CopuAdapt: A Causal Attentional Copula Adapter for Probabilistic Multivariate Time-Series Forecasting

BENJAMIN REDHEAD*, University of Edinburgh, United Kingdom
THOMAS LEE, University of Edinburgh, United Kingdom
CAMERON BARKER, University of Edinburgh, United Kingdom
ELLIOT CROWLEY, University of Edinburgh, United Kingdom
VICTOR ELVIRA, University of Edinburgh, United Kingdom
HENRY GOUK, University of Edinburgh, United Kingdom
AMOS STORKEY, University of Edinburgh, United Kingdom

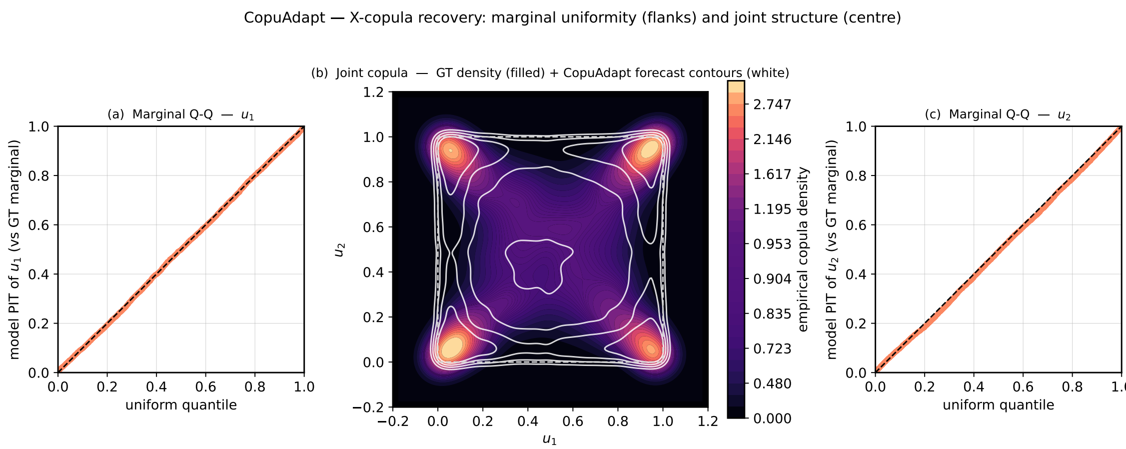

Fig. 1. Our proposed attentional copula adapter successfully recovers highly complex, multi-modal copula structures (centre) while maintaining impeccable marginal calibration (left & right).

Time-series foundation models (TSFMs) have achieved remarkable generalization capabilities across diverse forecasting tasks. However, to scale training efficiency and maximize data availability, the vast majority of state-of-the-art TSFMs are *univariate*, rendering them blind to cross-variate dependencies prevalent in real-world systems. Existing adapters fail to resolve this issue for probabilistic forecasting without introducing severe context length restrictions. To bridge this gap, we introduce *CopuAdapt*, a lightweight

Authors' Contact Information: Benjamin Redhead, s2817214@ed.ac.uk, University of Edinburgh, United Kingdom; Thomas Lee, T.L.Lee-1@sms.ed.ac.uk, University of Edinburgh, United Kingdom; Cameron Barker, C.B.Barker@sms.ed.ac.uk, University of Edinburgh, United Kingdom; Elliot Crowley, elliot.j.crowley@ed.ac.uk, University of Edinburgh, United Kingdom; Victor Elvira, victor.elvira@ed.ac.uk, University of Edinburgh, United Kingdom; Henry Gouk, henry.gouk@ed.ac.uk, University of Edinburgh, United Kingdom; Amos Storkey, a.storkey@ed.ac.uk, University of Edinburgh, United Kingdom.

