MARCO: PROPAGATING COVARIATE UNCERTAINTY FOR TIME SERIES FORECASTING WITH EXOGENOUS VARIABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Exogenous variables (i.e., covariates) often govern how a target time series evolves: electricity prices follow load and renewable generation, and wind power follows wind speed. Existing methods either forecast deterministically, encoding historical covariates as features or forecasting the covariates first and using that single trajectory, or forecast the target probabilistically while treating the covariates as fixed context. Neither models the future covariates, which are unknown when a forecast is issued, probabilistically, nor captures how their individual uncertainties combine into a joint distribution over covariate futures to which the target should respond. We propose MarCO, which treats the future covariates as a random vector: it forecasts their marginal distributions, couples the marginals into a joint distribution through a copula, draws coherent covariate scenarios from it, and lets every scenario modulate the target forecast, which is obtained as the mixture over scenarios. Using historical inputs only, MarCO ranks first in MSE on 11 of 12 real-world datasets against state-of-the-art deterministic and probabilistic methods and reduces their average MSE by 5.2% and 5.9%, while additionally quantifying the uncertainty of the covariates themselves.

1 INTRODUCTION

Time series forecasting underpins decisions in energy systems (Lago et al., 2021; Wang et al., 2024), traffic (Li et al., 2018; Wu et al., 2019), finance (Sezer et al., 2020), weather (Bi et al., 2023), public health (Cramer et al., 2022), and retail demand planning (Makridakis et al., 2022; Salinas et al., 2020). Most deep forecasters (Lim & Zohren, 2021; Nie et al., 2023; Liu et al., 2024) rely only on the historical values of the target series (i.e., endogenous variables), yet in many applications the target is accompanied by covariate series (i.e., exogenous variables) that carry information its own history does not: day-ahead electricity prices are largely determined by the load and the renewable generation of the same hours (Lago et al., 2021; Nowotarski & Weron, 2018), and the output of a wind farm is driven by wind speed rather than by its own past (Zhou et al., 2022; Pinson et al., 2009). Although some covariates, such as calendar features and scheduled quantities, are known over the forecast horizon (Lim et al., 2021), most are physical measurements that are observed only up to the forecast origin, so their future is as unknown as that of the target (Lago et al., 2021; Qiu et al., 2026). We focus on this setting, in which only the history of the target and of the covariates is available at forecast time.

Existing methods handle this setting in two ways (Figure 1). Most of them forecast the target directly from the covariate history and do not model the covariate future (Wang et al., 2024; Zhou et al., 2025; Chen et al., 2026; Arango et al., 2025) (Figure 1a). The covariates thus act only as additional historical features, and how they will evolve over the horizon, which often drives the target, is left unexploited. An emerging paradigm remedies this by first forecasting the covariate future and then using the forecast to assist target forecasting (Qiu et al., 2026; Li et al., 2026) (Figure 1b). This recovers the missing information, but only as a single predicted trajectory, whereas an unknown covariate future is a range of plausible outcomes; reducing it to one path before it reaches the target discards the alternatives, and with them the uncertainty that the target forecast should inherit. This observation motivates us to propose a new paradigm (Figure 1c): model the covariate future probabilistically, let the target respond to each of its outcomes, and average the target forecast over them,

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

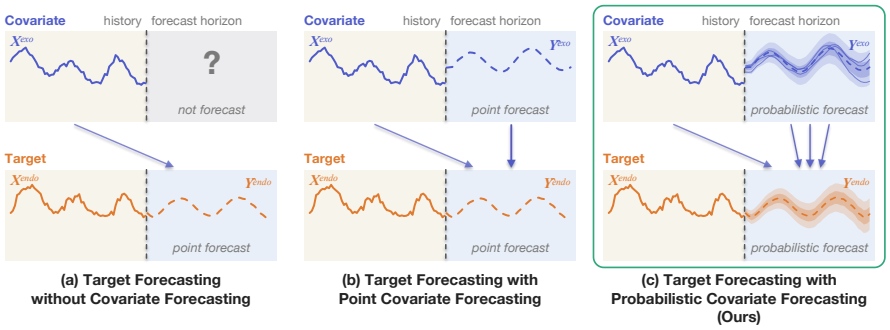

Figure 1: Three paradigms for forecasting with exogenous variables when only the history is available. $\mathbf{X}^{\text{endo}}/\mathbf{X}^{\text{exo}}$ denote the historical endogenous/exogenous variables and $\mathbf{Y}^{\text{endo}}/\mathbf{Y}^{\text{exo}}$ their future counterparts. Only (c), used by MarCO, propagates the uncertainty of the covariate future to the target.

so that the uncertainty of the covariates is quantified explicitly and propagated to the target at the same time.

Realizing this paradigm requires a distribution over the entire covariate future, i.e., every covariate at every horizon step. Its marginals are easy to obtain: the distribution of a single covariate at a single horizon step can be estimated with quantile forecasting (Koenker & Bassett, 1978; Wen et al., 2017). The difficulty lies in **how to model the joint distribution of the covariate future across covariates and horizon steps**. Future covariate values depend strongly on one another: the wind speed of one hour largely persists into the next (Zhou et al., 2022), and the loads of neighbouring regions rise and fall together (Lago et al., 2021). Marginals carry no information about such dependence, so sampling them independently yields incoherent scenarios, whereas learning a full joint density over all covariates at all horizon steps from limited data is difficult and forfeits the reliability of the marginals.

To close this gap, we propose MarCO, which couples **Mar**ginal quantile forecasts of the covariates through a **Co**pula for time series forecasting with exogenous variables. Following the paradigm in Figure 1c, MarCO models the covariate future probabilistically and forecasts the target as a mixture over its outcomes, using historical inputs only. To address the challenge above, it builds on Sklar's theorem (Sklar, 1959), by which any joint distribution decomposes into its marginals and a copula that encodes only the dependence: MarCO estimates the marginals of the covariate future with reliable one-dimensional quantile forecasting and learns the dependence across covariates and horizon steps separately with a low-rank Gaussian copula, without fitting a high-dimensional density. Coherent covariate scenarios drawn from the resulting joint distribution then shift and rescale a base target distribution through an identity-initialized modulator, and all components are trained jointly. Our contributions are summarized as follows:

- We propose MarCO, an end-to-end framework for time series forecasting with exogenous variables that models the covariate future probabilistically, through quantile marginals coupled by a low-rank Gaussian copula, and propagates its uncertainty into the target forecast as a mixture over coherent covariate scenarios, using historical inputs only.
- We introduce an identity-initialized location–scale modulation of the target quantile function that lets the marginals, the copula, and the target forecaster be trained jointly and stably.
- We conduct extensive experiments on 12 real-world datasets. Under the same historical-input setting, MarCO ranks first in MSE on 11 of 12 datasets and reduces the average MSE of the strongest deterministic and probabilistic methods by 5.2% and 5.9%, while additionally reporting predictive distributions for the covariates; ablations confirm that each component of the scenario mixture contributes. Code is available at `https://anonymous.4open.science/r/marco-EAA7`.

