RI-Loss: A Learnable Residual-Informed Loss for Time Series Forecasting

Jieting Wang[1,2]*, **Xiaolei Shang**[1], **Feijiang Li**[1,2], **Furong Peng**[1,2]

[1]Institute of Big Data Science and Industry, Shanxi University
[2]Key Laboratory of Evolutionary Science Intelligence of Shanxi Province, Taiyuan, Shanxi, China
jtwang@sxu.edu.cn

Abstract

Time series forecasting relies on predicting future values from historical data, yet most state-of-the-art approaches—including transformer and multilayer perceptron-based models—optimize using Mean Squared Error (MSE), which has two fundamental weaknesses: its point-wise error computation fails to capture temporal relationships, and it does not account for inherent noise in the data. To overcome these limitations, we introduce the Residual-Informed Loss (RI-Loss), a novel objective function based on the Hilbert-Schmidt Independence Criterion (HSIC). RI-Loss explicitly models noise structure by enforcing dependence between the residual sequence and a random time series, enabling more robust, noise-aware representations. Theoretically, we derive the first non-asymptotic HSIC bound with explicit double-sample complexity terms, achieving optimal convergence rates through Bernstein-type concentration inequalities and Rademacher complexity analysis. This provides rigorous guarantees for RI-Loss optimization while precisely quantifying kernel space interactions. Empirically, experiments across eight real-world benchmarks and five leading forecasting models demonstrate improvements in predictive performance, validating the effectiveness of our approach.

Code — https://github.com/shang-xl/RI-Loss
Appendix — https://arxiv.org/abs/2511.10130

Introduction

Time series data consist of chronologically ordered observations, where forecasting involves identifying latent patterns and trends from historical data to predict future values. In practice, temporal datasets typically demonstrate considerable complexity, manifesting as nonlinear relationships, non-stationary distributions, long-term dependencies, and significant noise contamination. Deep learning architectures excel at modeling these intricate temporal structures due to their exceptional capacity for nonlinear function approximation, allowing for more precise characterization of dynamic temporal patterns (Cao et al. 2018; Liu et al. 2022; Deng and Hooi 2021). These advanced modeling capabilities have enabled transformative applications across multiple domains, including financial market prediction, industrial process monitoring, disease spread forecasting, and meteorological condition estimation.

Recent advances in time series analysis have witnessed the emergence of transformer-based and multilayer perceptron (MLP)-based architectures as dominant approaches, with numerous studies demonstrating their state-of-the-art performance (Wu et al. 2021; Zeng et al. 2022; Liu et al. 2024). However, despite their widespread adoption of MSE as the standard loss function, this conventional choice presents two critical limitations for long-term forecasting tasks.

The first limitation of MSE lies in its point-wise error computation, which makes it susceptible to overfitting observational noise, thereby impairing model generalization. More critically, MSE's formulation disregards the temporal dynamics of sequential data, neglecting essential interdependencies across time steps. These inherent weaknesses hinder the model's capacity to capture and preserve long-range dependencies effectively.

To better capture temporal dependencies, we propose a novel loss function inspired by a fundamental principle from (Yi and Wang 2019): an ideal model should produce residuals that are statistically indistinguishable from random noise, indicating complete extraction of all predictable patterns. While directly quantifying the difference between residuals and random sequences is problematic due to noise's inherent stochasticity, we reformulate this intuition through statistical independence measures. Specifically, our approach maximizes the dependence between model residuals and random noise sequences, thereby compelling the model to extract more informative temporal features from the input data.

Building on these insights, we propose a novel forecasting framework that incorporates kernel-based nonparametric independence testing via a residual-informed loss function. Our approach centers on the Hilbert-Schmidt Independence Criterion (HSIC), which provides a theoretically sound measure of statistical dependence while remaining computationally efficient. The framework achieves two key innovations: it effectively captures nonlinear temporal dependencies between model residuals and random noise through kernel methods, and it establishes a new dependency-aware modeling paradigm that unifies statistical learning principles with deep neural forecasting architectures.

Theoretically, we establish rigorous learning guarantees

by analyzing the self-bounding properties of HSIC and deriving novel generalization bounds through double-sample Rademacher complexity analysis. These theoretical contributions achieve dual significance: they formally validate the learnability of HSIC-Loss, and they provide a generalizable analytical framework for assessing the generalization capacity of dependence-aware loss functions. Our comprehensive empirical evaluation demonstrates consistent performance gains across diverse real-world datasets.

The main contributions of this paper are as follows:

- Methodological Innovation: We propose RI-Loss, a novel residual-informed loss function that simultaneously addresses two fundamental limitations of MSE: susceptibility to observational noise and failure to capture temporal dependencies, establishing a new kernel-based learning paradigm for time series forecasting.
- Theoretical Advancement: We develop the first second-order Rademacher complexity bounds for the Hilbert-Schmidt Independence Criterion (HSIC) in time series analysis, providing rigorous generalization guarantees and creating a theoretical framework for analyzing dependence-aware loss functions.
- Empirical Validation: Through extensive experiments on eight diverse real-world benchmarks, we demonstrate that RI-Loss consistently enhances performance across state-of-the-art architectures (including Transformers and MLPs).

Complete proofs and additional experimental results are provided in the Appendix.

Related Work

Time Series Forecasting

Time series forecasting has undergone significant evolution, progressing from simple moving averages to classical statistical methods like Autoregressive Integrated Moving Average (ARIMA) (Box et al. 1970), through machine learning approaches such as Support Vector Regression (SVR) (Smola and Scholkopf 2004), to contemporary deep learning architectures. The deep learning revolution has introduced specialized models for temporal data processing: Recurrent Neural Networks (RNNs) effectively capture sequential dependencies through iterative time-step processing (Cao et al. 2018); Convolutional Neural Networks (CNNs) extract localized temporal patterns via convolutional kernels (Hewage et al. 2020; Liu et al. 2022); Graph Neural Networks (GNNs) model complex multivariate interactions (Deng and Hooi 2021; Cao et al. 2020); Transformers utilize self-attention mechanisms to identify both long-range dependencies and cross-variable correlations (Zhou et al. 2021; Wu et al. 2021; Liu et al. 2024); while Multilayer Perceptrons (MLPs) provide computationally efficient solutions for resource-constrained scenarios (Zeng et al. 2022; Xu, Zeng, and Xu 2024). Despite these architectural advances, most deep time series models continue to rely on Mean Squared Error (MSE) loss for parameter optimization, presenting limitations in handling noise and temporal relationships.

