KOONPRO: A VARIANCE-AWARE KOOPMAN PROBABILISTIC MODEL ENHANCED BY NEURAL PROCESSES FOR TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

The probabilistic forecasting of time series is a well-recognized challenge, particularly in disentangling correlations among interacting time series and addressing the complexities of distribution modeling. By treating time series as temporal dynamics, we introduce **KooNPro**, a novel probabilistic time series forecasting model that combines variance-aware deep **Koo**pman model with **N**eural **Pro**cesses. KooNPro introduces a variance-aware continuous spectrum using Gaussian distributions to capture complex temporal dynamics with improved stability. It further integrates the Neural Processes to capture fine dynamics, enabling enhanced dynamics capture and prediction. Extensive experiments on nine real-world datasets demonstrate that KooNPro consistently outperforms state-of-the-art baselines. Ablation studies highlight the importance of the Neural Process component and explore the impact of key hyperparameters. Overall, KooNPro presents a promising novel approach for probabilistic time series forecasting.

1 INTRODUCTION

Time series forecasting is a critical task with applications in supply chain management, finance, energy, and healthcare. Probabilistic forecasting is particularly valuable as it quantifies uncertainty, enabling robust decision-making for downstream tasks like inventory optimization and risk management. For example, in supply chain management, it helps businesses prepare for demand fluctuations by planning across multiple scenarios, reducing costs, and improving efficiency.

While traditional methods like Auto-Regressive Integrated Moving Average (ARIMA) (Said & Dickey, 1984) and the family of Kalman filters (Auger et al., 2013) struggle with the complexities of real-world time series, deep learning approaches such as Recurrent Neural Networks (Salehinejad et al., 2017), and subsequently to Long Short-Term Memory networks (**?**). The application of the attention mechanism (Vaswani et al., 2017) has led to the emergence of Transformer-based approaches like Zhou et al. (2021), Wu et al. (2021), and Zhou et al. (2022). Additionally, diffusion probabilistic models (Ho et al., 2020; Rasul et al., 2021; Li et al., 2022; Fan et al., 2024) have emerged as a promising paradigm for probabilistic forecasting, with recent works like Kollovieh et al. (2024) demonstrating their potential. Moreover, foundation models for time series forecasting, such as Ansari et al. (2024), have introduced the concept of learning a universal representation for diverse time series tasks. Despite these advancements, many deep learning models focus primarily on point estimation, which limits their ability to fully capture the uncertainty and dynamics inherent in complex systems.

State space models, such as Rangapuram et al. (2018a) Deep Factors Wang et al. (2019), and Paria et al. (2021) offer an alternative by modeling time series as low-dimensional latent decompositions. These methods excel at representing temporal dynamics in reduced spaces but often rely on fixed assumptions about state transitions, which may fail to generalize to highly nonlinear and non-stationary scenarios. Dynamic mode decomposition (DMD) is a powerful tool for modeling time series as dynamic systems (Kuttichira et al., 2017), aiming to identify the underlying dynamics and subsequently utilize them for prediction (Kou & Zhang, 2019; Yuan et al., 2021). However, DMD's reliance on discrete eigenvalues (dot spectrum) restricts its ability to describe the nonlinear patterns of complex systems. Koopman theory offers a powerful framework for analyzing nonlinear

systems by representing them in a linear but infinite-dimensional space. The ability to decompose complex dynamics into Koopman eigenfunctions has inspired many works to incorporate this theory into time series modeling. Combining the Koopman theory (Koopman, 1931) with deep learning (Lusch et al., 2018) enhances the ability to model temporal dynamics with greater sophistication, leading to more accurate predictions. For example, models such as Liu et al. and Wang et al. (2023) enhance predictive capabilities through deep Koopman theory by learning the continuous spectrum of dynamics. However, these models are often affected by spectrum pollution, leading to unstable convergence and reduced accuracy, especially in highly nonlinear systems (Colbrook & Townsend, 2024).

To address this, the concept of pseudospectra has been introduced. Pseudospectra accounts for the potential regions where eigenvalues may appear under different perturbation conditions, which makes them more robust to noise and better suited for capturing continuous variations. Inspired by the idea of perturbations in pseudospectra and drawing on the notion that probability distributions also characterize variables within a neighborhood by measuring their overall dispersion (Colbrook et al., 2024), we propose a variance-aware continuous spectrum and model it using Gaussian distributions. In this way, the perturbation of eigenvalues is not only determined by their local perturbations but also influenced by the Gaussian distribution over global dynamics. Moreover, to better capture the global patterns of dynamic systems, we utilize the latent representation of the DMD mean as a condition to guide the learning of the spectrum, thereby obtaining the final variance-aware continuous spectrum distribution. We introduce a learnable Koopman operator to model the global evolution of feature representations through variance-aware continuous spectrum analysis.

To solve the proposed variance-aware continuous spectrum distribution, we integrate it into the Neural Process (NP) framework (Garnelo et al., 2018b; **?**) to employ its probabilistic inference capabilities. Specifically, we introduce KooNPro, a novel probabilistic approach that integrates the latent feature representation obtained from NP with the representation learned through variance-aware continuous spectrum modeling to enhance the global modeling capability of the dynamic system. We employ variational inference to solve this new modeling framework and derive a new Evidence Lower Bound (ELBO) to optimize the combined model. By incorporating variance-aware continuous spectrum, KooNPro maintains high predictive accuracy and robustness in high-dimension and non-stationary tasks, achieving state-of-the-art performance. Our contributions are summarized as follows:

- We introduce KooNPro, a novel probabilistic prediction model that synergistically integrates the probabilistic Koopman model with NP. Drawing inspiration from perturbations in pseudospectra, KooNPro incorporates a variance-aware continuous spectrum to learn temporal dynamics effectively. Additionally, it utilizes NP to capture the underlying global dynamics that govern the entire time series, facilitating improved predictive capabilities.
- Extensive experiments conducted on diverse real-world datasets convincingly demonstrate the superiority of our proposed model. It consistently achieves state-of-the-art performance, significantly outperforming existing methods across two metrics.
- We perform a comprehensive ablation studies demonstrating the ability of NP to capture temporal dynamics, thereby enhancing prediction performance, and evaluate the impact of key hyperparameters on model efficacy. Additionally, we provide a detailed case study that visually illustrates the prediction performance of KooNPro, yielding intuitive and interpretable results.

2 Related work

This work engages with three key areas: probabilistic time series prediction, the probabilistic Koopman model, and Neural Processes. While each of these fields has received considerable attention, we limit our discussion to the most pertinent studies to ensure brevity. A more extensive review of the literature can be found in the AppendixA.

Probabilistic time series prediction: Recent developments in probabilistic forecasting for time series integrate deep learning, statistical approaches, and diffusion models. Rangapuram et al. (2018b), Salinas et al. (2020) and Li et al. (2021) combine state space models with deep learning. Feng et al. (2023) and Tang & Matteson (2021) developed attention-based mechanisms that enhance long-range dependencies for improved forecast accuracy. Gaussian Processes (GP), combined with temporal

decomposition, were used by Yan et al. (2021), Nguyen & Quanz (2021), and Salinas et al. (2019) to better model uncertainty in multivariate settings. The data distribution-based generative models proposed by (Gouttes et al., 2021) and the Kalman filter-based approaches of de Bézenac et al. (2020), blend probabilistic techniques for better scalability and robustness. Diffusion models, such as those proposed by Rasul et al. (2021), Li et al. (2022), and Fan et al. (2024), model forecasting as a denoising task, excelling in high-dimensional settings.

Probabilistic Koopman model: The Koopman theory has been developed in lots of fields, the most related work is the probabilistic Koopman model, originally proposed by Morton et al. (2019). Han et al. (2022) designed a stochastic Koopman neural network for control, in which the latent observables are represented through a Gaussian distribution. Colbrook et al. (2024) integrates the concept of variance into the Koopman framework by introducing variance-pseudospectra, thereby ensuring convergence within the model. For time series tasks, Naiman et al. (2023) utilized Koopman theory to represent the latent conditional prior dynamics via a linear map, while Mallen et al. (2024) introduced a framework enabling probabilistic forecasting for systems with periodically varying uncertainty.

Neural process: Neural Processes (NPs), first proposed by Garnelo et al. (2018a) as Conditional Neural Processes, bridge neural networks' scalability and Gaussian Processes' ability to model uncertainty. Garnelo et al. (2018b) provide flexible, probabilistic function approximations, making them efficient and adaptable across tasks. Attentive Neural Processes (Kim et al., 2019), Convolutional Conditional Neural Processes (Gordon et al., 2019b), Gaussian Neural Processes (Bruinsma et al., 2021) and autoregressive NP (Bruinsma et al., 2023), deploy different methods to probe the relation between input-output pairs. Other work extends NPs to meta-learning, such as Meta-Learning Stationary Stochastic Processes (Foong et al., 2020) and Meta-Learning Probabilistic Inference (Gordon et al., 2019a). Comprehensive surveys by Jha et al. (2022) have organized these developments, exploring their wide applications in uncertainty-aware learning.

3 BACKGROUND

3.1 KOOPMAN THEORY AND CONTINUOUS SPECTRUM

The spectral decomposition of the Koopman operator is a powerful tool for understanding the underlying dynamics. If the operator has a discrete spectrum, it can be expressed in terms of its eigenvalues λ_j and eigenfunctions ϕ_j, where $\mathcal{K}\phi_j = \lambda_j \phi_j$. This decomposition reveals how different modes contribute to the system's evolution. However, many complex systems exhibit a continuous spectrum, indicating chaotic or highly irregular behavior. In such cases, the Koopman operator's spectrum is not composed of isolated eigenvalues but rather a continuum, requiring a more nuanced analysis. The operator can be represented through a spectral measure σ and an integral over the continuous spectrum: $\mathcal{K}g(\mathbf{x}) = \int_\sigma e^{i\omega t} dE(\omega) g(\mathbf{x})$, where $E(\omega)$ is a projection-valued measure. This spectral approach provides a comprehensive framework for capturing both regular and chaotic components of a system's dynamics. While the Koopman operator effectively models dynamics for predictive purposes, it is prone to spectral pollution, which can impede convergence and stability. Inspired by perturbations in pseudospectra and the notion that probability distributions characterize variables through overall dispersion, we propose a variance-aware continuous spectrum and model it using Gaussian distribution. This method ensures that eigenvalue perturbations are influenced not only by local disturbances but also by the Gaussian distribution over the dynamics.

3.2 NEURAL PROCESS

A stochastic process can be viewed as a random function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ where inputs can be regarded as indexing the output random variables. With a relaxed use of notation, we employ $p(f)$ in denoting a stochastic process, where f map inputs $x \in \mathcal{X}$ to $y \in \mathcal{Y}$. When fulfilling exchangeability and consistency as stated by Kolmogorov's extension theorem (Oksendal, 2013), Neural Process (NP) cited by Garnelo et al. (2018a) is a stochastic process that describes the predictive distribution over the target set $(\boldsymbol{x}_D, \boldsymbol{y}_D) := (\boldsymbol{x}_i, \boldsymbol{y}_i)_{i \in D}$ given the context set $(\boldsymbol{x}_C, \boldsymbol{y}_C) := (\boldsymbol{x}_i, \boldsymbol{y}_i)_{i \in C}$. Garnelo et al. (2018b) prompts use the distribution of a high-dimensional random vector $\boldsymbol{S}$ to represent $p(f)$ as

$$p(\boldsymbol{y}_D | \boldsymbol{x}_D, \boldsymbol{x}_C, \boldsymbol{y}_C) := \int p(\boldsymbol{y}_D | \boldsymbol{x}_D, \boldsymbol{S}) p(\boldsymbol{S} | \boldsymbol{x}_C, \boldsymbol{y}_C) d\boldsymbol{S}. \tag{1}$$

NP can be divided into two components: an encoder that maps the input-output pair of the context set $(\boldsymbol{x}_C, \boldsymbol{y}_C)$ to $\boldsymbol{S}$ for representing pf, and a decoder combine $\boldsymbol{S}$ with $\boldsymbol{x}_D$ to generate $\boldsymbol{y}_D$. Due to the intractable log-likelihood, NPs adopt amortized variational inference as Kingma & Welling (2014) and maximize the evidence lower bound (ELBO) of the log-likelihood as follows

$$\log\left(\boldsymbol{y}_D|\boldsymbol{x}_D, \boldsymbol{y}_C, \boldsymbol{x}_C\right) \geq \mathbb{E}_{q\sim q_{(\boldsymbol{S}|\boldsymbol{x}_D,\boldsymbol{y}_D)}}\left[\log p\left(\boldsymbol{y}_D|\boldsymbol{x}_D, \boldsymbol{S}\right)\right] - D_{KL}\left(q\left(\boldsymbol{S}|\boldsymbol{x}_D, \boldsymbol{y}_D\right)||p\left(\boldsymbol{S}|\boldsymbol{x}_C, \boldsymbol{y}_C\right)\right). \tag{2}$$

4 METHOD

This section offers a detailed overview of our model, with the complete architecture illustrated in Fig.1. The central idea of our approach is to learn temporal dynamics for probabilistic future prediction by integrating Neural Process (NP) with the probabilistic deep Koopman model. Initially, NP captures the discrete spectrum of dynamics governing the entire time series which is shown by the downward arrows in Fig.1. Additionally, inspired by the concept of pseudospectra, we utilize the probabilistic deep Koopman model to refine these dynamics, obtaining a variance-aware continuous spectrum for prediction which is demonstrated by the shadowed box in Fig.1.