2 Related Work

2.1 Time Series Forecasting with Exogenous Variables

Using external information to explain a target series has a long history in statistics, where regression with autocorrelated errors conditions a linear forecast on known regressors (Hyndman & Athanasopoulos, 2018). Deep forecasters made this conditioning learnable and, when only the history of the covariates is observed, fall into the two paradigms of Figure 1. Direct methods forecast the target from the covariate history: TFT (Lim et al., 2021) gates and attends over static, historical, and known-future inputs; NBEATSx (Olivares et al., 2023) dedicates basis-expansion blocks to exogenous inputs; TiDE (Das et al., 2023) concatenates covariates with the target history in a dense encoder; and TimeXer (Wang et al., 2024), CrossLinear (Zhou et al., 2025), and ExoTST (Tayal et al., 2024) embed the covariates separately and let them interact with the target through attention. Covariate-forecasting methods forecast the covariate future first: DAG (Qiu et al., 2026) injects the temporal and channel correlations learned in doing so into the target forecaster, and GCGNet (Li et al., 2026) aligns the two forecasting problems through a shared graph structure inside a variational generator. As discussed in Section 1, the former leave the evolution of the covariates over the horizon unexploited and the latter reduce it to a single trajectory; MarCO instead models the covariate future as a distribution and forecasts the target as a mixture over its outcomes.

2.2 Probabilistic Modeling in Time Series

Probabilistic methods describe a future value through a distribution rather than a point estimate. Most of them model marginal distributions: DeepAR (Salinas et al., 2020) and deep state-space models (Rangapuram et al., 2018) parameterize a univariate likelihood for each series, and MQ-RNN (Wen et al., 2017) regresses a set of quantiles for each series and horizon step directly. Such marginals are reliable and cheap to learn, but they do not describe how the variables vary together. Joint dependence has been addressed in two ways. Copula-based methods separate it from the marginals: GP-Copula (Salinas et al., 2019) couples univariate marginals through a low-rank Gaussian copula, and TACTiS (Drouin et al., 2022; Ashok et al., 2024) learns an attentional copula over all series. Generative methods capture it implicitly, including normalizing flows (Rasul et al., 2021b), diffusion (Rasul et al., 2021a; Tashiro et al., 2021; Kollovieh et al., 2023), and flow matching (Kollovieh et al., 2025). In all of these methods, however, the joint distribution being modeled is that of the *target* series, and the covariates act as fixed context whose future is either assumed known or left unmodeled: most methods append them to the conditioning network (Salinas et al., 2020; Tashiro et al., 2021), and KITE (Cheng et al., 2026), the first to design covariate conditioning for a generative forecaster, modulates attention with statistical priors over the covariates without modeling their future. Outside deep learning, copulas have long been used to turn marginal wind-power forecasts into coherent trajectories (Pinson et al., 2009; Patton, 2012), but as a post-processing step decoupled from any downstream forecaster. MarCO brings the marginal–copula decomposition to the *covariate* future: it learns the marginals and the copula of the covariates jointly with the target forecaster, so that their joint distribution, and not only their marginals, is propagated to the target.

3 Preliminaries

3.1 Copulas and Sklar's Theorem

As motivated in Section 1, MarCO represents the covariate future through its marginals and a copula. Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_M)$ be a random vector with joint distribution function $H(\mathbf{t}) = P(\xi_1 \leq t_1, \ldots, \xi_M \leq t_M)$ and continuous marginal distribution functions $F_1, \ldots, F_M$. Sklar's theorem (Sklar, 1959) states that there exists a unique function $C : [0,1]^M \to [0,1]$, the *copula*, such that:

$$H(\mathbf{t}) = C\big(F_1(t_1), \ldots, F_M(t_M)\big), \tag{1}$$

and, conversely, that composing any copula with any set of marginals yields a valid joint distribution. Since the probability integral transforms $u_m = F_m(\xi_m)$ are uniform on $[0,1]$, the copula is their joint distribution and carries the dependence among the coordinates but nothing about their individual scales or shapes. Equation 1 thus separates the estimation of a joint distribution into two problems that can be solved with different tools: one-dimensional marginals and

a scale-free dependence structure. The *Gaussian copula* with correlation matrix $\mathbf{R}$ is $C_{\mathbf{R}}(\mathbf{u}) = \Phi_{\mathbf{R}}\big(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_M)\big)$, where Φ is the standard normal distribution function and $\Phi_{\mathbf{R}}$ that of $\mathcal{N}(\mathbf{0}, \mathbf{R})$, and its density is:

$$c_{\mathbf{R}}(\mathbf{u}) \;=\; |\mathbf{R}|^{-1/2} \exp\Big(-\tfrac{1}{2}\, \mathbf{z}^{\top}(\mathbf{R}^{-1} - \mathbf{I})\, \mathbf{z} \Big), \qquad \mathbf{z} = \Phi^{-1}(\mathbf{u}). \tag{2}$$

3.2 Problem Statement

We address the task of time series forecasting with exogenous variables when the future covariates are unavailable. Given the history of N endogenous variables $\mathbf{X}^{\text{endo}} \in \mathbb{R}^{N \times T}$ and of D exogenous covariates $\mathbf{X}^{\text{exo}} \in \mathbb{R}^{D \times T}$ over T time steps, our goal is to forecast the future endogenous values $\mathbf{Y}^{\text{endo}} \in \mathbb{R}^{N \times F}$ over a horizon of length F. Because the target depends on a covariate future that is itself unknown, we model it through a distribution $p_\theta(\mathbf{Y}^{\text{endo}} \mid \mathbf{X}^{\text{endo}}, \mathbf{X}^{\text{exo}})$ rather than a single estimate, and read off the point forecast as its median. Unlike settings that assume the future covariates to be known (Qiu et al., 2026; Cheng et al., 2026), the future covariates $\mathbf{Y}^{\text{exo}} \in \mathbb{R}^{D \times F}$ are not an input: they are unobserved when the forecast is issued and serve only as training labels. We therefore treat $\mathbf{Y}^{\text{exo}}$ as a latent random vector, additionally estimate its predictive distribution $p_\theta(\mathbf{Y}^{\text{exo}} \mid \mathbf{X}^{\text{endo}}, \mathbf{X}^{\text{exo}})$, and write the target distribution as an integral over it:

$$p_\theta(\mathbf{Y}^{\text{endo}} \mid \mathbf{h}) \;=\; \int p_\theta(\mathbf{Y}^{\text{endo}} \mid \mathbf{Y}^{\text{exo}}, \mathbf{h})\, p_\theta(\mathbf{Y}^{\text{exo}} \mid \mathbf{h})\, \mathrm{d}\mathbf{Y}^{\text{exo}}, \tag{3}$$

where $\mathbf{h} = (\mathbf{X}^{\text{endo}}, \mathbf{X}^{\text{exo}})$ abbreviates the historical inputs. Direct forecasting (Figure 1a) learns the left-hand side and leaves $\mathbf{Y}^{\text{exo}}$ implicit; point covariate forecasting (Figure 1b) replaces $p_\theta(\mathbf{Y}^{\text{exo}} \mid \mathbf{h})$ with a point mass, discarding the spread of the covariate future. MarCO estimates both factors and evaluates the integral explicitly. Throughout, $\mathcal{T} = \{\tau_1 < \cdots < \tau_Q\}$ denotes a fixed set of quantile levels and $Q(\tau)$ the quantile function of a one-dimensional distribution.