Recent work has proposed MSE alternatives via shape alignment (Cuturi and Blondel 2017; Le Guen and Thome 2019) and dependency modeling (Wang et al. 2025), but these face computational complexity or residual pattern neglect. Our RI-Loss addresses both issues through statistical independence, offering noise-robust temporal modeling via kernel dependency measurement, and provable learning guarantees through HSIC-based theory.

HSIC-based Learning Method

The Hilbert-Schmidt Independence Criterion (HSIC) is a powerful kernel-based independence measure widely used in machine learning. In representation learning, it enhances latent variable interpretability through deep generative model regularization (Li et al. 2021). For causal inference, HSIC enables nonlinear conditional independence testing, overcoming linearity constraints in causal discovery (Li et al. 2024; Hu, Sejdinovic, and Evans 2024). Transfer learning benefits from its synergistic use with maximum mean discrepancy for cross-domain feature alignment (Wang et al. 2020), while time series analysis leverages HSIC for nonlinear Granger causality detection in financial modeling (Ren, Li, and Han 2020). Despite these successful applications, HSIC's potential for time series forecasting remains largely untapped, offering significant opportunities to advance temporal dependency modeling.

Preliminaries

Time Series Forecasting Task

Time series forecasting aims to predict future values based on historical observations. In our framework, given a look-back window $\boldsymbol{X}_t = (X_t, X_{t-1}, ..., X_{t-w+1})^\top \in \mathbb{R}^w$ of w historical observations, we predict the H-step future trajectory $\boldsymbol{Y} = (Y_{t+1}, \ldots, Y_{t+H})^\top \in \mathbb{R}^H$ through $\hat{\boldsymbol{Y}} = f(\boldsymbol{X}_t)$, where $f : \mathbb{R}^w \to \mathbb{R}^H$ is the forecasting function. For multivariate series with d-dimensional observations $\boldsymbol{x}_t \in \mathbb{R}^d$, this extends to predicting $\boldsymbol{Y} = \boldsymbol{y}_{t+1}, \ldots, \boldsymbol{y}_{t+H} \in \mathbb{R}^{H\times d}$ from input $\boldsymbol{X}_t = \boldsymbol{x}_t, \boldsymbol{x}_{t-1}, ..., \boldsymbol{x}_{t-w+1} \in \mathbb{R}^{w\times d}$.

Generally, the forecasting quality is evaluated via the MSE:

$$\text{MSE} = \frac{1}{H}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|_F^2 = \frac{1}{H}\sum_{k=1}^{H}\|\boldsymbol{y}_{t+k} - \hat{\boldsymbol{y}}_{t+k}\|_2^2, \quad (1)$$

where $\|\cdot\|_F^2$ represents the square of the Frobenius norm for matrices. The optimal solution of MSE corresponds to the conditional expectation of the true data, which means minimizing MSE guarantees that the mean of the predicted sequence matches that of the true sequence.

Additive Noise Model for Time Series Observations

Conventional time series prediction frameworks typically adopt an additive model structure of the form: $\boldsymbol{Y} = h(\boldsymbol{X}_t) + \boldsymbol{\epsilon}$, where $\boldsymbol{Y}$ denotes the observation sequence, $h : \mathbb{R}^w \to \mathbb{R}^H$ represents the true underlying system dynamics that maps historical patterns to future values, and $\boldsymbol{\epsilon}$ is a zero-mean additive noise vector satisfying $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$. The noise process $\boldsymbol{\epsilon} = (\epsilon_{t+1}, \ldots, \epsilon_{t+H})^\top$ further exhibits two key

properties: (i) temporal uncorrelation with $\mathbb{E}[\epsilon_{t+i}\epsilon_{t+j}] = 0$ for all $i, j = 1, \ldots, H$ where $i \neq j$, and (ii) signal independence with $\epsilon_{t+i} \perp h(\boldsymbol{X}_t)$ for all prediction horizons i.

Although we define the loss within the above basic framework, it naturally generalizes to more complex scenarios, including heteroscedastic noise structures $\epsilon_{t+i} \sim \mathcal{N}(0, \sigma^2_{t+i})$ that accommodate time-varying volatility patterns commonly observed in financial and environmental time series, as well as state-dependent noise models of the form $Y_{t+i} = h(\boldsymbol{X}_t) + g(\boldsymbol{X}_t)\epsilon_{t+i}$ where the noise amplitude $g(\boldsymbol{X}_t)$ varies with the system state.

Hilbert-Schmidt Independence Criterion

The Hilbert-Schmidt Independence Criterion (HSIC) provides a kernel-based measure of statistical dependence between random variables. Formally, for variables $R \in \mathcal{R}$ and $S \in \mathcal{S}$ with joint distribution $\mathbb{P}_{R,S}$, using reproducing kernel Hilbert spaces (RKHS) $\mathcal{F}$ on $\mathcal{R}$ and $\mathcal{G}$ on $\mathcal{S}$, the HSIC is defined as the squared Hilbert-Schmidt norm of their cross-covariance operator.

As established by (Gretton et al. 2005; Greenfeld and Shalit 2020), the HSIC admits an important probabilistic interpretation: it represents the maximum achievable covariance between normalized functions in the respective RKHS. Mathematically, for $f \in \mathcal{F}$ and $g \in \mathcal{G}$, this is expressed as:

$$\mathbb{E}(\mathrm{HSIC}(R, S)) = \sup_{\|f\|_{\mathcal{F}} \leq 1, \|g\|_{\mathcal{G}} \leq 1} (\mathbb{COV}[f(R), g(S)])^2, \tag{2}$$

where $\mathbb{COV}$ denotes covariance. The HSIC possesses two key properties: $\mathrm{HSIC}(R, S) = 0$ if and only if R and S are independent, and the magnitude of HSIC reflects the strength of dependence. This formulation makes HSIC particularly powerful for detecting nonlinear dependencies that conventional correlation measures might miss, as it operates in high-dimensional feature spaces through the kernel trick.