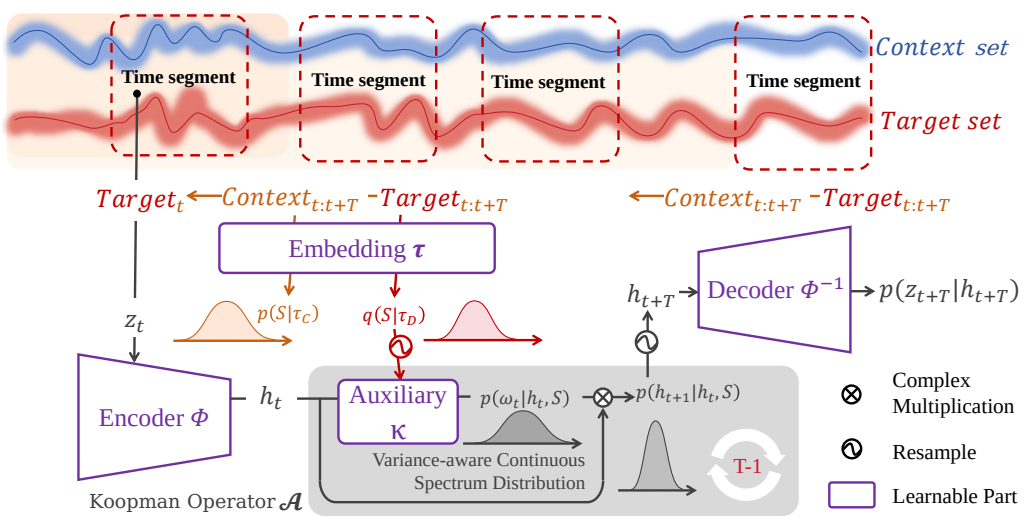

Figure 1: $p(\boldsymbol{S}|\tau_D)$ and $p(\boldsymbol{S}|\tau_C)$ are the latent representation of DMD mean over the target set and the context set respectively. Given the first time step $\boldsymbol{z}_1$ in the target set, the encoder Φ generates the latent variable $\boldsymbol{h}_1$. We apply Koopman operator $\mathcal{A}$ to $\boldsymbol{h}_1$ for $T-1$ steps then generate the distribution of $\boldsymbol{h}_{2:T}$. Finally, the decoder Φ^{-1} map $\boldsymbol{h}_{2:T}$ back to origin space to generate the distribution of $\boldsymbol{z}_{2:T}$.

4.1 CAPTURE TEMPORAL DYNAMICS BY NP

The proposed model first identifies the distribution of the latent variable $\boldsymbol{S} \in \mathbb{R}^s$ in Eq.1 through an embedding $\boldsymbol{\tau}$ which is presented in the left part of Fig.1. The latent variable integrates the underlying dynamics present in the time series, which allows the model to be more reactive to global features of the time series. In time series analysis, dynamic mode decomposition (DMD) is a classical method that elucidates the relationship between time series data and corresponding dynamics. However, DMD's reliance on the dot spectrum limits its capacity to capture the nonlinear patterns inherent in complex systems, thereby constraining its predictive power. To address these limitations, we integrate NP with DMD to generate a stochastic process that captures the dynamics of time series more effectively.

We utilize Takens' theory like Yuan et al. (2021), define a time series $\boldsymbol{z}_{1:T} \in \mathbb{R}^{T\times d\times k}$, where T denotes time length, d denotes features and k denotes delay embedding length, and let $\boldsymbol{x} = \boldsymbol{z}_{1:T-1}$, $\boldsymbol{y} = \boldsymbol{z}_{2:T}$. To estimate the dynamics governing time series in the context set, we employ an MLP τ

to embed $\hat{\boldsymbol{x}} \in \mathbb{R}^{(T-1)\times d\times 1}, \hat{\boldsymbol{y}} \in \mathbb{R}^{(T-1)\times d\times 1}$ as the initial component of delay embedding like

$$\tau_C := \tau(\boldsymbol{x}_C, \boldsymbol{y}_C) = \psi\left(\frac{1}{c}\sum_{i=1}^{c} vec\left(\hat{\boldsymbol{x}}^\dagger \hat{\boldsymbol{y}}\right)\right), \tag{3}$$

where ψ denotes a learnable multilayer perception (MLP), c represents the number of items in the context set, and the eigenvalue of $\hat{\boldsymbol{x}}^\dagger \hat{\boldsymbol{y}}$ can approximate the dot spectrum of dynamics. We model the distribution of $\boldsymbol{S}$ by a factorized Gaussian parametrised by τ_C, i.e.,

$$p(\boldsymbol{S}|\boldsymbol{x}_C, \boldsymbol{y}_C) = \mathcal{N}(\boldsymbol{S}; \mu(\tau_C), \sigma(\tau_C)). \tag{4}$$

Denote $p(\boldsymbol{S}|\tau_C) := p(\boldsymbol{S}|x_C, y_C)$ for any set C, thus $p(\boldsymbol{S})$ represents the distribution of underlying dynamics according to different time series sets. The ELBO in Eq.2 can be reformulated as follows

$$\log(\boldsymbol{y}_D|\boldsymbol{x}_D, \boldsymbol{y}_C, \boldsymbol{x}_C) \geq \mathbb{E}_{q\sim q_{(\boldsymbol{S}|\tau_D)}}[\log p(\boldsymbol{y}_D|\boldsymbol{x}_D, \boldsymbol{S})] - D_{KL}(q(\boldsymbol{S}|\tau_D), p(\boldsymbol{S}|\tau_C)). \tag{5}$$

The second term captures the difference in the distribution of $\boldsymbol{S}$, reflecting the diversity of temporal dynamics across different time series. Although we employ τ to embed the dot spectrum that characterizes temporal dynamics, it provides only a coarse approximation. Regarding the first term, we maximize it using the probabilistic deep Koopman model, which provides a more nuanced representation of temporal dynamics through the lens of the pseudospectra.

4.2 Probabilistic deep Koopman model

The probabilistic deep Koopman model shown in the shadowed box of Fig.1 concentrates on explaining local characteristics of time series, namely intricate temporal dynamics, with a variance-aware continuous spectrum. This approach offers a deeper understanding of temporal dynamics behavior over time, leading to improved predictive capabilities. In essence, estimating the true temporal dynamics is equivalent to maximize the expection $\mathbb{E}_{q_{(\boldsymbol{S}|\tau_D)}}[\log p(\boldsymbol{y}_D|\boldsymbol{x}_D, \boldsymbol{S})]$.

For simplicity, we discuss only the data in the target set and omit the index D. Firstly, we vectorized the delay embedding time series of size $(T-1)\times d\times k$ to $\boldsymbol{x} = \boldsymbol{z}_{1:T-1}$ of size $(T-1)\times(d\cdot k)\times 1$. Following Lusch et al. (2018), We employ three MLPs, an encoder $\phi : \mathbb{R}^{d\times k} \to \mathbb{R}^n$ and a decoder $\phi^{-1} : \mathbb{R}^n \to \mathbb{R}^{d\times k}$ to identify an appropriate linear space, along with an auxiliary network $\kappa : \mathbb{R}^{n+s} \to \mathbb{R}^n$ to learn the continuous spectrum of dynamics. For the encoder ϕ, we define $\boldsymbol{h}_{1:T-1} = \phi(\boldsymbol{x})$ and hypothesize the latent space created by ϕ possesses linear characteristics, thus the dynamics in such space can be described as follows

$$\boldsymbol{h}_{t+1} = e^{\boldsymbol{\lambda}_t \Delta t}\boldsymbol{h}_t. \tag{6}$$

We assume $\boldsymbol{h}_t$ represents the eigenvalue of the dynamics, consisting of a pair of conjugate numbers, and $\boldsymbol{\lambda}_t$ to be pure imaginary like Lange et al. (2021), let $\boldsymbol{\lambda}_t = j\boldsymbol{\omega}_t$, thus we have

$$\mathbf{Re}(\boldsymbol{h}_{t+1}) = \mathbf{Re}(\boldsymbol{h}_t)\odot\cos(\boldsymbol{\omega}_t\Delta t) - \mathbf{Im}(\boldsymbol{h}_t)\odot\sin(\boldsymbol{\omega}_t\Delta t), \tag{7}$$

$$\mathbf{Im}(\boldsymbol{h}_{t+1}) = \mathbf{Re}(\boldsymbol{h}_t)\odot\sin(\boldsymbol{\omega}_t\Delta t) + \mathbf{Im}(\boldsymbol{h}_t)\odot\cos(\boldsymbol{\omega}_t\Delta t), \tag{8}$$

where $\odot$ is the element-wise product. We can simplify Eq.7 and Eq.8 to

$$\boldsymbol{h}_{t+1} := \mathcal{A}(\boldsymbol{h}_t, \boldsymbol{\omega}_t). \tag{9}$$

In order to learn the continuous spectrum $\mathcal{A}$, which is conditioned on the global temporal dynamics representation $\boldsymbol{S}$, we utilize the auxiliary network κ to project the concatenation of $\boldsymbol{h}_t$ and $\boldsymbol{S}$ to κ_t,

$$\kappa : \begin{bmatrix}\boldsymbol{h}_t \\ \boldsymbol{S}\end{bmatrix} \mapsto \kappa_t. \tag{10}$$

Drawing on the theory of pseudospectra, we aim to learn dynamic information not only from the eigenvalues but also from their neighborhoods. To facilitate this, we employ an auxiliary network to generate the corresponding Gaussian distribution of $\boldsymbol{\omega}_t$ parameterized by κ_t as

$$p(\boldsymbol{\omega}_t|\boldsymbol{h}_t, \boldsymbol{S}) = \mathcal{N}(\boldsymbol{\omega}_t; \mu(\kappa_t), \sigma(\kappa_t)). \tag{11}$$

To simplify our model, we assume the independence of $\boldsymbol{h}_{1:T-1}$ on the linear space. Consequently, we can describe the uncertain dynamics as

$$p(\boldsymbol{\omega}_{1:T-1}|\boldsymbol{h}_{1:T-1}, \boldsymbol{S}) = \prod_{t=1}^{T-1}\mathcal{N}(\boldsymbol{\omega}_t; \mu(\kappa_t), \sigma(\kappa_t)). \tag{12}$$

Since $\boldsymbol{h}_t = \mathcal{A}(\boldsymbol{h}_{t-1}, \boldsymbol{\omega}_{t-1})$, it yields that $p(\boldsymbol{h}_T|\boldsymbol{\omega}_{1:T-1}, \boldsymbol{h}_{1:T-1}, \boldsymbol{S}) = 1$. Thus, the variance-aware continuous spectrum in the linear space can be characterized within the framework of the Gaussian distribution like

$$p(\boldsymbol{h}_{2:T}|\boldsymbol{h}_{1:T-1}, \boldsymbol{S}) = p(\boldsymbol{h}_T|\boldsymbol{\omega}_{1:T-1}, \boldsymbol{h}_{1:T-1}, \boldsymbol{S})p(\boldsymbol{\omega}_{1:T-1}|\boldsymbol{h}_{1:T-1}, \boldsymbol{S}) \tag{13}$$

$$= \prod_{t=1}^{T-1} \mathcal{N}(\boldsymbol{\omega}_t; \mu(\kappa_t), \sigma(\kappa_t)). \tag{14}$$

Utilizing the decoder ϕ^{-1}, we map the temporal dynamic evolution from the linear space back to the original space, as follows

$$p(\boldsymbol{z}_{2:T}|\boldsymbol{h}_{2:T}) = \prod_{t=2}^{T} \mathcal{N}(\boldsymbol{z}_t; \mu(\phi^{-1}(\boldsymbol{h}_t)), \sigma(\phi^{-1}(\boldsymbol{h}_t))). \tag{15}$$