Manuscript submitted to ACM

distribution-free adapter that lifts independent univariate TSFM samples into a calibrated multivariate joint distribution. CopuAdapt pairs non-parametric empirical marginals with a state-dependent sparse causal directed acyclic graph (DAG) learned over the look-back history. By invoking a causal Markov assumption over the DAG's topological ordering, we transform a dense autoregressive chain into a fully parallelized cross-attention decoder. Evaluated alongside the state-of-the-art flow-matching TSFM Sundial, CopuAdapt achieves an optimal overall rank of 1.44 out of 4 across standard multivariate benchmarks, significantly enhancing the capture of joint path geometry under tight look-back constraints while totalling only 0.1% of the base TSFM parameter size.

CCS Concepts: • **Computing Methodologies → Probabilistic Reasoning**.

Additional Key Words and Phrases: Time-series, Forecasting, Foundation Models, Adapters, Parameter-Efficient Fine-tuning

ACM Reference Format:
Benjamin Redhead, Thomas Lee, Cameron Barker, Elliot Crowley, Victor Elvira, Henry Gouk, and Amos Storkey. 2026. CopuAdapt: A Causal Attentional Copula Adapter for Probabilistic Multivariate Time-Series Forecasting. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Foundation models are gaining prominence due to their ability to learn underlying structures across diverse datasets, exhibiting strong generalisation capabilities. This capability has led to an increase in research on time-series foundation models (TSFMs), trained on a diverse corpus of time-series data to generalise to unseen tasks. However, among these foundation models there is a stark disparity between their design and the real-world data for which they are intended to be used. Most TSFMs, due to ease of construction or to increase the amount of training series available, design a *univariate* architecture and train on each channel of a time-series. This univariate design is often supported by an assumption of channel independence, however that assumption does not hold in many real-world scenarios, as demonstrated in Sec. 5.1. To solve this issue, without necessitating retraining a new TSFM, adapters have been proposed [3, 4, 11]. However, the majority of those proposed are limited to point-forecasting [4, 11], and the only adapter, to our knowledge, that can complete this task for probabilistic forecasters uses latent space mixing with restrictive context length requirements [3]. We propose *CopuAdapt*: a novel adaptor, inspired by attentional copulas [2, 6], that extracts a local causal DAG to enable efficient inference and modelling of the correlations between covariates of the time-series. This method results in significant improvement in the forecasting of path geometries, empirically verified by state-of-the-art signature kernel MMD values, while also improving marginal fidelity in low-dimensional settings with low seasonality.

To summarise our contributions are as follows:

(1) **Causal Copula Adaptation:** We introduce a novel, distribution-free adapter framework that leverages a two-phase curriculum learning strategy to construct a mathematically valid non-parametric attentional copula over frozen TSFM marginals.
(2) **Topologically Batched Cross-Attention:** We integrate state-dependent local causal discovery via sparse encoders to strictly mask the copula decoder. By partitioning the resulting DAG into independent topological levels and utilizing custom cross-attention, we bypass the computational bottlenecks of traditional auto-regressive approaches.

(3) **Empirical Validation:** We conduct a thorough empirical evaluation on highly correlated multivariate datasets, demonstrating that our adapter consistently outperforms alternative adaptation methods and standard channel-independent implementations across joint trajectory and path-distribution metrics.

2 Problem Formulation

To describe CopuAdapt, we first need to formulate the multivariate forecasting problem. In multivariate probabilistic forecasting, a time-series is a $t \in \mathbb{N}$ indexed sequence of values $\boldsymbol{x_t} \in \mathbb{R}^d$, where each dimension of $\boldsymbol{x}_t$ is known as a channel and d is the number of channels. We also denote, $\boldsymbol{x}^{(i)}_{t:t'} \in \mathbb{R}$ as the values of the ith channel for time steps t to t'.