By expanding the Hilbert-Schmidt norm through kernel-based inner products, we derive the following equivalent formulation of HSIC:

Definition 1 (Population HSIC)**.** *For random variables $R \in \mathcal{R}$ and $S \in \mathcal{S}$ with joint distribution $\mathbb{P}_{RS}$, let $k : \mathcal{R} \times \mathcal{R} \to \mathbb{R}$ and $l : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ be characteristic kernels, and let (R', S') be an independent copy of (R, S). The Hilbert-Schmidt Independence Criterion admits the following equivalent expression:*

$$\begin{aligned}\mathrm{HSIC}(R, S) = \mathbb{E}_{RR'SS'}[k(R, R')l(S, S')] \\ + \mathbb{E}_{RR'}[k(R, R')]\mathbb{E}_{SS'}[l(S, S')] \\ - 2\mathbb{E}_{RS}\left[\mathbb{E}_{R'}[k(R, R')]\mathbb{E}_{S'}[l(S, S')]\right].\end{aligned} \tag{3}$$

The HSIC statistic has range $[0, \infty)$. This formulation reveals that HSIC measures the discrepancy between the joint embedding and the product of marginal embeddings.

The kernel-based approach provides three key advantages: it is nonparametric (making no assumptions about the dependence structure), flexible (accommodating mixed data types through kernel choice), and universal (detecting any measurable dependence when using characteristic kernels).

We now present the finite-sample estimator of HSIC, which operationalizes the population measure for practical computation.

Definition 2 (Empirical HSIC Estimator)**.** *Given an i.i.d. sample $\{(r_i, s_i)\}_{i=1}^n$, the U-statistic estimator of HSIC is given by:*

$$\begin{aligned}\mathrm{HSIC} = \frac{1}{\binom{n}{2}} \sum_{i<j} k(r_i, r_j)l(s_i, s_j) \\ + \frac{1}{\binom{n}{4}} \sum_{i<j<k<l} k(r_i, r_j)l(s_k, s_l) \\ - \frac{2}{\binom{n}{3}} \sum_{i<j<k} \left[k(r_i, r_j)l(s_j, s_k) + k(r_j, s_k)l(s_i, s_j)\right],\end{aligned} \tag{4}$$

where all sums are over distinct index tuples, the binomial coefficient $\binom{n}{m}$ represents the number of ways to choose m elements from a set of n distinct items, and k, l are characteristic kernels.

Residual-Informed Loss

Decomposition of MSE Loss for Additive Model

Under the additive noise model described earlier, where the observed outputs $\boldsymbol{Y}$ are corrupted by additive noise $\boldsymbol{\epsilon}$ with mean zero such that $\boldsymbol{Y} = h(\boldsymbol{X}_t) + \boldsymbol{\epsilon}$, we can analyze the relationship between the true MSE and the observed MSE. The true MSE measures the prediction error with respect to the noise-free target $h(\boldsymbol{X}_t)$: $\mathrm{MSE}_{\mathrm{true}} = \frac{1}{H}\|h(\boldsymbol{X}_t) - \hat{\boldsymbol{Y}}\|_F^2$. The observed MSE quantifies the discrepancy between predictions and noisy measurements: $\mathrm{MSE}_{\mathrm{obs}} = \frac{1}{H}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|_F^2$.

Using the observation model $h(\boldsymbol{X}_t) = \boldsymbol{Y} - \boldsymbol{\epsilon}$, we can express the true MSE as:

$$\begin{aligned}\mathrm{MSE}_{\mathrm{true}} &= \frac{1}{H}\|(\boldsymbol{Y} - \boldsymbol{\epsilon}) - \hat{\boldsymbol{Y}}\|_F^2 \\ &= \frac{1}{H}\left(\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|_F^2 - 2\langle \boldsymbol{Y} - \hat{\boldsymbol{Y}}, \boldsymbol{\epsilon}\rangle + \|\boldsymbol{\epsilon}\|_F^2\right) \\ &= \mathrm{MSE}_{\mathrm{obs}} - \frac{2}{H}\langle \boldsymbol{Y} - \hat{\boldsymbol{Y}}, \boldsymbol{\epsilon}\rangle + \frac{1}{H}\|\boldsymbol{\epsilon}\|_F^2.\end{aligned} \tag{5}$$

Assuming the noise $\boldsymbol{\epsilon}$ is zero-mean and uncorrelated with the prediction error, taking expectations on both sides yields:

$$\mathbb{E}[\mathrm{MSE}_{\mathrm{true}}] = \mathbb{E}[\mathrm{MSE}_{\mathrm{obs}}] - \frac{1}{H}\mathbb{E}[\|\boldsymbol{\epsilon}\|_F^2], \tag{6}$$

where $\mathbb{E}[\|\boldsymbol{\epsilon}\|_F^2]$ denotes the noise variance like term. This shows that the observed MSE systematically overestimates the true prediction error by exactly the noise power. However, in general cases, since the prediction $\hat{\boldsymbol{Y}}$ is learned from noisy observations $\boldsymbol{Y}$, the prediction error $\hat{\boldsymbol{Y}} - \boldsymbol{Y}$ depends on the noisy $\boldsymbol{\epsilon}$, making the cross-term in Eq. (5) non-zero:

Theorem 1 (Cross-Term Expectation for Linear Projection)**.** *Let $\boldsymbol{Y} = (Y_{t+1}, \ldots, Y_{t+H})^\top \in \mathbb{R}^H$ be a noisy observation vector generated by: $\boldsymbol{Y} = h(\boldsymbol{X}_t) + \boldsymbol{\epsilon}$, where $h(\boldsymbol{X}_t) : \mathcal{X} \to \mathbb{R}^H$ is a deterministic mapping, and the noise vector $\boldsymbol{\epsilon} \in \mathbb{R}^H$ satisfies: $\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}, \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \sigma^2\boldsymbol{I}_H, \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_H)$. Here, $\boldsymbol{I}_H$ denotes the $H \times H$ identity matrix and $\sigma^2 > 0$*

is the noise variance. For any linear estimator $\hat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$ *with projection matrix* $\boldsymbol{P} \in \mathbb{R}^{H\times H}$*, the expected normalized cross-term evaluates to:*

$$\mathbb{E}\left[\frac{1}{H}\langle \boldsymbol{Y} - \hat{\boldsymbol{Y}}, \boldsymbol{\epsilon}\rangle\right] = \frac{\sigma^2}{H}\mathrm{tr}(\boldsymbol{I}_H - \boldsymbol{P}), \tag{7}$$

where $\mathrm{tr}(\cdot)$ *represents the trace of a matrix, defined as the sum of its diagonal elements.*

