff:512, d_model:256, e_layers:1, seq_len:720
TimeMixer	CD-96	lr:1.00e-05, d_ff:512, d_model:512, e_layers:1, seq_len:720
PatchTST	CI-192	lr:0.0001,e_layers:4,d_ff:256,d_model:128,dropout:0.469,fc_dropout:0.405,patch_size:8,stride:8,seq_len:512
PatchTST	CD-192	lr:0.0002,e_layers:1,d_ff:512,d_model:1.02e+03,dropout:0.010,fc_dropout:0.333,patch_size:16,stride:8,seq_len:192
TSMixer	CI-192	lr:0.0002,num_blocks:5,hidden_size:1.02e+03,dropout:0.247,activation:relu,seq_len:512
TSMixer	CD-192	lr:0.0010,num_blocks:1,hidden_size:64,dropout:0.033,activation:relu,seq_len:192
CrossFormer	CI-192	lr: 0.0002, e_layers: 6, d_ff: 512, d_model: 128, dropout: 0.266, seg_len: 8, baseline: 1, cross_factor: 5, seq_len: 336
CrossFormer	CD-192	lr: 0.0002, e_layers: 1, d_ff: 256, d_model: 128, dropout: 0.326, seg_len: 7, baseline: 0, cross_factor: 9, seq_len: 336
DLinear	CI-192	lr: 0.0077, seq_len: 720
iTransformers	CD-192	lr:1.60e-05, e_layers:9, d_ff:2048, d_model:1024, dropout:0.012, seq_len:192
TimeMixer	CI-192	lr 0.00008, d_ff 512, d_model 512, e_layers 4, seq_len 192
TimeMixer	CD-192	lr:0.0005, d_ff:512, d_model:128, e_layers:8, seq_len:720
PatchTST	CI-336	lr:0.0007,e_layers:8,d_ff:1024,d_model:512,dropout:0.150,fc_dropout:0.021,patch_size:8,stride:8,seq_len:720
PatchTST	CD-336	lr:0.0007,e_layers:8,d_ff:1024,d_model:512,dropout:0.150,fc_dropout:0.021,patch_size:8,stride:8,seq_len:720
TSMixer	CI-336	lr:0.0002,num_blocks:5,hidden_size:1.02e+03,dropout:0.265,activation:relu,seq_len:512
TSMixer	CD-336	lr:0.0001,num_blocks:6,hidden_size:256,dropout:0.334,activation:gelu,seq_len:512
CrossFormer	CI-336	lr: 0.0002, e_layers: 6, d_ff: 512, d_model: 128, dropout: 0.266, seg_len: 8, baseline: 1, cross_factor: 5, seq_len: 336
CrossFormer	CD-336	OOM
DLinear	CI-336	lr: 0.0014, seq_len: 720
iTransformers	CD-336	lr:1.60e-05, e_layers:9, d_ff:2048, d_model:1024, dropout:0.012, seq_len:192
TimeMixer	CI-336	lr:0.0001, d_ff:256, d_model:128, e_layers:2, seq_len:720
TimeMixer	CD-336	lr:1.00e-05, d_ff:512, d_model:128, e_layers:1, seq_len:720
PatchTST	CI-720	lr:0.0005,e_layers:7,d_ff:256,d_model:256,dropout:0.767,fc_dropout:0.878,patch_size:16,stride:8,seq_len:720
PatchTST	CD-720	lr:6.07e-06,e_layers:3,d_ff:256,d_model:1.02e+03,dropout:0.006,fc_dropout:0.642,patch_size:16,stride:8,seq_len:192
TSMixer	CI-720	lr:0.0003,num_blocks:2,hidden_size:1.02e+03,dropout:0.108,activation:relu,seq_len:512
TSMixer	CD-720	lr:0.0002,num_blocks:8,hidden_size:32,dropout:0.281,activation:relu,seq_len:512
CrossFormer	CI-720	lr: 0.0002, e_layers: 1, d_ff: 512, d_model: 128, dropout: 0.228, seg_len: 6, baseline: 0, cross_factor: 17, seq_len: 336
CrossFormer	CD-720	lr: 0.0080, e_layers: 1, d_ff: 128, d_model: 128, dropout: 0.189, seg_len: 8, baseline: 0, cross_factor: 9, seq_len: 192
DLinear	CI-720	lr: 0.0030, seq_len: 720
iTransformers	CD-720	lr:0.0001, e_layers:8, d_ff:1024, d_model:1024, dropout:0.356, seq_len:192
TimeMixer	CI-720	lr:0.0001, d_ff:256, d_model:512, e_layers:2, seq_len:720
TimeMixer	CD-720	lr:1.00e-05, d_ff:512, d_model:512, e_layers:1, seq_len:720

