WHEN EXOGENOUS FEATURES HURT: RETHINKING FEATURE SELECTION IN TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature selection plays a critical role in machine learning in improving generalization and interpretability. In time series forecasting, this problem is particularly challenging: predicting a target variable often relies on a diverse set of exogenous features, yet recent studies have shown that incorporating certain features can unexpectedly degrade performance. This counterintuitive phenomenon has led to the growing adoption of univariate or channel-independent approaches, raising a fundamental question: when and why are exogenous features indeed beneficial?

In this work, we argue that effective feature selection should be guided not by static relevance (e.g., correlation), but by the temporal stability of the conditional relationship between features and the target. We introduce *Conditional Distribution Difference* (CDD), a tractable proxy that characterizes temporal instability in the conditional relationship through a bin-wise representation. Based on this representation, we derive CDD-std, the standard deviation of CDD across time, as a feature selection metric that quantifies conditional stability. Our key insight is that features with high CDD-std and unstable conditional relationships are difficult to generalize and tend to harm forecasting performance. We validate this insight on controlled synthetic experiments and real-world datasets, showing that CDD-std provides a more effective indicator of feature usefulness than correlation-based metrics. We develop simple yet effective feature selection strategies building on CDD-std, including hard filtering and TIMELESS, achieving consistent performance improvement with state-of-the-art forecasting models across diverse settings.

1 INTRODUCTION

Time-series forecasting finds wide applications in economics, urban computing, and epidemiology (Zhu & Shasha, 2002; Zheng et al., 2014; Deb et al., 2017; Mathis et al., 2024). One challenge in forecasting systems is selecting useful exogenous features, which provide complementary information beyond the target variable and can capture rapid or irregular changes induced by external factors. For example, promotions and holidays can cause abrupt demand peaks that univariate models fail to anticipate (Eisenach et al., 2020; Zhao et al., 2025d). However, exogenous features are not inherently beneficial: including irrelevant or unstable features can instead degrade forecasting performance.

As illustrated in Figure 1(a), **incorporating exogenous features does not consistently improve forecasting accuracy**. Existing forecasting studies primarily focus on designing increasingly advanced architectures (Wang et al., 2024b; Liu et al., 2023a; Chen et al., 2023), while the question of which exogenous features are actually useful remains less explored. This observation has also motivated univariate or channel-independent forecasting approaches (Nie et al., 2023; Zeng et al., 2023; Tan et al., 2024). We argue that this phenomenon is fundamentally a feature selection problem: the availability of exogenous information does not imply that all such information is beneficial.

Despite extensive feature selection studies in general machine learning (Dash & Liu, 1997; Li et al., 2017), **reliable and scalable feature selection remains challenging for time-series forecasting**. Correlation-based approaches (Yu & Liu, 2003; 2004; Qiu et al., 2024; Ni et al., 2025) typically rely on static relationships and may fail under temporal distribution shifts or spurious correlations. As shown in Figure 1(b), stronger feature-target correlation does not necessarily imply greater forecasting utility. Alternatively, directly evaluating the performance change caused by adding or removing features (Tran et al., 2015; Thabtah et al., 2020; Crabbé & Van Der Schaar, 2021) is more direct

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

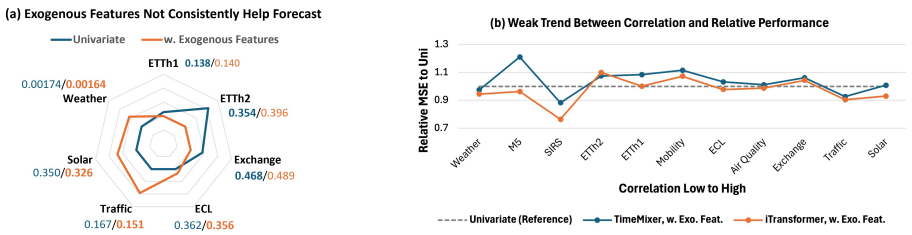


Figure 1: Exogenous features are not inherently beneficial. **(a)** Including exogenous features does not consistently improve accuracy (with iTransformer). **(b)** Inconsistent relationship between feature-target correlation and performance (values < 1 indicate improvement when including exogenous features); higher correlation does not necessarily refer to exogenous features being more useful.

but becomes computationally expensive for high-dimensional datasets. These limitations motivate a scalable method that assesses feature effectiveness while accounting for temporal dynamics.

To this end, we propose *Conditional Distribution Difference* (CDD), an algorithm that captures bin-wise discrepancies in the conditional relationship between an exogenous feature and the target. The key distinction is therefore not whether a feature is statistically dependent on the target, but whether this relationship remains temporally stable enough to be useful for forecasting. Based on CDD, we introduce CDD-std, the standard deviation of CDD across temporal bins, as a feature selection metric that quantifies the temporal stability of this relationship. Intuitively, features with higher instability are more likely to degrade forecasting performance when their relationships with the target shift across time. Our contributions are as follows:

- **Feature Selection Metric.** We propose *Conditional Distribution Difference* (CDD) algorithm, a distribution-based procedure for quantifying temporal discrepancies in feature-target relationships, and introduce CDD-std as a metric of their temporal stability.
- **Empirical Insight.** We demonstrate that CDD-std is a strong and reliable metric of feature selection through extensive experiments: lower CDD-std is consistently associated with greater performance gains when incorporating exogenous features.
- **Practical Methods.** We develop two CDD-std-guided feature selection strategies, including hard filtering and TIMELESS, achieving consistent improvements over using all exogenous inputs, with consistent performance gains across various forecasting settings.

The rest of this paper is organized around three main questions. First, we introduce CDD-std as a practical metric for measuring temporal instability in feature-target relationships. Second, we investigate the relationship between CDD-std and the forecasting utility of exogenous features. Third, we show how CDD-std can be used for feature selection to improve forecasting performance. Related works on time-series forecasting and feature selection are discussed in Appendix A.

2 OUR FEATURE SELECTION METRIC

2.1 OVERVIEW

In this section, we introduce *Conditional Distribution Difference* (CDD) and its corresponding algorithm (Algorithm 1). Our goal is to characterize the temporal instability of the conditional relationship between exogenous features and the target. However, directly estimating and comparing the full conditional distributions across temporal periods, can be challenging in practice. We therefore provide a tractable proxy for this instability by measuring distributional differences within divided temporal bins for each exogenous feature. We provide the algorithm and justify this proxy in the main text, with the complete theoretical analysis and interpretations in Appendix B.

Based on the feature-by-temporal-bin representation, we derive statistics that capture different aspects of temporal conditional instability. In particular, we use CDD-std, the standard deviation of CDD across temporal bins, as a feature selection metric. Intuitively, a larger CDD-std indicates that the feature-target relationship varies more substantially across time, suggesting potentially reduced generalizability. We examine this interpretation in controlled synthetic settings and then investigate its practical utility for feature selection on real-world forecasting datasets.

Algorithm 1 Conditional Distribution Difference (CDD)

Require: Target $Y \in \mathbb{R}^T$, exogenous feature $X \in \mathbb{R}^{T\times d}$, window w, step s, horizon h, bins K.
Ensure: Bin-level CDD matrix $\mathbf{D} \in \mathbb{R}^{K\times d}$.

```
Divide the sequence into K bins with boundaries {b_0, b_1, ..., b_K}.
for k = 1 to K do
    Extract bin data (X_b, Y_b), such that X_b = X[b_{k-1} : b_k − h] and Y_b = Y[b_{k-1} + h : b_k].
    Normalize (X_b, Y_b) ~ N(0, 1) .
    Construct reference distribution R_k = {X_{b,j} · Y_b | j = 1, ..., d}.
end for
for t ∈ [0, T − w − h, s] do
    Identify bin index k at time step t.
    Select window data X_L ∈ [t, t + w), Y_H ∈ [t + w, t + w + h).
    Normalize (X_L, Y_H) ~ N(0, 1).
    Align window lengths by (X_L, Y_H) ← (Down(X_L, m), Down(Y_H, m)), m = min(w, h)[1].
    for j = 1 to d do
        Compute WD(X_{L,j} · Y_H, R_k[j]), append to d_{k,j}[2].
    end for
end for
D[k, j] ← mean(d_{k,j}), ∀k ∈ [1, K], j ∈ [1, d]
```

2.2 CDD Algorithm

In time-series forecasting with exogenous features, the predictive task can be formulated as modeling the conditional distribution $P(Y_{\mathrm{H}} \mid X_{\mathrm{L}})$, where X_{L} denotes the historical exogenous inputs and Y_{H} denotes the future target. The effectiveness of an exogenous feature, therefore, depends on whether this conditional relationship can be reliably learned and generalized. In practice, both marginal distributions $P(X_{\mathrm{L}})$, $P(Y_{\mathrm{H}})$, and conditional distribution $P(Y_{\mathrm{H}} \mid X_{\mathrm{L}})$, may evolve over time. While recent studies focus on mitigating marginal distribution shifts (Kim et al., 2021; Liu et al., 2023b; Fan et al., 2023), capturing a stable conditional relationship remains challenging, as it requires models to generalize across changing temporal patterns, which consequently impacts the effectiveness of exogenous features. Instead of relying on complex designs (Liu et al., 2024a; Zhao et al., 2025a), we take a complementary feature selection perspective and focus on identifying exogenous features whose predictive relationships remain sufficiently stable over time. To characterize this temporal instability, we introduce *Conditional Distribution Difference* (CDD) as the discrepancy between the conditional distributions across different time periods:

$$\mathrm{CDD} \;=\; \mathbb{E}_{t,t'}\left[\mathcal{D}(P_t(Y_{\mathrm{H}} \mid X_{\mathrm{L}}),\; P_{t'}(Y_{\mathrm{H}} \mid X_{\mathrm{L}}))\right], \tag{1}$$

where $\mathcal{D}(\cdot,\cdot)$ denotes a discrepancy measure between conditional distributions, and P_t denotes the distribution at time period t. However, directly measuring $P(Y_{\mathrm{H}} \mid X_{\mathrm{L}})$ is challenging in practice. Therefore, we adopt two practical approximations to estimate the conditional distributions:

(1) Joint distribution as a proxy. Directly estimating the conditional distribution $P(Y_{\mathrm{H}} \mid X_{\mathrm{L}})$ is challenging in practice. Therefore, we approximate it by measuring the corresponding joint distribution $P(Y_{\mathrm{H}}, X_{\mathrm{L}})$, which provides a tractable proxy for the underlying conditional relationship. The rationale of this proxy relies on applying instance normalization, which stabilizes the marginal distributions of X_{L} and Y_{H} into a normalized space. From the relation $P(Y_{\mathrm{H}} \mid X_{\mathrm{L}}) = \frac{P(Y_{\mathrm{H}}, X_{\mathrm{L}})}{P(X_{\mathrm{L}})}$, when $P(X_{\mathrm{L}})$ is stabilized, the joint distribution is reflective of the underlying conditional relationship, where marginal scale differences are minimized. We formally characterize this relationship as:

Proposition 2.1 (Joint distribution as a proxy for conditional stability)**.** *Consider two temporal segments t and t' with joint distributions $P_t(X, Y)$ and $P_{t'}(X, Y)$. If the residual discrepancy between the marginal distributions $P_t(X)$ and $P_{t'}(X)$ after instance normalization is small, then the discrepancy in the conditional relationship $P(Y \mid X)$ is bounded by the discrepancy in the joint distribution $P(X, Y)$ plus this residual marginal discrepancy.*

[1] $\mathrm{Down}(X, m)$ denotes a downsampling that partitions X into m segments and averages each segment.
[2] $\mathcal{WD}(\cdot,\cdot)$ denotes the Wasserstein Distance.

Proposition 2.1 shows that the joint distribution is a controlled proxy for the conditional relationship: As instance normalization reduces marginal distribution variations, changes in the joint distribution primarily reflect changes in the conditional relationship. The formal derivation is in Appendix B.

(2) Bin-wise representation with local comparison. We partition the time series into K temporal bins and compute an aggregated joint distribution within each bin to represent the relationship between the feature and the target in that period. This design relies on a mild local stationarity assumption (Gama et al., 2014), under which the conditional relationship does not change drastically within a short temporal segment. The bin-level distribution provides a reasonable reference for characterizing local behavior. This bin-wise formulation can be viewed as a tractable approximation of the pairwise temporal discrepancy in Eq. 1. Instead of explicitly computing discrepancies between all pairs of time steps (t, t'), we approximate the expectation over t' using a bin-level reference distribution, which serves as an aggregated surrogate for nearby conditional distributions.

With these approximations, we provide Algorithm 1, which computes a bin-wise representation $\mathbf{D} \in \mathbb{R}^{K \times d}$ of CDD by measuring how local joint distributions deviate from bin-level references over time. Each entry $\mathbf{D}[k, j]$ captures the discrepancy for feature j within bin k. We provide additional theoretical justifications and interpretations in Appendix B.