The likelihood $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{S})$ in Eq.5, where $\boldsymbol{x} = \boldsymbol{z}_{1:T-1}$ and $\boldsymbol{y} = \boldsymbol{z}_{2:T}$, can be derived as follows

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{S}) = p(\boldsymbol{z}_{2:T}|\boldsymbol{h}_{1:T-1}, \boldsymbol{S}) = p(\boldsymbol{z}_{2:T}|\boldsymbol{h}_{2:T})p(\boldsymbol{h}_{2:T}|\boldsymbol{h}_{1:T-1}, \boldsymbol{S}), \tag{16}$$

given that the encoder ϕ is a deterministic function. Insert Eq.14 and Eq.15 to Eq.16, we can draw the expectation as

$$\mathbb{E}_{q\sim q_{(\boldsymbol{S}|\tau_D)}}[\log p(\boldsymbol{y}_D|\boldsymbol{x}_D, \boldsymbol{S})] = \mathbb{E}_{q\sim q_{(\boldsymbol{S}|\tau_D)}}[\sum_{t=2}^{T} \log(p(\boldsymbol{z}_t|\boldsymbol{h}_t)) + \sum_{t=1}^{T-1} \log(p(\boldsymbol{h}_{t+1}|\boldsymbol{h}_t, \boldsymbol{S}))]. \tag{17}$$

The first term can be interpreted as the prediction loss, while the second term represents the linear loss, as described in Lusch et al. (2018). By maximizing the ELBO in Eq.5, we derive $\boldsymbol{S}_C$ approximates the dynamics for each time series present during the training phase. Consequently, our predictions in the test phase commence with the initial $\boldsymbol{z}_1 \in \mathbb{R}^{1\times d\times k}$ like

$$\boldsymbol{z}_{\tilde{T}} = \phi^{-1}(\mathcal{A}^{\tilde{T}-1}(\phi(\boldsymbol{z}_1), \boldsymbol{S}_C)). \tag{18}$$

It is crucial to note that the time length T during the training stage is distinct from the prediction horizon $\tilde{T}$, with the input history length solely determined by the delay embedding length k. This allows us to train a single model to forecast future values of arbitrary length.

5 Experiments

In this section, we conduct comprehensive experiments on nine real-world datasets to evaluate the performance of KooNPro against state-of-the-art baselines. We demonstrate the capacity of Neural Process to enhance and investigate the influence of key hyperparameters of KooNPro through ablation studies. Finally, we visualize the prediction results for the `Solar` dataset and analyze the relationship between prediction error and variance.

5.1 Settings

Datasets. We consider nine real-world datasets characterized by a range of temporal dynamics, namely `ETTs`, `Solar`, `Electricity`, `Traffic`, `Taxi`, and `KDD-cup`. The data is recorded at intervals of 15 minutes, 30 minutes, 1 hour, or 1 day frequencies. Refer to Appendix B.3 for details. All datasets are split chronologically and adopt the same train/validation/test ratios, *i.e.*, 7:1:2.

Baselines. We assess the predictive performance of KooNPro in comparison with multivariate time series forecasting models, including GP-Copula (Salinas et al., 2019), Transformer-MAF (Rasul et al., 2021), TimeGrad (Rasul et al., 2021), TACTiS (Ashok et al., 2023), D^{3}VAE (Li et al., 2023), DPK (Mallen et al., 2024), and MG-TSD (Fan et al., 2024).

Evaluation Metrics. Following previous work (Fan et al., 2024), We assess our model and all baselines using $\mathbf{CRPS_{sum}}$ (Continuous Ranked Probability Score), a widely used metric for probabilistic time series forecasting, as well as $\mathbf{NRMSE_{sum}}$ (Normalized Root Mean Squared Error). The details of metrics is shown in Appendix C.

Implementation details. The training process is early stopped within 5 epochs using the Adam optimizer with a fixed learning rate of 10^{-5}. We set the mini-batch size to 128. Additional hyperparameters, such as time length T, delay embedding length k, and layers of MLPs are detailed in Appendix B.2. All models are trained and tested on a single NVIDIA RTX4070Ti 12GB GPU.

5.2 RESULTS

Table 1: Comparison of $\mathbf{CRPS_{sum}}$ (denoted as C-s, smaller is better) and $\mathbf{NRMSE_{sum}}$ (denoted as N-s, smaller is better) across nine real-world datasets. The means and standard errors are based on 10 independent runs of retraining and evaluation. The best performances are in red and the second are in blue.

Model	Metric	ETTh1	ETTh2	ETTm1	ETTm2	Solar	Electricity	Traffic	Taxi	Cup
GP-Copula	C-s	$0.537_{\pm 0.019}$	$0.264_{\pm 0.023}$	$0.241_{\pm 0.043}$	$0.147_{\pm 0.019}$	$0.305_{\pm 0.024}$	$0.078_{\pm 0.035}$	$0.199_{\pm 0.008}$	$0.286_{\pm 0.066}$	$0.217_{\pm 0.000}$
	N-s	$0.835_{\pm 0.032}$	$0.383_{\pm 0.032}$	$0.424_{\pm 0.088}$	$0.172_{\pm 0.027}$	$0.671_{\pm 0.034}$	$0.122_{\pm 0.058}$	$0.293_{\pm 0.018}$	$0.412_{\pm 0.115}$	$0.346_{\pm 0.010}$
Trans-MAF	C-s	$0.800_{\pm 0.049}$	$0.223_{\pm 0.012}$	$0.379_{\pm 0.029}$	$0.228_{\pm 0.040}$	$0.964_{\pm 0.013}$	$0.144_{\pm 0.046}$	$0.477_{\pm 0.021}$	$0.403_{\pm 0.046}$	$0.257_{\pm 0.013}$
	N-s	$1.285_{\pm 0.188}$	$0.315_{\pm 0.024}$	$0.577_{\pm 0.082}$	$0.356_{\pm 0.065}$	$1.665_{\pm 0.022}$	$0.245_{\pm 0.079}$	$0.762_{\pm 0.031}$	$0.598_{\pm 0.029}$	$0.390_{\pm 0.010}$
Timegrid	C-s	$0.547_{\pm 0.022}$	$0.241_{\pm 0.002}$	$0.227_{\pm 0.029}$	$0.212_{\pm 0.015}$	$0.594_{\pm 0.011}$	$0.044_{\pm 0.010}$	$0.455_{\pm 0.020}$	$0.327_{\pm 0.058}$	$0.271_{\pm 0.087}$
	N-s	$0.889_{\pm 0.039}$	$0.325_{\pm 0.000}$	$0.363_{\pm 0.072}$	$0.291_{\pm 0.018}$	$1.081_{\pm 0.015}$	$0.072_{\pm 0.014}$	$0.560_{\pm 0.016}$	$0.498_{\pm 0.095}$	$0.341_{\pm 0.097}$
TACTIS	C-s	$0.601_{\pm 0.004}$	$0.208_{\pm 0.002}$	$0.634_{\pm 0.009}$	$0.142_{\pm 0.015}$	$1.871_{\pm 0.022}$	$0.254_{\pm 0.012}$	$0.456_{\pm 0.003}$	$0.981_{\pm 0.013}$	$0.276_{\pm 0.008}$
	N-s	$0.907_{\pm 0.010}$	$0.320_{\pm 0.006}$	$1.013_{\pm 0.010}$	$0.320_{\pm 0.006}$	$2.309_{\pm 0.528}$	$0.391_{\pm 0.016}$	$1.871_{\pm 0.003}$	$1.170_{\pm 0.013}$	$0.403_{\pm 0.013}$
D^3VAE	C-s	$0.445_{\pm 0.023}$	$0.266_{\pm 0.016}$	$0.219_{\pm 0.009}$	$0.177_{\pm 0.017}$	$0.312_{\pm 0.035}$	$0.198_{\pm 0.020}$	$0.265_{\pm 0.027}$	$0.257_{\pm 0.008}$	$0.243_{\pm 0.035}$
	N-s	$0.662_{\pm 0.029}$	$0.479_{\pm 0.027}$	$0.257_{\pm 0.058}$	$0.263_{\pm 0.007}$	$0.642_{\pm 0.070}$	$0.253_{\pm 0.104}$	$0.926_{\pm 0.103}$	$0.391_{\pm 0.038}$	$0.501_{\pm 0.005}$
DPK	C-s	$0.718_{\pm 0.011}$	$0.471_{\pm 0.024}$	$0.556_{\pm 0.018}$	$0.341_{\pm 0.062}$	$0.753_{\pm 0.035}$	$0.784_{\pm 0.008}$	$0.827_{\pm 0.007}$	$0.843_{\pm 0.009}$	$0.728_{\pm 0.029}$
	N-s	$1.026_{\pm 0.012}$	$0.725_{\pm 0.040}$	$0.887_{\pm 0.023}$	$0.391_{\pm 0.157}$	$1.130_{\pm 0.040}$	$1.062_{\pm 0.005}$	$1.160_{\pm 0.007}$	$1.165_{\pm 0.010}$	$1.147_{\pm 0.056}$
MG-TSD	C-s	$0.430_{\pm 0.038}$	$0.174_{\pm 0.009}$	$0.254_{\pm 0.054}$	$0.129_{\pm 0.009}$	$0.298_{\pm 0.025}$	$0.107_{\pm 0.055}$	$0.528_{\pm 0.057}$	$0.250_{\pm 0.073}$	$0.323_{\pm 0.015}$
	N-s	$0.693_{\pm 0.083}$	$0.220_{\pm 0.017}$	$0.394_{\pm 0.076}$	$0.292_{\pm 0.046}$	$0.623_{\pm 0.026}$	$0.155_{\pm 0.063}$	$0.710_{\pm 0.058}$	$0.347_{\pm 0.078}$	$0.638_{\pm 0.056}$
KooNPro	C-s	$0.328_{\pm 0.037}$	$0.149_{\pm 0.051}$	$0.165_{\pm 0.057}$	$0.081_{\pm 0.020}$	$0.211_{\pm 0.033}$	$0.057_{\pm 0.006}$	$0.184_{\pm 0.022}$	$0.226_{\pm 0.041}$	$0.204_{\pm 0.017}$
	N-s	$0.520_{\pm 0.045}$	$0.224_{\pm 0.065}$	$0.225_{\pm 0.028}$	$0.122_{\pm 0.034}$	$0.313_{\pm 0.044}$	$0.095_{\pm 0.012}$	$0.289_{\pm 0.025}$	$0.330_{\pm 0.078}$	$0.308_{\pm 0.030}$

Tab.1 presents the $\mathbf{CRPS_{sum}}$ and $\mathbf{NRMSE_{sum}}$ values, averaged over 10 independent runs. The results demonstrate that our model exhibits superior performance in both $\mathbf{CRPS_{sum}}$ and $\mathbf{NRMSE_{sum}}$, consistently outperforming all baseline models across the nine datasets.

In comparison, traditional baseline models (e.g., GP-Copula and Timegrid) have higher $\mathbf{CRPS_{sum}}$ values on most datasets, which indicates their difficulty in handling complex time series data. Compared to modern deep generative models (e.g., D^3VAE and TACTIS), KooNPro achieves a better balance between predictive stability and accuracy. Although MG-TSD shows a slight advantage in $\mathbf{NRMSE_{sum}}$ on certain datasets (e.g., `ETTs` series), its $\mathbf{CRPS_{sum}}$ fluctuates significantly, suggesting an inadequate characterization of distributional uncertainty. Our KooNPro introduces variance-aware spectra to enhance global pattern capturing, avoiding spectral instability caused by local perturbations, thus achieving higher stability and accuracy in handling complex nonlinear patterns. This demonstrates its effectiveness in high-dimensional time series forecasting.

5.3 ABLATION STUDY

In this section, we conduct an ablation study to identify the factors contributing to the success of KooNPro in prediction tasks. First, we delve into the contribution of Neural Process (NP), whose primary function is to learn the underlying dynamics across various time segments and represent these dynamics within $\boldsymbol{S}$. To evaluate the effectiveness of NP, we record the KL-divergence between $\boldsymbol{S}_D$ and $\boldsymbol{S}_C$ during the training stage as depicted in the upper section of Fig.2. It can be observed that the KL divergence in the training set decreases over time, converging to a similar level as that of the validation set around the 70th epoch. This suggests that the model is effectively learning the underlying dynamics of time series. To test the generalization capability of the learned $\boldsymbol{S}_C$, We draw 150-time segments from the test set and calculate the KL divergence with $\boldsymbol{S}_C$ as showcased in the lower section of Fig.2. The low-level KL-divergence indicates that $\boldsymbol{S}_C$ can approximate temporal dynamics even for data instances that were not present in the training set. To assess the contribution of NP, we train a probabilistic Koopman model, following the procedure detailed in Sec.4.2 excluding

$\boldsymbol{S}_D$. Thus the prediction process is described as follows

$$\boldsymbol{z}_{\tilde{T}} = \phi^{-1}(\mathcal{A}^{\tilde{T}-1}(\phi(\boldsymbol{z}_1))). \tag{19}$$

In Tab.2, we compare the prediction performance of KooNPro and without NP version, we observe a degradation in accuracy and an increase in deviation, resulting in collapsed and unstable performance.