4 Methodology

4.1 Structure Overview

Figure 2 illustrates the overall architecture of MarCO, which evaluates the integral in Equation 3 by following the decomposition of Equation 1: it forecasts a one-dimensional marginal for every future value, couples the covariate marginals with a copula, and forecasts the target as a mixture over scenarios drawn from the resulting joint distribution. The architecture consists of a marginal quantile forecaster followed by the Copula Scenario Mixture (CSM) module, which comprises two stages. Specifically, we first perform *Marginal Quantile Forecasting*. The covariate history $\mathbf{X}^{\text{exo}}$ and the target history $\mathbf{X}^{\text{endo}}$ are normalized per variable and embedded as one token per variable, together with calendar tokens, so that covariates and targets become objects of the same kind. To capture the dependence between variables, a channel-wise Transformer applies attention across these variate tokens, which yields one token $\mathbf{h}_d^{\text{exo}}$ per covariate and $\mathbf{h}_n^{\text{endo}}$ per target variable in which the covariate and target histories have already informed each other. To obtain a preliminary distribution for every future value, a single quantile head, shared by all covariate and target tokens, then maps each token to a quantile function: on the covariate tokens it gives the marginals $\hat{Q}^{\text{exo}}$ of the future covariates, and on the target tokens it gives $\hat{Q}^{\text{endo}}$, the distribution of the target before any covariate future is taken into account (Section 4.2). Next, since the marginals describe each future covariate value in isolation, *Joint Exogenous Sampling* couples the $D \cdot F$ covariate marginals through a low-rank Gaussian copula and draws S covariate scenarios $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}$ by inverse quantile sampling, so that every scenario is coherent across covariates and horizon steps (Section 4.3). Then, to let the target respond to each covariate future, *Scenario-Conditioned Target Modulation* shifts and rescales $\hat{Q}^{\text{endo}}$ once per scenario, conditioned on the tokens and on $\mathbf{x}^{(s)}$, into a scenario-conditioned quantile function $\hat{Q}^{(s)}$; the target forecast $\hat{p}(\mathbf{Y}^{\text{endo}} \mid \mathbf{h})$ is the equal-weight mixture of the S scenario-conditioned distributions, and its median is the point forecast $\hat{\mathbf{Y}}^{\text{endo}}$ (Section 4.4). Finally, we combine the quantile losses of the covariate and target marginals ($\mathcal{L}_{\text{exo}}$, $\mathcal{L}_{\text{endo}}$), the likelihood of the copula ($\mathcal{L}_{\text{cop}}$), and the CRPS of the mixture ($\mathcal{L}_{\text{crps}}$) as the overall loss function, enabling end-to-end training (Section 4.5); at inference, only the forward pass of Figure 2a is executed. We detail each stage in the following subsections.

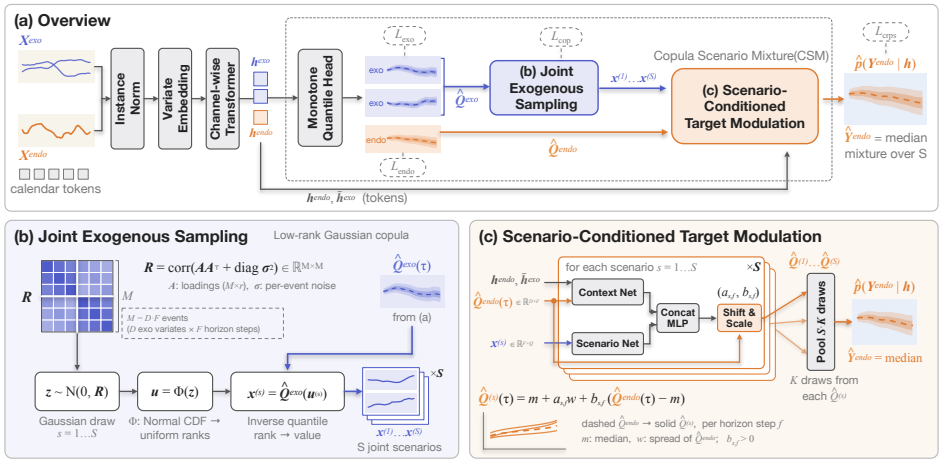

Figure 2: Overall architecture of MarCO. (a) The histories are normalized, embedded as variate tokens, and encoded by a channel-wise Transformer; a quantile head shared by all tokens yields the covariate marginals $\hat{Q}^{\text{exo}}$ and the covariate-agnostic target quantile function $\hat{Q}^{\text{endo}}$. The CSM module then (b) couples the covariate marginals through a low-rank Gaussian copula and draws S joint covariate scenarios, and (c) shifts and rescales $\hat{Q}^{\text{endo}}$ once per scenario and pools K draws from each scenario-conditioned distribution into the target forecast. Blue marks covariate quantities, orange target quantities, and dashed pills the four training-only objectives.

4.2 Marginal Quantile Forecasting

This stage forecasts the marginal distribution of every future covariate and target value; it consists of instance normalization, variate embedding, a channel-wise Transformer, and a quantile head (Figure 2a).

Instance normalization and variate embedding. Covariates and targets live on different scales and drift between windows. Each variable is therefore normalized with its own mean and standard deviation over the T look-back steps, and all outputs of the model are de-normalized with the statistics of the variable they describe, so the covariate marginals and the target quantiles are expressed in their own units. Following iTransformer (Liu et al., 2024), the whole normalized history of a variable is embedded as one variate token by a linear map $\mathbb{R}^T \to \mathbb{R}^d$, and the calendar features are embedded as additional tokens in the same way.

Channel-wise Transformer. To capture the dependence between variables, L standard Transformer blocks apply self-attention across the variate tokens, i.e., along the channel dimension rather than along time, and produce $\mathbf{h}_n^{\text{endo}} \in \mathbb{R}^d$ for the N target variables and $\mathbf{h}_d^{\text{exo}} \in \mathbb{R}^d$ for the D covariates; the calendar tokens inform the others through attention and are dropped afterwards. Channel-wise attention is the only place where covariate and target histories interact before forecasting, so the token of a covariate already reflects the target history and vice versa.

Quantile head. To obtain a preliminary distribution for every future value, a single head maps a token $\mathbf{h}$ to F quantile functions on the levels $\mathcal{T}$. The head is shared by all $N + D$ tokens: the same projections forecast the covariates and the target, so the covariate marginals are trained with the same capacity as the target and the number of parameters is independent of D. Writing τ_{med} for the median level, it predicts the median directly and the other levels as cumulative positive increments:

$$\hat{Q}(\tau_{\text{med}}) = \mathbf{W}_{\text{med}}\mathbf{h}, \qquad \hat{Q}(\tau_{\text{med}\pm j}) = \hat{Q}(\tau_{\text{med}}) \pm \sum_{i \le j} \text{softplus}\big(\mathbf{W}_i^{\pm}\mathbf{h}\big), \tag{4}$$

where $j = 1, \ldots, (Q-1)/2$, and $\mathbf{W}_{\text{med}}$ and $\mathbf{W}_i^{\pm}$ are learnable projections. Applied to the covariate tokens, the head gives the marginals $\hat{Q}_{d,f}^{\text{exo}}$ of every covariate d at every horizon step f; applied to

the target tokens, it gives $\hat{Q}^{\text{endo}}_{n,f}$, which describes the target before any covariate scenario is taken into account and which we refer to as the *base* quantile function. The monotone construction keeps every quantile function strictly increasing, which CSM relies on when evaluating $\hat{Q}^{\text{exo}}$ at sampled levels (Section 4.3) and transforming $\hat{Q}^{\text{endo}}$ (Section 4.4).

4.3 Joint Exogenous Sampling

This stage couples the $M = D \cdot F$ covariate marginals into the joint distribution $p_\theta(\mathbf{Y}^{\text{exo}} \mid \mathbf{h})$ as in Equation 1 and draws coherent scenarios from it (Figure 2b). Treating every covariate at every horizon step as one coordinate lets a single copula capture dependence across covariates and across horizon steps alike.

