Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

VDSA: Calibrated Financial Forecasting via Few-Step Parameter-Space Diffusion in Selective State Space Models

Lihong Ye
Geely University of China
Yanqing Yang
Chengdu Jincheng College
Liming She
iwanna2026@163.com
Geely University of China
Shuai Yang
Geely University of China

Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11096562/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

VDSA: Calibrated Financial Forecasting via Few-Step Parameter-Space Diffusion in Selective State Space Models

Lihong Ye[1], Yanqing Yang[2], Liming She[1,*], and Shuai Yang[1,*]

[1] Geely University of China, Chengdu, Sichuan 641423, China
[2] Chengdu Jincheng College, Chengdu, Sichuan 611731, China
* corresponding authors: iwanna2026@163.com (Liming She); ycsdev@163.com (Shuai Yang)

ABSTRACT

Financial and macroeconomic time series exhibit non-stationary dynamics, regime switches, and heavy-tailed shocks, making calibrated uncertainty as important as point accuracy. Existing probabilistic state space sequence models often represent dynamics uncertainty with unimodal Gaussian reparameterizations, which can be mis-specified when the posterior over latent dynamics is multi-modal or heavy-tailed. We therefore developed the Variational Diffusion Spectral Adapter (VDSA), a parameter-space generative reparameterization that models the posterior over state transition eigenvalues as a context-conditioned diffusion process. VDSA is integrated with a selective state space model (SSM) backbone: the backbone encodes context while the diffusion module generates context-conditioned spectral samples, inducing a predictive distribution through SSM rollouts. To make diffusion inference deployable, we introduce few-step deterministic sampling and teacher-student distillation that compress sampling to 2-8 denoising steps while preserving calibration. We evaluate VDSA on Exchange Rate, Monash Archive datasets, and a synthetic regime-switching benchmark using point error, proper scoring rules, interval calibration, and tail-risk diagnostics. On Exchange Rate at horizon 30, VDSA reduced CRPS from 0.0364 for a Gaussian spectral posterior to 0.0351 and improved 80% interval coverage from 0.78 to 0.79, while maintaining MAE comparable to deterministic SSM baselines. VaR exceedance rates also moved closer to the nominal 5% level (0.052 vs. 0.061). Few-step inference achieved substantially higher throughput than full diffusion (relative throughput 0.35 at 4 steps vs. 0.12 at 50 steps). The empirical design employs a "baseline evidence mapping table" that maps each comparison method to the corresponding results table, paired statistical tests, VaR backtests, and details regarding reproducibility. These results indicate that parameter space diffusion is a practical mechanism for achieving calibrated forecasts under institutional uncertainty. They also demonstrate that the primary advantage of the method proposed in this paper lies in its probabilistic calibration capability, rather than in an absolute lead in point forecast error.

Introduction

Financial forecasting is not merely a matter of minimising the mean error, as actual decision-making is often influenced by risk constraints, tail events and distributional robustness. While the forecast mean may appear reasonably accurate, the forecast distribution may suffer from a lack of calibration. This lack of calibration is further amplified when market regimes shift or volatility clusters emerge. Classical models, such as autoregressive integrated moving average (ARIMA) models, vector autoregressive (VAR) models, sparse regularised linear models, and volatility models from the GARCH family, are commonly used as interpretable baseline methods; however, their distributional assumptions may be overly restrictive in the presence of structural breaks and non-Gaussian tails[Durbin et al., 2012,Tibshirani, 1996,Bollerslev, 1986,Engle, 2002]. Modern deep forecasting models, including decomposition-based linear models, patch-based Transformers, multiscale mixture models, and foundation model-based predictors, generally exhibit superior accuracy and scalability; however, their probabilistic outputs may still be under-calibrated if uncertainty and distributional drift are not explicitly addressed[Cai et al., 2024,Zhou et al., 2021,Wu et al., 2021,Oreshkin et al.(2020),Nie et al., 2023,Wu et al., 2023,Liu e Consequently, probabilistic forecasting emphasises appropriate scoring rules and calibration intervals, rather than focusing solely on point prediction errors[Gneiting et al., 2007,Hyndman et al., 2006].