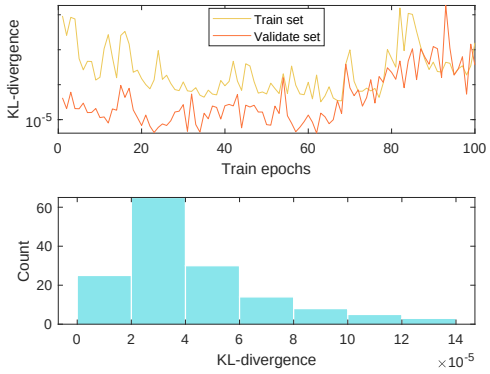

Figure 2: The upper section illustrates the KL-divergence between $\boldsymbol{S}_D$ and $\boldsymbol{S}_C$ during the training phase, with epochs of training stages on the x-axis and an exponential scale on the y-axis. The lower section analyzes 150 test dataset time segments, displaying their KL-divergence with $\boldsymbol{S}_C$, with the x-axis using an exponential scale and the y-axis showing time series counts.

Table 2: Comparison of $\mathbf{CRPS_{sum}}$ (denoted as C-s, smaller is better) and $\mathbf{NRMSE_{sum}}$ (denoted as N-s, smaller is better) between KooNPro and without NP version (denoted as without-NP). The smaller mean and standard error indicate better performance.

Model	Metric	ETTh1	ETTh2	ETTm1	ETTm2	Solar	Electricity	Traffic	Taxi	Cup
KooNPro	C-s	$0.328_{\pm 0.037}$	$0.149_{\pm 0.051}$	$0.165_{\pm 0.057}$	$0.081_{\pm 0.020}$	$0.211_{\pm 0.033}$	$0.057_{\pm 0.006}$	$0.184_{\pm 0.022}$	$0.226_{\pm 0.041}$	$0.204_{\pm 0.017}$
	N-s	$0.520_{\pm 0.045}$	$0.224_{\pm 0.065}$	$0.225_{\pm 0.028}$	$0.122_{\pm 0.034}$	$0.313_{\pm 0.044}$	$0.095_{\pm 0.012}$	$0.289_{\pm 0.025}$	$0.330_{\pm 0.078}$	$0.308_{\pm 0.030}$
without-NP	C-s	$0.390_{\pm 0.083}$	$0.218_{\pm 0.120}$	$0.332_{\pm 0.068}$	$0.155_{\pm 0.042}$	$0.341_{\pm 0.059}$	$0.093_{\pm 0.027}$	$0.313_{\pm 0.051}$	$0.252_{\pm 0.068}$	$0.674_{\pm 0.224}$
	N-s	$0.609_{\pm 0.117}$	$0.318_{\pm 0.169}$	$0.511_{\pm 0.103}$	$0.243_{\pm 0.056}$	$0.531_{\pm 0.065}$	$0.144_{\pm 0.036}$	$0.590_{\pm 0.095}$	$0.380_{\pm 0.092}$	$1.058_{\pm 0.341}$

Next, we investigate the impact of the hyperparameter time length T and delay embedding length k on the prediction performance. In Pic.3, we investigate those hyperparameters work in the `ETTh1` dataset. The orange and transparent orange curves report the change of metric $\mathbf{CRPS_{sum}}$ and $\mathbf{NRMSE_{sum}}$ along with the T increase. We find the prediction accuracy of KooNPro increases dramatically with increasing T until $T = 30$. After that, the accuracy can't improve even if we pay more resources. The same picture appeared when we ablate k shown in the blue and transparent blue curves. This phenomenon may suggest that temporal dynamics behind time series can't be unveiled by simply improving the learning time segment. Additional ablation studies across various datasets are presented in the AppendixD.

5.4 Case Study

The predictive capability of KooNPro arises from its ability to capture the underlying dynamics of time series data effectively, we illustrate this using the `Solar` dataset, which comprises hourly measurements from 137 solar plants located in Alabama state. The dataset exhibits a clear diurnal pattern: values are non-zero from 6:00 to 18:00 and zero at night. In robust probabilistic prediction, large errors often correlate with high variance, reflecting uncertainty. We evaluate this by examining the relationship between prediction variance and accuracy, measured by the mean absolute error (MAE) of the predicted mean relative to the true value. A scatter plot of MAE against prediction variance, presented in the right panel of Fig.4, reveals a correlation of 0.94 with a significance level near zero, demonstrating that KooNPro effectively captures underlying temporal dynamics and

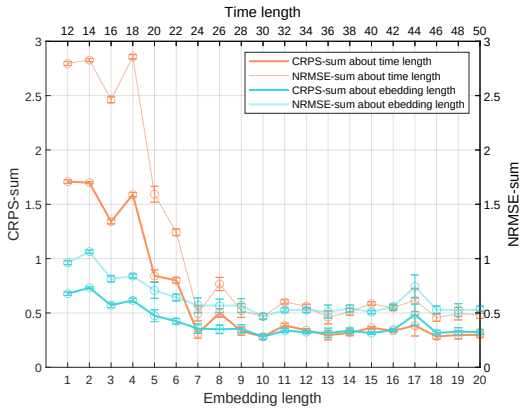

Figure 3: The top axis shows the increase in time length T, while the bottom axis indicates the increase in delay embedding length k. The orange and transparent orange curves represent the $\mathbf{CRPS_{sum}}$ and $\mathbf{NRMSE_{sum}}$ for increasing T. The blue and transparent blue curves correspond to these metrics for increasing k. Error bars reflect the standard error from 10 independent retraining and evaluation runs.

provides reliable predictions. The left panel visualizes the ground truth and predictions for the first eight plants over 24 hours. The results demonstrate that KooNPro accurately captures the temporal fluctuations in `Solar` energy generation: the predicted mean closes to zero during nighttime and rises to varying peaks during daylight hours for each plant. Furthermore, the prediction intervals widen during peak sunlight periods, reflecting the reliability of the predictive result.

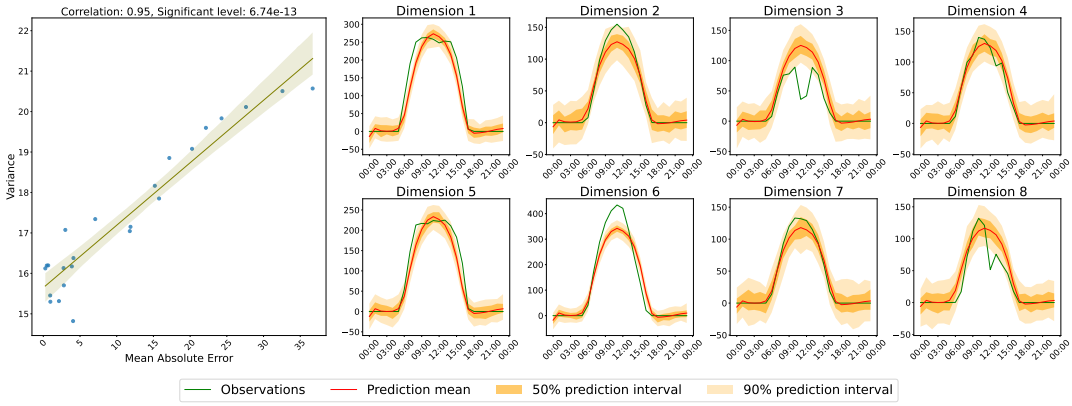

Figure 4: The left panel illustrates the correlation between the MAE of the prediction means relative to the true values and the prediction variance. The title indicates that the correlation achieves 0.95 and the significant level is $6.74 \times e - 13$. The right panel visualizes the changes in the first eight dimensions of the Solar dataset over 24 hours.

6 Conclusion

This paper presents KooNPro, a novel probabilistic forecasting model that treats time series as temporal dynamics. KooNPro employs a variance-aware continuous spectrum, inspired by pseudospectra concepts, to capture underlying dynamics in time series. It integrates Neural Processes to comprehend global dynamics across the entire series, enhancing its learning capabilities. Extensive experiments on nine real-world datasets demonstrate KooNPro's superior performance compared to state-of-the-art methods. Comprehensive ablation studies explore the origins of KooNPro's predictive power, while visualizations of `Solar` dataset predictions showcase its accuracy and reliability.

REFERENCES

Ian Abraham, Gerardo De La Torre, and Todd D Murphey. Model-based control using koopman operators. *arXiv preprint arXiv:1709.01568*, 2017.

Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, et al. Gluonts: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, 21(116):1–6, 2020.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Nicolas Chapados, and Alexandre Drouin. Tactis-2: Better, faster, simpler attentional copulas for multivariate time series. *arXiv preprint arXiv:2310.01327*, 2023.

François Auger, Mickael Hilairet, Josep M Guerrero, Eric Monmasson, Teresa Orlowska-Kowalska, and Seiichiro Katsura. Industrial applications of the kalman filter: A review. *IEEE Transactions on Industrial Electronics*, 60(12):5458–5471, 2013.

Omri Azencot, Wotao Yin, and Andrea Bertozzi. Consistent dynamic mode decomposition. *SIAM Journal on Applied Dynamical Systems*, 18(3):1565–1585, 2019.

Omri Azencot, N Benjamin Erichson, Vanessa Lin, and Michael Mahoney. Forecasting sequential data using consistent koopman autoencoders. In *International Conference on Machine Learning*, pp. 475–485. PMLR, 2020.

Wessel P. Bruinsma, James Requeima, Andrew Y. K. Foong, Jonathan Gordon, and Richard E. Turner. The Gaussian Neural Process, January 2021.

Wessel P Bruinsma, Stratis Markou, James Requiema, Andrew YK Foong, Tom R Andersson, Anna Vaughan, Anthony Buonomo, J Scott Hosking, and Richard E Turner. Autoregressive conditional neural processes. *arXiv preprint arXiv:2303.14468*, 2023.

Steven L. Brunton, Bingni W. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Koopman Invariant Subspaces and Finite Linear Representations of Nonlinear Dynamical Systems for Control. *PLOS ONE*, 11(2):e0150171, February 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0150171.

Steven L Brunton, Marko Budišić, Eurika Kaiser, and J Nathan Kutz. Modern koopman theory for dynamical systems. *arXiv preprint arXiv:2102.12086*, 2021.

Matthew J Colbrook and Alex Townsend. Rigorous data-driven computation of spectral properties of koopman operators for dynamical systems. *Communications on Pure and Applied Mathematics*, 77(1):221–283, 2024.

Matthew J Colbrook, Qin Li, Ryan V Raut, and Alex Townsend. Beyond expectations: residual dynamic mode decomposition and variance for stochastic dynamical systems. *Nonlinear Dynamics*, 112(3):2037–2061, 2024.

Emmanuel de Bézenac, Syama Sundar Rangapuram, Konstantinos Benidis, Michael Bohlke-Schneider, Richard Kurle, Lorenzo Stella, Hilaf Hasson, Patrick Gallinari, and Tim Januschowski. Normalizing Kalman Filters for Multivariate Time Series Analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2995–3007. Curran Associates, Inc., 2020.

Akshunna S Dogra and William Redman. Optimizing neural networks via koopman operator theory. *Advances in Neural Information Processing Systems*, 33:2087–2097, 2020.

Xinyao Fan, Yueying Wu, Chang Xu, Yuhao Huang, Weiqing Liu, and Jiang Bian. MG-TSD: Multi-Granularity Time Series Diffusion Models with Guided Learning Process, March 2024.

Shibo Feng, Chunyan Miao, Ke Xu, Jiaxiang Wu, Pengcheng Wu, Yang Zhang, and Peilin Zhao. Multi-scale Attention Flow for Probabilistic Time Series Forecasting, July 2023.

Andrew Foong, Wessel Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard Turner. Meta-learning stationary stochastic process prediction with convolutional neural processes. *Advances in Neural Information Processing Systems*, 33:8284–8295, 2020.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional Neural Processes. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1704–1713. PMLR, July 2018a.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural Processes, July 2018b.

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E. Turner. Meta-Learning Probabilistic Inference For Prediction, August 2019a.

Jonathan Gordon, Wessel P Bruinsma, Andrew YK Foong, James Requeima, Yann Dubois, and Richard E Turner. Convolutional conditional neural processes. *arXiv preprint arXiv:1910.13556*, 2019b.