Low-rank Gaussian copula. We use the Gaussian copula of Equation 2, whose dependence is summarized by one correlation matrix. A full $M \times M$ correlation matrix has $O(M^2)$ parameters, too many to learn from limited data, so we parameterize it through learnable loadings $\mathbf{A} \in \mathbb{R}^{M \times r}$ with rank $r \ll M$ and a positive idiosyncratic vector $\boldsymbol{\sigma} \in \mathbb{R}^M_{>0}$, i.e., $\mathbf{R} = \operatorname{corr}(\mathbf{A}\mathbf{A}^\top + \operatorname{diag}\boldsymbol{\sigma}^2)$. The r columns of $\mathbf{A}$ act as common factors, such as a weather front that moves all wind covariates over consecutive hours, and $\boldsymbol{\sigma}$ controls how much of each coordinate is idiosyncratic. The correlated Gaussian variable is drawn by reparameterization and mapped to probability levels:

$$\mathbf{z} \;=\; \boldsymbol{\Delta}^{-1/2}\big(\mathbf{A}\boldsymbol{\epsilon}_c + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}_i\big), \qquad \mathbf{u} \;=\; \underline{u} + (\bar{u} - \underline{u})\,\Phi(\mathbf{z}), \tag{5}$$

where $\boldsymbol{\epsilon}_c \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r)$, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M)$, $\boldsymbol{\Delta} = \operatorname{diag}(\mathbf{A}\mathbf{A}^\top + \operatorname{diag}\boldsymbol{\sigma}^2)$ rescales each coordinate to unit variance so that $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ without ever forming $\mathbf{R}$, and $[\underline{u}, \bar{u}] \subset (0, 1)$ keeps the levels away from the extremes. The reparameterization makes the scenarios differentiable with respect to $(\mathbf{A}, \boldsymbol{\sigma})$, so the copula can also be trained through the target objective in Section 4.5. We use one $\mathbf{R}$ shared across windows and initialize it near the identity, so that training starts from independent sampling and learns the dependence from data.

Inverse quantile sampling. Each level is mapped through its own marginal quantile function, $x^{(s)}_m = \hat{Q}^{\text{exo}}_m(u_m)$ for $m = 1, \ldots, M$, which interpolates linearly between the Q predicted quantile nodes and extrapolates beyond the outermost nodes linearly in the normal-score space $\Phi^{-1}(u)$, i.e., with Gaussian tails. Only Q levels are predicted, and Gaussian tails give the extreme levels a plausible reach without breaking monotonicity. Repeating the two steps S times yields the scenarios $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)} \in \mathbb{R}^{D \times F}$. Since each u_m is uniform whatever $\mathbf{R}$ is, every coordinate of every scenario follows its marginal $\hat{Q}^{\text{exo}}_m$, so the copula only changes how the coordinates move together. Independent sampling ($\mathbf{A} = \mathbf{0}$) and plugging in the median forecast ($\mathbf{x}^{(s)} \equiv \hat{Q}^{\text{exo}}(\tau_{\text{med}})$, the paradigm of Figure 1b) are two degenerate cases of this stage.

4.4 Scenario-Conditioned Target Modulation

Instead of predicting a new target distribution for every scenario, which would require S full forecasts and could produce invalid quantiles, this stage lets each scenario shift and rescale the base quantile function (Figure 2c) $\hat{Q}^{\text{endo}}$: the base already carries the uncertainty of the target given its own history, and a scenario should only move and widen or narrow it.

Context and scenario networks. For each target variable n, a context network encodes $[\mathbf{h}^{\text{endo}}_n \,\|\, \bar{\mathbf{h}}^{\text{exo}} \,\|\, \hat{Q}^{\text{endo}}_n]$, where $\bar{\mathbf{h}}^{\text{exo}}$ is the mean of the covariate tokens, and a scenario network encodes $[\mathbf{x}^{(s)} \,\|\, \hat{Q}^{\text{exo}}(\tau_{\text{med}}) \,\|\, \mathbf{x}^{(s)} - \hat{Q}^{\text{exo}}(\tau_{\text{med}})]$; passing the deviation from the median covariate forecast explicitly lets the response depend on how far a scenario departs from it rather than on the level of the covariates alone. Both networks are two-layer MLPs whose outputs are concatenated and mapped by a third MLP to a shift $a_{s,f}$ and a raw scale $\tilde{b}_{s,f}$ for every horizon step f.

Location–scale transform. The scenario-conditioned quantile function is:

$$\hat{Q}^{(s)}_f(\tau) \;=\; m_f + a_{s,f}\, w_f + b_{s,f}\big(\hat{Q}^{\text{endo}}_f(\tau) - m_f\big), \qquad b_{s,f} = \operatorname{softplus}\big(\tilde{b}_{s,f} + \log(e-1)\big), \tag{6}$$

where $m_f = \hat{Q}_f^{\text{endo}}(\tau_{\text{med}})$ is the base median and $w_f = \hat{Q}_f^{\text{endo}}(\tau_Q) - \hat{Q}_f^{\text{endo}}(\tau_1)$ its spread. Expressing the shift in units of the spread makes the same raw output mean the same relative displacement in calm and in volatile windows, and $b_{s,f} > 0$ preserves the ordering of the quantiles, so every $\hat{Q}^{(s)}$ is a valid quantile function. The output layer of the third MLP is zero-initialized, which gives $a_{s,f} = 0$ and $b_{s,f} = 1$, i.e., $\hat{Q}^{(s)} \equiv \hat{Q}^{\text{endo}}$ for every scenario: MarCO starts as a covariate-agnostic quantile forecaster and learns the response to covariate scenarios as a perturbation of it, which we found necessary for stable joint training of the marginals, the copula, and the modulator.

Scenario mixture. The target forecast is the equal-weight mixture of the scenario-conditioned distributions:

$$\hat{p}(\mathbf{Y}^{\text{endo}} \mid \mathbf{h}) \;=\; \frac{1}{S}\sum_{s=1}^{S} p^{(s)}(\mathbf{Y}^{\text{endo}}), \qquad p^{(s)} \text{ the density with quantile function } \hat{Q}^{(s)}, \tag{7}$$

which is the Monte Carlo form of Equation 3 with $p_\theta(\mathbf{Y}^{\text{exo}} \mid \mathbf{h})$ replaced by the scenario distribution of Section 4.3. Since the mixture quantiles have no closed form, we draw K samples from each $\hat{Q}^{(s)}$ by inverse transform sampling and pool the $S \cdot K$ samples with equal weights; the empirical median of the pool is the point forecast $\hat{\mathbf{Y}}^{\text{endo}}$ and its empirical quantiles provide prediction intervals. Equation 7 makes the propagation of covariate uncertainty explicit: the variance of the mixture is the average within-scenario variance, which reflects the uncertainty of the target given the covariates, plus the variance of the scenario-conditioned centers, which is inherited from the covariate future through the modulator. Collapsing the covariate future to one trajectory sets the second term to zero, which is precisely the limitation of Figure 1b.