2.3 Interpreting CDD Statistics

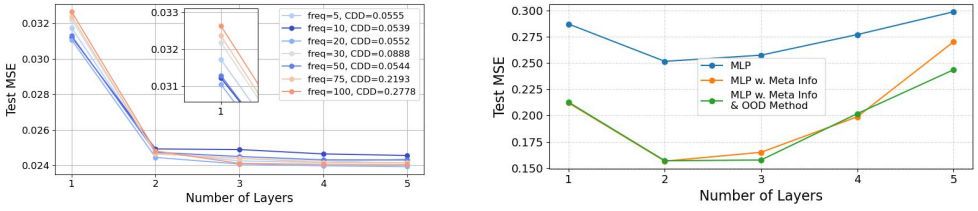
Algorithm 1 produces a bin-wise matrix $\mathbf{D} \in \mathbb{R}^{K \times d}$, where each column captures the conditional discrepancy of an exogenous feature across temporal bins. We consider two statistics of $\mathbf{D}$: the mean (CDD-mean) and standard deviation (CDD-std), which characterize the overall magnitude and temporal variability of the conditional discrepancy, respectively.

High CDD-mean reflects modeling complexity. A larger CDD-mean indicates that the conditional relationship is more complex on average and may require greater model capacity to approximate. To examine its practical implication, we construct a controlled synthetic setting where the complexity of the underlying relationship varies with signal frequency.

Example 2.1. Consider example that the covariate and target follow deterministic nonlinear patterns:

$$X_f(t) = \cos\left(\frac{\pi t}{f}\right) + \mathcal{N}(0, \sigma^2), \quad Y_f(t) = \sin\left(\frac{\pi t}{f}\right) + \mathcal{N}(0, \sigma^2). \tag{2}$$

By varying the frequency f, we control the complexity of the underlying relationship, which corresponds to different CDD-mean values. The forecasting objective is to predict Y_f using historical observations of X_f for varying f. Additional experimental details are provided in Appendix C.1.



(a) Forecasting errors with varying CDD-mean (b) Forecasting errors with high CDD-std

Figure 2: **(a):** Forecasting errors under different CDD-means. Signals with higher CDD-mean yield larger prediction errors for shallow models, and can be mitigated by increasing model capacity. **(b):** Forecasting errors under high CDD-std with regular forecaster, with frequency labels, and with OOD method (IRM (Arjovsky et al., 2019)). Performance is rarely improved by increasing model capacity.

Figure 2a shows the forecasting error under different frequencies and model capacities. For a fixed frequency, increasing model capacity consistently reduces forecasting error, indicating that sufficiently expressive models can capture complex conditional relationships even when CDD-mean is large. Meanwhile, with shallow models, signals with lower CDD-mean tend to achieve smaller forecasting errors. Thus, high CDD-mean is associated with greater modeling difficulty for low-capacity predictors, but this difficulty can be mitigated by increasing model capacity. Given the

expressive capacity and overparameterization of modern neural forecasting models, we argue that such modeling complexity is generally not the primary bottleneck for exploiting exogenous features.

High CDD-std reflects temporal instability. In contrast, CDD-std captures how consistently the conditional relationship is maintained across temporal bins. A larger CDD-std indicates that the relationship between an exogenous feature and the target varies substantially across time, so a mapping learned from one period may not generalize to another. To isolate this effect, we construct a synthetic setting where the conditional relationship alternates between two regimes:

Example 2.2. Consider example that the relationship between covariate and target varies over time:

$$f(t) = \begin{cases} f_{\text{low}}, & \text{if } \lfloor t/P \rfloor \text{ is even,} \\ f_{\text{high}}, & \text{otherwise.} \end{cases} \tag{3}$$

$$X(t) = \cos\left(\frac{\pi t}{f(t)}\right) + \mathcal{N}(0, \sigma^2), \quad Y(t) = \sin\left(\frac{\pi t}{f(t)}\right) + \mathcal{N}(0, \sigma^2), \tag{4}$$

where the alternating frequency induces a time-varying conditional relationship and hence a high CDD-std. Additional experimental details are provided in Appendix C.2.

Figure 2b shows that, unlike high CDD-mean, increasing model capacity provides limited improvement when the conditional relationship varies across time. In contrast, providing information about the underlying regime substantially improves forecasting performance, indicating that the main difficulty arises from temporal inconsistency rather than model capacity. Since such regime information is generally unavailable in real-world forecasting, **we use CDD-std as the feature selection metric to identify exogenous features with potentially unreliable temporal relationships.**

2.4 CDD-STD AS A METRIC IN PRACTICE

To validate our claim that the usefulness of exogenous features is governed by the stability of their conditional relationship with the target, i.e., CDD-std, we conduct experiments on multiple real-world forecasting datasets. Our goal is to examine whether CDD-std can serve as a reliable metric for determining when exogenous features are beneficial. We hypothesize that (i) when CDD-std is large, indicating greater temporal inconsistency in the feature-target relationship, incorporating exogenous features is more likely to degrade performance; and (ii) when CDD-std is small, indicating more consistent relationships over time, exogenous features are more likely to improve forecasting accuracy. To enable fair comparison across datasets, we normalize CDD-std by the magnitude of its mean, i.e., $\frac{\sigma}{|\mu|}$. For simplicity, we refer to this normalized quantity as CDD-std in the remainder of the paper.

Our evaluation covers diverse domains and forecasting horizons, including both short-term and long-term settings. Detailed experimental setups are provided in Appendix C.3. We visualize partial results in Figure 3, where datasets are sorted by increasing CDD-std along the x-axis. Full numerical results are reported in Table 9 & 10 in Appendix D.2. We conduct a hyperparameter study, showing that dataset ranking induced by CDD-std is robust to the choice of K, detailed in Table 8 in Appendix D.1.

Cross-Dataset Analysis. To examine whether CDD-std better captures the usefulness of exogenous information across different datasets than conventional static relevance measures, we plot the mean feature-level CDD-std within each dataset with the corresponding performance difference between univariate forecasting and forecasting with exogenous features. As shown in Figure 3, we observe a clearer monotonic relationship between the mean CDD-std and the performance gap across datasets. Specifically, datasets with smaller mean CDD-std tend to benefit more from incorporating exogenous features, whereas datasets with larger mean CDD-std often exhibit degraded performance when such features are included. In contrast, correlation-based measures

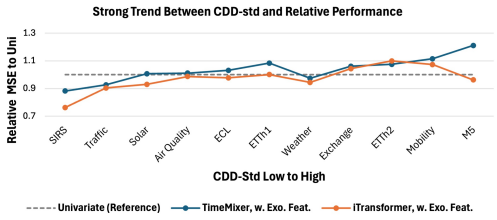


Figure 3: Consistent relationship between CDD-std and performance. Lower CDD-std refers to greater gains from exogenous features, higher CDD-std refers to limited or negative gains.

exhibit no significant association with the corresponding performance gap, as shown earlier in Figure 1(b). This cross-dataset analysis suggests that temporal instability captured by CDD-std is more informative of the overall usefulness of exogenous information than static correlation.

3 METRIC-GUIDED METHODOLOGY

Having established CDD-std as a practical metric for assessing the effectiveness of exogenous features, we next investigate how to leverage this insight for feature selection. Since large CDD-std indicates temporal inconsistency that is difficult to mitigate through model capacity alone, features with high CDD-std may be undesirable for forecasting. This motivates two CDD-std-guided selection strategies with different trade-offs between efficiency and flexibility. The first applies an empirically derived hard filtering to remove unreliable features, while the second uses CDD-std as feature-level meta-information to adaptively learn feature importance. We detail the two approaches below.

3.1 CDD-STD HARD FILTERING

Given the relationship between CDD-std and feature effectiveness, we use CDD-std as a practical criterion for feature selection. Features with smaller CDD-std tend to provide more stable predictive information, whereas larger CDD-std is associated with greater instability. Since a theoretically justified boundary for feature filtering is difficult to derive, we obtain an empirical operating point from the observed relationship between CDD-std and performance gains. Specifically, we regress performance gains on CDD-std using previous development analysis, aggregated across datasets and forecasting horizons, and define the empirical threshold yields CDD-std ≈ 0.223, which is fixed for all subsequent evaluations rather than re-estimated for each dataset or model. Thus, 0.223 is used as a data-driven operating point, not as a universal or theoretically optimal bound. As shown in later experiments, **w. Threshold** improves forecasting performance on both seen and unseen datasets.

More generally, feature selection in practice often operates under a fixed feature budget, where the goal is to retain the top-K features rather than determine an absolute cutoff. In this setting, the empirical threshold is unnecessary: we simply rank candidate features by CDD-std and select those with the smallest values, namely, **w. Ranking**. This relies only on the relative ordering of features and is therefore less sensitive to the choice of an absolute boundary. Under fixed feature budgets, w. Ranking also outperforms other feature-selection strategies in later experiments.

3.2 TIMELESS, CDD-STD-GUIDED META LEARNING

While hard filtering provides a simple and effective mechanism for removing unreliable features, it applies a fixed decision rule that may not adapt well to different datasets or forecasting settings. In practice, the impact of exogenous features can be data-dependent, making a hard selection criterion potentially suboptimal. To address this limitation, we propose TIMELESS, which leverages CDD-std as feature-level meta-information and learn to adaptively determine how to select each feature. TIMELESS learns a data-driven weighting over exogenous features based on their estimated reliability. We use the CDD-std of each feature as input to a lightweight MLP, which produces feature-wise scores. These scores are combined with the corresponding CDD-std and normalized using a softmax function, yielding continuous feature weights. We then retain features whose learned weights are higher than the average and suppress the rest. The procedure can be written as:

$$\begin{aligned} \mathbf{F}_{\text{score}} &= \text{softmax}\left(\text{MLP}(\text{CDD-std})\right), \\ \mathrm{F}_{\text{sel}} &= \mathrm{F} \times \left(\mathbb{1}_{\mathrm{F}_{\text{score}} \geq \mu(\mathrm{F}_{\text{score}})} \cdot \mathrm{F}_{\text{score}}\right). \end{aligned} \tag{5}$$

where $\mathrm{F} \in \mathbb{R}^{D-1}$ are the input exogenous features, and $\mathrm{F}_{\text{score}}$ are continuous feature weights produced by the MLP. The MLP and softmax operations are fully differentiable, and $\mathrm{F}_{\text{score}}$ is jointly optimized with the forecasting objective through backpropagation. The subsequent binary comparison is used only to construct the hard selection mask and is not involved in gradient computation. Specifically, the mask determines which features are retained, while the continuous scores remain the learnable quantities through which the selection module receives gradients. Unselected features are zeroed out for dimensional consistency. Compared to hard filtering, TIMELESS provides greater flexibility across datasets and forecasting settings, at the cost of additional training overhead.

Table 1: Averaged error comparisons between forecasting with all exogenous features (Base), CDD-std threshold (w. Threshold), and TIMELESS. Our feature selection approaches improve performance in 81 out of 96 evaluations, highlighting the necessity of feature selection in time-series forecasting and the effectiveness of both CDD-std–based methods. Detailed results are reported in Table 11 & 12.