The decomposition of MSE in Eq.(5) and Theorem 1 motivate our design of a loss function that explicitly accounts for the dependency between prediction residuals and inherent noise. To explicitly penalize both prediction accuracy and residual-noise dependency, we first propose a naive loss combining the empirical MSE with a linear residual-noise alignment term: $\mathcal{L}_{\text{naive}} = \text{MSE}_{\text{obs}} - \lambda\langle \boldsymbol{Y} - \hat{\boldsymbol{Y}}, \boldsymbol{\epsilon}\rangle$, where λ is the trade-off parameter. However, this linear term only captures first-order dependencies.

Definition of Residual-Informed Loss

To generalize $\langle \boldsymbol{Y} - \hat{\boldsymbol{Y}}, \boldsymbol{\epsilon}\rangle$ to nonlinear relationships, we replace the inner product with the HSIC. We analyze their relationship. If R and S are centered random variables, that is, $\mathbb{E}[R] = 0$ and $\mathbb{E}[S] = 0$, then their covariance simplifies to $\mathbb{COV}(R, S) = \mathbb{E}[RS]$. At this time, we can observe that $\mathbb{E}[RS] \leq \mathbb{E}(\text{HSIC}(R, S))$. This inequality holds because if $\mathcal{F}$ and $\mathcal{G}$ contain the identity functions (i.e., $f(R) = R \in \mathcal{F}$, $g(S) = S \in \mathcal{G}$), then $\mathbb{E}(\text{HSIC}(R, S)) = \sup_{f\in\mathcal{F}, g\in\mathcal{G}} \mathbb{COV}[f(R), g(S)] \geq \mathbb{COV}[R, S]$. Thus, assuming the residual expectation is zero ($\mathbb{E}[\boldsymbol{Y} - \hat{\boldsymbol{Y}}] = \boldsymbol{0}$) and noise expectation is zero ($\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}$), we have $\mathbb{E}(\langle \boldsymbol{Y} - \hat{\boldsymbol{Y}}, \boldsymbol{\epsilon}\rangle) \leq \mathbb{E}(\text{HSIC}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}, \boldsymbol{\epsilon}))$. This yields the Residual-Informed (RI) loss:

Definition 3. *For the observed output series* $\boldsymbol{Y}$*, the predicted series* $\hat{\boldsymbol{Y}}$ *and a noise series* $\boldsymbol{\epsilon}$*, RI loss is defined as:*

$$\text{RI}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \text{MSE}_{\text{obs}} + \lambda \exp(-\tau \cdot \text{HSIC}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}, \boldsymbol{\epsilon})), \tag{8}$$

where the exponential transformation is applied to ensure that the loss function remains positive-valued and λ*,* τ *are the trade-off parameter and transformation parameter, respectively.*

RI-Loss in Eq. (8) requires the model to preserve interpretable signals by minimizing the MSE to ensure prediction accuracy, while explicitly separating the noise structure by maximizing the dependence between the residuals and the noise, thereby forcing the model to push unexplainable variation (noise structure) into the residuals. Mathematically, when the residual $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$ is highly correlated with the noise $\boldsymbol{\epsilon}$, it indicates that the model has excluded as much of the noise-distributed components as possible from the prediction $\hat{\boldsymbol{Y}}$, ensuring that the prediction result $\hat{\boldsymbol{Y}}$ contains only the true signal that is independent of the noise.

RI Advantage

We investigate how RI loss varies with increasing noise ratios $\rho \in [0, 1]$, comparing its behavior to conventional MSE.

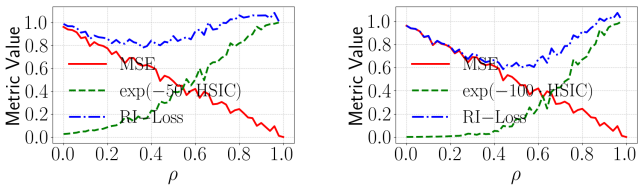


Figure 1: The RI-Loss varies with the noise ratio.

The results is demonstrated in Figure 1. Our experiment generates a sinusoidal signal with noise $\boldsymbol{y}_{\text{true}} = \sin(\boldsymbol{x}) + \boldsymbol{\epsilon}$ where $\boldsymbol{x} \in [0, 2\pi]$ contains 500 points and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$ represents baseline Gaussian noise. Crucially, for each ρ value (51 linearly spaced steps), we corrupt the underlying noise-free sinusoidal component by adding $\mathcal{N}(0, 1)$ noise to randomly selected $\rho \times 100\%$ of the $\sin(\boldsymbol{x})$ values, producing $\boldsymbol{y}_{\text{noisy}}$. The metrics include $\text{MSE} = \mathbb{E}[(\boldsymbol{y}_{\text{true}} - \boldsymbol{y}_{\text{noisy}})^2]$ and $\text{RI} = \text{MSE} + \exp(-\tau \cdot \text{HSIC}(\boldsymbol{y}_{\triangle}, \boldsymbol{\epsilon}))$ ($\tau = 50, 100$, $\lambda = 1$), where $\boldsymbol{y}_{\triangle} = \boldsymbol{y}_{\text{true}} - \boldsymbol{y}_{\text{noisy}}$. For HSIC computation, we employ Gaussian kernel matrices $k_{ij} = \exp(-|\boldsymbol{y}_{\triangle,i} - \boldsymbol{y}_{\triangle,j}|^2)$ for residual and $l_{ij} = \exp(-|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j|^2)$ for noise.

As shown in Figure 1, the RI-Loss demonstrates a convex dependence on the noise ratio ρ, reflecting the fundamental trade-off between signal reconstruction and noise suppression. For about $\rho < 0.5$ in the left panel ($\rho < 0.6$ in the right panel), the system prioritizes signal fitting, where controlled noise retention improves reconstruction fidelity (manifested in decreasing MSE). When about $\rho > 0.5$ in the left panel ($\rho > 0.6$ in the right panel), the exponential HSIC penalty ($\exp(-\tau \cdot \text{HSIC}) \to 1$) becomes dominant, enforcing noise rejection at the expense of elevated RI-Loss. The characteristic minimum at $\rho \approx 0.5$ in the left panel ($\rho \approx 0.6$ in the right panel) naturally emerges as the optimal operating point that balances these competing requirements.