Given the above notation, the problem of multivariate probabilistic forecasting can be stated as learning the joint-conditional distribution $P(\boldsymbol{x}_{t:t+h}|\boldsymbol{x}_{0:t})$. Most time-series foundation models (TSFMs) assume *channel independence* [1, 5, 8, 9]; they learn a univariate distribution $p_i = P(x^{(i)}_{t:t+h}|x^{(i)}_{0:t})$ for each channel i to approximate $P(\boldsymbol{x}_{t:t+h}|\boldsymbol{x}_{0:t})$. Hence, the problem of reconstructing the joint-conditional distribution via an adapter for TSFMs reduces to:

$$P(\boldsymbol{x}_{t:t+h}|\boldsymbol{x}_{0:t}) = \phi(p_1, p_2, ..., p_d). \tag{1}$$

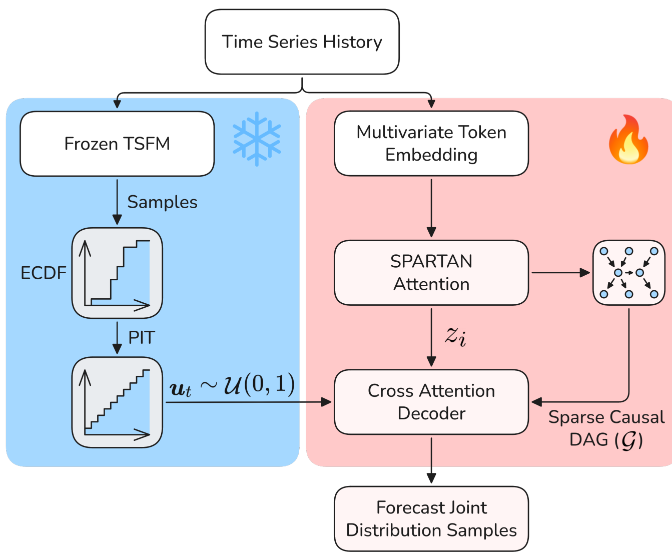

Fig. 2. CopuAdapt architecture showing the frozen TSFM (left) providing univariate marginals, and the learned adapter (right) which learns a non-parametric copula fitted to the multivariate joint-distribution.

3 Methodology

3.1 Inference Approach

In order to produce samples of this joint-conditional distribution using samples of the univariate marginals from a TSFM, we take an auto-regressive modelling approach where for a given time-step $t + j$ the copula output is given by

$$c(\boldsymbol{u}_{t+j}|c(\boldsymbol{u})_{t:t+j-1}) = \prod_{i=1}^{d} f(u^{(i)}_{t+j}|\boldsymbol{u}^{(1,s<i)}_{t:t+j}, \boldsymbol{x}^{(1,s\leq i)}_{0:t}). \tag{2}$$

However, auto-regression over these variates is computationally inefficient and can cause the network to learn spurious correlations which negatively impact performance in non-stationary or noisy environments. Thus we make a causal Markov assumption applied over a topological sort of a learned DAG $\mathcal{G}$. This assumption results in the updated formulation

$$c(\boldsymbol{u}_{t+j}|c(\boldsymbol{u})_{t:t+j-1}) = \prod_{i=1}^{d} f(u_{t+j}^{(i)}|\boldsymbol{u}_{t:t+j}^{(\mathrm{Pa}(i))}, \boldsymbol{x}_{0:t}^{(\mathrm{Pa}(i)\cup\{i\})}). \tag{3}$$

In our proposed adapter architecture $\mathcal{G}$ is learned via SPARTAN attention [7], the copula function ϕ is learned via a cross-attention decoder, and the probabilities $u_t^{(i)}$ are obtained from the ECDF of TSFM output samples:

$$u_t^{(i)} = P_i(\boldsymbol{x}_t^{(i)}|\boldsymbol{x}_{0:t-1}) \approx \frac{1}{K}\sum_{j=1}^{K} \mathbf{1}(x_j^{(i)} \leq x_{\mathrm{true}}^{(i)}) \sim \mathcal{U}(0,1). \tag{4}$$

3.2 Training Objective and Optimization

To learn the multivariate dependency structure without altering the foundational univariate representations, we define a joint optimization objective comprising a continuous-density copula log-likelihood and a structural sparsity regularization term.