4.5 Loss Function

The training objective of MarCO consists of four components, one per intermediate output in Figure 2a, all computed from the training labels $\mathbf{Y}^{\text{endo}}$ and $\mathbf{Y}^{\text{exo}}$. 1) $\mathcal{L}_{\text{exo}}$ and $\mathcal{L}_{\text{endo}}$ train the marginals of Section 4.2 with the pinball loss, $\mathcal{L}_{\text{pin}}(\hat{Q}, y) = \frac{1}{FQ}\sum_{f,q} \max\{\tau_q(y_f - \hat{Q}_f(\tau_q)), (\tau_q - 1)(y_f - \hat{Q}_f(\tau_q))\}$, averaged over the D covariates for $\mathcal{L}_{\text{exo}} = \mathcal{L}_{\text{pin}}(\hat{Q}^{\text{exo}}, \mathbf{Y}^{\text{exo}})$ and over the N targets for $\mathcal{L}_{\text{endo}} = \mathcal{L}_{\text{pin}}(\hat{Q}^{\text{endo}}, \mathbf{Y}^{\text{endo}})$. 2) $\mathcal{L}_{\text{cop}}$ trains the copula of Section 4.3 by maximum likelihood on the pseudo-ranks of the observed covariate future: we map $\mathbf{Y}^{\text{exo}}$ through the predicted marginal distribution functions to $\hat{\mathbf{u}}$, set $\hat{\mathbf{z}} = \Phi^{-1}(\hat{\mathbf{u}})$, and minimize the negative log density of Equation 2:

$$\mathcal{L}_{\text{cop}} \;=\; \tfrac{1}{2}\Big[\log|\mathbf{R}| + \hat{\mathbf{z}}^\top \mathbf{R}^{-1}\hat{\mathbf{z}} - \hat{\mathbf{z}}^\top\hat{\mathbf{z}}\Big], \tag{8}$$

evaluated in $O(Mr^2)$ time through the Woodbury identity and the matrix determinant lemma. The marginals are detached when computing $\hat{\mathbf{u}}$, so that $\mathcal{L}_{\text{cop}}$ updates only $(\mathbf{A}, \boldsymbol{\sigma})$ and cannot improve the likelihood by distorting the marginals, in the spirit of two-stage inference for copula models (Patton, 2012). 3) $\mathcal{L}_{\text{crps}}$ trains the mixture of Section 4.4 with the CRPS (Gneiting & Raftery, 2007) estimated from its samples. Because the K draws within a scenario share $\mathbf{x}^{(s)}$, we apply the fair ensemble CRPS to the S samples that share a within-scenario index k and average over k:

$$\mathcal{L}_{\text{crps}} \;=\; \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{S}\sum_{s=1}^{S}\left|y^{(s,k)} - y\right| - \frac{1}{S(S-1)}\sum_{s<s'}\left|y^{(s,k)} - y^{(s',k)}\right|\right], \tag{9}$$

which is an unbiased estimate of the CRPS of the mixture for any $S \geq 2$. The total loss is:

$$\mathcal{L} \;=\; \lambda_{\text{endo}}\,\mathcal{L}_{\text{endo}} + \lambda_{\text{exo}}\,\mathcal{L}_{\text{exo}} + \lambda_{\text{cop}}\,\mathcal{L}_{\text{cop}} + \lambda_{\text{crps}}\,\mathcal{L}_{\text{crps}}, \tag{10}$$

where the weights balance the four terms; their values and the training schedule are given in Appendix A.2.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on 12 real-world datasets from the TSF-X benchmark of DAG (Qiu et al., 2026): five hourly electricity price datasets, NP, PJM, BE, FR, and DE (Lago et al., 2021), whose

Table 1: Forecasting results on 12 datasets in the historical-input setting, where the inputs are $\mathbf{X}^{\text{endo}}$ and $\mathbf{X}^{\text{exo}}$ ($T \to F = 168 \to 24$; $60 \to 10$ for Colbun and Rapel). **Red**: the best, Blue: the second best. Baseline results are from Table 6 of DAG and Table 9 of KITE; results including TFT are in Table 6.

Models	MarCO		DAG		KITE		GCGNet		TimeXer		TiDE		DUET		CrossLinear		Amplifier		TimeKAN		PatchTST	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
NP	**0.215**	**0.257**	0.237	0.259	0.220	**0.257**	0.240	0.260	0.270	0.292	0.297	0.309	0.262	0.271	0.246	0.282	0.302	0.321	0.336	0.344	0.283	0.296
PJM	**0.068**	0.164	0.072	0.162	**0.068**	**0.160**	0.072	0.164	0.096	0.191	0.105	0.213	0.082	0.178	0.096	0.198	0.093	0.195	0.130	0.236	0.106	0.212
BE	**0.347**	**0.239**	0.376	0.244	0.355	0.255	0.365	0.259	0.392	0.253	0.492	0.322	0.376	0.253	0.389	0.253	0.417	0.283	0.424	0.280	0.405	0.264
FR	**0.337**	0.196	0.345	**0.190**	**0.337**	0.197	0.352	0.199	0.366	0.195	0.415	0.254	0.359	0.208	0.397	0.204	0.429	0.250	0.448	0.250	0.397	0.223
DE	**0.397**	**0.401**	0.422	0.411	0.418	0.402	0.413	0.402	0.501	0.445	0.465	0.433	0.424	0.403	0.438	0.423	0.493	0.455	0.504	0.462	0.503	0.450
Energy	0.109	0.254	0.112	0.257	0.103	0.247	0.112	0.259	0.138	0.293	0.117	0.265	0.108	0.254	**0.102**	**0.246**	0.108	0.254	0.109	0.254	0.108	0.254
Sdwpfm1	**0.485**	**0.485**	0.532	0.512	0.534	**0.485**	0.537	0.509	0.558	0.533	0.561	0.540	0.550	0.495	0.549	0.553	0.579	0.557	0.598	0.579	0.557	0.557
Sdwpfm2	**0.565**	0.533	0.609	0.548	0.623	**0.527**	0.630	0.547	0.627	0.570	0.636	0.582	0.612	0.537	0.635	0.586	0.669	0.600	0.684	0.621	0.648	0.610
Sdwpfh1	**0.632**	0.571	0.647	0.592	0.650	**0.570**	0.651	0.578	0.651	0.587	0.719	0.641	0.673	0.575	0.687	0.652	1.111	0.816	0.755	0.679	0.669	0.629
Sdwpfh2	**0.739**	0.654	**0.739**	0.635	0.801	**0.632**	0.798	0.645	0.820	0.677	0.783	0.674	0.975	0.694	0.810	0.718	0.834	0.691	0.892	0.740	0.769	0.680
Colbun	**0.060**	0.096	0.072	0.103	0.069	0.098	0.070	**0.094**	0.084	0.146	0.095	0.140	0.073	0.106	0.068	0.101	0.077	0.120	0.091	0.129	0.083	0.125
Rapel	**0.176**	0.222	0.195	**0.215**	0.212	0.229	0.217	0.221	0.301	0.308	0.238	0.270	0.221	0.225	0.201	0.224	0.240	0.259	0.226	0.238	0.215	0.232
1st Count	11	4	1	2	2	6	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0

two covariates are day-ahead load and generation forecasts; Energy, hourly thermoelectric generation with the generation of five other sources as covariates; four wind power datasets, Sdwpfm1, Sdwpfm2 (30-minute), Sdwpfh1, and Sdwpfh2 (hourly) (Zhou et al., 2022), with six weather covariates such as wind speed and direction; and two daily hydropower datasets, Colbun and Rapel, whose target is the reservoir water level and whose covariates are inflow and precipitation. Every dataset has one target variable. All methods forecast from the histories and calendar features only, as defined in Section 3.2; MarCO uses $\mathbf{Y}^{\text{exo}}$ as training labels only. Each dataset is split chronologically into training, validation, and test segments in the ratio 7:1:2. Following DAG, the look-back T and horizon F are $168 \to 24$ for the hourly and 30-minute datasets and $60 \to 10$ for Colbun and Rapel. Dataset statistics are given in Appendix A.1.

Baselines. We compare with eleven methods: DAG (Qiu et al., 2026) and KITE (Cheng et al., 2026), the state-of-the-art deterministic and probabilistic methods for forecasting with exogenous variables; the TSF-X methods GCGNet (Li et al., 2026), TimeXer (Wang et al., 2024), CrossLinear (Zhou et al., 2025), TFT (Lim et al., 2021), and TiDE (Das et al., 2023); and the multivariate methods DUET (Qiu et al., 2025), Amplifier (Fei et al., 2025), TimeKAN (Huang et al., 2025), and PatchTST (Nie et al., 2023). All baseline results are in the historical-input setting of Section 3.2 and are taken from Table 6 of DAG and Table 9 of KITE. The point forecast of KITE is the mean of its samples, as in its paper, and that of MarCO is the median of the scenario mixture.