Dataset	iTransformer						TimeMixer					
	Base		w. Threshold		w. TIMELESS		Base		w. Threshold		w. TIMELESS	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.139	0.289	0.141	0.289	0.139	0.286	0.151	0.302	0.149	0.299	0.137	0.287
			-0.9%	+0.0%	+0.0%	+0.8%			+1.0%	+0.9%	+9.1%	+4.7%
Weather	0.00164	0.0300	0.00166	0.0300	0.00163	0.0297	0.00162	0.0298	0.00164	0.0299	0.00168	0.0304
			-1.4%	+0.0%	+0.6%	+0.9%			-0.8%	-0.3%	-3.2%	-1.8%
Exchange	0.489	0.482	0.462	0.469	0.461	0.466	0.504	0.492	0.441	0.468	0.475	0.481
			+5.6%	+2.7%	+5.7%	+3.2%			+12.5%	+4.9%	+5.7%	+2.3%
ETTh2	0.392	0.491	0.383	0.483	0.355	0.465	0.392	0.489	0.394	0.489	0.370	0.472
			+2.2%	+1.5%	+9.4%	+5.3%			-0.6%	+0.0%	+5.6%	+3.5%
Mobility	0.131	0.287	0.122	0.279	0.119	0.275	0.147	0.306	0.132	0.289	0.136	0.293
			+6.9%	+2.9%	+9.0%	+4.3%			+10.4%	+5.6%	+8.0%	+4.1%
M5	0.457	0.529	0.475	0.541	0.445	0.529	0.500	0.567	0.413	0.517	0.455	0.540
			-3.9%	-2.3%	+2.6%	+0.0%			+17.4%	+8.9%	+9.1%	+4.7%
Dataset	**CrossLinear**						**TimeFilter**					
	Base		w. Threshold		w. TIMELESS		Base		w. Threshold		w. TIMELESS	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.147	0.294	0.145	0.293	0.145	0.293	0.139	0.289	0.147	0.295	0.142	0.291
			+1.1%	+0.4%	+1.2%	+0.6%			-2.8%	-1.2%	-1.7%	-0.7%
Weather	0.00174	0.0304	0.00171	0.0303	0.00171	0.0302	0.00174	0.0306	0.00177	0.0310	0.00174	0.0306
			+1.5%	+0.5%	+1.6%	+0.5%			-1.8%	-1.3%	+0.0%	+0.0%
Exchange	0.461	0.469	0.458	0.468	0.456	0.465	0.517	0.488	0.479	0.474	0.474	0.468
			+0.6%	+0.4%	+1.2%	+0.8%			+7.4%	+2.8%	+8.5%	+4.2%
ETTh2	0.348	0.460	0.345	0.459	0.339	0.454	0.393	0.489	0.357	0.465	0.361	0.468
			+0.8%	+0.4%	+2.3%	+1.5%			+9.1%	+5.0%	+8.3%	+4.4%
Mobility	0.117	0.272	0.115	0.269	0.116	0.270	0.123	0.278	0.115	0.270	0.118	0.272
			+1.5%	+1.2%	+0.9%	+0.8%			+6.0%	+2.8%	+4.1%	+2.1%
M5	0.432	0.529	0.426	0.524	0.422	0.520	0.424	0.513	0.413	0.505	0.423	0.511
			+1.4%	+1.0%	+2.4%	+1.8%			+2.6%	+1.6%	+0.4%	+0.5%

4 EVALUATION

We evaluate the proposed feature selection approaches across a diverse set of datasets and agnostic forecasting models, with comprehensive scopes. First, we assess both hard filtering and TIMELESS on the datasets studied in Section 2.4, demonstrating that CDD-std–guided feature selection improves forecasting performance across agnostic models, including iTransformer (Liu et al., 2023a), TimeMixer (Wang et al., 2024a), CrossLinear (Zhou et al., 2025), and TimeFilter (Hu et al., 2025). Second, we compare our approaches with commonly used feature selection methods, showing that CDD-std-based selection more effectively identifies reliable exogenous features under a fixed feature budget. Third, we further examine the generalization and robustness of our approaches through evaluations on unseen datasets and models, ablations disentangling the contribution from the additional MLP in TIMELESS, evaluations on stable datasets to verify no performance degradation via feature selection, and applications to foundation model Chronos-2 (Ansari et al., 2025).

To show the observed improvements are systematic rather than random, we conduct paired t-tests across model-dataset-horizon in our evaluations, treating per model, per dataset, and per horizon combination as a paired comparison. Detailed experimental setups are in Appendix C.4. Reproducible codes are anonymously at `https://anonymous.4open.science/r/cdd_iclr-4CFD`.

4.1 PERFORMANCE GAIN VIA FEATURE SELECTION

We evaluate the effectiveness of our CDD-std–based feature selection approaches, including both thresholding and TIMELESS. Our results show that applying feature selection improves forecasting performance compared to using all exogenous features without selection, particularly on datasets with higher CDD-std values. Summarized results are reported in Table 1. The results show performance improvements in 81 out of 96 evaluations, demonstrating the robustness of the proposed approaches and the practical effectiveness of both methods. Furthermore, performance gains are more pronounced on datasets with larger CDD-std (datasets in Table 1 are sorted in increasing order of CDD-std), providing strong empirical evidence that CDD-std captures feature-level reliability and serves as an effective signal for guiding feature selection. We further perform paired t-tests across all models

and datasets, showing threshold-based filtering achieves a statistically significant improvement ($p = 1.2 \times 10^{-4}$), while TIMELESS yields an even stronger significance ($p = 4.5 \times 10^{-8}$).

4.2 COMPARISON WITH FEATURE SELECTION APPROACHES

We compare CDD-std-guided strategies with feature selection baselines, including Spearman correlation (Hauke & Kossowski, 2011), Mutual Information (MI) (Vergara & Estévez, 2014), and minimum Redundancy Maximum Relevance (mRMR) (Peng et al., 2005). These methods are common statistical criteria for measuring feature-target dependence and feature selection. For each baseline, features are ranked according to the corresponding selection criterion, with a fixed feature budget of the top 1/3 features for a fair comparison. We slightly modify the selection criterion of TIMELESS to select the top 1/3 features according to the learned scores. We use same forecasting models as in Section 4.1 on the Exchange and ETTh2 datasets, with the summarized results reported in Table 2.

The results demonstrate the advantages of CDD-std-guided strategies. With ranking-based filtering, our method achieves 0.90%, 2.38%, and 1.81% lower average error (MSE+MAE) than Spearman correlation, MI, and mRMR, respectively. With TIMELESS, the average error is further reduced by 1.90%, 3.29%, and 2.72% compared with Spearman correlation, MI, and mRMR, respectively, with statistically significant improvements in all three comparisons ($p = 0.004$, 7.4×10^{-5}, and 9.8×10^{-5}), demonstrating the effectiveness of incorporating CDD-std as feature-level meta-information for adaptive feature selection across different datasets and forecasting scenarios.

Table 2: Averaged error comparisons between CDD-std-guided feature selection (w. Ranking and w. TIMELESS) and feature selection baselines (Spearman, MI, and mRMR) under fixed budgets. Our approaches outperform baseline feature selection criteria. Detailed results are reported in Table 13.

Dataset	Model	Spearman		MI		mRMR		w. Ranking		w. TIMELESS	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh2	iTransformer	0.365	0.472	0.398	0.496	0.398	0.496	0.386	0.483	**0.356**	**0.467**
	TimeMixer	**0.373**	**0.476**	0.394	0.491	0.394	0.491	0.380	0.478	0.380	0.480
	CrossLinear	0.358	0.467	0.358	0.467	0.358	0.467	0.354	0.464	**0.351**	**0.462**
	TimeFilter	0.372	0.475	0.400	0.493	0.400	0.493	0.380	0.470	**0.352**	**0.462**
Exchange	iTransformer	0.505	0.483	0.455	0.468	0.440	0.462	0.446	0.464	**0.437**	**0.457**
	TimeMixer	0.480	0.482	0.465	0.474	0.490	0.477	**0.462**	**0.472**	0.466	**0.472**
	CrossLinear	0.451	0.471	0.451	0.471	0.451	0.470	**0.447**	**0.467**	0.448	**0.467**
	TimeFilter	0.480	0.474	0.488	0.476	0.457	**0.464**	**0.456**	**0.464**	0.461	0.466

4.3 GENERALIZATION AND ABLATION ANALYSIS

Generalization on Unseen Datasets and Models. We examine whether CDD-std and the proposed feature selection strategies generalize beyond the datasets and models used in Section 2.4. Since the CDD-std threshold is determined from the development datasets in Section 2.4, we fix the threshold and evaluate both approaches on completely unseen datasets and models to assess their generalization. Specifically, we consider TimeXer (Wang et al., 2024b) and TSMixer (Chen et al., 2023) on two unseen datasets, Neuro and SP500, from Time-HD (Ni et al., 2025). Results averaged across prediction horizons are reported in Table 3. Both threshold-based filtering and TIMELESS consistently improve forecasting performance, including on the high-frequency Neuro and irregular SP500 datasets, where unstable exogenous relationships are more likely to arise. Paired t-tests across model–dataset–horizon evaluations further support that these improvements are systematic, with $p = 2.6 \times 10^{-4}$ for threshold-based filtering and $p = 7.0 \times 10^{-8}$ for TIMELESS.

Ablation Analysis on Metadata Guidance. To verify that the improvements of TIMELESS arise from the CDD-std-guided feature selection rather than the additional MLP itself, we conduct an ablation study by replacing the CDD-std metadata input to the MLP with a constant value of 1 while keeping the network architecture unchanged. This ablated variant isolates the effect of metadata guidance from that of the MLP module. The summarized results are reported in Table 15. The results demonstrate that the performance gains of TIMELESS primarily come from the informative CDD-std metadata. Compared with the Base setting, TIMELESS with CDD-std input reduces the average MSE and MAE by 3.38% and 1.86%, respectively, with statistically significant improvements ($p = 3.6 \times 10^{-6}$ and 3.9×10^{-8}). Replacing the CDD-std metadata with a constant value of 1 still yields regressions of 1.34% in MSE and 0.42% in MAE relative to Base, while neither is statistically

Table 3: Averaged error comparisons between forecasting with all exogenous features (Base), CDD-std threshold feature selection (w. Threshold), and TIMELESS on unseen datasets and models. Both methods demonstrate performance improvements. Detailed results are reported in Table 14.

Dataset	TimeXer						TSMixer					
	Base		w. Threshold		w. TIMELESS		Base		w. Threshold		w. TIMELESS	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Neuro	0.0175	0.0327	0.0157	0.0326	0.0164	0.0312	0.0329	0.0414	0.0240	0.0370	0.0135	0.0295
			+10.2%	+0.2%	+6.2%	+4.5%			+26.9%	+10.6%	+58.8%	+28.7%
SP500	2.8060	1.1558	2.6188	1.1383	2.5373	1.1125	3.4483	1.2890	2.7563	1.2050	2.4920	1.1260
			+6.7%	+1.5%	+9.6%	+3.7%			+20.1%	+6.5%	+27.7%	+12.6%

significant ($p = 0.18$ and 0.31). More importantly, directly comparing the CDD-std input with the constant-1 input shows that CDD-std guidance further reduces the average MSE and MAE by 3.62% and 2.02%, respectively, with statistically significant improvements ($p = 9.4 \times 10^{-4}$ and 1.3×10^{-5}). These results indicate that the additional MLP alone does not account for the performance gains of TIMELESS, while incorporating informative CDD-std metadata provides the key benefit.

Evaluation on Stable Datasets. To examine whether the proposed feature selection approaches introduce performance degradation when feature instability is already low, we further evaluate threshold-based filtering and TIMELESS on two relatively stable datasets, Solar and Traffic. Detailed results are in Table 16. Both approaches achieve performance comparable to the Base setting. Compared with Base, threshold-based filtering increases the average MSE by 0.39%, while TIMELESS increases it by 0.92%; neither difference is statistically significant ($t = 1.42$, $p = 0.16$ and $t = 1.09$, $p = 0.29$, respectively). The slight average increase is expected, as low-instability datasets contain few, if any, unstable features to remove. Consequently, feature selection may occasionally discard mildly informative features, resulting in small and statistically insignificant performance changes that remain within the noise level of repeated runs. These results indicate that TIMELESS does not materially degrade forecasting performance when feature instability is already low.

Generalization to Foundation Models. We evaluate CDD-std thresholding with the time-series foundation model Chronos-2 (Ansari et al., 2025) on Exchange and Neuro. We exclude TIMELESS as it requires training from scratch. As in Table 4, CDD-std thresholding improves zero-shot forecasting by up to 10%, suggesting its effectiveness can also benefit foundation-model forecasting. A more systematic study of feature selection, such as more comprehensive evaluations and analysis on how feature selection can impact zero-shot forecasting performance, remains promising and is left for future work, as our study primarily focuses on supervised forecasting scenarios.

Table 4: Averaged error comparison between Base and CDD-std threshold-based feature selection on Chronos-2. Detailed results in Table 17.

Dataset	Base		w. Threshold	
	MSE	MAE	MSE	MAE
Exchange	0.4425	0.4690	0.4168	0.4505
			+5.8%	+3.9%
Neuro	0.0147	0.0239	0.0146	0.0238
			+1.1%	+0.3%

5 CONCLUSION

We propose CDD-std, a simple yet effective metric for quantifying the reliability of exogenous features in time-series forecasting. Through both controlled synthetic experiments and real-world evaluations, we demonstrate that CDD-std is a reliable metric to assess the effectiveness of exogenous features across diverse datasets and models. Building on this insight, we develop two practical feature selection strategies with different cost–flexibility trade-offs, both of which improve forecasting performance. Our results highlight that, beyond model design, principled feature selection is essential for effectively leveraging exogenous information in modern forecasting systems. Future work includes extending CDD-std to more adaptive settings and combining it with other time-series properties, such as stationarity and seasonality (Qiu et al., 2024; Zhao et al., 2025c), to enable more comprehensive analysis. Additionally, we discuss the limitations and future works of our study in Appendix E.

AI USE STATEMENT

In this work, we used generative AI tools for improving the clarity, grammar, and overall presentation of the manuscript, as well as for assisting with LaTeX formulation and formatting. We have not used generative AI tools for research ideation, problem formulation, method design, experimental design, experimental execution, data analysis, interpretation of experimental results, or the development of the scientific claims and conclusions of this work. All such aspects of the research were conducted by the authors.