Adèle Gouttes, Kashif Rasul, Mateusz Koren, Johannes Stephan, and Tofigh Naghibi. Probabilistic Time Series Forecasting with Implicit Quantile Networks, July 2021.

Zongyu Guo, Cuiling Lan, Zhizheng Zhang, Yan Lu, and Zhibo Chen. Versatile neural processes for learning implicit neural representations. *arXiv preprint arXiv:2301.08883*, 2023.

Minghao Han, Jacob Euler-Rolle, and Robert K Katzschmann. DESKO: STABILITY-ASSURED ROBUST CONTROL OF NONLINEAR SYSTEMS WITH A DEEP STOCHASTIC KOOPMAN OPERATOR. 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Saurav Jha, Dong Gong, Xuesong Wang, Richard E Turner, and Lina Yao. The neural process family: Survey, applications and perspectives. *arXiv preprint arXiv:2209.00517*, 2022.

Eurika Kaiser, J Nathan Kutz, and Steven L Brunton. Data-driven discovery of koopman eigenfunctions for control. *Machine Learning: Science and Technology*, 2(3):035023, 2021.

Makoto Kawano, Wataru Kumagai, Akiyoshi Sannai, Yusuke Iwasawa, and Yutaka Matsuo. Group equivariant conditional neural processes. *arXiv preprint arXiv:2102.08759*, 2021.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive Neural Processes, July 2019.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL `http://arxiv.org/abs/1312.6114`.

Marcel Kollovieh, Abdul Fatir Ansari, Michael Bohlke-Schneider, Jasper Zschiegner, Hao Wang, and Yuyang Bernie Wang. Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting. *Advances in Neural Information Processing Systems*, 36, 2024.

Bernard O Koopman. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

Milan Korda and Igor Mezić. Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. *Automatica*, 93:149–160, 2018.

Jiaqing Kou and Weiwei Zhang. Dynamic mode decomposition with exogenous input for data-driven modeling of unsteady flows. *Physics of fluids*, 31(5), 2019.

Deepthi Praveenlal Kuttichira, EA Gopalakrishnan, Vijay Krishna Menon, and KP Soman. Stock price prediction using dynamic mode decomposition. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 55–60. IEEE, 2017.

Henning Lange, Steven L Brunton, and J Nathan Kutz. From fourier to koopman: Spectral methods for long-term time series prediction. *Journal of Machine Learning Research*, 22(41):1–38, 2021.

Hyungi Lee, Eunggu Yun, Giung Nam, Edwin Fong, and Juho Lee. Martingale posterior neural processes. *arXiv preprint arXiv:2304.09431*, 2023.

Longyuan Li, Junchi Yan, Xiaokang Yang, and Yaohui Jin. Learning Interpretable Deep State Space Model for Probabilistic Time Series Forecasting, January 2021.

Yan Li, Xinjiang Lu, Yaqing Wang, and Dejing Dou. Generative time series forecasting with diffusion, denoise, and disentanglement. *Advances in Neural Information Processing Systems*, 35: 23009–23022, 2022.

Yanhong Li, Jack Xu, and David C. Anastasiu. An Extreme-Adaptive Time Series Prediction Model Based on Probability-Enhanced LSTM Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8684–8691, June 2023. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v37i7.26045.

Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning Non-stationary Time Series Dynamics with Koopman Predictors.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):4950, November 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07210-0.

Alex T Mallen, Henning Lange, and J Nathan Kutz. Deep probabilistic koopman: long-term time-series forecasting under periodic uncertainties. *International Journal of Forecasting*, 40(3): 859–868, 2024.

Iva Manojlović, Maria Fonoberova, Ryan Mohr, Aleksandr Andrejčuk, Zlatko Drmač, Yannis Kevrekidis, and Igor Mezić. Applications of koopman mode analysis to neural networks. *arXiv preprint arXiv:2006.11765*, 2020.

Peiman Mohseni and Nick Duffield. Spectral convolutional conditional neural processes. *arXiv preprint arXiv:2404.13182*, 2024.

Jeremy Morton, Freddie D Witherden, and Mykel J Kochenderfer. Deep variational koopman models: Inferring koopman observations for uncertainty-aware dynamics modeling and control. *arXiv preprint arXiv:1902.09742*, 2019.

Ilan Naiman and Omri Azencot. A koopman approach to understanding sequence neural models. *arXiv preprint arXiv:2102.07824*, 2021.

Ilan Naiman, N Benjamin Erichson, Pu Ren, Michael W Mahoney, and Omri Azencot. Generative modeling of regular and irregular time series data via koopman vaes. *arXiv preprint arXiv:2310.02619*, 2023.

Abhinav Narasingam, Sang Hwan Son, and Joseph Sang-Il Kwon. Data-driven feedback stabilisation of nonlinear systems: Koopman-based model predictive control. *International Journal of Control*, 96(3):770–781, 2023.

Nam Nguyen and Brian Quanz. Temporal Latent Auto-Encoder: A Method for Probabilistic Multivariate Time Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35 (10):9117–9125, May 2021. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v35i10.17101.

Tung Nguyen and Aditya Grover. Transformer neural processes: Uncertainty-aware meta learning via sequence modeling. *arXiv preprint arXiv:2207.04179*, 2022.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Biswajit Paria, Rajat Sen, Amr Ahmed, and Abhimanyu Das. Hierarchically regularized deep forecasting. *arXiv preprint arXiv:2106.07630*, 2021.

Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. *Advances in neural information processing systems*, 31, 2018a.

Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep State Space Models for Time Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8857–8868. PMLR, July 2021.

William T Redman, Maria Fonoberova, Ryan Mohr, Ioannis G Kevrekidis, and Igor Mezic. An operator theoretic view on pruning deep neural networks. *arXiv preprint arXiv:2110.14856*, 2021.

Said E Said and David A Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607, 1984.

Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*, 2017.

David Salinas, Michael Bohlke-Schneider, Laurent Callot, Roberto Medico, and Jan Gasthaus. High-dimensional multivariate forecasting with low-rank Gaussian Copula Processes. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, July 2020. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2019.07.001.

Lifeng Shen and James Kwok. Non-autoregressive Conditional Diffusion Models for Time Series Prediction. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 31016–31029. PMLR, July 2023.

Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning koopman invariant subspaces for dynamic mode decomposition. *Advances in neural information processing systems*, 30, 2017.

Binh Tang and David S Matteson. Probabilistic Transformer For Time Series Analysis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23592–23608. Curran Associates, Inc., 2021.

Kshitij Tayal, Arvind Renganathan, Rahul Ghosh, Xiaowei Jia, and Vipin Kumar. Koopman invertible autoencoder: Leveraging forward and backward dynamics for temporal modeling. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 588–597. IEEE, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Rui Wang, Yihe Dong, Sercan Ö Arik, and Rose Yu. Koopman Neural Forecaster for Time Series with Temporal Distribution Shifts, February 2023.

Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. Deep factors for forecasting. In *International conference on machine learning*, pp. 6607–6617. PMLR, 2019.

Haomin Wen, Youfang Lin, Yutong Xia, Huaiyu Wan, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. DiffSTG: Probabilistic Spatio-Temporal Graph Forecasting with Denoising Diffusion Models, March 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430. Curran Associates, Inc., 2021.

Tijin Yan, Hongwei Zhang, Tong Zhou, Yufeng Zhan, and Yuanqing Xia. ScoreGrad: Multivariate Probabilistic Time Series Forecasting with Continuous Energy-based Generative Models, June 2021.

Yuan Yuan, Kaiwen Zhou, Wenwu Zhou, Xin Wen, and Yingzheng Liu. Flow prediction using dynamic mode decomposition with time-delay embedding based on local measurement. *Physics of Fluids*, 33(9), 2021.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, May 2021. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v35i12.17325.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 27268–27286. PMLR, June 2022.

A APPENDIX: RELATED WORK

This work engages with three key areas: probabilistic time series prediction, the probabilistic Koopman model, and Neural Processes. While each of these fields has received considerable attention in recent literature. Here we talk only most related work:

Probabilistic time series prediction: Recent developments in probabilistic forecasting for time series integrate deep learning, statistical approaches, and diffusion models. State space models have been significantly enhanced by deep learning, as seen in (Rangapuram et al. (2018b)), (Salinas et al. (2020)) and (Li et al. (2021)), where they effectively capture temporal dynamics and handle missing data. (Feng et al. (2023)) and (Tang & Matteson (2021)) developed attention-based mechanisms that enhance long-range dependencies for improved forecast accuracy. Gaussian Processes (GP), combined with temporal decomposition, were used by (Yan et al. (2021)) and (Nguyen & Quanz (2021)) to better model uncertainty in multivariate settings. (Salinas et al. (2019)) proposed a Gaussian copula process, and (Ashok et al. (2023)) combine Gaussian copula and attention mechanism for capturing dependencies in high-dimensional multivariate forecasting. The data distribution-based generative models proposed by (Gouttes et al. (2021)) and the Kalman filter-based approaches of (de Bézenac et al. (2020)), blend probabilistic techniques for better scalability and robustness. Diffusion models, such as those proposed by (Rasul et al. (2021)), (Li et al. (2022)), and (Fan et al. (2024)), model forecasting as a denoising task, excelling in high-dimensional settings. For faster training and prediction, (Shen & Kwok (2023)) introduced non-autoregressive diffusion models. (Wen et al. (2024)) introduced spatio-temporal diffusion models, extending these methods to spatial dependencies.

Probabilistic Koopman model: Over the past two decades, Koopman techniques have garnered substantial attention, with applications spanning analysis (Brunton et al. (2016)), (Takeishi et al. (2017)), (Lusch et al. (2018)), (Azencot et al. (2019)); control Abraham et al. (2017), Korda & Mezić (2018), Kaiser et al. (2021), (Narasingam et al. (2023)); optimization (Dogra & Redman (2020)), (Manojlović et al. (2020)), (Naiman & Azencot (2021)), (Redman et al. (2021)), and forecasting Azencot et al. (2020), Lange et al. (2021), (Wang et al. (2023)), (Liu et al.), (Tayal et al. (2023)). (Brunton et al. (2021)) comprehensively discusses these advances and highlights future research directions. The most closely related work to ours is the probabilistic Koopman model, originally proposed by (Morton et al. (2019)). (Han et al. (2022)) designed a stochastic Koopman neural network for control, in which the latent observables are represented through a Gaussian distribution. (Colbrook et al., 2024) integrates the concept of variance into the Koopman framework by introducing variance-pseudospectra, thereby ensuring convergence within the model. For time series tasks, (Naiman et al. (2023)) utilized Koopman theory to represent the latent conditional prior dynamics via a linear map, while (Mallen et al. (2024)) introduced a framework enabling probabilistic forecasting for systems with periodically varying uncertainty.

Neural process: Neural Processes (NPs), first proposed by (Garnelo et al. (2018a)) as Conditional Neural Processes, bridge neural networks' scalability and Gaussian Processes' ability to model uncertainty. (Garnelo et al. (2018b)) provide flexible, probabilistic function approximations, making them efficient and adaptable across tasks, but struggled with long-range dependencies and flexible function distributions. Various models have since emerged to address these limitations. Attentive Neural Processes (Kim et al. (2019)) introduced attention mechanisms to handle long-range dependencies better. Convolutional Conditional Neural Processes (Gordon et al. (2019b)) adapted convolutional layers to process images and time series data more effectively. Gaussian Neural Processes (Bruinsma et al. (2021)) combined NPs with Gaussian inference, improving performance in regression tasks by leveraging Gaussian uncertainty. (Lee et al. (2023)) refined uncertainty quantification of NP, providing better posterior estimation. Other work extends NPs to meta-learning, such as Meta-Learning Stationary Stochastic Processes (Foong et al. (2020)) and Meta-Learning Probabilistic Inference (Gordon et al. (2019a)), both addressing generalization to unseen tasks. Recent work focuses on expanding NP applications to complex data structures. Group Equivariant Conditional Neural Processes (Kawano et al. (2021))and Versatile Neural Processes (Guo et al. (2023)) improved the handling of symmetries and representations, enhancing NPs' capacity to model equivariant and implicit functions. To improve generation ability furtherly, (Nguyen & Grover (2022)) and (Mohseni & Duffield (2024)) combine NPs with transformer and neural operator respectively. Comprehensive surveys by (Jha et al. (2022)) have organized these developments, exploring their wide applications in uncertainty-aware learning.

B APPENDIX: IMPLEMENTATION DETAILS

B.1 REPRODUCIBILITY

Our code will be released in an open repository once this paper is accepted.