Theoretical Analysis of RI

The major purpose of this paper is presenting a novel approach that enhances standard MSE optimization by incorporating a HSIC regularization term between model residuals and random noise. Hence, we establish a theoretical framework to investigate HSIC's generalization properties.

The current theoretical analysis operates under the baseline assumption of independent and identically distributed (i.i.d.) residuals and noise terms. This initial configuration provides crucial insights into HSIC's generalization properties. In future work, we will investigate the generalization properties of both HSIC and MSE under more complex dependency structures.

The generalization gap measures how well empirical training error approximates the true expected error (Bartlett and Mendelson 2003). A smaller gap indicates better generalization performance from training data to the underlying distribution. Theoretical bounds on this gap ensure reliable model behavior on unseen data. Our theorem shows that the empirical HSIC converges to the population HSIC as $n \to \infty$, providing generalization guarantees.

Theorem 2. *Let* $\sigma_1, ..., \sigma_n$ *are independent Rademacher random*

variables ($\mathbb{P}(\sigma_i = \pm 1) = \frac{1}{2}$), where n is the number of samples. For random variables $R \in \mathcal{R}$ and $S \in \mathcal{S}$ with joint distribution $\mathbb{P}_{RS}$, let $k : \mathcal{R} \times \mathcal{R} \to \mathbb{R}$ and $l : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ be characteristic kernels. Let (R', S') be an independent and identically distributed (i.i.d.) copy of (R, S). Suppose that $c_1(k) \leq k \leq c_2(k)$ and $c_1(l) \leq l \leq c_2(l)$, for $\delta > 0$, then the following holds:

$$\left|\mathrm{HSIC}(\{(r_i, s_i)\}_{i=1}^n) - \mathbb{E}[\mathrm{HSIC}(R, S)]\right| \leq 3 \sum_{j=k,l} c_2(j) \sum_{m=1}^{3} \gamma_m(j). \tag{9}$$

Specifically, the above key functions for $f \in \{k, l\}$ are defined as:

$$\begin{aligned}
&\gamma_1(n, \delta; f) \\
&= \frac{10(c_2(f) - c_1(f))}{n} \ln \frac{2}{\delta} \\
&\gamma_2(n, \delta, R_\sigma; f) \\
&= 2(c_2(f) - c_1(f)) \sqrt{2 \ln \frac{2}{\delta} \left(\frac{\mathcal{R}_\sigma}{2n} + \frac{1}{n^2} \ln \frac{2}{\delta} \right)} \\
&\gamma_3(n, \delta, W_{\sigma,\sigma}, W_{\sigma,\alpha}, W_\sigma, F; f) \\
&= 4 \Bigg(\ln \frac{2}{\delta} \Bigg(\frac{C_0}{n-1} (\mathbb{E}W_{\sigma,\sigma} + \sqrt{2}\mathbb{E}W_{\sigma,\alpha} \\
&+ \frac{2(\mathbb{E}W_\sigma + F)}{n} + \frac{\sqrt{8}Fn^{1/2}}{n^2} + \frac{4F}{n^2}) + \frac{1}{n^2} \ln \frac{2}{\delta} \Bigg) \Bigg)^{\frac{1}{2}}.
\end{aligned} \tag{10}$$

In the above functions, the first-order and second-order sample Rademacher complexities are:

$$\mathcal{R}_\sigma = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f_1(X_i) \right], \tag{11}$$

$$W_{\sigma,\sigma} = \sup_{f \in \mathcal{F}} \frac{1}{n^2} \left| \sum_{i,j} \sigma_i \sigma_j f_2(X_i, X_j) \right|, \tag{12}$$

$$W_{\sigma,\alpha} = \sup_{f \in \mathcal{F}} \sup_{\alpha : \|\alpha\|_2 \leq 1} \frac{1}{n^2} \sum_{i,j} \sigma_i \alpha_j f_2(X_i, X_j), \tag{13}$$

$$W_\sigma = \sup_{f \in \mathcal{F}, k=1,\ldots,n} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i f_2(X_i, X_k) \right|, \tag{14}$$

$$F = sup_{f \in \mathcal{F}} \|f_2\|_\infty, \tag{15}$$

where f_1 and f_2 respectively represent the centered kernels:

$$\begin{aligned}
f_1(X_i) &= \mathbb{E}[f(X_i, X)|X_i] - \mathbb{E}[f(X_i, X_j)], \\
f_2(X_i, X_j) &= f(X_i, X_j) - f_1(X_i) - f_1(X_j) \\
&\quad - \mathbb{E}[f(X_i, X_j)].
\end{aligned} \tag{16}$$

Theorem 2 makes a significant breakthrough in the non-asymptotic analysis of HSIC statistics, establishing for the first time a convergence upper bound that explicitly incorporates double-sample Rademacher complexities. Through an innovative decomposition of the error into three terms with clear statistical significance: the γ_1 term controls the deviation with typical $\mathcal{O}(1/n)$ convergence rate ($\mathcal{O}(\cdot)$ characterizes the asymptotic order); the γ_2 term's square root structure reflects classical sub-exponential convergence properties; and the most innovative γ_3 term employs sophisticated probability inequality techniques to constrain the influence of double-sample complexities (including $W_{\sigma,\sigma}$ and $W_{\sigma,\alpha}$) at the $\mathcal{O}(1/\sqrt{n})$ level.

This theoretical achievement not only maintains optimal convergence rates but, more importantly, rigorously quantifies higher-order interaction effects in kernel function spaces in HSIC analysis, providing a new theoretical framework for high-dimensional nonparametric independence tests and offering important guidance for kernel method selection and sample complexity estimation. In the RI loss framework, r denotes the residual vector $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$, s denotes the noise vector $\boldsymbol{\epsilon}$, and n is the sequence length. Theorem 2 establishes that the empirical $\mathrm{HSIC}(r, s)$ estimator exhibits finite-sample convergence with error decaying as $\mathcal{O}(1/\sqrt{n})$, dependent on kernel complexities $\mathcal{R}_\sigma$,$W_{\sigma,\sigma}$,$W_{\sigma,\alpha}$,W_σ, and achieves asymptotic consistency: $\lim_{n\to\infty} \mathrm{HSIC}_n(r, s) \overset{a.s.}{=} \mathbb{E}_{R,S}[\mathrm{HSIC}(R, S)]$. This result provides fundamental guarantees for RI loss optimization, ensuring both theoretical soundness and practical reliability in large-sample regimes.