For a given ground-truth realization $y_{d,t}$ (indexed by variate d and forecast horizon t), we apply the probability integral transform (PIT) via the empirical cumulative distribution function $\widehat{F}_{d,t}$, which is constructed from K univariate samples generated by the frozen time-series foundation model:

$$u_{d,t} = \widehat{F}_{d,t}(y_{d,t}). \tag{5}$$

Crucially, this marginal transformation operates entirely outside the computational gradient graph. The empirical marginals remain frozen during each forward pass, ensuring that optimization gradients flow exclusively through the copula head and the structural causal encoder.

To flexibly model the conditional copula density, we discretize the unit interval $[0,1]$ into n_{bins} uniform segments. The uniform innovation $u_{d,t}$ is mapped to a discrete bin index $\mathrm{bin}(u) = \min(\lfloor u \cdot n_{\mathrm{bins}} \rfloor, n_{\mathrm{bins}} - 1)$. The network subsequently outputs a categorical distribution $\hat{p}_{d,t}(\cdot \mid \mathrm{Pa}(d))$ over these histogram bins. This distribution is explicitly conditioned on $\mathrm{Pa}(d)$, the localized DAG-parent set dictated by the sparse causal encoder.

Because we are modeling a continuous density rather than a strictly categorical classification, the discrete probability mass must be scaled by the bin width $(1/n_{\mathrm{bins}})$. We incorporate a $+\log n_{\mathrm{bins}}$ correction term to map the categorical mass to a continuous density function. This ensures that numerical loss values remain comparable and strictly invariant to the choice of the n_{bins} hyper-parameter. The resulting negative log-likelihood (NLL) over the spatial dimension D and temporal horizon H is:

$$\mathcal{L}_{\mathrm{NLL}} = -\frac{1}{DH}\sum_{d=1}^{D}\sum_{t=1}^{H}\left[\log \hat{p}_{d,t}(\mathrm{bin}(u_{d,t}) \mid \mathrm{Pa}(d)) + \log n_{\mathrm{bins}}\right]. \tag{6}$$

To enforce the discovery of a sparse and interpretable causal structure, we apply an L_1-style norm penalty to the off-diagonal entries of the soft adjacency matrix $\boldsymbol{M}$, which is emitted dynamically by the SPARTAN encoder over the context window:

$$\mathcal{L}_{\mathrm{sp}} = \sum_{i \neq j} \|M_{ij}\|. \tag{7}$$

The complete, end-to-end training objective is then defined as the linear combination of the density-calibrated copula likelihood and the topological sparsity constraint, balanced by a regularization coefficient λ:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{NLL}} + \lambda \mathcal{L}_{\text{sp}}. \tag{8}$$

4 Architecture

To implement this adapter we utilise a two-stage curriculum learning approach. This works via a frozen TSFM producing samples to form our empirical marginal CDFs at each time-step which our adapter then uses to learn a non-parametric copula better matching the joint distribution. To do this we generate an ECDF following Eq. 4, then use the probability integral transform to map this to a uniform distribution. The resulting uniform values are used as inputs to a cross-attention decoder. The DAG required for inference using this decoder is generated via embedding the forecast context into one token per variate, then using these variate tokens as inputs to the SPARTAN attention mechanism to extract local causal structure, represented as a DAG, and a representation z_i. This context and representation are then used in the cross-attention to get keys and values. This decoder is then run with context and representation inputs according to the inference approach described in Sec. 3.1. Our architecture totals $200,000$ parameters 0.1% of the parameters of the base Sundial TSFM used in our evaluations. A diagram of the architecture is shown in Figure 2.