Implementation details. Following DAG, we adopt MSE and MAE on the normalized target as metrics and use the TFB framework (Qiu et al., 2024) with a rolling forecast of stride one and no drop-last. MarCO forecasts with $S = 128$ scenarios and $K = 64$ draws each; training uses $S = 32$, $K = 1$. It is trained with AdamW and gradient-norm clipping 1.0 for 50 epochs with early stopping on the validation segment, with quantile levels $\mathcal{T} = \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The model width, depth, feed-forward width, dropout, learning rate, batch size, copula rank r, modulator width, and λ_{exo} are selected per dataset by grid search; the grid and the selected values are listed in Appendix A.2. All results use a single seed (2021). Experiments are run on an NVIDIA H200 GPU.

5.2 Main Results

Table 1 reports the results; we have the following observations. 1) MarCO ranks first in MSE on 11 of 12 datasets (tied with KITE on PJM and FR and with DAG on Sdwpfh2) and lowers the MSE averaged over all datasets from 0.363 (DAG) and 0.366 (KITE) to 0.344, i.e., by 5.2% and 5.9%; the reduction against every other baseline is at least 7.3%. Energy is the exception, where CrossLinear and KITE are ahead. 2) In MAE, MarCO is the best on four datasets and on par with KITE (0.339 vs. 0.338) while ahead of DAG (0.344); its median is trained under pinball and CRPS objectives rather than tuned to MAE. 3) The gains are largest where the future covariates are informative yet uncertain: on the four wind power datasets MarCO lowers the MSE of DAG and KITE by up to 9.3%, and on the hydropower datasets Colbun and Rapel by 9.7–17.0%. There the target is driven

Table 2: Ablation of the CSM module on the five electricity price datasets ($168 \to 24$). MSE and CRPS of the target ($\times 10^3$) and the energy score of the covariate scenarios on the validation segment, one shared configuration for all rows ($S = 128$, $K = 64$); G. Mean: geometric mean over datasets.

Variant	MSE						CRPS						Energy
	BE	DE	FR	NP	PJM	G. Mean	BE	DE	FR	NP	PJM	G. Mean	G. Mean
w/o CSM (base $\hat{Q}^{\text{endo}}$)	976.6	117.2	55.3	44.5	53.3	108.5	352.9	151.6	128.8	105.3	104.0	149.8	–
Deterministic (median scenario)	910.6	116.2	**52.9**	43.5	48.8	103.5	318.4	146.9	124.1	101.0	97.0	141.6	1.544
Independent ($\mathbf{A} = \mathbf{0}$)	909.5	115.5	53.0	43.0	49.0	103.2	317.7	146.9	124.1	100.5	96.8	141.3	1.162
Conditional copula	909.1	115.5	53.5	**42.2**	48.4	102.8	318.1	147.0	124.6	**100.2**	96.5	141.3	1.115
MarCO (static copula)	**906.4**	**114.4**	53.0	42.4	**47.5**	**102.1**	**316.0**	**146.6**	**124.0**	100.3	**95.7**	**140.7**	1.123

Table 3: Ablation of the training objective under the protocol of Table 2: MSE of the target ($\times 10^3$).

Variant	BE	DE	FR	NP	PJM	G. Mean
MarCO	911.6	**111.7**	52.0	41.9	48.2	**101.3**
+ point loss on the median	**907.0**	114.4	53.0	43.0	**47.7**	102.4
$\lambda_{\text{endo}} = 1$	913.2	112.6	52.6	**41.5**	49.6	102.2
Pinball on the mixture instead of CRPS	911.7	112.3	**51.9**	41.6	48.5	101.4
w/o $\mathcal{L}_{\text{crps}}$	1014.7	115.7	52.3	44.8	51.2	107.1

by weather-dependent covariates (wind speed; inflow and precipitation) whose future spans a range of outcomes, where a mixture over scenarios rather than a single trajectory is expected to matter most.

5.3 Model Analyses

Ablation on CSM. Table 2 removes or replaces the two stages of CSM. *w/o CSM* forecasts the target with the base quantile function $\hat{Q}^{\text{endo}}$, i.e., without responding to the covariate future (Figure 1a); *Deterministic* feeds the median covariate forecast as the only scenario (Figure 1b); *Independent* samples the marginals without the copula; *Conditional copula* predicts the loadings $\mathbf{A}$ from the tokens of each window instead of sharing one $\mathbf{R}$. We observe: 1) letting the target respond to the covariate future is the largest single gain, reducing MSE by 4.6% and CRPS by 5.5% over *w/o CSM*; 2) sampling scenarios instead of plugging in the median helps the target slightly but improves the covariate scenarios substantially, lowering their energy score from 1.544 to 1.162; 3) coupling the scenarios with the copula further reduces MSE by 1.1%, CRPS on all five datasets, and the energy score to 1.123, so the joint structure of the covariate future, not only its spread, matters for the target; 4) the more flexible conditional copula is worse than the shared one on four of five datasets, so a static dependence structure suffices at this scale.

Ablation on the objective. Table 3 varies the loss of Section 4.5. Removing $\mathcal{L}_{\text{crps}}$ leaves the modulator untrained and raises MSE by 5.7%, confirming that the scenario mixture must be trained as a whole. Adding a point loss on the median or raising λ_{endo} to 1 pulls the base quantile function toward a point forecast and worsens MSE by about 1%, and replacing the CRPS by a pinball loss on the mixture quantiles gives the same MSE; we keep the CRPS, which scores the whole predictive distribution rather than Q fixed levels.

6 Conclusion

We propose MarCO for forecasting with exogenous variables whose future is unknown. It treats the future covariates as a random vector, forecasting their marginals, coupling them through a Gaussian copula, and forecasting the target as a mixture over coherent covariate scenarios. On 12 real-world datasets, MarCO ranks first in MSE on 11 of 12 against state-of-the-art deterministic and probabilistic methods, and additionally provides predictive distributions for the covariates. Extending the copula beyond the Gaussian family and to longer horizons is future work.

AI USE STATEMENT

Generative AI tools were used in this work for language polishing of the manuscript, for drafting LaTeX tables from result spreadsheets, and as a coding assistant during implementation. All AI-assisted text, tables, and code were reviewed and verified by the authors, who checked every reported number against the experiment logs. The research questions, the method, the experimental design, and the analysis of the results were developed by the authors, who take full responsibility for the content of this paper.

ETHICS STATEMENT

This work is methodological research on time series forecasting with exogenous variables. All experiments use publicly available benchmark datasets of electricity prices, energy consumption, wind power, and hydropower, which contain no personal data, and involve no human subjects. Forecasts from MarCO are probabilistic and, like those of any forecasting model, should be used with an understanding of their uncertainty in decision making; we are not aware of ethical risks specific to the proposed method.

REPRODUCIBILITY STATEMENT

The problem setting is defined in Section 3.2, the architecture, sampling procedure, and training objective of MarCO in Section 4, and the datasets, baselines, metrics, and fixed hyper-parameters in Section 5.1. Appendix A.1 gives the dataset statistics and splits, Appendix A.2 the remaining implementation details, the hyper-parameter search space, and the per-dataset configurations, and Appendix A.3 the full results. The source code and the scripts that reproduce all experiments are available at `https://anonymous.4open.science/r/marco-EAA7`.