Selective State Space Models (SSMs) provide a powerful backbone for modelling long-term context and, thanks to their linear-time recursion capability, have in recent years emerged as a competitive alternative to attention mechanisms for sequence modelling[Gu et al., 2023,Gu et al., 2022]. Such models represent dynamic processes through structured state transitions, whose parameters can be interpreted as a compact set of spectral measures governing damping and oscillation. When the dynamic process is stable and unimodal, a common practice is to impose Gaussian uncertainty on these parameters via reparameterisation, thereby obtaining a manageable variational approximation. However, in financial time series, effective dynamics may switch between low-volatility mean-reverting states and high-volatility stress states, and the posterior distribution

of spectral parameters may exhibit multimodal or heavy-tailed characteristics. If a unimodal Gaussian posterior is enforced, the approximation may collapse to an averaged state that represents neither the steady-state period nor the stress period, resulting in systematic under-calibration of the forecast distribution during the most critical periods.

Diffusion models are powerful density models that learn to map Gaussian noise to complex distributions through a sequence of denoising steps[Ho et al., 2020, Song et al., 2021]. Diffusion has been applied to time series generation, imputation, and probabilistic forecasting in the data space[Tashiro et al., 2021, Rasul et al., 2021, Lopez Alcaraz et al., 2022]. Yet data-space diffusion typically samples the observed sequence itself, which can be expensive at inference time and may provide limited interpretability about latent dynamics. Our starting point is that, for state space forecasting, uncertainty is often more meaningful when it is placed on dynamics parameters rather than directly on observations. Parameter-space uncertainty can also be constrained to ensure stability, and it can be summarized as spectral posterior shifts that correspond to interpretable regime changes. We therefore propose to move diffusion from the data space into the parameter space, focusing on the eigenvalues that parameterize the transition operator. We treat the spectral vector as a random variable conditioned on context and model its posterior using a conditional diffusion process. The backbone remains deterministic and linear-time, and uncertainty is injected through spectral sampling. This choice is motivated by the observation that the parameter dimension is small compared to the data dimension, so parameter-space diffusion can be more efficient than sequence-space diffusion. Still, naive diffusion sampling is iterative, so it may be too slow for rolling forecasts and large-scale evaluations. This motivates a second design goal: the diffusion posterior must be deployable, meaning it should support few-step (and optionally single-step) sampling without sacrificing calibration.

Our method, the Variational Diffusion Spectral Adapter (VDSA), extends selective state space forecasting with three components. First, a context encoder produces a summary representation that conditions the spectral posterior. Second, a diffusion posterior network generates spectral samples through a learned denoising process. Third, a forecasting head produces predictive samples by rolling out the state space recurrence with sampled spectra. To reduce inference cost, VDSA uses deterministic sub-sampling of diffusion steps in the spirit of DDIM[Song et al., 2021], and it trains a student denoiser via teacher–student distillation to match multi-step samples with 2–4 steps and optionally with one amortized mapping. The overall training objective combines forecast loss, diffusion loss, and distillation alignment, so the learned posterior remains task-relevant and calibrated.

We evaluated our approach on controlled and real-world settings. The first was a synthetic regime-switching benchmark where the ground-truth dynamics were explicitly bimodal, which let us test whether the learned spectral posterior became multi-modal as expected. The second was the Exchange Rate dataset, a public multivariate daily benchmark with 8 currency series and long temporal coverage, which exhibited volatility clustering and structural shifts. The third was a transfer evaluation on three Monash Archive datasets to test whether the method remained competitive outside the Exchange Rate benchmark. We reported mean absolute error (MAE) and symmetric mean absolute percentage error (sMAPE) for point accuracy, and we emphasized CRPS, pinball loss, MSIS, and coverage metrics for probabilistic calibration[Gneiting et al., 2007, Hyndman et al., 2006]. We also included tail-risk diagnostics based on VaR exceedance rates and formal VaR backtests, because the usefulness of calibration is often judged in the tails[Kupiec, 1995, Christoffersen, 1998].

Despite these advances, a gap remains between expressive uncertainty modeling and deployable SSM inference. Data-space diffusion can be costly at test time when used for rolling forecasts, and it does not directly expose interpretable dynamics parameters. Gaussian reparameterizations over spectral parameters are efficient but can be mis-specified under regime-switching and heavy-tailed dynamics. These gaps led to the following research question: Can a context-conditioned, parameter-space diffusion posterior over SSM spectral dynamics improve probabilistic calibration under regime changes while remaining deployable via few-step sampling?