RI-based Time Series Forecasting

Our proposed RI-Loss demonstrates broad compatibility with existing deep time series architectures, as illustrated in Figure 2. The modular design enables seamless integration with both Transformer-based (e.g., Autoformer, Informer) and MLP-based (e.g., DLinear) forecasting frameworks. We employ a uniform distribution over the interval $[-1, 1]$ as the noise distribution. Complete implementation details are provided in Algorithm 1 of Appendix B.

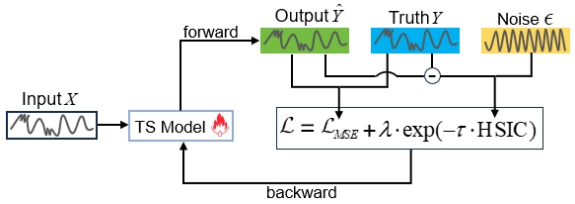


Figure 2: RI-Loss Based Time Series Model.

Experimental Analysis

We conduct a comprehensive evaluation of the proposed RI-Loss. Additional experimental results, including the effects of the lookback window size, visualization of correlation structures, and hyperparameter sensitivity analysis, are provided in the appendix.

Experiments Settings

Datasets We conduct extensive experiments on eight widely used real-world datasets, including ETT (ETTh1, ETTh2, ETTm1, ETTm2), Exchange, ILI, Weather and Electricity. Please refer to Appendix C for data descriptions.

Backbones We use five long-term time series forecasting models as our backbone models to comprehensively evaluate the effectiveness of RI-Loss. These include three Transformer-based models: Informer (Zhou et al. 2021), Autoformer (Wu et al. 2021), and iTransformer (Liu et al. 2024) and two MLP-based models: Decomposition-Linear (Dlinear) (Zeng et al. 2022) and Retrieval Augmented Time Series Forecasting (RAFT) (Han et al. 2025).

Models		Informer(2021)		Autoformer(2021)		DLinear(2022)		iTransformer(2024)		RAFT(2025)	
Metric		MSE	Ours	MSE	Ours	MSE	Ours	MSE	Ours	MSE	Ours
ETTh1	96	0.878	**0.875**	0.456	**0.440**	0.384	**0.346**	0.387	**0.382**	0.387	**0.378**
	192	1.013	**0.970**	0.492	**0.474**	0.443	**0.404**	0.441	**0.437**	0.423	**0.422**
	336	**1.172**	1.225	**0.506**	0.522	0.447	**0.443**	0.491	**0.490**	**0.458**	0.460
	720	1.175	**1.169**	**0.496**	0.544	0.472	**0.471**	0.508	**0.490**	0.463	**0.462**
ETTh2	96	3.340	**2.299**	0.375	**0.336**	0.297	**0.279**	0.301	**0.296**	0.296	**0.293**
	192	5.684	**3.949**	0.451	**0.417**	0.373	**0.355**	0.380	**0.374**	0.384	**0.377**
	336	4.418	**3.412**	0.477	**0.444**	0.450	**0.412**	0.424	**0.419**	0.426	**0.415**
	720	3.249	**3.094**	0.476	**0.445**	0.696	**0.582**	0.430	**0.422**	0.434	**0.417**
ETTm1	96	0.678	**0.536**	0.478	**0.466**	0.302	**0.294**	0.342	**0.322**	0.329	**0.316**
	192	0.784	**0.610**	0.549	**0.534**	0.337	**0.333**	0.383	**0.376**	0.363	**0.357**
	336	1.011	**0.807**	0.516	**0.480**	0.371	**0.369**	0.418	**0.412**	0.391	**0.385**
	720	1.037	**0.978**	0.526	**0.523**	0.426	**0.423**	0.487	**0.478**	**0.444**	0.447
ETTm2	96	0.464	**0.343**	0.242	**0.216**	0.170	**0.164**	0.186	**0.175**	0.177	**0.174**
	192	0.811	**0.504**	0.300	**0.269**	0.226	**0.220**	0.254	**0.243**	0.243	**0.238**
	336	1.286	**1.124**	0.336	**0.325**	0.292	**0.276**	0.316	**0.306**	0.302	**0.297**
	720	3.949	**3.269**	0.432	**0.415**	0.406	**0.375**	0.414	**0.407**	0.402	**0.396**
Exchange	96	1.022	**0.901**	0.154	**0.138**	0.086	**0.084**	0.086	**0.085**	0.089	**0.084**
	192	1.204	**1.176**	0.297	**0.297**	0.199	**0.163**	0.181	**0.180**	0.192	**0.177**
	336	**1.577**	1.674	**0.457**	0.511	0.345	**0.265**	0.338	**0.337**	0.370	**0.343**
	720	**2.206**	2.574	**1.079**	1.188	**0.872**	1.128	**0.853**	0.863	1.133	**1.058**
ILI	24	6.158	**5.972**	3.427	**3.322**	2.205	**2.199**	**2.695**	2.739	2.040	**2.015**
	36	5.930	**5.739**	3.596	**3.425**	2.392	**2.106**	2.563	**2.531**	2.085	**1.962**
	48	**5.185**	5.237	3.493	**3.178**	2.298	**1.984**	2.567	**2.455**	2.069	**1.884**
	60	5.296	**5.091**	**2.847**	3.019	2.410	**1.970**	2.625	**2.480**	2.065	**1.859**
Weather	96	0.496	**0.479**	**0.224**	**0.224**	0.144	**0.143**	0.174	**0.169**	**0.189**	**0.189**
	192	0.582	**0.575**	0.305	**0.283**	0.188	**0.184**	0.224	**0.219**	0.239	**0.234**
	336	0.643	**0.609**	0.353	**0.347**	0.240	**0.234**	0.283	**0.278**	0.291	**0.283**
	720	0.651	**0.637**	0.456	**0.419**	0.317	**0.309**	0.359	**0.357**	0.366	**0.358**
Electricity	96	0.289	**0.281**	0.202	**0.197**	**0.140**	0.141	**0.148**	0.149	**0.174**	0.176
	192	0.296	**0.295**	0.218	**0.210**	**0.154**	0.155	0.166	**0.165**	0.172	**0.170**
	336	**0.297**	0.302	**0.231**	**0.231**	**0.169**	0.170	**0.178**	0.180	0.184	**0.181**
	720	**0.341**	0.367	0.257	**0.244**	**0.204**	**0.204**	**0.209**	**0.209**	**0.218**	0.220
1^{st} Count		6	26	7	27	5	28	7	26	5	28