REFERENCES

Sebastian Pineda Arango, Pedro Mercado, Shubham Kapoor, Abdul Fatir Ansari, Lorenzo Stella, Huibin Shen, Hugo Senetaire, Caner Turkmen, Oleksandr Shchur, Danielle C. Maddix, Michael Bohlke-Schneider, Yuyang Wang, and Syama Sundar Rangapuram. ChronosX: Adapting pre-trained time series models with exogenous variables. In *International Conference on Artificial Intelligence and Statistics*, 2025.

Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Nicolas Chapados, and Alexandre Drouin. TACTiS-2: Better, faster, simpler attentional copulas for multivariate time series. In *International Conference on Learning Representations (ICLR)*, 2024.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:533–538, 2023.

Xinyang Chen, Huidong Jin, Yu Huang, and Zaiwen Feng. XLinear: A lightweight and accurate MLP-based model for long-term time series forecasting with exogenous inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Hanyin Cheng, Jingrong Zhou, Yang Shu, and Chenjuan Guo. KITE: Knowledge-guided probabilistic modeling for time series forecasting with exogenous variables. In *International Conference on Machine Learning (ICML)*, 2026.

Estee Y. Cramer, Evan L. Ray, Velma K. Lopez, Johannes Bracher, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119, 2022. TODO(bib): complete author list.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with TiDE: Time-series dense encoder. *Transactions on Machine Learning Research*, 2023.

Alexandre Drouin, Étienne Marcotte, and Nicolas Chapados. TACTiS: Transformer-attentional copulas for time series. In *International Conference on Machine Learning (ICML)*, 2022.

Jingru Fei, Kun Yi, Wei Fan, Qi Zhang, and Zhendong Niu. Amplifier: Bringing attention to neglected low-energy components in time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11645–11653, 2025.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Songtao Huang, Zhen Zhao, Can Li, and Lei Bai. TimeKAN: KAN-based frequency decomposition learning architecture for long-term time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2025.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2nd edition, 2018.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Marcel Kollovieh, Abdul Fatir Ansari, Michael Bohlke-Schneider, Jasper Zschiegner, Hao Wang, and Yuyang Wang. Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with Gaussian process priors for probabilistic time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2025.

Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations (ICLR)*, 2018.

Zhengyu Li, Xiangfei Qiu, Yuhan Zhu, Xingjian Wu, Jilin Hu, Chenjuan Guo, and Bin Yang. GCGNet: Graph-consistent generative network for time series forecasting with exogenous variables. In *International Conference on Learning Representations (ICLR)*, 2026.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.

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

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

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

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

Jakub Nowotarski and Rafał Weron. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81:1548–1568, 2018.

Kin G. Olivares, Cristian Challu, Grzegorz Marcjasz, Rafał Weron, and Artur Dubrawski. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. *International Journal of Forecasting*, 39(2):884–900, 2023.

Andrew J. Patton. A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110:4–18, 2012.

Pierre Pinson, Henrik Madsen, Henrik Aa. Nielsen, George Papaefthymiou, and Bernd Klöckl. From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy*, 12(1):51–62, 2009.

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

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. DUET: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025.

Xiangfei Qiu, Yuhan Zhu, Zhengyu Li, Xingjian Wu, Bin Yang, and Jilin Hu. DAG: A dual correlation network for time series forecasting with exogenous variables. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2509.14933.

Syama Sundar Rangapuram, Matthias 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 (NeurIPS)*, 2018.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning (ICML)*, 2021a.

Kashif Rasul, Abdul-Saboor Sheikh, Ingmar Schuster, Urs Bergmann, and Roland Vollgraf. Multivariate probabilistic time series forecasting via conditioned normalizing flows. In *International Conference on Learning Representations (ICLR)*, 2021b.

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 (NeurIPS)*, 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, 2020.

Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, 90:106181, 2020.

M. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Kshitij Tayal, Arvind Renganathan, Rahul Ghosh, Xiaowei Jia, and Vipin Kumar. ExoTST: Exogenous-aware temporal sequence transformer for time series prediction. In *IEEE International Conference on Data Mining (ICDM)*, 2024.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A multi-horizon quantile recurrent forecaster. In *NeurIPS Time Series Workshop*, 2017. arXiv:1711.11053.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph WaveNet for deep spatial-temporal graph modeling. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

Jingbo Zhou, Xinjiang Lu, Yixiong Xiao, Jiantao Su, Junfu Lyu, Yanjun Ma, and Dejing Dou. SDWPF: A dataset for spatial dynamic wind power forecasting challenge at KDD cup 2022. *arXiv preprint arXiv:2208.04360*, 2022.

Pengfei Zhou et al. CrossLinear: Plug-and-play cross-correlation embedding for time series forecasting with exogenous variables. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025. TODO(bib): complete author list.

A Experimental Details

A.1 Datasets

Table 4 summarizes the 12 datasets. All are taken from the TSF-X benchmark released with DAG (Qiu et al., 2026) and split chronologically into training, validation, and test segments in the ratio 7:1:2. Each dataset has a single target variable; the covariates are the remaining columns of the file. NP, PJM, BE, FR, and DE are the day-ahead electricity price markets of Lago et al. (2021), whose covariates are the two day-ahead forecasts published by the market operator (load and wind or generation forecasts). Energy records hourly thermoelectric generation, with battery, geothermal, hydroelectric, solar, and wind generation as covariates. The four SDWPF datasets (Zhou et al., 2022) record the active power of a wind farm at 30-minute (Sdwpfm1, Sdwpfm2) and hourly (Sdwpfh1, Sdwpfh2) resolution, with 2-m temperature, surface pressure, relative humidity, wind speed, wind direction, and total precipitation as covariates. Colbun and Rapel are daily reservoir water levels of two Chilean hydropower plants, with inflow and precipitation as covariates. Calendar features are derived from the timestamps (hour of day, day of week, day of month, and day of year for hourly data, plus minute of hour for 30-minute data; day of week, month, and year for daily data) and enter MarCO as additional variate tokens (Section 4.2).

Table 4: Statistics of the 12 datasets. Length is the number of time steps; the split gives the lengths of the training, validation, and test segments; $T \to F$ is the look-back and horizon.

Dataset	Domain	Frequency	Length	#Covariates	Split (train/val/test)	$T \to F$
NP, PJM, BE, FR, DE	Electricity price	1 hour	52,416	2	36,690 / 5,242 / 10,484	$168 \to 24$
Energy	Power generation	1 hour	13,064	5	9,144 / 1,307 / 2,613	$168 \to 24$
Sdwpfm1, Sdwpfm2	Wind power	30 min	29,281	6	20,496 / 2,928 / 5,857	$168 \to 24$
Sdwpfh1, Sdwpfh2	Wind power	1 hour	14,641	6	10,248 / 1,464 / 2,929	$168 \to 24$
Colbun	Hydropower	1 day	2,958	2	2,070 / 296 / 592	$60 \to 10$
Rapel	Hydropower	1 day	3,366	2	2,355 / 337 / 674	$60 \to 10$

A.2 Implementation Details

Model. Variate tokens are produced by a linear map of the normalized history of each variable and of each calendar feature, as in iTransformer (Liu et al., 2024); the calendar tokens take part in attention and are dropped before the quantile head. The quantile head of Equation 4 predicts the median and softplus increments, so no post-hoc sorting is needed. In the inverse quantile sampling of Section 4.3, the probability levels are clipped to $[\underline{u}, \bar{u}] = [0.001, 0.999]$ and the tails beyond the outermost predicted quantiles are extrapolated linearly in normal-score space. The copula likelihood of Equation 8 is evaluated with the Woodbury identity and the matrix determinant lemma, so its cost is $O(Mr^2)$. The output layer of the modulator is zero-initialized, giving $a_{s,f} = 0$ and $b_{s,f} = 1$ at initialization. The Monte Carlo sizes are $S = 32$, $K = 1$ during training, $S = 64$, $K = 16$ for validation and model selection, and $S = 128$, $K = 64$ at test time, with three independent random streams for the copula factors, the idiosyncratic noise, and the within-scenario draws.