Table 27: Hyperparameter settings for the weather dataset. The best variant for each model is shown.

Model	Setting	Best Hyper. Params
PatchTST	CI-96	
PatchTST	CD-96	
TSMixer	CI-96	lr:0.0097,num_blocks:9,hidden_size:32,dropout:0.271,activation:relu,seq_len:512
TSMixer	CD-96	lr:0.0097,num_blocks:5,hidden_size:32,dropout:0.311,activation:relu,seq_len:512
CrossFormer	CI-96	lr': 0.000515701560401399, 'e_layers': 7, 'd_ff': 128, 'd_model': 256, 'dropout': 0.25730612136617825, 'seg_len': 11, 'baseline': 0, 'cross_factor': 10, 'seq_len': 96, 'affine': 1
CrossFormer	CD-96	lr': 5.2122436504604685e-05, 'e_layers': 3, 'd_ff': 256, 'd_model': 512, 'dropout': 0.12303593071645767, 'seg_len': 8, 'baseline': 0, 'cross_factor': 7, 'seq_len': 336, 'affine': 0
DLinear	CI-96	lr: 0.0078, seq_len: 512
iTransformers	CD-96	OOM
TimeMixer	CI-96	OOM
TimeMixer	CD-96	From TimeMixer Paper
PatchTST	CI-192	
PatchTST	CD-192	
TSMixer	CI-192	lr:0.0067,num_blocks:9,hidden_size:1.02e+03,dropout:0.318,activation:relu,seq_len:512
TSMixer	CD-192	lr:0.0097,num_blocks:5,hidden_size:32,dropout:0.311,activation:relu,seq_len:512
CrossFormer	CI-192	OOM
CrossFormer	CD-192	OOM
DLinear	CI-192	lr: 0.0078, seq_len: 512
iTransformers	CD-192	OOM
TimeMixer	CI-192	OOM
TimeMixer	CD-192	From TimeMixer Paper
PatchTST	CI-336	
PatchTST	CD-336	
TSMixer	CI-336	lr:0.0067,num_blocks:9,hidden_size:1.02e+03,dropout:0.318,activation:relu,seq_len:512
TSMixer	CD-336	lr:0.0097,num_blocks:5,hidden_size:32,dropout:0.311,activation:relu,seq_len:512
CrossFormer	CI-336	OOM
CrossFormer	CD-336	OOM
DLinear	CI-336	lr: 0.0038, seq_len: 336
iTransformers	CD-336	OOM
TimeMixer	CI-336	OOM
TimeMixer	CD-336	From TimeMixer Paper
PatchTST	CI-720	
PatchTST	CD-720	
TSMixer	CI-720	
TSMixer	CD-720	lr:0.0097,num_blocks:5,hidden_size:32,dropout:0.311,activation:relu,seq_len:512
CrossFormer	CI-720	OOM
CrossFormer	CD-720	OOM
DLinear	CI-720	lr: 0.0070, seq_len: 336
iTransformers	CD-720	OOM
TimeMixer	CI-720	OOM
TimeMixer	CD-720	From TimeMixer Paper

Table 28: Hyperparameter settings for the Traffic dataset. The best variant for each model is shown.