B.2 DETAILS OF MODEL

In Tab.3, we show the hyperparameters of KooNPro include time length T, Embedding length k, layers of the encode ϕ, layers of decoder ϕ^{-1}, layers of the auxiliary network κ, layers of ψ employed in Eq.3. Note that the choice of T and k is based on the ablation study showcased in AppendixD.

Table 3: Hyperparameters of KooNPro

Name	T	k	ϕ	ϕ^{-1}	κ	ψ
ETTh1	30	10	4	4	4	2
ETTh2	20	10	4	4	4	2
ETTm1	20	10	4	4	4	2
ETTm1	20	10	4	4	4	2
Solar	30	15	8	8	8	3
Electricity	30	10	4	4	4	2
Traffic	20	10	4	4	4	2
Taxi	30	15	6	6	6	3
KDD-cup	10	10	4	4	4	2

B.3 BENCHMARK DATASETS

For our experiments, we use `ETTs`, `Solar`, `Electricity`, `Traffic`, `Taxi`, and `KDD-cup` open-source datasets, with their properties listed in Tab.B.3. The dataset can be obtained through the links below.

(i) `ETTs`: https://github.com/zhouhaoyi/ETDataset

(ii) `Solar`: https://www.nrel.gov/grid/solar-power-data.html

(iii) `Electricity`: https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014

(iv) `Traffic`: https://pems.dot.ca.gov

(v) `Taxi`: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

(vi) `KDD-cup`: https://www.kdd.org/kdd2018/kdd-cup

C APPENDIX: EVALUATION METRIC

We consider two metrics: $\mathbf{CRPS_{sum}}$ and $\mathbf{NRMSE_{sum}}$, the first one can describe the predictive distribution, and the second can describe the distance between truth value and prediction mean, more details can be found in Gluonts documentation (Alexandrov et al. (2020)).

$\mathbf{CRPS_{sum}}$: $\mathbf{CRPS}$ is a univariate, strictly proper scoring rule that quantifies the compatibility of a cumulative distribution function F with an observed value $x \in \mathbb{R}$ as:

$$\mathbf{CRPS} = \int_{\mathbb{R}} (\mathbf{F(y)} - \mathbb{I}(\mathbf{x} \leq \mathbf{y}))^{\mathbf{2}} \mathbf{dy}, \tag{20}$$

where $\mathbb{I}(x \leq y)$ denotes the indicator function. The CRPS achieves the minimum value when predictive prediction F same as the data distribution. $\mathbf{CRPS}$ can be extend to $\mathbf{CPRS_{sum}}$ to

evaluate multivariate distribution:

$$\mathbf{CRPS_{sum}} = \mathbb{E}_t[\mathbf{CPRS}(F_{sum}^{-1}, \sum_i x_t^i)], \tag{21}$$

where F_{sum}^{-1} is computed by aggregating samples across dimensions and subsequently sorting them to obtain quantiles. A smaller $\mathbf{CRPS_{sum}}$ indicates more accurate predictions.

$\mathbf{NRMSE_{sum}}$: $\mathbf{NRMSE_{sum}}$ is an adaptation of the Root Mean Squared Error (RMSE) that accounts for the scale of the target values. It is defined as follows:

$$\mathbf{NRMSE_{sum}} = \sqrt{\frac{\text{mean}((\hat{\mathbf{Y}} - \mathbf{Y})^{\mathbf{2}})}{\text{mean}(|\mathbf{Y}|)}}, \tag{22}$$

where $\hat{Y}$ represents the predicted time series, and Y represents the true target time series. $\mathbf{NRMSE_{sum}}$ quantifies the average squared difference between predictions and true values across all dimensions, normalized by the mean absolute magnitude of the target values. A smaller $\mathbf{NRMSE_{sum}}$ indicates more accurate predictions.

D APPENDIX: ABLATION STUDY

We conduct ablation studies on the time length T and delay embedding length k across several datasets. The results indicate that merely extending the learning time segment is insufficient to reveal the temporal dynamics underlying the time series, as discussed in Sec.5.3. The choices for T and k presented in Tab.3 are informed by the results of this ablation study.

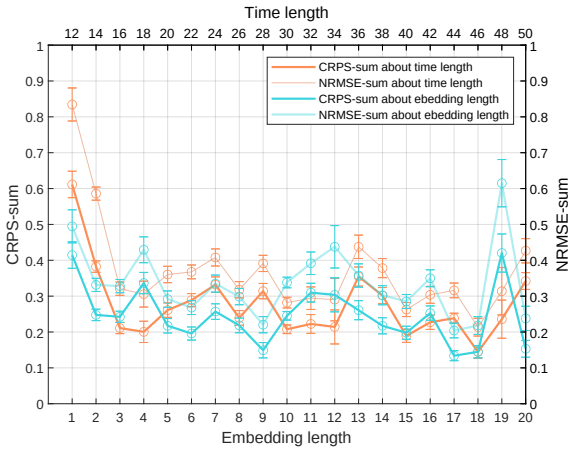

Figure 5: Ablation study on `ETTh2` dataset.

E FORECAST FRACTIONAL STEPS DATA

To showcase the efficacy of KooNPro in fractional steps scenarios, we conducted predictions on fMRI data, with time intervals of sample equal to 0.72 seconds. Figure 12 illustrates KooNPro's outstanding predictive capability in the fMRI case, demonstrating its ability to handle various time resolutions effectively.

F SUBSTITUTION NEURAL PROCESS BY ATTENTION NEURAL PROCESS AND GAUSSIAN PROCESS

Followed by the Attention Neural Process (ANP) (Kim et al., 2019) and combine the scene of our work, we employ the attention mechanism to generate $\boldsymbol{S}_D$ which governs temporal dynamics of the

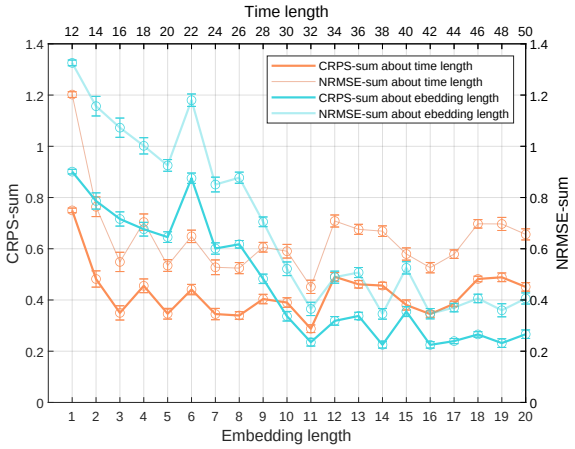

Figure 6: Ablation study on `ETTm1` dataset.

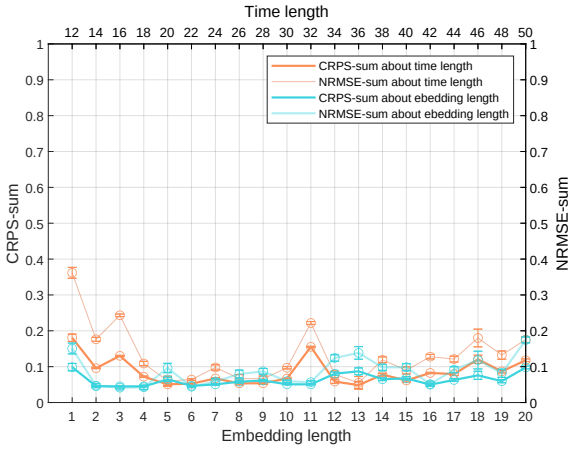

Figure 7: Ablation study on `ETTm2` dataset.

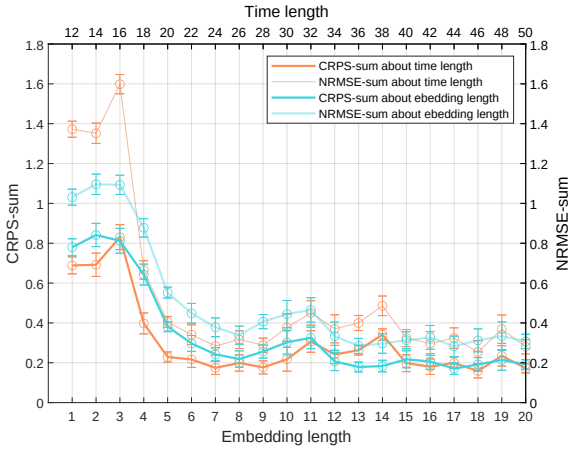

Figure 8: Ablation study on `Solar` dataset.

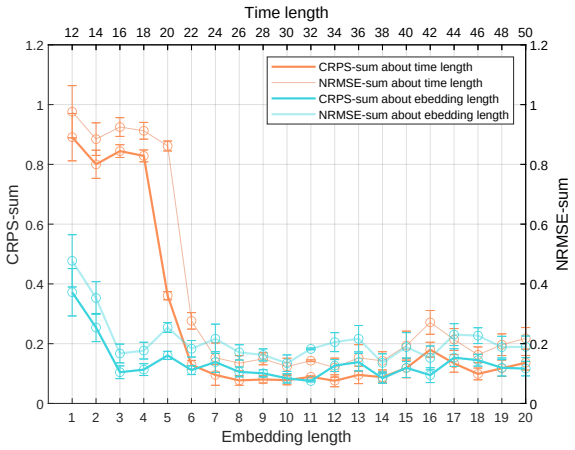

Figure 9: Ablation study on `Electricity` dataset.

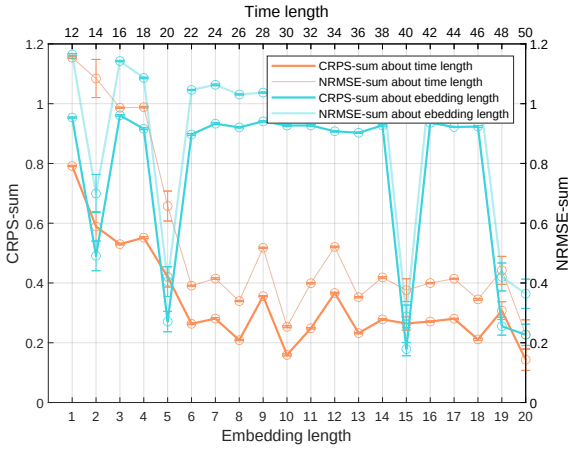

Figure 10: Ablation study on `Taxi` dataset.

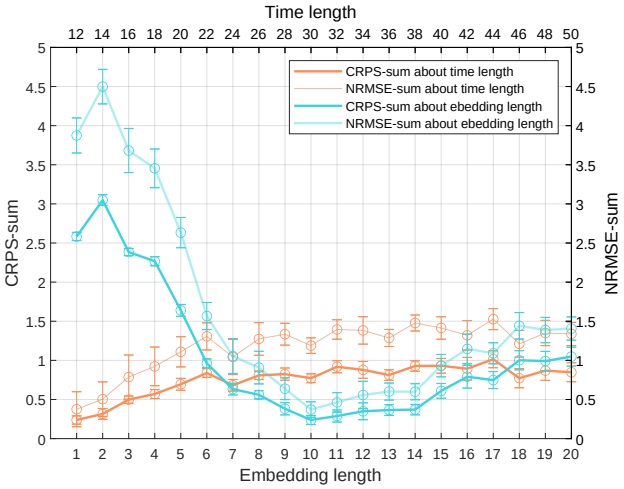

Figure 11: Ablation study on `KDD-cup` dataset.

Table 4: Datasets detail

Name	Frequency	Dimensions	Context length	Prediction length
ETTh1	1 hour	7	10	24
ETTh2	1 hour	7	10	24
ETTm1	15 min	7	10	24
ETTm2	15 min	7	10	24
Solar	1 hour	137	15	24
Electricity	1 hour	370	10	24
Traffic	1 hour	862	10	24
Taxi	30 min	1214	15	24
KDD-cup	1 hour	270	10	48

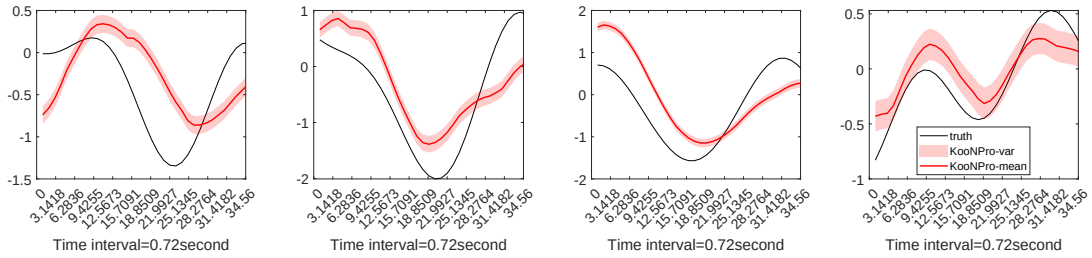

Figure 12: Predict fMRI signals for 48-time points, and the time interval is 0.72 seconds. The unit of the x-axis is second. KooNPro predicts both a mean and a variance of the signal. KooNPro is capable of accurately capturing the wave patterns in fMRI data.

whole time series. To complete the abovementioned process, we set the Key K as $\boldsymbol{x}_C$, the value V as $\boldsymbol{S}_C$ and the query Q as $\boldsymbol{x}_D$, then calculate the $\boldsymbol{S}_D$.