5 Experiments

5.1 Dataset selection

Let $\boldsymbol{x} \in \mathbb{R}^{T\times d}$ denote the realization of a given dataset on its training split. We characterize the cross-sectional dependency within each dataset by computing the contemporaneous (lag-0) Pearson cross-correlation matrix $\boldsymbol{R} \in [-1, 1]^{d\times d}$, defined element-wise by:

$$R_{ij} = \frac{\frac{1}{T-1} \sum_{t=1}^{T} (x_t^{(i)} - \mu_i)(x_t^{(j)} - \mu_j)}{\sigma_i \, \sigma_j}, \tag{9}$$

where $\mu_i = \frac{1}{T} \sum_{t=1}^{T} x_t^{(i)}$ and $\sigma_i^2 = \frac{1}{T-1} \sum_{t=1}^{T} (x_t^{(i)} - \mu_i)^2$ represent the empirical mean and variance of channel i, respectively. In compact matrix form, this matches $\boldsymbol{R} = \boldsymbol{D}_\sigma^{-1} \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} \, \boldsymbol{D}_\sigma^{-1} / (T-1)$, where $\tilde{\boldsymbol{X}}$ is the column-centred data matrix and $\boldsymbol{D}_\sigma$ is the diagonal matrix of variate standard deviations.

The statistic R_{ij} isolates the strength of linear lag-0 correlations between channels. While it intentionally omits lagged cross-correlations and non-linear dependencies, this restricted definition aligns precisely with the structure that our attentional copula decoder must reconstruct at any given forward forecast time-step t.

Inclusion Criterion. A dataset is included in our core benchmark if and only if it satisfies a strict two-part dependency threshold:

$$\bar{\rho} > 0.5 \quad \text{and} \quad \rho_{50} > 0.5.$$

The median constraint ($\rho_{50} > 0.5$) guarantees that a strict majority of the variate pairs exhibit robust co-movement, ensuring that the global summary statistic is not artificially inflated by a small, highly correlated cluster within an otherwise disconnected dataset.

Screening Results. Table 1 reports the dependency summary computed across standard time-series benchmarking datasets. The clear mathematical break point validates our selection: *illness*, *traffic*, and *exchange* comfortably clear both thresholds, forming our core high-dependence evaluation suite. Conversely, *electricity*, the *ETT* variants, and *weather* fail the criteria and are excluded from the main track.

Table 1. Cross-variate dependence summary across standard time-series datasets. D indicates the number of variates, and T is the temporal length of the training split. $\bar{\rho}$ is the mean absolute off-diagonal correlation (Eq. (9)), while $\rho_{50}, \rho_{90}, \rho_{99}$ are the corresponding empirical quantiles. datasets above the rule satisfy both inclusion conditions ($\bar{\rho} > 0.5, \rho_{50} > 0.5$) and comprise the main benchmark; datasets below are excluded.

Dataset	D	T	$\bar{\rho}$	ρ_{50}	ρ_{90}	ρ_{99}
illness	7	966	**0.708**	**0.805**	0.963	0.988
traffic	862	17,544	**0.564**	**0.590**	0.783	0.880
exchange	8	7,588	**0.513**	**0.554**	0.890	0.914
electricity	321	26,304	0.489	0.463	0.863	0.917
ETTm2	7	69,680	0.324	0.200	0.669	0.914
ETTh2	7	11,520	0.288	0.174	0.612	0.895
weather	21	52,696	0.184	0.112	0.498	0.762

5.2 Results

5.3 Main Forecasting Results

We evaluate the performance of `CopuAdapt` against three representative baselines: the baseline channel-independent application of the underlying TSFM, an ICA-based linear recombination framework (*ICA-Sundial*), and the latent-space mixing adapter *AdaPTS* [3]. All methods utilize the state-of-the-art transformer-based flow-matching model *Sundial* [9] as the frozen core. Crucially, while *Sundial* generates high-fidelity marginal posteriors, its inherent channel-independent assumption prevents it from natively capturing cross-sectional interactions. Joint-forecast profiles are tracked using three orthogonal probabilistic criteria: Signature MMD (*Sig-MMD*) to measure global path trajectory geometry [10], normalized CRPS (*nCRPS*) to measure pointwise calibration, and RBF MMD to evaluate spatial joint marginal alignment. Table 2 reports these metrics aggregated across five forecast horizons per dataset split.