Additionally, generative AI tools were used for language-level writing improvement and LaTeX-related assistance. We have reviewed and verified all AI-assisted content and made the final decisions regarding the wording, mathematical formulations, and presentation. AI-generated suggestions were treated as writing assistance rather than as a source of scientific knowledge or research contributions. We take responsibility for the final content of this work, including all text, claims, mathematical formulations, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide an anonymized code repository containing the implementation of our proposed method as supplementary material. Detailed experimental settings, dataset descriptions, preprocessing procedures, and implementation details are provided in the appendix. Complete proofs of the theoretical results are also included in the appendix.

REFERENCES

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Yongli Cai, Yun Kang, Malay Banerjee, and Weiming Wang. A stochastic sirs epidemic model with infectious force under intervention strategies. *Journal of Differential Equations*, 259(12): 7463–7502, 2015.

Caglar Cakan, Nikola Jajcay, and Klaus Obermayer. neurolib: A simulation framework for whole-brain neural mass modeling. *Cognitive Computation*, 15(4):1132–1152, 2023.

Vitor Cerqueira, Luis Torgo, and Carlos Soares. Machine learning vs statistical methods for time series forecasting: Size matters. *arXiv preprint arXiv:1909.13316*, 2019.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O Arik, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting. *arXiv preprint arXiv:2303.06053*, 2023.

Xueyi Cheng. A comprehensive study of feature selection techniques in machine learning models. *Available at SSRN 5154947*, 2024.

Jonathan Crabbé and Mihaela Van Der Schaar. Explaining time series predictions with dynamic masks. In *International Conference on Machine Learning*, pp. 2166–2177. PMLR, 2021.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent data analysis*, 1(1-4): 131–156, 1997.

Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74:902–924, 2017.

Carson Eisenach, Yagna Patel, and Dhruv Madeka. Mqtransformer: Multi-horizon forecasts with context dependent and feedback-aware attention. *arXiv preprint arXiv:2009.14799*, 2020.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-ts: a general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 7522–7529, 2023.

Cong Feng, Mingjian Cui, Bri-Mathias Hodge, and Jie Zhang. A data-driven multi-model methodology with deep feature selection for short-term wind forecasting. *Applied Energy*, 190:1245–1257, 2017.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. *arXiv preprint arXiv:2105.06643*, 2021.

Google. Google covid-19 community mobility reports. `https://www.google.com/covid19/mobility/`, 2020.

Ning Gui, Danni Ge, and Ziyin Hu. Afs: An attention-based mechanism for supervised feature selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3705–3713, 2019.

Jan Hauke and Tomasz Kossowski. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93, 2011.

Julien Herzen, Francesco Lässig, Samuele Giuliano Piazzetta, Thomas Neuer, Léo Tafti, Guillaume Raille, Tomas Van Pottelbergh, Marek Pasieka, Andrzej Skrodzki, Nicolas Huguenin, et al. Darts: User-friendly modern machine learning for time series. *Journal of Machine Learning Research*, 23 (124):1–6, 2022.

Siu Lau Ho and Min Xie. The use of arima models for reliability forecasting and analysis. *Computers & industrial engineering*, 35(1-2):213–216, 1998.

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. Timefilter: Patch-specific spatial-temporal graph filtration for time series forecasting. *arXiv preprint arXiv:2501.13041*, 2025.

Jaeseok Jang, Chanyoung Jung, and Hail Jung. Optimized time series feature selection for manufacturing ai: Reducing complexity and improving classification accuracy. *IEEE Access*, 2025.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*, 2021.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical transactions of the royal society a: mathematical, physical and engineering sciences*, 379(2194), 2021.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4): 1748–1764, 2021.

Haoxin Liu, Harshavardhan Kamarthi, Lingkai Kong, Zhiyuan Zhao, Chao Zhang, and B Aditya Prakash. Time-series forecasting for out-of-distribution generalization using invariant learning. In *Forty-first International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=SMUXPVKUBg.

Haoxin Liu, Zhiyuan Zhao, Jindong Wang, Harshavardhan Kamarthi, and B Aditya Prakash. Lst-prompt: Large language models as zero-shot time series forecasters by long-short-term prompting. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 7832–7840, 2024b.

Haoxin Liu, Harshavardhan Kamarthi, Zhiyuan Zhao, Shangqing Xu, Shiyu Wang, Qingsong Wen, Tom Hartvigsen, Fei Wang, and B Aditya Prakash. How can time series analysis benefit from multiple modalities? a survey and outlook. *arXiv preprint arXiv:2503.11835*, 2025a.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023a.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=LO7ciRpjI5.

Zhiding Liu, Mingyue Cheng, Zhi Li, Zhenya Huang, Qi Liu, Yanhu Xie, and Enhong Chen. Adaptive normalization for non-stationary time series forecasting: A temporal slice perspective. *Advances in Neural Information Processing Systems*, 36:14273–14292, 2023b.

Ioannis E Livieris, Emmanuel Pintelas, and Panagiotis Pintelas. A cnn–lstm model for gold price time-series forecasting. *Neural computing and applications*, 32(23):17351–17360, 2020.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International journal of forecasting*, 38(4):1346–1364, 2022.

Sarabeth M Mathis, Alexander E Webber, Tomás M León, Erin L Murray, Monica Sun, Lauren A White, Logan C Brooks, Alden Green, Addison J Hu, Roni Rosenfeld, et al. Evaluation of flusight influenza forecasting in the 2021–22 and 2022–23 seasons with a new target laboratory-confirmed influenza hospitalizations. *Nature communications*, 15(1):6289, 2024.

Juntong Ni, Shiyu Wang, Zewen Liu, Xiaoming Shi, Xinyue Zhong, Zhou Ye, and Wei Jin. U-cast: Learning hierarchical structures for high-dimensional time series forecasting. *arXiv preprint arXiv:2507.15119*, 2025.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Tong Niu, Jianzhou Wang, Haiyan Lu, Wendong Yang, and Pei Du. Developing a deep learning framework with two-stage feature selection for multivariate financial time series forecasting. *Expert Systems with Applications*, 148:113237, 2020.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods. *Proc. VLDB Endow.*, 17(9):2363–2377, 2024.

Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. The performance of lstm and bilstm in forecasting time series. In *2019 IEEE International conference on big data (Big Data)*, pp. 3285–3292. IEEE, 2019.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 914–921, 2020.

Mingtian Tan, Mike Merrill, Vinayak Gupta, Tim Althoff, and Tom Hartvigsen. Are language models actually useful for time series forecasting? *Advances in Neural Information Processing Systems*, 37:60162–60191, 2024.

Fadi Thabtah, Firuz Kamalov, Suhel Hammoud, and Seyed Reza Shahamiri. Least loss: A simplified filter method for feature selection. *Information Sciences*, 534:1–15, 2020.

HD Tran, Nitin Muttil, and Bimalka J Chris Perera. Selection of significant input variables for time series forecasting. *Environmental Modelling & Software*, 64:156–163, 2015.

Jorge R Vergara and Pablo A Estévez. A review of feature selection methods based on mutual information. *Neural computing and applications*, 24(1):175–186, 2014.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 38626–38652, 2024a.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in Neural Information Processing Systems*, 37:469–498, 2024b.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024.

Xin Wu, Fei Teng, Xingwang Li, Ji Zhang, Qiang Duan, and Tianrui Li. Out-of-distribution generalization in time series: A survey. *Information Fusion*, pp. 104336, 2026.

Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 856–863, 2003.

Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research*, 5(Oct):1205–1224, 2004.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Zhiyuan Zhao, Haoxin Liu, and B Aditya Prakash. Tackling time-series forecasting generalization via mitigating concept drift. *arXiv preprint arXiv:2510.14814*, 2025a.

Zhiyuan Zhao, Haoxin Liu, Alexander Rodríguez, and B Aditya Prakash. Performative time-series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3968–3979, 2025b.

Zhiyuan Zhao, Juntong Ni, Shangqing Xu, Haoxin Liu, Wei Jin, and B Aditya Prakash. Timerecipe: A time-series forecasting recipe via benchmarking module level effectiveness. *arXiv preprint arXiv:2506.06482*, 2025c.

Zhiyuan Zhao, Sitan Yang, Kin G Olivares, Boris N Oreshkin, Stan Vitebsky, Michael W Mahoney, B Aditya Prakash, and Dmitry Efimov. Tat: Temporal-aligned transformer for multi-horizon peak demand forecasting. *arXiv preprint arXiv:2507.10349*, 2025d.

Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):1–55, 2014.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Pengfei Zhou, Yunlong Liu, Junli Liang, Qi Song, and Xiangyang Li. Crosslinear: Plug-and-play cross-correlation embedding for time series forecasting with exogenous variables. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4120–4131, 2025.

Yunyue Zhu and Dennis Shasha. Statstream: Statistical monitoring of thousands of data streams in real time. In *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*, pp. 358–369. Elsevier, 2002.

A Related Work

A.1 Time-Series Forecasting

Traditional time-series forecasting often relies on statistical models such as ARIMA and exponential smoothing, which typically assume linear dynamics and relatively limited temporal dependence (Ho & Xie, 1998; Cerqueira et al., 2019). With the rise of deep learning, neural models have become a common choice because they can capture nonlinear and multi-scale patterns. Early deep architectures for time series include convolutional neural networks and recurrent models (Siami-Namini et al., 2019; Livieris et al., 2020).

More recently, Transformer-based architectures have become a mainstream approach for long-horizon forecasting, mainly due to their ability to model long-range dependencies and flexible attention-based sequence modeling (Lim et al., 2021; Nie et al., 2023; Liu et al., 2023a). Most recently, several time-series foundation models have been proposed. By pretraining on large-scale time-series corpora, these models can achieve strong zero-shot and few-shot forecasting performance across domains (Ansari et al., 2025; Das et al., 2024; Woo et al., 2024; Liu et al., 2025b). Meanwhile, reusing large language models for time-series forecasting has also become an active research topic (Jin et al., 2023; Liu et al., 2024b; 2025a).

A.2 Feature Selection

Time-series feature selection aims to identify a subset of variables (including exogenous features) that improves forecasting accuracy and robustness while controlling model complexity (Lim & Zohren, 2021). Classic filter and wrapper methods are often designed for i.i.d. data; they may overlook temporal dependence, and can be computationally expensive in high-dimensional and long-horizon settings (Cheng, 2024; Jang et al., 2025).

Recent work integrates feature selection with deep networks and evaluates selection quality on industrial applications (Feng et al., 2017; Niu et al., 2020). In addition, some deep models provide embedded selection signals, such as using attention-related weights as a proxy for feature importance (Gui et al., 2019).

A.3 Out-of-Distribution Generalization

Real-world time series are often non-stationary, so the training and test distributions can differ substantially, which makes out-of-distribution (OOD) generalization a central challenge (Liu et al., 2024a; Zhao et al., 2025b). Existing studies commonly fall into categories such as robust optimization, domain generalization or adaptation, and meta-learning (Wu et al., 2026).

Different from most existing works, in this work, we do not propose a new forecasting architecture. Instead, we propose to measure the potential value of exogenous features from a data perspective, and use this signal to improve the performance of different forecasting models. In particular, we use CDD to identify and select exogenous features with more stable conditional relationships over time. Our solutions are model-agnostic, also different from popular model-internal embedded approaches in conventional feature selection.

B Theoretical Analysis

B.1 Proof of Proposition 2.1

Proof. By the definition of conditional probability,

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)}.$$

Taking logarithms and applying the identity to temporal segments t and t' gives

$$\begin{aligned}\log P_t(Y \mid X) - \log P_{t'}(Y \mid X) = &[\log P_t(X,Y) - \log P_{t'}(X,Y)] \\ &- [\log P_t(X) - \log P_{t'}(X)]\,.\end{aligned}$$

By the triangle inequality,

$$\begin{aligned}\big|\log P_t(Y \mid X) - \log P_{t'}(Y \mid X)\big| \le &\big|\log P_t(X,Y) - \log P_{t'}(X,Y)\big| \\ &+ \big|\log P_t(X) - \log P_{t'}(X)\big|.\end{aligned}$$

Let

$$\delta_X = \big|\log P_t(X) - \log P_{t'}(X)\big|$$

denote the residual marginal discrepancy after instance normalization. Then

$$\big|\Delta_{\text{Conditional}}\big| \le \big|\Delta_{\text{Joint}}\big| + \delta_X.$$

Therefore, when the marginal discrepancy δ_X is small, the joint-distribution discrepancy provides a controlled approximation to the change in the conditional relationship. □

Interpretation. The proposition separates the change in the conditional relationship into two components: the change in the joint distribution and the residual change in the feature marginal. Instance normalization suppresses the latter by reducing temporal variations in the marginal distribution of X. Consequently, the joint-distribution discrepancy becomes a closer proxy for changes in $P(Y \mid X)$. This provides the theoretical motivation for using joint feature–target interactions in CDD rather than explicitly estimating the conditional distribution.