When employing the Gaussian process to predict $\boldsymbol{S}_D$ at the target set $\boldsymbol{x}_D$ using context set $\boldsymbol{x}_C$ and their corresponding outputs $\boldsymbol{S}_C$ by modeling a joint Gaussian distribution. The procedure can be summarized as follows

$$\mathbf{K}_C = \mathrm{k}(z_C, z_C) + \sigma_n^2, \tag{23}$$

$$\mathbf{K}_D = \mathrm{K}(z_D, z_D), \tag{24}$$

$$\mathbf{K}_{CD} = \mathrm{k}(z_C, z_D), \tag{25}$$

where k we choose the radial basis function kernel. Consequently, the distribution of $\boldsymbol{S}_D$ $\mathcal{N}(\mu, \Sigma)$ can be calculated by

$$\mu = K_{CD}^\top K_C^{-1} S_C, \tag{26}$$

$$\Sigma = K_D - K_{CD}^\top K_C^{-1} K_{CD}, \tag{27}$$

where μ is mean and Σ is covariance.

In Tab.5 we compare different approaches for generating $\boldsymbol{S}D$, including cases without it. For clarity, we refer to the method utilizing ANP as **with-ANP** and the method utilizing GP as **with-GP**. The overall performance of with-ANP is comparable to KooNPro, but ANP exhibits higher computational complexity than NP, which computational complexity is raised from O(n + m) to O(n(n + m)) (Kim et al., 2019). The attention mechanism tends to perform well on high-dimensional data. As shown in Table 5, the predictive performance of with-ANP improves as data dimensionality increases. Notably, in the `taxi` dataset, which has the highest dimensionality (1214) among all datasets, with-ANP outperforms KooNPro in both metrics.

With-GP demonstrates significantly degraded performance, particularly on `ETTs`. This phenomenon may be attributed to the critical influence of prior knowledge in defining the kernel k and its corresponding hyperparameters. Furthermore, the presence of non-learnable parameters may impede $\boldsymbol{S}_D$ from adequately capturing the temporal dynamics governing the entire time series.

Table 5: Prediction performance when replacing Neural Process (NP) with Attention Neural Process (ANP), Gaussian Process (GP), or removing NP version (denoted as with-ANP, with-GP, and without-NP, respectively).

Model	Metric	ETTh1	ETTh2	ETTm1	ETTm2	Solar	Elec.	Traffic	Taxi	Cup
KooNPro	C-s	$0.328_{\pm 0.037}$	$0.149_{\pm 0.051}$	$0.165_{\pm 0.057}$	$0.081_{\pm 0.020}$	$0.211_{\pm 0.033}$	$0.057_{\pm 0.006}$	$0.184_{\pm 0.022}$	$0.226_{\pm 0.041}$	$0.204_{\pm 0.017}$
	N-s	$0.520_{\pm 0.045}$	$0.224_{\pm 0.065}$	$0.225_{\pm 0.028}$	$0.122_{\pm 0.034}$	$0.313_{\pm 0.044}$	$0.095_{\pm 0.012}$	$0.289_{\pm 0.025}$	$0.330_{\pm 0.078}$	$0.308_{\pm 0.030}$
with-ANP	C-s	$0.334_{\pm 0.061}$	$0.194_{\pm 0.063}$	$0.195_{\pm 0.029}$	$0.107_{\pm 0.019}$	$0.201_{\pm 0.062}$	$0.067_{\pm 0.022}$	$0.197_{\pm 0.016}$	$0.213_{\pm 0.056}$	$0.306_{\pm 0.008}$
	N-s	$0.541_{\pm 0.084}$	$0.300_{\pm 0.093}$	$0.350_{\pm 0.061}$	$0.182_{\pm 0.033}$	$0.371_{\pm 0.052}$	$0.127_{\pm 0.033}$	$0.308_{\pm 0.030}$	$0.314_{\pm 0.081}$	$0.497_{\pm 0.008}$
with-GP	C-s	$0.912_{\pm 0.067}$	$0.613_{\pm 0.056}$	$0.579_{\pm 0.023}$	$0.438_{\pm 0.067}$	$0.304_{\pm 0.052}$	$0.125_{\pm 0.033}$	$0.231_{\pm 0.023}$	$0.349_{\pm 0.029}$	$0.413_{\pm 0.159}$
	N-s	$1.210_{\pm 0.061}$	$0.791_{\pm 0.058}$	$0.861_{\pm 0.029}$	$0.622_{\pm 0.075}$	$0.467_{\pm 0.067}$	$0.205_{\pm 0.052}$	$0.388_{\pm 0.016}$	$0.501_{\pm 0.033}$	$0.655_{\pm 0.252}$
without-NP	C-s	$0.390_{\pm 0.083}$	$0.218_{\pm 0.120}$	$0.332_{\pm 0.068}$	$0.155_{\pm 0.042}$	$0.341_{\pm 0.059}$	$0.093_{\pm 0.027}$	$0.313_{\pm 0.051}$	$0.252_{\pm 0.068}$	$0.674_{\pm 0.224}$
	N-s	$0.609_{\pm 0.117}$	$0.318_{\pm 0.169}$	$0.511_{\pm 0.103}$	$0.243_{\pm 0.056}$	$0.531_{\pm 0.065}$	$0.144_{\pm 0.036}$	$0.590_{\pm 0.095}$	$0.380_{\pm 0.092}$	$1.058_{\pm 0.341}$

G APPENDIX: LONG-TERM PREDICTION

To evaluate KooNPro's ability to predict by capturing the temporal dynamics embedded in time series, we test the performance of various models under increasing prediction lengths, as detailed in Tab.6. According to the results in Tab.7, KooNPro exhibits the least degradation across both metrics compared to other methods. This outcome highlights the effectiveness of KooNPro in learning and leveraging the temporal dynamics of time series for accurate predictions.

Table 6: Datasets detail

Name	Frequency	Dimensions	Context length	Prediction length
ETTh1	1 hour	7	10	48
ETTh2	1 hour	7	10	48
ETTm1	15 min	7	10	48
ETTm2	15 min	7	10	48
Solar	1 hour	137	15	48
Electricity	1 hour	370	10	48
Traffic	1 hour	862	10	48
Taxi	30 min	1214	15	48
KDD-cup	1 hour	270	10	72

H APPENDIX: ROBUSTNESS OF PERFORMANCE

To evaluate the robustness of KooNPro's predictive performance, we test it under varying Signal-to-Noise Ratio (SNR) conditions (20dB/40dB/60dB). During training, KooNPro is provided with ground truth data. At the testing stage, Gaussian noise is added to the input data, and the predictions are compared against the ground truth. As presented in Tab.8, the performance degradation with decreasing SNR remains within acceptable limits, demonstrating KooNPro's robust predictive capability across varying noise levels.

I APPENDIX: THE NUMBER OF PARAMETERS OF LEARNED MODEL

The backbone of KooNPro is built on MLPs, so the number of parameters is determined by the input data dimensions, as well as the depth and width of the MLP. Table 9 provides the memory consumption of KooNPro's parameters across different datasets.

Table 7: Comparison of $\mathbf{CRPS_{sum}}$ (denoted as C-s, smaller is better) and $\mathbf{NRMSE_{sum}}$ (denoted as N-s, smaller is better) across nine real-world datasets. The means and standard errors are based on 10 independent runs of retraining and evaluation. The best performances are in red and the second are in blue. The block with a '-' denotes a numerical issue encountered during model training with longer prediction lengths.

Model	Metric	ETTh1	ETTh2	ETTm1	ETTm2	Solar	Electricity	Traffic	Taxi	Cup
GP-Copula	C-s	$0.611_{\pm 0.031}$	$0.381_{\pm 0.034}$	$0.646_{\pm 0.056}$	$0.558_{\pm 0.054}$	$0.465_{\pm 0.089}$	$0.234_{\pm 0.047}$	$0.529_{\pm 0.006}$	$1.007_{\pm 0.025}$	$0.731_{\pm 0.025}$
	N-s	$0.909_{\pm 0.033}$	$0.610_{\pm 0.062}$	$0.762_{\pm 0.138}$	$0.844_{\pm 0.093}$	$0.709_{\pm 0.085}$	$0.382_{\pm 0.103}$	$0.713_{\pm 0.014}$	$1.464_{\pm 0.057}$	$1.073_{\pm 0.056}$
Trans-MAF	C-s	$1.271_{\pm 0.051}$	$0.507_{\pm 0.014}$	$1.045_{\pm 0.085}$	$0.279_{\pm 0.007}$	-	$0.198_{\pm 0.069}$	$0.596_{\pm 0.031}$	$0.769_{\pm 0.046}$	$0.410_{\pm 0.073}$
	N-s	$1.571_{\pm 0.144}$	$0.780_{\pm 0.027}$	$1.797_{\pm 0.122}$	$0.485_{\pm 0.127}$	-	$0.397_{\pm 0.127}$	$0.872_{\pm 0.046}$	$0.936_{\pm 0.010}$	$0.515_{\pm 0.083}$
Timegrid	C-s	$0.796_{\pm 0.084}$	$0.477_{\pm 0.007}$	$0.458_{\pm 0.059}$	$0.346_{\pm 0.010}$	$0.886_{\pm 0.036}$	$0.263_{\pm 0.028}$	$0.726_{\pm 0.050}$	$0.791_{\pm 0.021}$	$0.421_{\pm 0.059}$
	N-s	$0.953_{\pm 0.102}$	$0.697_{\pm 0.009}$	$0.588_{\pm 0.104}$	$0.455_{\pm 0.014}$	$1.243_{\pm 0.056}$	$0.423_{\pm 0.077}$	$0.932_{\pm 0.055}$	$0.971_{\pm 0.195}$	$0.522_{\pm 0.098}$
TACTIS	C-s	$0.752_{\pm 0.004}$	$0.401_{\pm 0.001}$	$1.331_{\pm 0.013}$	$0.261_{\pm 0.023}$	$3.786_{\pm 1.708}$	$0.360_{\pm 0.004}$	$0.552_{\pm 0.067}$	$1.368_{\pm 0.014}$	$0.390_{\pm 0.018}$
	N-s	$0.943_{\pm 0.004}$	$0.522_{\pm 0.002}$	$1.853_{\pm 0.030}$	$0.422_{\pm 0.023}$	$5.615_{\pm 2.168}$	$0.498_{\pm 0.002}$	$0.751_{\pm 0.023}$	$1.591_{\pm 0.020}$	$0.505_{\pm 0.022}$
D^3VAE	C-s	$0.916_{\pm 0.036}$	$0.626_{\pm 0.044}$	$0.598_{\pm 0.021}$	$0.737_{\pm 0.064}$	$0.725_{\pm 0.064}$	$0.408_{\pm 0.048}$	$0.704_{\pm 0.069}$	$0.814_{\pm 0.035}$	-
	N-s	$1.265_{\pm 0.078}$	$0.861_{\pm 0.041}$	$0.768_{\pm 0.067}$	$0.951_{\pm 0.031}$	$0.919_{\pm 0.141}$	$0.601_{\pm 0.051}$	$1.126_{\pm 0.139}$	$1.308_{\pm 0.164}$	-
DPK	C-s	$0.891_{\pm 0.027}$	$0.744_{\pm 0.071}$	$0.824_{\pm 0.040}$	$0.519_{\pm 0.092}$	$0.938_{\pm 0.004}$	$0.997_{\pm 0.012}$	$1.131_{\pm 0.002}$	$0.969_{\pm 0.006}$	$0.900_{\pm 0.015}$
	N-s	$1.262_{\pm 0.032}$	$0.998_{\pm 0.138}$	$1.349_{\pm 0.070}$	$0.592_{\pm 0.202}$	$1.301_{\pm 0.004}$	$1.263_{\pm 0.013}$	$1.478_{\pm 0.005}$	$1.213_{\pm 0.005}$	$1.309_{\pm 0.024}$
MG-TSD	C-s	$0.619_{\pm 0.056}$	$0.435_{\pm 0.099}$	$0.371_{\pm 0.085}$	$0.269_{\pm 0.005}$	$1.000_{\pm 0.001}$	$0.174_{\pm 0.027}$	$0.617_{\pm 0.045}$	$0.409_{\pm 0.051}$	$0.590_{\pm 0.091}$
	N-s	$0.967_{\pm 0.071}$	$0.627_{\pm 0.117}$	$0.539_{\pm 0.109}$	$0.318_{\pm 0.050}$	$1.610_{\pm 1.286}$	$0.283_{\pm 0.035}$	$0.882_{\pm 0.053}$	$0.621_{\pm 0.068}$	$0.767_{\pm 0.073}$
KooNPro	C-s	$0.488_{\pm 0.025}$	$0.376_{\pm 0.025}$	$0.365_{\pm 0.018}$	$0.227_{\pm 0.002}$	$0.417_{\pm 0.021}$	$0.165_{\pm 0.017}$	$0.401_{\pm 0.014}$	$0.396_{\pm 0.023}$	$0.327_{\pm 0.014}$
	N-s	$0.746_{\pm 0.044}$	$0.541_{\pm 0.035}$	$0.564_{\pm 0.023}$	$0.343_{\pm 0.034}$	$0.657_{\pm 0.040}$	$0.287_{\pm 0.017}$	$0.715_{\pm 0.030}$	$0.658_{\pm 0.042}$	$0.457_{\pm 0.021}$

Table 8: KooNPro represents the case where the input data is noise-free, while 20dB, 40dB, and 60dB indicate scenarios where Gaussian noise is added to the input, resulting in SNR of 20dB, 40dB, and 60dB, respectively.