Table 1: The table reports multivariate forecasting results on eight datasets and compares them with several baseline models. All experiments use the same prediction length settings as the corresponding backbone models. Ours indicates models trained with RI-Loss under identical architectures. The best results are shown in bold. The last row indicates the number of top-1 times of each model.

ID	-3dB		3dB		10dB	
	MSE	Ours	MSE	Ours	MSE	Ours
1	0.485	**0.473**	0.481	**0.457**	0.481	**0.456**
2	0.693	**0.579**	0.649	**0.558**	0.613	**0.551**
3	0.437	**0.433**	0.428	**0.423**	0.425	**0.422**
4	0.471	**0.391**	0.459	**0.380**	0.455	**0.375**
6	0.340	**0.329**	0.326	**0.319**	0.325	**0.320**

Table 2: MSE comparison of DLinear at 720 steps under different noise levels. The dataset IDs follow the same order as in Table 1.

Evaluation Metrics Two widely used evaluation metrics in long-term time series forecasting, MSE and Mean Absolute Error (MAE), are employed to validate the proposed loss function. Lower values indicate better performance.

Method	Dlinear	iTransformer	RAFT
MSE	16.3	32.5	1386
RI-Loss	30.5	40.2	1393

Table 3: Training time comparison (ms/iter) between MSE and RI-Loss at forecasting horizon 720 on the Weather dataset.

Implementation Details We conduct our experiments using the official implementations of each backbone model from their respective repositories. To ensure a fair comparison, we adopt the original experimental setups and hyperparameter configurations when incorporating RI-Loss to enhance the performance of the backbone models. The hyperparameter λ in RI-Loss is set to 10, and τ is set to 1. The kernel function is implemented as a Gaussian kernel with a bandwidth of 1. To mitigate the impact of randomness, all

Models	Loss	ETTh2			
		96	192	336	720
Informer	RI-Loss	**2.299**	**3.949**	**3.412**	3.094
	MAE	2.444	3.995	3.713	**3.083**
	MSE	3.340	5.684	4.418	3.249
	Pearson+MSE	2.569	4.193	4.406	3.123
Autoformer	RI-Loss	**0.336**	**0.417**	**0.444**	0.445
	MAE	0.337	0.420	0.449	**0.443**
	MSE	0.375	0.451	0.477	0.476
	Pearson+MSE	0.364	0.421	0.449	0.448
DLinear	RI-Loss	**0.279**	**0.355**	**0.412**	**0.582**
	MAE	**0.279**	0.358	0.420	0.590
	MSE	0.297	0.373	0.450	0.696
	Pearson+MSE	0.305	0.404	0.494	0.714
iTrans.	RI-Loss	**0.296**	**0.374**	**0.419**	**0.422**
	MAE	0.298	0.377	0.421	0.427
	MSE	0.301	0.380	0.424	0.430
	Pearson+MSE	0.299	0.381	0.424	0.435
RAFT	RI-Loss	**0.293**	**0.377**	**0.415**	**0.417**
	MAE	0.294	0.379	0.420	0.422
	MSE	0.296	0.384	0.426	0.434
	Pearson+MSE	0.300	0.386	0.426	0.431

Table 4: Comparison of different loss functions across five backbone models on the ETTh2 dataset. "iTrans." denotes the iTransformer model.

reported results are averaged over five independent runs.

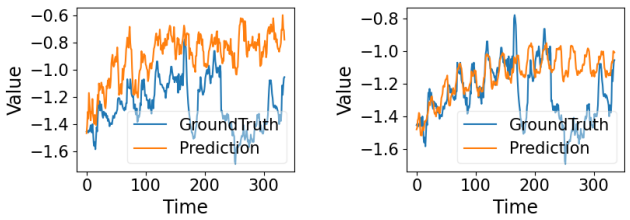
Main Results

Table 1 presents a comparison of model performance using RI-Loss versus the traditional MSE Loss function. Overall, RI-Loss yields better results. We analyzed the prediction error reduction across all datasets and found that RI-Loss leads to significant improvements in both the Informer and DLinear models. Specifically, it achieves an average reduction of 9.4% in MSE for Informer, and 5.2% in MSE for DLinear. Other models also show similar gains. Across 160 test cases, RI-Loss outperforms MSE loss in 130 instances, further demonstrating its robustness and broad applicability. Results under the MAE metric are provided in Appendix F.1.

The Friedman test is conducted to evaluate the statistical significance of performance differences between the two loss functions, showing that all five models achieve varying degrees of ranking improvement after incorporating RI-Loss. Further details are given in Appendix E.

Noise Robustness

To evaluate the robustness of our method under noisy conditions, we add zero-mean Gaussian noise to the input data at varying Signal-to-Noise Ratios (SNRs) and conduct experiments using DLinear as the backbone model. The results, summarized in Table 2, show that our loss consistently outperforms the conventional MSE across all SNR levels. While MSE suffers significant performance degradation as noise increases, our method effectively mitigates the impact of noise, demonstrating better generalization and practical applicability for noisy time series forecasting.



(a) MSE (b) RI-Loss

Figure 3: Forecasting results on ETTh1 with input-336-predict-336, comparing RI-Loss and MSE objectives. Blue denotes ground truth, and orange denotes predictions.

Running Cost

To evaluate the computational overhead of RI-Loss, we adopt DLinear, iTransformer, and RAFT as backbone models and measure per-iteration training time (ms/iter) on the Weather (forecasting horizon 720). As shown in Table 3, RI-Loss adds a small computational overhead because HSIC is costlier than MSE, but it consistently boosts backbone accuracy and remains effective for long-term forecasting on large datasets. Additional results are provided in Appendix F.2.