B.2 Theoretical Justification for Algorithm 1

The theoretical justification for using the product interaction $X_L Y_H$ and Wasserstein distance as the discrepancy measure in Algorithm 1. In particular, under bounded normalized variables, the product mapping $(X, Y) \mapsto XY$ is Lipschitz, which implies that a large discrepancy in the product-interaction distribution provides a sufficient signal of non-negligible instability in the underlying joint distribution. Therefore, the proposed CDD can be viewed as a practical and conservative proxy for temporal instability in the feature–target relationship, rather than a direct estimator of the conditional-distribution discrepancy.

Proposition B.1 (Product interaction as a conservative proxy)**.** *Consider two temporal segments t and t' with joint distributions $P_t(X,Y)$ and $P_{t'}(X,Y)$. Assume that the normalized variables satisfy $|X| \le M_X$ and $|Y| \le M_Y$ almost surely. Let P_t^{XY} and $P_{t'}^{XY}$ denote the distributions induced by the product interaction XY. Then*

$$W_1\big(P_t^{XY}, P_{t'}^{XY}\big) \le (M_X + M_Y) W_1(P_t(X,Y), P_{t'}(X,Y))\,.$$

Consequently, a sufficiently large discrepancy in the product interaction implies a non-negligible discrepancy in the underlying joint distribution, although the converse does not necessarily hold.

Proof of Proposition B.1. Let P_t and $P_{t'}$ denote the joint distributions of (X, Y) in two temporal segments, and let

$$f(x, y) = xy.$$

The corresponding product-interaction distributions are the pushforward distributions

$$P_t^{XY} = f_{\#}P_t, \qquad P_{t'}^{XY} = f_{\#}P_{t'}.$$

Under the boundedness assumption $|X| \leq M_X$ and $|Y| \leq M_Y$, the product mapping f is Lipschitz. For any (x, y) and (x', y'),

$$\begin{aligned} |f(x, y) - f(x', y')| &= |xy - x'y'| \\ &= |y(x - x') + x'(y - y')| \\ &\leq |y||x - x'| + |x'||y - y'| \\ &\leq M_Y|x - x'| + M_X|y - y'| \\ &\leq (M_X + M_Y)\|(x, y) - (x', y')\|_1. \end{aligned}$$

Therefore, f is $(M_X + M_Y)$-Lipschitz with respect to the ℓ_1 metric.

Let π be any coupling between P_t and $P_{t'}$. By the definition of the pushforward distributions and the Wasserstein distance,

$$\begin{aligned} W_1\big(P_t^{XY}, P_{t'}^{XY}\big) &\leq \mathbb{E}_{((X,Y),(X',Y'))\sim\pi}\left[|XY - X'Y'|\right] \\ &\leq (M_X + M_Y)\mathbb{E}_\pi\left[\|(X, Y) - (X', Y')\|_1\right]. \end{aligned}$$

Taking the infimum over all couplings π gives

$$W_1\big(P_t^{XY}, P_{t'}^{XY}\big) \leq (M_X + M_Y)W_1(P_t, P_{t'}).$$

This proves Proposition B.1. □

Corollary B.2 (Conservative instability screening)**.** *Under the conditions of Proposition B.1, if*

$$W_1\big(P_t^{XY}, P_{t'}^{XY}\big) > \epsilon,$$

then

$$W_1(P_t(X, Y), P_{t'}(X, Y)) > \frac{\epsilon}{M_X + M_Y}.$$

Therefore, a large discrepancy in the product interaction provides a sufficient signal of non-negligible joint-distribution instability, whereas a small product discrepancy does not imply stability of the full joint distribution.

Proof. The result follows directly from Proposition B.1 by contradiction. If

$$W_1(P_t(X, Y), P_{t'}(X, Y)) \leq \frac{\epsilon}{M_X + M_Y},$$

then Proposition B.1 implies

$$W_1\big(P_t^{XY}, P_{t'}^{XY}\big) \leq \epsilon,$$

which contradicts the assumption. Therefore,

$$W_1(P_t(X, Y), P_{t'}(X, Y)) > \frac{\epsilon}{M_X + M_Y}.$$

□

Interpretation. Proposition B.1 provides a one-sided guarantee for using the product interaction in practice. The mapping $(X, Y) \mapsto XY$ is not information-preserving and therefore cannot fully characterize the joint distribution. However, its temporal instability provides a conservative signal of joint instability: if the distribution of XY changes substantially across temporal segments, then the underlying joint distribution must also have changed. Thus, CDD should be interpreted as a practical proxy for conditional-relation instability rather than a direct estimator of conditional-distribution discrepancy.

The theoretical motivation for CDD follows a sequential chain from conditional-relation instability to a tractable feature-selection signal. We first use the joint distribution as a proxy for the conditional relationship. As shown in Proposition 2.1, the change in the conditional distribution can be decomposed into the change in the joint distribution and the change in the marginal distribution of the feature. Since instance normalization suppresses temporal variation in the feature marginal, the joint-distribution discrepancy provides a controlled approximation to the underlying conditional-relation change. Thus, the joint distribution serves as the first proxy for conditional instability.

We then need a tractable representation of the joint distribution itself. We use the product interaction XY as a further proxy because Proposition B.1 establishes a one-sided relationship between the Wasserstein discrepancy of the product interaction and that of the underlying joint distribution. In particular, a large discrepancy in the distribution of XY necessarily implies a non-negligible discrepancy in the joint distribution, although the converse does not hold. Therefore, temporal changes in the product interaction provide a conservative signal for detecting instability in the joint distribution, and consequently in the underlying conditional relationship. This gives the chain:

$$\text{Conditional Instability} \;\rightarrow\; \text{Joint-Distribution Instability} \;\rightarrow\; \text{One-Sided Guarantee}.$$

CDD operationalizes this chain by measuring the Wasserstein discrepancy of the product interaction across temporal bins. We then summarize its temporal variation using CDD-std and use this signal to identify exogenous features with unstable feature-target relationships for feature selection.

Moreover, the use of Wasserstein distance is a practical choice rather than an essential component of the proposed framework. Wasserstein distance is a well-established measure of distributional discrepancy with desirable theoretical properties, which enables the above analysis and provides a principled way to quantify temporal changes. More generally, our key idea is to explicitly characterize the temporal instability of feature-target relationships and use its variation across temporal segments for feature selection. Other discrepancy measures or statistics could potentially be incorporated into the same framework. For example, one could compute the correlation between each feature and the target within each temporal bin and then characterize the cross-bin variability of these correlations. Exploring such alternative instantiations and their relative effectiveness is left for future work.

C Additional Experimental Setup

For computational resources, all our experiments are performed on NVIDIA Tesla V100 and A100 GPUs. All CDD-related values in all experiments are computed on training period only.

C.1 Setup for Example 2.1

We construct Example 2.1 as follows. We generate a time series over the range $t \in [0, 2000]$, resulting in 2,000 time steps. The first 70% of the sequence is used for training, the final 20% for testing, and the remaining 10% in the middle is reserved for validation.

We set the noise level to $\sigma^2 = 0.15$. For the frequency parameter, we consider $f \in \{5, 10, 20, 30, 50, 75, 100\}$, and the corresponding CDD values are listed in Table 5. In this example, we treat all training time steps as a single bin to obtain a unified reference distribution, as the goal is to understand how intra-bin CDD values are related with pattern complexities.

Remark C.1. As Table 5 shows, larger f does not necessarily refer to a lower CDD. For $f = 75, 100$, the CDD increases because the lookback window fails to include a full signal period, producing a distribution that deviates from the reference distribution.

Table 5: CDD values at different frequencies f.

f	5	10	20	30	50	75	100
CDD	0.055	0.054	0.055	0.089	0.054	0.219	0.278

For the predictor, we employ a simple multilayer perceptron (MLP) with the number of hidden layers varying from 1 to 5. All hidden layers use a fixed hidden dimension of 16 (due to the simplicity of the task) and ReLU activations. Each configuration is evaluated under four random seeds, and we report the mean prediction error averaged over these four runs.

C.2 Setup for Example 2.2

We construct Example 2.2 as follows. We generate a time series over the range $t \in [0, 2000]$, resulting in 2,000 time steps. The first 70% of the sequence is used for training, the final 20% for testing, and the remaining 10% in the middle is reserved for validation.

We consider two frequencies, $f = 5$ and $f = 50$, both of which yield low CDD values in the non-time-varying setting by Table 5, indicating that the conditional distribution is locally stable. We set $P = 200$ to control the switch of frequency. We set the noise variance to $\sigma^2 = 0.15$. In this experiment, we partition the sequence into bins based on the varying time intervals induced by the two frequencies, and compute the mean and standard deviation of the CDD within each bin. This setup allows us to examine how inter-bin CDD patterns correlate with distribution shifts and the resulting challenges for domain adaptation.

For the regular predictor, we employ a simple multilayer perceptron (MLP) with the number of hidden layers varying from 1 to 5. All hidden layers use a fixed dimension of 16, given the simplicity of the task, and ReLU activations. As a basic domain adaptation baseline, we use the same MLP architecture while augmenting the input by concatenating a one-hot domain indicator with the original X (MLP w. Meta Info). We further incorporate the same one-hot domain indicator into the invariant risk minimization (IRM) (Arjovsky et al., 2019) framework, denoted as MLP w. Meta Info & OOD Method. Each configuration is evaluated using four random seeds, and we report the mean prediction error averaged across the four runs.

C.3 Setup for Generalization Analysis

We consider forecasting tasks spanning both short-term and long-term horizons across multiple domains. The datasets employed in our experiments are summarized in Table 6. In particular, we incorporate Time-HD (Ni et al., 2025) as a complementary benchmark datasets to conventional long-term time-series forecasting (LTSF) datasets (Zhou et al., 2021). We omit other popular used datasets, such as PEMS (Song et al., 2020), since they primarily represent the traffic domain, which is already covered by LTSF benchmarks. We also exclude Monash (Godahewa et al., 2021) and Darts (Herzen et al., 2022), as they are univariate time-series datasets. Furthermore, within the Time-HD benchmark, we omit certain datasets whose domains overlap with LTSF (e.g., Weather), as well as datasets with unstable convergence with existing methods (e.g., Wiki), which may introduce bias into the evaluation.

Regarding the choice of forecasting horizons, for the LTSF datasets, we adopt commonly used horizons of 96, 192, 336, and 720 with a fixed lookback window of 96. For the Time-HD datasets, we consider shorter horizons of 12, 24, 48, and 96, while also fixing the lookback window to 96, thereby placing more emphasis on short-term forecasting scenarios. In addition, for the Time-HD datasets, we preprocess the input features by selecting a subset of variables. This is motivated by the high dimensionality of the Time-HD datasets, where only a subset of features is expected to be informative for predicting the target. For instance, the Air Quality dataset in Time-HD contains measurements from hundreds of cities, many of which are geographically distant from the target location. Selecting a subset of features, as a simple form of prior knowledge, helps reduce redundancy and remove variables that are irrelevant to the forecasting target.

For model selection in the validation experiments, we choose TimeMixer (Wang et al., 2024a), iTransformer (Liu et al., 2023a), CrossLinear (Zhou et al., 2025), and TimeFilter (Hu et al., 2025),

Table 6: Description of the datasets used in our evaluations.

Datasets		Domain	Description
LTSF	ECL	Energy	Electricity consumption load.
	ETTh1/h2	Energy	Electricity transformer temperature measurements.
	Exchange	Finance	Foreign exchange rates across multiple countries.
	Solar	Energy	Solar photovoltaic power generation from the National Renewable Energy Laboratory.
	Traffic	Social	California freeway road occupancy rates.
	Weather	Environment	Meteorological data collected by the Max Planck Institute for Biogeochemistry.
Time-HD	Air Quality	Environment	Air quality measurements across multiple cities in China.
	M5	Sale	Item-level retail sales data from CA, TX, and WI (Makridakis et al., 2022).
	Mobility	Social	Community mobility patterns used for COVID-19 analysis (Google, 2020).
	Neurolib	NeuralScience	Simulated neural dynamics generated using the Neurolib framework (Cakan et al., 2023).
	SIRS	Epidemiology	Simulations extending the classical stochastic SIR epidemic model (Cai et al., 2015).
	SP500	Finance	Stock market index maintained by S&P Dow Jones Indices.

which are recent advanced supervised forecasting models that support both univariate forecasting and forecasting with exogenous features. This capability is essential for a fair comparison when evaluating the utility of exogenous features. We exclude other recent models for the following reasons. PatchTST (Nie et al., 2023) and PAttn (Tan et al., 2024) are primarily univariate and channel-independent, and therefore do not support forecasting with exogenous features. TimeXer (Wang et al., 2024b) only supports forecasting with exogenous features but does not admit a univariate setting. We also exclude foundation models (Aksu et al., 2024; Ansari et al., 2025), as our goal is to analyze feature utility within a supervised forecasting setting.