Model	Metric	ETTh1	ETTh2	ETTm1	ETTm2	Solar	Elec.	Traffic	Taxi	Cup
KooNPro	C-s	$0.328_{\pm 0.037}$	$0.149_{\pm 0.051}$	$0.165_{\pm 0.057}$	$0.081_{\pm 0.020}$	$0.211_{\pm 0.033}$	$0.057_{\pm 0.006}$	$0.184_{\pm 0.022}$	$0.226_{\pm 0.041}$	$0.204_{\pm 0.017}$
	N-s	$0.520_{\pm 0.045}$	$0.224_{\pm 0.065}$	$0.225_{\pm 0.028}$	$0.122_{\pm 0.034}$	$0.313_{\pm 0.044}$	$0.095_{\pm 0.012}$	$0.289_{\pm 0.025}$	$0.330_{\pm 0.078}$	$0.308_{\pm 0.030}$
60dB	C-s	$0.367_{\pm 0.023}$	$0.195_{\pm 0.029}$	$0.183_{\pm 0.017}$	$0.111_{\pm 0.012}$	$0.229_{\pm 0.032}$	$0.074_{\pm 0.008}$	$0.192_{\pm 0.023}$	$0.248_{\pm 0.017}$	$0.217_{\pm 0.055}$
	N-s	$0.586_{\pm 0.023}$	$0.284_{\pm 0.032}$	$0.380_{\pm 0.021}$	$0.132_{\pm 0.015}$	$0.323_{\pm 0.042}$	$0.118_{\pm 0.014}$	$0.333_{\pm 0.029}$	$0.337_{\pm 0.021}$	$0.339_{\pm 0.105}$
40dB	C-s	$0.385_{\pm 0.028}$	$0.221_{\pm 0.021}$	$0.271_{\pm 0.033}$	$0.152_{\pm 0.001}$	$0.243_{\pm 0.039}$	$0.192_{\pm 0.052}$	$0.213_{\pm 0.019}$	$0.274_{\pm 0.015}$	$0.229_{\pm 0.019}$
	N-s	$0.608_{\pm 0.049}$	$0.317_{\pm 0.024}$	$0.565_{\pm 0.048}$	$0.162_{\pm 0.003}$	$0.359_{\pm 0.038}$	$0.283_{\pm 0.061}$	$0.344_{\pm 0.029}$	$0.380_{\pm 0.012}$	$0.352_{\pm 0.031}$
20dB	C-s	$0.432_{\pm 0.021}$	$0.251_{\pm 0.010}$	$0.311_{\pm 0.018}$	$0.194_{\pm 0.017}$	$0.269_{\pm 0.047}$	$0.287_{\pm 0.058}$	$0.268_{\pm 0.016}$	$0.285_{\pm 0.026}$	$0.249_{\pm 0.020}$
	N-s	$0.687_{\pm 0.034}$	$0.331_{\pm 0.028}$	$0.491_{\pm 0.026}$	$0.245_{\pm 0.019}$	$0.385_{\pm 0.056}$	$0.415_{\pm 0.049}$	$0.369_{\pm 0.024}$	$0.384_{\pm 0.012}$	$0.395_{\pm 0.036}$

Table 9: Comparison of memory usage for different models' parameters across datasets (in MB).

Model	ETTh1	ETTh2	ETTm1	ETTm2	Solar	Elec.	Traffic	Taxi	Cup
Trans-MAF	9.69	9.69	9.69	9.69	12.21	16.73	26.26	33.08	14.79
Timegrid	4.22	4.22	4.22	4.22	25.93	29.32	37.83	45.06	27.82
TACTIS	7.48	7.48	7.48	7.48	7.49	7.49	7.50	7.51	7.49
D^3VAE	58.41	58.41	58.41	58.41	58.72	59.29	60.50	61.36	59.05
DPK	0.17	0.17	0.17	0.17	0.25	0.40	0.70	0.92	0.34
MG-TSD	2.41	2.41	2.41	2.41	3.05	6.33	14.60	21.65	5.87
KooNPro	1.73	1.73	1.73	1.73	10.23	7.18	14.42	17.50	5.60

Fig.13 illustrates the relationship between the number of parameters and the predictive performance of each model on `ETTm2`. The results show that KooNPro achieves the best performance with the fewest parameters.

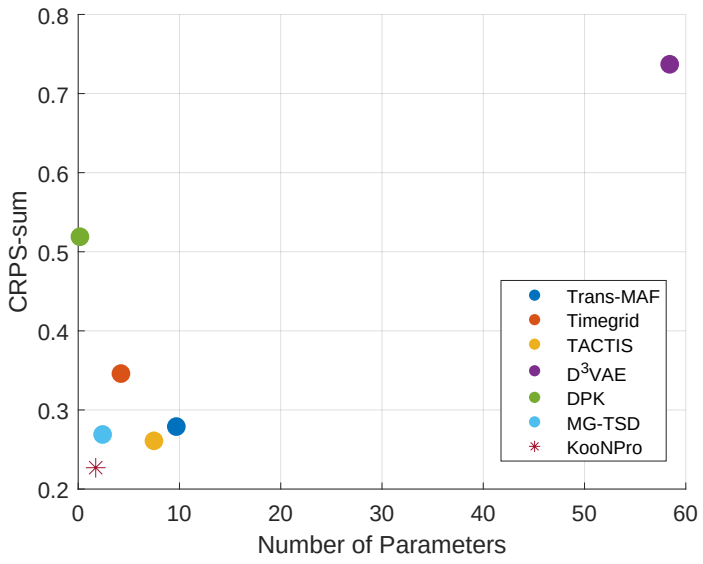

Figure 13: The relationship between the number of parameters and prediction performance of different models on `ETTm2` is shown. The x-axis represents the number of parameters, while the y-axis denotes $\mathbf{CRPS_{sum}}$. Performance improves as you move closer to the bottom-left corner of the plot.

J Compare to DeepAR/MQ-CNN

DeepAR and MQ-CNN are univariate probabilistic prediction methods. Therefore, we predict each dimension separately and report the mean performance across all dimensions in Table 10.

K Details on training baselines

We train baselines by open code which is reported in corresponding papers, and follow the default setting. The code for the baseline methods is obtained from the following sources.

Table 10: Comparison of $\mathbf{CRPS_{sum}}$ (denoted as C-s, smaller is better) and $\mathbf{NRMSE_{sum}}$ (denoted as N-s, smaller is better) across nine real-world datasets. The means and standard errors are based on 10 independent runs of retraining and evaluation. The best performances are in red and the second are in blue.

Model	Metric	ETTh1	ETTh2	ETTm1	ETTm2	Solar	Elec.	Traffic	Taxi	Cup
kooNPro	C-s	$0.328_{\pm 0.037}$	$0.149_{\pm 0.051}$	$0.165_{\pm 0.057}$	$0.081_{\pm 0.020}$	$0.211_{\pm 0.033}$	$0.057_{\pm 0.006}$	$0.184_{\pm 0.022}$	$0.226_{\pm 0.041}$	$0.204_{\pm 0.017}$
	N-s	$0.520_{\pm 0.045}$	$0.224_{\pm 0.065}$	$0.225_{\pm 0.028}$	$0.122_{\pm 0.034}$	$0.313_{\pm 0.044}$	$0.095_{\pm 0.012}$	$0.289_{\pm 0.025}$	$0.330_{\pm 0.078}$	$0.308_{\pm 0.030}$
DeepAR	C-s	$0.431_{\pm 0.010}$	$0.194_{\pm 0.063}$	$0.482_{\pm 0.031}$	$0.611_{\pm 0.073}$	$0.602_{\pm 0.032}$	$0.214_{\pm 0.014}$	$0.289_{\pm 0.018}$	$0.521_{\pm 0.101}$	$0.882_{\pm 0.066}$
	N-s	$0.577_{\pm 0.019}$	$0.300_{\pm 0.093}$	$0.742_{\pm 0.058}$	$0.925_{\pm 0.102}$	$1.138_{\pm 0.042}$	$0.378_{\pm 0.032}$	$0.422_{\pm 0.035}$	$0.914_{\pm 0.151}$	$1.329_{\pm 0.038}$
MQ-CNN	C-s	$0.507_{\pm 0.060}$	$0.377_{\pm 0.024}$	$0.427_{\pm 0.061}$	$0.486_{\pm 0.068}$	$0.755_{\pm 0.023}$	$0.155_{\pm 0.023}$	$0.753_{\pm 0.117}$	$0.673_{\pm 0.069}$	$1.136_{\pm 0.043}$
	N-s	$0.771_{\pm 0.076}$	$0.741_{\pm 0.094}$	$0.956_{\pm 0.062}$	$0.872_{\pm 0.062}$	$1.323_{\pm 0.107}$	$0.214_{\pm 0.028}$	$1.012_{\pm 0.159}$	$1.035_{\pm 0.160}$	$1.658_{\pm 0.207}$

- GP-Copula: https://github.com/mbohlkeschneider/gluon-ts/tree/mvrelease
- Transfomer-MAF:https://github.com/zalandoresearch/pytorch-ts/tree/master/pts/model/transformertempflow
- Timegrad: https://github.com/zalandoresearch/pytorch-ts
- TACTIS: https://github.com/servicenow/tactis
- D^{3}VAE: https://github.com/ramber1836/d3vae
- DPK: https://github.com/AlexTMallen/koopman-forecasting
- MG-TSD: https://github.com/Hundredl/MG-TSD

1.GP-Copula (2019): A method combining Gaussian Processes with Copulas to model complex dependencies between variables in multivariate time series, providing uncertainty estimates and capturing nonlinear correlations.

2. Transformer-MAF (2021) Combines Transformers with Masked Autoregressive Flow (MAF) to model long-term dependencies and the conditional probability distribution of time series data effectively.

3. TimeGrad (2021) A diffusion model-based approach for time series forecasting, which progressively generates samples to capture complex dynamics and uncertainty in the data.

4. TACTiS (2023) A probabilistic autoregressive model leveraging Transformers to handle non-stationary time series, focusing on dynamic structures and probabilistic predictions.

5. D3VAE (2023) A deep variational autoencoder (VAE)-based model designed for time series, featuring a dynamic decoder to effectively capture and predict complex temporal structures.

6. DPK (2024) Dynamic Probabilistic Kernel (DPK) models probabilistic dependencies in time series using a dynamic kernel-based approach, balancing flexibility and efficiency for multivariate data.

7. MG-TSD (2024) Multi-Granularity Time Series Decomposition (MG-TSD) decomposes time series into components of varying frequencies or trends, modeling each with a probabilistic framework to capture multi-scale patterns.

L BACKGROUND

Koopman operator theory provides a linear perspective on nonlinear dynamical systems by focusing on the evolution of observables, which are functions of the system state. For a dynamical system described by $\mathbf{x}_{n+1} = \mathbf{F}(\mathbf{x}_n)$, where $\mathbf{x}$ represents the state vector, the Koopman operator $\mathcal{K}$ acts on an observable g such that $\mathcal{K}g(\mathbf{x}) = g(\mathbf{F}(\mathbf{x}))$. This approach transforms the problem into analyzing the linear operator $\mathcal{K}$ in the space of observables, despite the underlying dynamics being nonlinear. The advantage of this transformation lies in applying linear operator theory, particularly spectral theory, to gain insights into the system's behavior.