Ablation Study

To analyze the impact of different components in our loss, we conducted four ablation experiments on ETTh2 dataset: RI-Loss, MAE, MSE, and RI-Loss with HSIC replaced by the Pearson Correlation (PC). As shown in Table 4, replacing HSIC with PC reduces performance, since PC captures only linear relationships and cannot model nonlinear dependencies in complex time series. RI-Loss also outperforms MSE and MAE, confirming the effectiveness of our design. Further details are in Appendix F.3.

Visualization of Prediction Results

We use DLinear as the backbone model to visualize the prediction results. As shown in Figure 3, compared with the MSE loss, RI-Loss significantly improves the alignment with the ground truth. Predictions with RI-Loss align more closely with the ground truth's statistical properties, yielding results with higher structural similarity. More visualizations are provided in Appendix F.4.

Conclusion

Time series forecasting is crucial across many domains; however, most existing models rely on MSE, which struggles with temporal dependencies and inherent noise. To address this, we propose RI-Loss, an HSIC-based loss function that enhances temporal pattern extraction while mitigating noise interference. We further establish a learning bound for HSIC to support its learnability. Experiments on diverse benchmarks demonstrate the superior performance of RI-Loss for time series forecasting. Future work will explore more complex noise correlations and residual dependencies.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. T2495251, 62306170, U24A20253, 62136005, 62476160, 62441239,62276162), the Science and Technology Major Project of Shanxi (No. 202201020101006), the Special Fund for Science and Technology Innovation Teams of Shanxi Province (No. 202304051001001), and the Key Research and Development Project of Shanxi Province (No. 202402020101004).

References

Bartlett, P. L.; and Mendelson, S. 2003. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3: 463–482.

Box, G. E. P.; Jenkins, G. M.; Reinsel, G. C.; and Ljung, G. M. 1970. *Time Series Analysis: Forecasting and Control*.

Cao, D.; Wang, Y.; Duan, J.; Zhang, C.; Zhu, X.; Huang, C.; Tong, Y.; Xu, B.; Bai, J.; Tong, J.; and Zhang, Q. 2020. Spectral temporal graph neural network for multivariate time-series forecasting. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 17766–17778.

Cao, W.; Wang, D.; Li, J.; Zhou, H.; Li, L.; and Li, Y. 2018. BRITS: Bidirectional Recurrent Imputation for Time Series. In *Neural Information Processing Systems*, 6776–6786.

Cuturi, M.; and Blondel, M. 2017. Soft-DTW: a differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 894–903.

Deng, A.; and Hooi, B. 2021. Graph Neural Network-Based Anomaly Detection in Multivariate Time Series. In *AAAI Conference on Artificial Intelligence*.

Greenfeld, D.; and Shalit, U. 2020. Robust Learning with the Hilbert-Schmidt Independence Criterion. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 3759–3768. PMLR.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In Jain, S.; Simon, H. U.; and Tomita, E., eds., *Algorithmic Learning Theory*, 63–77. Berlin, Heidelberg: Springer Berlin Heidelberg.

Han, S.; Lee, S.; Cha, M.; Arik, S. O.; and Yoon, J. 2025. Retrieval Augmented Time Series Forecasting. In *Forty-second International Conference on Machine Learning*.

Hewage, P.; Behera, A.; Trovati, M.; Pereira, E.; Ghahremani, M.; Palmieri, F.; and Liu, Y. 2020. Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. *Soft Comput.*, 24(21): 16453–16482.

Hu, R.; Sejdinovic, D.; and Evans, R. J. 2024. A Kernel Test for Causal Association via Noise Contrastive Backdoor Adjustment. *Journal of Machine Learning Research*, 1–56.

Le Guen, V.; and Thome, N. 2019. Shape and Time Distortion Loss for Training Deep Time Series Forecasting Models. In *Advances in Neural Information Processing Systems*, 4191–4203.

Li, J.; Qian, Y.; Wang, J.; and Liu, S. 2024. PHSIC against Random Consistency and Its Application in Causal Inference. In *In Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, 2108–2116.

Li, Y.; Pogodin, R.; Sutherland, D. J.; and Gretton, A. 2021. Self-Supervised Learning with Kernel Dependence Maximization. In *Neural Information Processing Systems*, volume 34, 15543–15556.

Liu, M.; Zeng, A.; Chen, M.-H.; Xu, Z.; Lai, Q.; Ma, L.; and Xu, Q. 2022. SCINet: Time Series Modeling and Forecasting with Sample Convolution and Interaction. In *Neural Information Processing Systems*, volume 35, 5816–5828.

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

Ren, W.; Li, B.; and Han, M. 2020. A novel Granger causality method based on HSIC-Lasso for revealing nonlinear relationship between multivariate time series. *Physica A: Statistical Mechanics and its Applications*.

Smola, A.; and Scholkopf, B. 2004. A tutorial on support vector regression. *Statistics and Computing*, 14: 199–222.

Wang, H.; Pan, L.; Shen, Y.; Chen, Z.; Yang, D.; Yang, Y.; Zhang, S.; Liu, X.; Li, H.; and Tao, D. 2025. FreDF: Learning to Forecast in the Frequency Domain. In *The Thirteenth International Conference on Learning Representations*.

Wang, S.; Zhang, L.; Zuo, W.; and Zhang, B. 2020. Class-Specific Reconstruction Transfer Learning for Visual Recognition Across Domains. *IEEE Transactions on Image Processing*, 29: 2424–2438.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 22419–22430.

Xu, Z.; Zeng, A.; and Xu, Q. 2024. FITS: Modeling Time Series with 10k Parameters. In *The Twelfth International Conference on Learning Representations*.

Yi, D.; and Wang, Y. 2019. *Applied Time Series Analysis*. Beijing: Renmin University of China Press, 5nd edition.

Zeng, A.; Chen, M.-H.; Zhang, L.; and Xu, Q. 2022. Are Transformers Effective for Time Series Forecasting? In *AAAI Conference on Artificial Intelligence*, 11121–11128.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *AAAI Conference on Artificial Intelligence*, 11106–11115.