C.4 Setup for Evaluation

We evaluate the performance of our proposed CDD-std-based feature selection approaches in comparison with conventional forecasting using all exogenous features. The experimental setup largely follows Appendix C.3, and additionally includes the two unseen datasets introduced in Section 4.3. For instance, we adopt a fixed lookback length of 96, and all results are averaged over four random seeds.

For the thresholding method, we report the remaining feature dimensions after filtering in Table 7, where the reported feature dimensions after CDD-std–based hard thresholding are averaged across four forecasting horizons. As a result, the resulting feature dimensions may be non-integer values, as observed for the Traffic dataset. It is worth noting that for the Mobility and M5 datasets, the threshold-based filtering removes all exogenous features, reducing the forecasting task to univariate forecasting. This result reflects the adaptive nature of our feature selection strategy: when the exogenous features exhibit sufficiently high temporal instability, the method determines that none of them provides reliable predictive information and falls back to the target's historical observations alone. Importantly, this does not introduce an additional forecasting method, but rather represents a possible outcome of the proposed feature selection procedure.

Table 7: Feature dimensions before and after CDD-std threshold filtering. Datasets with higher CDD-std retain fewer features in general.

Dataset	Original	Filtered	Ratio
SIRS	20	20	1.00
Traffic	862	734.5	0.85
Solar	137	137	1.00
Air Quality	20	18	0.90
ECL	321	229	0.71
ETTh1	7	5	0.71
Weather	21	13	0.62
Exchange	8	2	0.25
ETTh2	7	3	0.43
Mobility	20	1	0.05
M5	20	1	0.05

Table 8: Hyperparameter sensitivity analysis of the CDD Algorithm (Algorithm 1) with respect to the number of bins K. Spearman rank correlation tests show that the dataset-wise rankings induced by CDD-std remain highly consistent across different choices of K, indicating that both the algorithm and the CDD-std metric are robust and largely insensitive to K within a reasonable range.

Dataset	$K = 8$ (Ref.)		$K = 4$		$K = 6$		$K = 10$		$K = 12$	
	Rank	CDD-std	Rank	CDD-std	Rank	CDD-std	Rank	CDD-std	Rank	CDD-std
SIRS	1	0.037	1	0.017	1	0.039	1	0.044	3	0.089
Traffic	2	0.057	2	0.029	2	0.046	2	0.054	1	0.061
Solar	3	0.071	4	0.065	3	0.069	3	0.088	2	0.073
ECL	4	0.105	3	0.062	4	0.113	4	0.112	4	0.099
Air Quality	5	0.131	5	0.103	6	0.115	6	0.147	6	0.112
ETTh1	6	0.150	7	0.170	5	0.114	7	0.155	7	0.140
Weather	7	0.157	8	0.198	8	0.186	9	0.173	5	0.101
Exchange	8	0.172	6	0.132	7	0.129	5	0.141	8	0.185
ETTh2	9	0.189	9	0.214	9	0.211	8	0.167	9	0.192
Mobility	10	0.408	10	0.312	10	0.292	10	0.314	10	0.319
M5	11	0.803	11	0.624	11	0.752	11	0.898	11	0.865
Spearman	–		$\rho = 0.963$ $p = 1.85 \times 10^{-6}$		$\rho = 0.982$ $p = 8.40 \times 10^{-8}$		$\rho = 0.927$ $p = 3.97 \times 10^{-5}$		$\rho = 0.945$ $p = 1.12 \times 10^{-5}$	

Moreover and notably, the evaluation in Section 4.1 chooses exhibit instability, where feature selection is expected to be more important. In contrast, for datasets with smaller CDD-std values, our approaches, especially the thresholding strategy, tend to retain most exogenous features. As a result, the performance differences become marginal and are more likely attributable to experimental noise rather than substantive gains from feature selection. We further validate this through analysis in Table 16.

Statistical Analysis. We use paired statistical analysis to assess whether the observed performance differences between methods are statistically significant. Specifically, we perform paired comparisons across model–dataset–horizon configurations, using the performance averaged over multiple random seeds for each configuration. This analysis treats each model–dataset–horizon configuration as a paired comparison rather than treating individual random seeds as independent samples. We report the resulting p-values as supplementary statistical evidence, while primarily considering the consistency and magnitude of performance changes across configurations. For experiments on stable datasets, we similarly use paired analysis to assess whether the observed performance changes indicate statistically significant degradation.

D Additional Results

D.1 Hyperparameter Sensitivity

We conduct a hyperparameter sensitivity study to examine the robustness of CDD with respect to the number of bins K. Using the datasets from the generalization analysis (Section 2.4), we vary K from 4 to 12 with a step size of 2. We aim to demonstrate that the CDD-std values and the relative ranking of datasets exhibit minimal variations across different choices of K, which reflects the overall utility of exogenous features. The detailed ranking results are reported in Table 8.

The evaluation results demonstrate the robustness of the proposed CDD algorithm and its associated metric, CDD-std. In particular, across a wide range of choices for K, the dataset-wise rankings induced by CDD-std remain highly stable, indicating that CDD-std is largely insensitive to reasonable variations in this hyperparameter.

We also observe a systematic trend: smaller K (coarser bins) tends to produce lower CDD-std values due to stronger temporal averaging, while larger K (finer bins) leads to higher CDD-std values by capturing more localized variations. This reflects a bias–variance trade-off in estimating conditional patterns over time. In practice, we use $K = 8$ for all experiments, and note that the threshold used in Section 3.1 may need to be recalibrated when K changes.

D.2 Detailed Results for Generalization Analysis

We provide the detailed numerical results corresponding to the CDD-std generalization analysis in Section 2.4. The results summarized in Figure 3 are reported in Tables 9 and 10. Performance changes are computed as the average relative difference between forecasting with exogenous features and univariate forecasting. Performance gains, corresponding to lower forecasting errors when incorporating exogenous features, are highlighted in green with a "+" symbol, while performance degradations, corresponding to higher forecasting errors, are highlighted in red with a "–" symbol. Figure 3 visualizes the corresponding changes in MSE across datasets.

D.3 Detailed Results for Performance Gain via Feature Selection

We provide the detailed results corresponding to the experiments of Table 1 in Tables 11 and 12. Tables 11 and 12 report the forecasting errors for each prediction horizon, providing a more detailed view of the performance of the proposed feature selection approaches.

D.4 Detailed Results for Comparison with Feature Selection Approaches

We provide the detailed results corresponding to the experiments of Table 2 in Table 13. Table 13 reports the forecasting errors for each prediction horizon for the proposed CDD-std-guided approaches and the compared feature selection methods.

D.5 Detailed Results for Generalization to Unseen Datasets and Models

We provide the detailed results corresponding to the experiments of Table 3 in Table 14. Table 14 reports the forecasting errors for each prediction horizon across the unseen datasets and forecasting models, providing a more detailed evaluation of the generalization ability of CDD-std-guided feature selection.

D.6 Detailed Results for Ablation Analysis on Metadata Guidance

We provide the detailed results of the ablation study in Table 15. Table 15 reports the forecasting errors for each prediction horizon, allowing a detailed comparison between the original CDD-std-guided approach and the constant-metadata ablation.

D.7 Detailed Results for Evaluation on Stable Datasets

We provide the detailed results on stable datasets in Table 16. Table 16 reports the forecasting errors for each prediction horizon, showing marginal performance difference between CDD-std-based feature selections methods than without feature selection on stable datasets.

D.8 Detailed Results for Generalization to Foundation Models

We provide the detailed results corresponding to the experiments of Table 4 in Table 17. Table 17 reports the forecasting errors for each prediction horizon. The results provide further evidence that CDD-std-guided feature selection can benefit zero-shot forecasting with pre-trained time-series foundation models.

E Limitation and Future Work

While our findings consistently show that temporally unstable exogenous features are less effective for forecasting, the current study is primarily empirical. Providing a more rigorous theoretical characterization of how temporal instability in conditional relationships affects generalization remains an important direction for future work.

Moreover, extending the proposed CDD-based metric and feature selection strategies to large-scale foundation models presents additional challenges. In such settings, feature representations are often rich enough that may alter the notion of conditional stability. Understanding how CDD interacts with

Table 9: Performance comparison between univariate forecasting and forecasting with exogenous features. Datasets are ordered by increasing CDD-std. Incorporating exogenous features does not always improve performance: gains are shown in green, while degradations are shown in red.

Dataset	Horizon	iTransformer				TimeMixer			
		Univariate		w. Exo. Feat.		Univariate		w. Exo. Feat.	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
SIRS	12	0.00139	0.0297	0.00105	0.0254	0.00057	0.0184	0.00054	0.0178
	24	0.00134	0.0277	0.00099	0.0245	0.00074	0.0215	0.00096	0.0241
	48	0.00125	0.0275	0.00099	0.0258	0.00159	0.0308	0.00176	0.0328
	96	0.00102	0.0242	0.00078	0.0217	0.00428	0.0488	0.00308	0.0436
				+23.7%	+12.4%			+11.8%	+10.6%
Traffic	96	0.169	0.247	0.150	0.228	0.236	0.329	0.203	0.301
	192	0.161	0.241	0.146	0.225	0.210	0.303	0.200	0.299
	336	0.158	0.244	0.144	0.230	0.198	0.294	0.192	0.294
	720	0.180	0.266	0.164	0.252	0.225	0.318	0.211	0.311
				+9.6%	+6.4%			+7.3%	+3.3%
Solar	96	0.307	0.311	0.287	0.296	0.297	0.322	0.310	0.338
	192	0.346	0.331	0.330	0.321	0.338	0.344	0.345	0.366
	336	0.375	0.349	0.346	0.336	0.370	0.358	0.367	0.376
	720	0.373	0.346	0.340	0.334	0.367	0.356	0.359	0.369
				+7.0%	+3.7%			-0.7%	-4.9%
Air	12	0.0107	0.0728	0.0101	0.0708	0.0105	0.0724	0.0105	0.0729
	24	0.0120	0.0777	0.0116	0.0770	0.0114	0.0761	0.0116	0.0778
	48	0.0122	0.0784	0.0123	0.0791	0.0118	0.0775	0.0120	0.0786
	96	0.0121	0.0784	0.0124	0.0797	0.0117	0.0779	0.0118	0.0784
				+1.3%	+0.3%			-1.2%	-1.2%
ECL	96	0.300	0.390	0.299	0.397	0.317	0.405	0.337	0.430
	192	0.319	0.396	0.320	0.406	0.337	0.413	0.381	0.460
	336	0.375	0.433	0.369	0.437	0.391	0.451	0.402	0.472
	720	0.458	0.494	0.432	0.484	0.457	0.498	0.429	0.491
				+2.3%	-0.7%			-3.2%	-4.9%
ETTh1	96	0.107	0.246	0.104	0.246	0.105	0.247	0.109	0.255
	192	0.129	0.274	0.127	0.274	0.128	0.274	0.135	0.284
	336	0.135	0.287	0.136	0.289	0.131	0.285	0.143	0.297
	720	0.186	0.340	0.190	0.345	0.192	0.346	0.216	0.370
				-0.1%	-0.5%			-8.4%	-4.7%
Weather	96	0.00135	0.0267	0.00131	0.0263	0.00126	0.0258	0.00126	0.0258
	192	0.00161	0.0299	0.00149	0.0287	0.00154	0.0291	0.00150	0.0287
	336	0.00175	0.0315	0.00163	0.0303	0.00167	0.0306	0.00163	0.0302
	720	0.00223	0.0359	0.00212	0.0347	0.00218	0.0354	0.00210	0.0346
				+5.5%	+3.2%			+2.6%	+1.4%
Exchange	96	0.103	0.234	0.105	0.237	0.098	0.230	0.114	0.254
	192	0.220	0.346	0.241	0.367	0.211	0.351	0.224	0.362
	336	0.424	0.491	0.465	0.507	0.422	0.492	0.444	0.510
	720	1.128	0.809	1.146	0.815	1.169	0.829	1.232	0.842
				-4.3%	-2.4%			-6.1%	-3.5%
ETTh2	96	0.259	0.385	0.269	0.395	0.265	0.391	0.276	0.403
	192	0.309	0.435	0.318	0.446	0.319	0.440	0.344	0.461
	336	0.344	0.466	0.395	0.502	0.361	0.475	0.391	0.495
	720	0.512	0.576	0.586	0.619	0.515	0.575	0.556	0.598
				-9.9%	-5.4%			-7.5%	-4.0%
Mobility	12	0.104	0.255	0.113	0.259	0.127	0.283	0.135	0.292
	24	0.114	0.269	0.118	0.273	0.136	0.292	0.146	0.304
	48	0.124	0.282	0.126	0.285	0.128	0.284	0.138	0.295
	96	0.147	0.310	0.168	0.332	0.137	0.295	0.170	0.332
				-7.3%	-3.1%			-11.5%	-6.0%
M5	12	0.368	0.466	0.372	0.470	0.348	0.462	0.411	0.511
	24	0.396	0.490	0.412	0.502	0.372	0.489	0.424	0.520
	48	0.533	0.582	0.464	0.539	0.416	0.525	0.474	0.560
	96	0.603	0.626	0.580	0.603	0.515	0.590	0.690	0.677
				+3.8%	+2.4%			-21.1%	-9.8%

representation learning and evaluating its effectiveness in these settings is a promising direction for future research.