Low-Dimensional Dominance. On the low-dimensional datasets, *Exchange* ($D = 8$) and *Illness* ($D = 7$), `CopuAdapt` completely dominates the benchmark, achieving the best (lowest) values across every tracked metric. For instance, on *Exchange*, our adapter delivers a clean sweep across path geometry and pointwise tracking, culminating in an Average Rank of **1.33**. On *Illness*, the separation is even more pronounced: `CopuAdapt` provides a substantial lift in univariate calibration, achieving an nCRPS of **0.13866** compared to 0.15948 for the unadapted channel-independent core, while nearly halving the spatial distribution error. The linear mixing approach of *ICA-Sundial* collapses severely in this regime due to unconstrained inverse scale transformations, proving the necessity of modelling localized copula structures over rigid linear combinations.

High-Dimensional Paths vs. Marginal Calibration. In the high-dimensional regime represented by the *Traffic* dataset ($D = 862$), the empirical results expose a compelling structural trade-off. `CopuAdapt` achieves the top value for path geometry, outperforming channel independence. This verifies that our dynamic DAG discovery mechanism effectively reintroduces realistic cross-sectional co-movements over the forecast horizon that channel independence fails to capture.

However, on this specific dataset, the unadapted *Channel Independent* baseline retains the best scores for nCRPS and RBF-MMD. This behaviour validates our architectural boundary conditions: *Traffic* exhibits intense temporal auto-correlation, which the base TSFM marginal core models with high precision. Because our framework invokes a within-variate conditional independence relaxation for more efficient inference, it accepts a minor inflation in pointwise calibration metrics. In exchange, it delivers a highly accurate joint path distribution and more efficient inference.

Global Performance Summary. Aggregated across all datasets and forecast horizons, `CopuAdapt` establishes a definitive state-of-the-art standard, securing the top overall Mean Rank of **1.44**. In comparison, *AdaPTS* struggles under short-context regimes, posting a degraded mean rank of 3.78 and failing entirely to compile under our default look-back budget ($L = 96$). Our framework proves highly robust, operating efficiently under tight look-back constraints to inject structured dependencies into isolated foundation model outputs without requiring any fine-tuning of the primary foundation model weights.

Table 2. Joint-forecast metrics on the three high-dependence datasets, aggregated by arithmetic mean across five forecast horizons per dataset ($h \in \{32, 64, 96, 192, 336\}$). nCRPS is the normalised CRPS (CRPS / mean$|y|$ computed on the same test windows). The *Avg. Rank* column is the mean per-horizon metric rank. Best (lowest) value per (dataset, metric) cell in bold. AdaPTS is reported at $L = 512$ because its preprocessing requires $L \geq 512$; all other methods use $L = 96$.

Method	**Dataset**	Sig-MMD	nCRPS	RBF-MMD	Avg. Rank
Ch. Indep.	Exchange	0.01675	0.02567	0.02007	2.80
	Illness	0.39992	0.15948	0.15792	1.80
	Traffic	1.12266	**0.20326**	**0.02508**	**1.60**
	Mean	—	—	—	2.07
ICA-Sundial	Exchange	0.01768	0.02448	0.01556	1.87
	Illness	0.57168	0.51477	0.95164	3.67
	Traffic	1.08780	0.41047	0.04916	2.60
	Mean	—	—	—	2.71
CopuAdapt	Exchange	**0.01431**	**0.02416**	**0.01553**	**1.33**
	Illness	**0.39210**	**0.13866**	**0.08458**	**1.20**
	Traffic	**1.02174**	0.27538	0.04176	1.80
	Mean	—	—	—	**1.44**
AdaPTS†	Exchange	0.05175	0.04945	0.16835	4.00
	Illness	1.04784	0.48281	0.65485	3.33
	Traffic	1.74731	0.60923	0.25167	4.00
	Mean	—	—	—	3.78

†AdaPTS uses a 512-step lookback; the other three methods use a 96-step lookback. The AdaPTS pipeline fails at $L = 96$ because its internal scaler receives an empty window, so a like-for-like short-context comparison is not available for this baseline.