Training. All models are trained with AdamW and gradient-norm clipping at 1.0 for 50 epochs, with early stopping on the validation loss and the checkpoint with the best validation score restored. $\mathcal{L}_{\text{crps}}$ is the only term that reaches the modulator; it also reaches the copula through the reparameterization in Equation 5 and the covariate marginals through the quantile interpolation of Section 4.3. The gradient of $\mathcal{L}_{\text{crps}}$ with respect to the covariate marginals is switched off for the first two epochs and ramped in linearly over the next two, so that the marginals calibrate before the target objective influences them; the marginals are detached when computing the pseudo-ranks for $\mathcal{L}_{\text{cop}}$ throughout. The optimizer uses weight decay 0.01, patience 5, and a learning rate that is constant for the first three epochs and then decays by a factor 0.9 per epoch. The loss weights are $\lambda_{\text{endo}} = 0.2$, $\lambda_{\text{crps}} = 0.1$, $\lambda_{\text{cop}} = 10^{-3}$, and $\lambda_{\text{exo}} \in \{0.2, 0.4\}$; all attention layers use four heads. The remaining hyper-parameters are selected per dataset by grid search over $d \in \{64, 128, 256, 512\}$, $L \in \{1, 2, 3\}$, $d_{\text{ff}} \in \{128, 256, 512\}$, dropout $\in \{0, 0.1, 0.2\}$, learning rate $\in \{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$, batch size $\in \{32, 64\}$, $r \in \{2, 4, 8\}$, and $d_{\text{mod}} \in \{64, 128\}$; Table 5 lists the selected values.

Table 5: Selected hyper-parameters of MarCO per dataset: model width d, depth L, feed-forward width d_{ff}, dropout (do), learning rate (lr), batch size (bs), copula rank r, modulator width d_{mod}, λ_{exo}, number of parameters, and training time.

Dataset	$T \to F$	d	L	d_{ff}	do	lr	bs	r	d_{mod}	λ_{exo}	#Params	Time (s)
NP	168→24	128	3	256	0.1	1e-4	64	2	128	0.2	0.54M	766
PJM	168→24	512	1	512	0.1	1e-3	64	2	64	0.2	1.83M	192
BE	168→24	256	3	128	0	1e-4	32	8	64	0.4	1.13M	419
FR	168→24	512	3	256	0.1	1e-4	64	4	64	0.4	4.20M	332
DE	168→24	128	3	512	0.2	1e-4	64	4	128	0.4	0.74M	683
Energy	168→24	512	1	128	0	1e-4	64	2	64	0.4	1.45M	42
Sdwpfm1	168→24	64	1	256	0	1e-4	64	2	64	0.4	0.13M	141
Sdwpfm2	168→24	64	2	128	0	3e-4	64	4	128	0.4	0.21M	95
Sdwpfh1	168→24	64	3	128	0	1e-3	32	2	64	0.4	0.18M	74
Sdwpfh2	168→24	64	3	512	0.2	1e-3	32	2	64	0.4	0.33M	76
Colbun	60→10	128	2	256	0	1e-3	64	8	128	0.2	0.36M	10
Rapel	60→10	128	3	512	0	1e-3	32	4	128	0.2	0.69M	24

A.3 Full Results

Table 6 reports the results of MarCO and all twelve baselines, including TFT, which is omitted from Table 1. The baseline numbers are taken from Table 6 of DAG (Qiu et al., 2026) and Table 9 of KITE (Cheng et al., 2026), which use the historical-input setting of Section 3.2 with the same splits and horizons.

Table 6: Full results on 12 datasets in the historical-input setting, where the inputs are $\mathbf{X}^{\text{endo}}$ and $\mathbf{X}^{\text{exo}}$ ($T \to F = 168 \to 24$; $60 \to 10$ for Colbun and Rapel). **Red**: the best, Blue: the second best.

Models	MarCO		DAG		KITE		GCGNet		TimeXer		TFT		TiDE		DUET		CrossLinear		Amplifier		TimeKAN		PatchTST	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
NP	**0.215**	**0.257**	0.237	0.259	0.220	**0.257**	0.240	0.260	0.270	0.292	0.330	0.321	0.297	0.309	0.262	0.271	0.246	0.282	0.302	0.321	0.336	0.344	0.283	0.296
PJM	**0.068**	0.164	0.072	0.162	**0.068**	**0.160**	0.072	0.164	0.096	0.191	0.129	0.231	0.105	0.213	0.082	0.178	0.096	0.198	0.093	0.195	0.130	0.236	0.106	0.212
BE	**0.347**	**0.239**	0.376	0.244	0.355	0.255	0.365	0.259	0.392	0.253	0.400	0.267	0.492	0.322	0.376	0.253	0.389	0.253	0.417	0.283	0.424	0.280	0.405	0.264
FR	**0.337**	0.196	0.345	**0.190**	**0.337**	0.197	0.352	0.199	0.366	0.195	0.445	0.231	0.415	0.254	0.359	0.208	0.397	0.204	0.429	0.250	0.448	0.250	0.397	0.223
DE	**0.397**	**0.401**	0.422	0.411	0.418	0.402	0.413	0.402	0.501	0.445	0.576	0.452	0.465	0.433	0.424	0.403	0.438	0.423	0.493	0.455	0.504	0.462	0.503	0.450
Energy	0.109	0.254	0.112	0.257	0.103	0.247	0.112	0.259	0.138	0.293	0.352	0.463	0.117	0.265	0.108	0.254	**0.102**	**0.246**	0.108	0.254	0.109	0.254	0.108	0.254
Sdwpfm1	**0.485**	**0.485**	0.532	0.512	0.534	**0.485**	0.537	0.509	0.558	0.533	0.596	0.531	0.561	0.540	0.550	0.495	0.549	0.553	0.579	0.557	0.598	0.579	0.557	0.557
Sdwpfm2	**0.565**	0.533	0.609	0.548	0.623	**0.527**	0.630	0.547	0.627	0.570	0.749	0.657	0.636	0.582	0.612	0.537	0.635	0.586	0.669	0.600	0.684	0.621	0.648	0.610
Sdwpfh1	**0.632**	0.571	0.647	0.592	0.650	**0.570**	0.651	0.578	0.651	0.587	0.754	0.670	0.719	0.641	0.673	0.575	0.687	0.652	1.111	0.816	0.755	0.679	0.669	0.629
Sdwpfh2	**0.739**	0.654	**0.739**	0.635	0.801	**0.632**	0.798	0.645	0.820	0.677	0.811	0.714	0.783	0.674	0.975	0.694	0.810	0.718	0.834	0.691	0.892	0.740	0.769	0.680
Colbun	**0.060**	0.096	0.072	0.103	0.069	0.098	0.070	**0.094**	0.084	0.146	0.449	0.279	0.095	0.140	0.073	0.106	0.068	0.101	0.077	0.120	0.091	0.129	0.083	0.125
Rapel	**0.176**	0.222	0.195	**0.215**	0.212	0.229	0.217	0.221	0.301	0.308	0.238	0.274	0.238	0.270	0.221	0.225	0.201	0.224	0.240	0.259	0.226	0.238	0.215	0.232
1st Count	11	4	1	2	2	6	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0