Table 10: Performance comparison between univariate forecasting and forecasting with exogenous features. Datasets are ordered by increasing CDD-std. Incorporating exogenous features does not always improve performance: gains are shown in green, while degradations are shown in red.

Dataset	Horizon	CrossLinear				TimeFilter			
		Univariate		w. Exo. Feat.		Univariate		w. Exo. Feat.	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
SIRS	12	0.00005	0.0056	0.00005	0.0056	0.00019	0.0106	0.00018	0.0104
	24	0.00007	0.0066	0.00007	0.0066	0.00025	0.0121	0.00024	0.0121
	48	0.00010	0.0080	0.00010	0.0079	0.00028	0.0130	0.00026	0.0128
	96	0.00012	0.0089	0.00013	0.0089	0.00025	0.0123	0.00024	0.0119
				+0.6%	+0.3%			+5.3%	+1.7%
Traffic	96	0.146	0.225	0.146	0.230	0.164	0.247	0.156	0.235
	192	0.149	0.228	0.149	0.228	0.158	0.238	0.154	0.233
	336	0.150	0.233	0.145	0.229	0.154	0.239	0.154	0.239
	720	0.168	0.256	0.172	0.255	0.181	0.265	0.179	0.265
				+0.1%	+0.0%			+2.3%	+1.7%
Solar	96	0.291	0.303	0.282	0.298	0.296	0.311	0.290	0.299
	192	0.341	0.334	0.320	0.316	0.335	0.328	0.333	0.326
	336	0.346	0.336	0.337	0.327	0.355	0.342	0.352	0.336
	720	0.346	0.335	0.343	0.334	0.348	0.337	0.347	0.338
				+3.7%	+2.5%			+0.9%	+1.5%
Air	12	0.0106	0.0725	0.0107	0.0728	0.0106	0.0720	0.0105	0.0718
	24	0.0117	0.0772	0.0116	0.0762	0.0116	0.0769	0.0116	0.0764
	48	0.0122	0.0793	0.0122	0.0792	0.0122	0.0791	0.0122	0.0782
	96	0.0121	0.0787	0.0122	0.0795	0.0119	0.0783	0.0120	0.0791
				+0.0%	+0.0%			-0.2%	+0.2%
ECL	96	0.308	0.398	0.289	0.388	0.308	0.399	0.276	0.379
	192	0.336	0.408	0.316	0.399	0.328	0.406	0.315	0.399
	336	0.386	0.439	0.367	0.431	0.386	0.443	0.372	0.435
	720	0.461	0.495	0.445	0.487	0.449	0.491	0.421	0.478
				+0.5%	+0.2%			+5.9%	+2.8%
ETTh1	96	0.108	0.248	0.107	0.247	0.108	0.250	0.105	0.246
	192	0.139	0.282	0.133	0.276	0.132	0.278	0.131	0.278
	336	0.149	0.300	0.150	0.301	0.142	0.293	0.141	0.294
	720	0.198	0.352	0.197	0.352	0.191	0.346	0.180	0.336
				+1.2%	+0.5%			+2.7%	+1.2%
Weather	96	0.00148	0.0274	0.00146	0.0275	0.00143	0.0271	0.00135	0.0264
	192	0.00164	0.0296	0.00158	0.0291	0.00165	0.0298	0.00167	0.0300
	336	0.00173	0.0307	0.00175	0.0308	0.00177	0.0312	0.00177	0.0311
	720	0.00214	0.0342	0.00215	0.0343	0.00224	0.0354	0.00216	0.0348
				+0.1%	+0.0%			+2.0%	+1.1%
Exchange	96	0.106	0.239	0.104	0.238	0.104	0.237	0.110	0.242
	192	0.213	0.347	0.213	0.345	0.209	0.341	0.209	0.340
	336	0.435	0.495	0.420	0.491	0.413	0.483	0.418	0.483
	720	1.117	0.807	1.107	0.802	1.224	0.848	1.332	0.885
				+1.3%	+0.6%			-6.1%	-2.2%
ETTh2	96	0.259	0.388	0.258	0.385	0.266	0.391	0.270	0.394
	192	0.303	0.430	0.307	0.434	0.307	0.433	0.319	0.444
	336	0.337	0.460	0.347	0.468	0.349	0.469	0.378	0.490
	720	0.485	0.558	0.479	0.554	0.515	0.577	0.605	0.629
				-0.5%	-0.3%			-9.4%	-4.7%
Mobility	12	0.091	0.236	0.095	0.243	0.096	0.242	0.096	0.241
	24	0.111	0.265	0.109	0.264	0.114	0.267	0.118	0.272
	48	0.122	0.277	0.122	0.278	0.121	0.279	0.124	0.281
	96	0.136	0.299	0.141	0.304	0.133	0.294	0.152	0.316
				-1.5%	-1.1%			-5.6%	-2.6%
M5	12	0.342	0.457	0.358	0.471	0.353	0.456	0.352	0.445
	24	0.374	0.490	0.381	0.495	0.367	0.472	0.373	0.477
	48	0.430	0.534	0.437	0.539	0.419	0.516	0.439	0.531
	96	0.558	0.615	0.552	0.611	0.517	0.579	0.532	0.588
				-1.4%	-0.9%			-2.5%	-1.4%

Table 11: Detail performance comparison between forecasting with all exogenous features (Base), CDD-std threshold feature selection (w. Threshold), and TIMELESS, with iTransformer and TimeMixer. Our feature selection approaches improve performance in 39 out of 48 evaluations.

Dataset	Horizon	iTransformer						TimeMixer					
		Base		w. Threshold		w. TIMELESS		Base		w. Threshold		w. TIMELESS	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.104	0.246	0.106	0.246	0.108	0.247	0.109	0.255	0.108	0.252	0.107	0.251
	192	0.127	0.274	0.128	0.275	0.130	0.274	0.135	0.284	0.131	0.279	0.130	0.277
	336	0.136	0.289	0.136	0.288	0.135	0.286	0.143	0.297	0.144	0.297	0.137	0.289
	720	0.190	0.345	0.192	0.345	0.184	0.338	0.216	0.370	0.214	0.367	0.174	0.332
				-0.9%	+0.0%	+0.0%	+0.8%			+1.0%	+0.9%	+9.1%	+4.7%
Weather	96	0.00131	0.0263	0.00136	0.0268	0.00132	0.0265	0.00126	0.0258	0.00128	0.0259	0.00126	0.0258
	192	0.00149	0.0287	0.00155	0.0290	0.00151	0.0286	0.00150	0.0287	0.00153	0.0290	0.00158	0.0294
	336	0.00163	0.0303	0.00166	0.0301	0.00163	0.0300	0.00163	0.0302	0.00163	0.0302	0.00167	0.0307
	720	0.00212	0.0347	0.00207	0.0341	0.00205	0.0338	0.00210	0.0346	0.00210	0.0345	0.00219	0.0356
				-1.4%	+0.0%	+0.6%	+0.9%			-0.8%	-0.3%	-3.2%	-1.8%
Exchange	96	0.105	0.237	0.108	0.240	0.101	0.233	0.114	0.254	0.104	0.238	0.103	0.245
	192	0.241	0.367	0.220	0.346	0.212	0.343	0.224	0.362	0.211	0.351	0.213	0.353
	336	0.465	0.507	0.377	0.468	0.420	0.487	0.444	0.510	0.416	0.493	0.424	0.496
	720	1.146	0.815	1.143	0.820	1.112	0.802	1.232	0.842	1.032	0.789	1.160	0.829
				+5.6%	+2.7%	+5.7%	+3.2%			+12.5%	+4.9%	+5.7%	+2.3%
ETTh2	96	0.269	0.395	0.262	0.388	0.260	0.386	0.276	0.403	0.271	0.399	0.275	0.399
	192	0.318	0.446	0.319	0.445	0.310	0.435	0.344	0.461	0.331	0.451	0.333	0.451
	336	0.395	0.502	0.354	0.471	0.346	0.466	0.391	0.495	0.373	0.481	0.360	0.471
	720	0.586	0.619	0.598	0.628	0.504	0.571	0.556	0.598	0.602	0.626	0.512	0.568
				+2.2%	+1.5%	+9.4%	+5.3%			-0.6%	+0.0%	+5.6%	+3.5%
Mobility	12	0.113	0.259	0.104	0.255	0.0985	0.247	0.135	0.292	0.127	0.283	0.121	0.277
	24	0.118	0.273	0.114	0.269	0.111	0.265	0.146	0.304	0.136	0.292	0.132	0.289
	48	0.126	0.285	0.124	0.282	0.124	0.282	0.138	0.295	0.128	0.284	0.130	0.287
	96	0.168	0.332	0.147	0.310	0.144	0.306	0.170	0.332	0.137	0.295	0.159	0.320
				+6.9%	+2.9%	+9.0%	+4.3%			+10.4%	+5.6%	+8.0%	+4.1%
M5	12	0.372	0.470	0.368	0.466	0.356	0.456	0.411	0.511	0.348	0.462	0.370	0.486
	24	0.412	0.502	0.396	0.490	0.393	0.494	0.424	0.520	0.372	0.489	0.400	0.504
	48	0.464	0.539	0.533	0.582	0.456	0.543	0.474	0.560	0.416	0.525	0.446	0.539
	96	0.580	0.603	0.603	0.626	0.576	0.621	0.690	0.677	0.515	0.590	0.602	0.632
				-3.9%	-2.3%	+2.6%	+0.0%			+17.4%	+8.9%	+9.1%	+4.7%

Table 12: Detail performance comparison between forecasting with all exogenous features (Base), CDD-std threshold feature selection (w. Threshold), and TIMELESS, with CrossLinear and TimeFilter. Our feature selection approaches improve performance in 42 out of 48 evaluations.

Dataset	Horizon	CrossLinear						TimeFilter					
		Base		w. Threshold		w. TIMELESS		Base		w. Threshold		w. TIMELESS	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.107	0.247	0.107	0.247	0.107	0.247	0.105	0.246	0.105	0.246	0.105	0.246
	192	0.133	0.276	0.136	0.280	0.135	0.278	0.131	0.278	0.129	0.276	0.129	0.276
	336	0.150	0.301	0.146	0.298	0.146	0.298	0.141	0.294	0.142	0.294	0.142	0.294
	720	0.197	0.352	0.192	0.347	0.192	0.347	0.180	0.336	0.197	0.352	0.191	0.347
				+1.1%	+0.4%	+1.2%	+0.6%			-2.8%	-1.2%	-1.7%	-0.7%
Weather	96	0.00146	0.0275	0.00140	0.0271	0.00139	0.0270	0.00135	0.0264	0.00142	0.0271	0.00135	0.0266
	192	0.00158	0.0291	0.00164	0.0297	0.00164	0.0296	0.00167	0.0300	0.00168	0.0300	0.00166	0.0299
	336	0.00175	0.0308	0.00168	0.0303	0.00168	0.0303	0.00177	0.0311	0.00176	0.0314	0.00173	0.0309
	720	0.00215	0.0343	0.00212	0.0341	0.00212	0.0342	0.00216	0.0348	0.00222	0.0355	0.00220	0.0352
				+1.5%	+0.5%	+1.6%	+0.5%			-1.8%	-1.3%	+0.0%	+0.0%
Exchange	96	0.104	0.238	0.101	0.235	0.101	0.235	0.110	0.242	0.105	0.238	0.103	0.235
	192	0.213	0.345	0.219	0.348	0.209	0.339	0.209	0.340	0.216	0.344	0.215	0.344
	336	0.420	0.491	0.422	0.491	0.422	0.491	0.418	0.483	0.432	0.491	0.429	0.489
	720	1.107	0.802	1.091	0.796	1.091	0.796	1.332	0.885	1.163	0.823	1.147	0.802
				+0.6%	+0.4%	+1.2%	+0.8%			+7.4%	+2.8%	+8.5%	+4.2%
ETTh2	96	0.258	0.385	0.258	0.386	0.258	0.386	0.270	0.394	0.264	0.387	0.266	0.392
	192	0.307	0.434	0.309	0.435	0.307	0.434	0.319	0.444	0.308	0.434	0.314	0.438
	336	0.347	0.468	0.339	0.463	0.332	0.455	0.378	0.490	0.346	0.466	0.347	0.466
	720	0.479	0.554	0.474	0.551	0.457	0.539	0.605	0.629	0.511	0.573	0.515	0.575
				+0.8%	+0.4%	+2.3%	+1.5%			+9.1%	+5.0%	+8.3%	+4.4%
Mobility	12	0.095	0.243	0.091	0.236	0.091	0.235	0.096	0.241	0.095	0.242	0.096	0.241
	24	0.109	0.264	0.111	0.265	0.110	0.265	0.118	0.272	0.113	0.267	0.113	0.268
	48	0.122	0.278	0.122	0.277	0.122	0.278	0.124	0.281	0.120	0.277	0.122	0.278
	96	0.141	0.304	0.136	0.299	0.140	0.303	0.152	0.316	0.133	0.293	0.139	0.300
				+1.5%	+1.2%	+0.9%	+0.8%			+6.0%	+2.8%	+4.1%	+2.1%
M5	12	0.358	0.471	0.342	0.457	0.343	0.456	0.352	0.456	0.353	0.455	0.355	0.457
	24	0.381	0.495	0.374	0.490	0.366	0.483	0.373	0.477	0.366	0.471	0.379	0.478
	48	0.437	0.539	0.430	0.534	0.434	0.535	0.439	0.531	0.418	0.515	0.423	0.518
	96	0.552	0.611	0.558	0.615	0.544	0.606	0.532	0.588	0.516	0.579	0.533	0.589
				+1.4%	+1.0%	+2.4%	+1.8%			+2.6%	+1.6%	+0.4%	+0.5%

Table 13: Detailed forecasting errors of our hard filtering and TIMELESS compared with Spearman correlation, mutual information (MI), and mRMR, with fixed feature selection budget (top 1/3).