6 Conclusion and Future Work

In this work, we addressed a critical structural mismatch in modern time-series forecasting: the reliance of large-scale foundation models on the assumption of channel independence when applied to highly correlated multivariate systems. Rather than undertaking the computationally prohibitive task of retraining these models from scratch, we presented `CopuAdapt`, a lightweight, distribution-free adapter framework that constructs a mathematically valid non-parametric attentional copula over frozen univariate marginals. By replacing temporal autoregression with a sparse causal DAG, our architecture maps generation over parallelized topological levels.

Our empirical validation alongside the state-of-the-art flow-matching TSFM *Sundial* confirms that `CopuAdapt` establishes a superior joint trajectory standard cross high-dependence real-world applications under highly restricted look-back constraints ($L = 96$). Furthermore, our dependency screening protocol provides a clear boundary for when cross-sectional copula adjustment provides an improvement over channel independence.

While our conditional independence relaxation improves inference efficiency, our experiments on the *Traffic* dataset reveal a trade-off, where omitting temporal auto-correlation within the copula space can lead to a slight degradation in pointwise calibration. To expand upon this framework, future work will explore integrating state-space models (SSMs) or lightweight causal convolutions, directly into the parallelized cross-attention decoder to efficiently reintroduce within-series temporal autoregression. Additionally, we intend to study the integration of this attentional copula mechanism natively *inside* the pre-training loops of future TSFM architectures, embedding interpretable, rigorous cross-variate modelling into the architecture.

References

[1] Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., and Wang, Y. Chronos: Learning the Language of Time Series, Nov. 2024. arXiv:2403.07815 [cs.LG].

[2] Ashok, A., Marcotte, , Zantedeschi, V., Chapados, N., and Drouin, A. TACTiS-2: Better, Faster, Simpler Attentional Copulas for Multivariate Time Series.

[3] Benechehab, A., Feofanov, V., Paolo, G., Thomas, A., Filippone, M., and Kégl, B. AdaPTS: Adapting Univariate Foundation Models to Probabilistic Multivariate Time Series Forecasting.

[4] Cheng, H., Wu, X., Shu, Y., Rao, Z., Pan, L., Yang, B., and Guo, C. CoRA: Boosting Time Series Foundation Models for Multivariate Forecasting through Correlation-aware Adapter.

[5] Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting, Apr. 2024. arXiv:2310.10688 [cs.CL].

[6] Drouin, A., Marcotte, , and Chapados, N. TACTiS: Transformer-Attentional Copulas for Time Series. In *Proceedings of the 39th International Conference on Machine Learning* (June 2022), PMLR, pp. 5447–5493.

[7] Lei, A., Schölkopf, B., and Posner, I. SPARTAN: A Sparse Transformer World Model Attending to What Matters.

[8] Liu, C., Aksu, T., Liu, J., Liu, X., Yan, H., Pham, Q., Savarese, S., Sahoo, D., Xiong, C., and Li, J. Moirai 2.0: When Less Is More for Time Series Forecasting, Nov. 2025.

[9] Liu, Y., Qin, G., Shi, Z., Chen, Z., Yang, C., Huang, X., Wang, J., and Long, M. Sundial: A Family of Highly Capable Time Series Foundation Models.

[10] Redhead, B. R., Lee, T. L., Gu, P., Elvira, V., and Storkey, A. Signature-Kernel Based Evaluation Metrics for Robust Probabilistic and Tail-Event Forecasting, Feb. 2026. arXiv:2602.10182 [cs.LG].

[11] Zetlin, Z., Moharreri, K., and Safi, M. TopoPrimer: The Missing Topological Context in Forecasting Models, May 2026. arXiv:2605.15035 [cs.LG].

Received 31 May 2026