Model	Dataset	Horizon	Spearman		MI		mRMR		w. Ranking		w. TIMELESS	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	**ETTh2**	96	0.278	0.404	0.280	0.405	0.280	0.405	0.276	0.401	0.266	0.395
		192	0.324	0.445	0.319	0.447	0.319	0.447	0.322	0.446	0.316	0.441
		336	0.357	0.475	0.385	0.497	0.385	0.497	0.360	0.475	0.360	0.478
		720	0.502	0.566	0.606	0.636	0.606	0.636	0.584	0.609	0.481	0.554
	Exchange	96	0.101	0.236	0.106	0.243	0.102	0.239	0.099	0.232	0.100	0.234
		192	0.202	0.345	0.209	0.349	0.203	0.346	0.207	0.348	0.205	0.341
		336	0.403	0.482	0.403	0.486	0.386	0.474	0.376	0.468	0.381	0.471
		720	1.312	0.870	1.101	0.795	1.069	0.788	1.101	0.806	1.063	0.783
TimeMixer	**ETTh2**	96	0.269	0.394	0.270	0.397	0.270	0.397	0.264	0.390	0.269	0.395
		192	0.330	0.450	0.326	0.448	0.326	0.448	0.327	0.446	0.324	0.445
		336	0.359	0.470	0.401	0.503	0.401	0.503	0.355	0.468	0.377	0.486
		720	0.534	0.589	0.580	0.616	0.580	0.616	0.572	0.607	0.550	0.593
	Exchange	96	0.100	0.233	0.103	0.238	0.105	0.241	0.101	0.235	0.102	0.236
		192	0.218	0.355	0.216	0.353	0.215	0.352	0.213	0.350	0.216	0.352
		336	0.429	0.494	0.430	0.494	0.428	0.493	0.425	0.489	0.422	0.489
		720	1.172	0.846	1.113	0.809	1.211	0.824	1.110	0.812	1.126	0.811
CrossLinear	**ETTh2**	96	0.261	0.387	0.261	0.387	0.261	0.387	0.259	0.385	0.260	0.384
		192	0.315	0.439	0.315	0.439	0.315	0.439	0.312	0.438	0.313	0.440
		336	0.356	0.474	0.355	0.474	0.355	0.474	0.352	0.471	0.351	0.470
		720	0.500	0.567	0.503	0.570	0.503	0.570	0.493	0.563	0.480	0.555
	Exchange	96	0.110	0.249	0.109	0.248	0.108	0.247	0.107	0.246	0.109	0.247
		192	0.216	0.352	0.216	0.352	0.216	0.352	0.211	0.344	0.210	0.344
		336	0.415	0.487	0.415	0.488	0.415	0.488	0.409	0.485	0.411	0.486
		720	1.064	0.794	1.064	0.794	1.064	0.794	1.062	0.793	1.060	0.789
TimeFilter	**ETTh2**	96	0.271	0.395	0.277	0.401	0.277	0.401	0.275	0.400	0.258	0.383
		192	0.320	0.442	0.321	0.445	0.321	0.445	0.322	0.443	0.303	0.429
		336	0.365	0.481	0.374	0.485	0.374	0.485	0.365	0.477	0.347	0.467
		720	0.531	0.584	0.626	0.640	0.626	0.640	0.559	0.560	0.500	0.567
	Exchange	96	0.100	0.233	0.106	0.241	0.101	0.233	0.099	0.232	0.102	0.235
		192	0.221	0.348	0.212	0.344	0.208	0.340	0.208	0.340	0.205	0.335
		336	0.429	0.490	0.428	0.491	0.420	0.485	0.405	0.476	0.424	0.488
		720	1.169	0.825	1.205	0.830	1.097	0.796	1.110	0.808	1.113	0.804

Table 14: Performance comparison between forecasting with all exogenous features (Base), CDD-std threshold feature selection (w. Threshold), and TIMELESS on unseen datasets and models. Both feature selection approaches consistently improve performance across evaluations, demonstrating the generalization ability and practical effectiveness of CDD-std-based methods.

Dataset	Horizon	TimeXer						TSMixer					
		Base		w. Threshold		w. TIMELESS		Base		w. Threshold		w. TIMELESS	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Neuro	12	0.00109	0.0098	0.00109	0.0098	0.00089	0.0089	0.00098	0.0093	0.00098	0.0093	0.00067	0.00767
	24	0.00263	0.0159	0.00263	0.0159	0.00232	0.0143	0.00325	0.0177	0.00325	0.0177	0.00209	0.0140
	48	0.0100	0.0315	0.0086	0.0314	0.0097	0.0284	0.0192	0.0432	0.0163	0.0405	0.0108	0.0326
	96	0.0562	0.0734	0.0505	0.0732	0.0527	0.0731	0.1080	0.0952	0.0755	0.0804	0.0406	0.0637
				+10.2%	+0.2%	+6.2%	+4.5%			+26.9%	+10.6%	+58.8%	+28.7%
SP500	12	0.783	0.660	0.741	0.661	0.760	0.647	0.915	0.724	0.866	0.704	0.829	0.683
	24	1.446	0.906	1.417	0.886	1.366	0.870	1.698	0.987	1.708	0.985	1.554	0.937
	48	3.033	1.291	2.768	1.256	2.626	1.224	3.824	1.458	3.276	1.379	2.653	1.243
	96	5.962	1.766	5.549	1.750	5.397	1.709	7.356	1.987	5.175	1.752	4.932	1.641
				+6.7%	+1.5%	+9.6%	+3.7%			+20.1%	+6.5%	+27.7%	+12.6%

Table 15: Ablation study on metadata inputs. Averaged forecasting error across all prediction horizons. **Base** denotes the standard multivariate model, **Constant** replaces the CDD-std metadata input with a constant value of 1, and **CDD-std** uses the standard metadata input in TIMELESS. Best results within each model are **bold**, while second-best results are underlined.

Dataset	Model	Base		Constant		CDD-std	
		MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	iTransformer	0.139	0.289	**0.138**	0.287	0.139	**0.286**
	TimeMixer	0.151	0.302	0.152	0.302	**0.137**	**0.287**
	CrossLinear	0.147	0.294	0.148	0.296	**0.145**	**0.293**
	TimeFilter	**0.139**	**0.289**	0.146	0.295	0.142	0.291
Weather	iTransformer	0.00164	0.0300	0.00167	0.0303	**0.00163**	**0.0297**
	TimeMixer	0.00162	**0.0298**	**0.00161**	**0.0298**	0.00168	0.0304
	CrossLinear	0.00174	0.0304	0.00174	0.0305	**0.00171**	**0.0302**
	TimeFilter	0.00174	0.0306	**0.00166**	**0.0300**	0.00174	0.0306
Exchange	iTransformer	0.489	0.482	0.544	0.493	**0.461**	**0.466**
	TimeMixer	0.504	0.492	0.552	0.511	**0.475**	**0.481**
	CrossLinear	0.461	0.469	0.462	0.469	**0.456**	**0.465**
	TimeFilter	0.517	0.488	0.539	0.494	**0.474**	**0.468**
ETTh2	iTransformer	0.392	0.491	0.405	0.498	**0.355**	**0.465**
	TimeMixer	0.392	0.489	0.387	0.486	**0.370**	**0.472**
	CrossLinear	0.348	0.460	0.350	0.461	**0.339**	**0.454**
	TimeFilter	0.393	0.489	0.366	0.471	**0.361**	**0.468**
Mobility	iTransformer	0.131	0.287	0.137	0.291	**0.119**	**0.275**
	TimeMixer	0.147	0.306	0.144	0.303	**0.136**	**0.293**
	CrossLinear	0.117	0.272	0.120	0.277	**0.116**	**0.270**
	TimeFilter	0.123	0.278	0.127	0.282	**0.118**	**0.272**
M5	iTransformer	0.457	0.529	**0.440**	**0.524**	0.445	0.529
	TimeMixer	0.500	0.567	0.528	0.580	**0.455**	**0.540**
	CrossLinear	0.432	0.529	**0.422**	0.521	**0.422**	**0.520**
	TimeFilter	0.424	0.513	0.428	0.515	**0.423**	**0.511**

Table 16: Performance comparison on stable datasets. Both threshold-based filtering and TIMELESS do not significantly degrade or improve forecasting performance in general when exogenous features are stable.

Dataset	Horizon	iTransformer						TimeMixer					
		Base		w. Threshold		w. TIMELESS		Base		w. Threshold		w. TIMELESS	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Solar	96	0.287	0.296	0.287	0.296	0.285	0.293	0.310	0.338	0.310	0.338	0.253	0.293
	192	0.330	0.321	0.330	0.321	0.330	0.322	0.345	0.366	0.345	0.366	0.336	0.353
	336	0.346	0.336	0.346	0.336	0.363	0.337	0.367	0.376	0.367	0.376	0.364	0.365
	720	0.340	0.334	0.340	0.334	0.360	0.346	0.359	0.369	0.359	0.369	0.360	0.355
Traffic	96	0.150	0.228	0.147	0.230	0.157	0.237	0.203	0.301	0.204	0.302	0.201	0.300
	192	0.146	0.225	0.144	0.226	0.158	0.240	0.200	0.299	0.203	0.302	0.197	0.298
	336	0.144	0.230	0.143	0.231	0.149	0.235	0.192	0.294	0.199	0.298	0.210	0.314
	720	0.164	0.252	0.164	0.253	0.178	0.265	0.211	0.311	0.210	0.309	0.214	0.317
Dataset	**Horizon**	**CrossLinear**						**TimeFilter**					
		Base		w. Threshold		w. TIMELESS		Base		w. Threshold		w. TIMELESS	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Solar	96	0.282	0.298	0.282	0.298	0.275	0.289	0.290	0.299	0.290	0.299	0.288	0.312
	192	0.320	0.316	0.320	0.316	0.327	0.318	0.333	0.326	0.333	0.326	0.333	0.328
	336	0.337	0.327	0.337	0.327	0.341	0.332	0.352	0.336	0.352	0.336	0.362	0.346
	720	0.343	0.334	0.343	0.334	0.344	0.340	0.347	0.338	0.347	0.338	0.347	0.346
Traffic	96	0.146	0.230	0.147	0.227	0.147	0.229	0.156	0.235	0.162	0.241	0.159	0.242
	192	0.149	0.228	0.149	0.229	0.149	0.228	0.154	0.233	0.158	0.235	0.159	0.242
	336	0.145	0.229	0.149	0.231	0.145	0.229	0.154	0.239	0.161	0.243	0.157	0.244
	720	0.172	0.255	0.167	0.252	0.164	0.249	0.179	0.265	0.179	0.268	0.181	0.264

Table 17: Error comparison between forecasting with all exogenous features (Base) and with selection on Chronos-2. Our CDD-based method yields consistent gains, up to 10%.

Dataset	Horizon	Chronos-2			
		Base		w. Threshold	
		MSE	MAE	MSE	MAE
Exchange	96	0.101	0.239	0.091	0.226
	192	0.216	0.361	0.217	0.353
	336	0.418	0.493	0.379	0.470
	720	1.035	0.783	0.980	0.753
				+5.8%	+3.9%
Neuro	12	0.00020	0.0038	0.00020	0.0038
	24	0.00110	0.0086	0.00110	0.0086
	48	0.00761	0.0225	0.00758	0.0224
	96	0.0500	0.0605	0.0494	0.0603
				+1.1%	+0.3%