We addressed this question with the Variational Diffusion Spectral Adapter (VDSA). The first contribution is a parameter-space diffusion posterior over SSM spectral parameters, which targets multi-modal dynamics uncertainty rather than observation noise alone. The second contribution is a deployable inference strategy based on few-step deterministic sampling and teacher-student distillation, which narrows the computational gap between full diffusion and deterministic SSM inference. The third contribution is a set of spectral diagnostics that relate posterior changes to volatility regimes without treating a single diagnostic plot as causal evidence. The fourth contribution is an empirical protocol that connects each baseline family to the specific evidence it is meant to provide, using paired predictive-accuracy tests, moving-block bootstrap confidence intervals, interval scores, and VaR backtesting to separate robust calibration gains from small numerical differences.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes VDSA and the few-step inference strategy. Section 4 reports experimental results. Section 5 discusses implications and limitations.

Related Work

State Space Models in Time Series

Structured State Space Models (SSMs) such as S4[Gu et al., 2022] and Mamba[Gu et al., 2023] have emerged as powerful alternatives to Transformers for long-sequence modeling, offering linear scaling with sequence length. In time series forecasting, SSM

variants such as MambaTS have been used to capture long-range dependencies efficiently[Cai et al., 2024]. At the same time, recent non-SSM forecasters have become strong baselines. DLinear and NLinear showed that carefully designed decomposition and linear models can be highly competitive on long-horizon forecasting benchmarks[Zeng et al., 2023]. PatchTST converted time series into patch tokens and improved Transformer efficiency for long contexts[Nie et al., 2023]. TimesNet represented temporal variation through a 2D tensorization of period-aligned patterns[Wu et al., 2023], iTransformer inverted the tokenization axis so that variables rather than time steps served as tokens[Liu et al., 2024], and TimeMixer used decomposable multiscale mixing to combine short- and long-term temporal information[Wang et al., 2024]. N-HiTS and Temporal Fusion Transformers also remain important reference points for hierarchical interpolation and interpretable multi-horizon probabilistic forecasting[Challu et al., 2023,Lim et al., 2021]. These methods define the contemporary benchmark context. In the empirical section, we restrict quantitative claims to baselines that are reported under the same chronological splits and evaluation metrics, and we use the remaining recent methods as literature context rather than as unreported evidence of superiority.

Diffusion Models for Forecasting

Diffusion models have been applied to time series forecasting and imputation. TimeGrad[Rasul et al., 2021] uses an autoregressive diffusion model in the data space, and CSDI[Tashiro et al., 2021] uses conditional score-based diffusion for imputation. Other diffusion-based time-series models have studied structured SSM backbones, sequence generation, and score-based probabilistic forecasting[Lopez Alcaraz et al., 2022,Yuan et al., 2024]. More broadly, latent-variable and generative modeling provide several routes to flexible posterior approximation, including Bayesian latent-variable modeling, variational autoencoders, richer variational priors, adversarial generation, and score-based density estimation[MacKay, 2003,Kingma et al., 2014,Tomczak et al., 2018,Arjovsky et al., 2017,Song et al., 2019]. These approaches motivate expressive uncertainty modeling but usually operate either in observation space or in generic latent spaces. In contrast, VDSA operates in the parameter space of spectral dynamics, reducing the diffusion dimensionality and enabling spectral diagnostics for regime shifts. This design makes the comparison to data-space diffusion essential rather than optional, so our empirical study treats data-space diffusion as a core probabilistic baseline and reports inference throughput alongside forecast quality.

Uncertainty Calibration in Finance

In financial contexts, well-calibrated uncertainty is of paramount importance. DeepAR[Salinas et al., 2020] and MQ-RNN[Wen et al., 2017] are capable of providing probabilistic forecasts, but they typically rely on unimodal likelihoods or independently estimated quantiles. Classical risk models such as GARCH, DCC-GARCH, and CAViaR remain important because they directly model the behavior of volatility and value at risk (VaR)[Bollerslev, 1986,Engle, 2002,Engle et al., 2004]. However, these models generally lack flexibility in nonlinear multivariate forecasting and may be vulnerable during regime shifts. Research on neural network calibration also indicates that even highly predictive models may suffer from poor calibration if they lack explicit calibration-aware objectives or diagnostic mechanisms[Guo et al., 2017]. This paper focuses on the inherent cognitive uncertainty of the process itself, allowing the predictive distribution to vary through spectral parameters obtained via sampling, rather than relying solely on variance adjustments at the observation level. This motivation also implies that calibration assessment should not rely solely on visual reliability curves but should also employ appropriate scoring rules, interval scoring, and VaR coverage tests.

Methods

Theoretical Motivation: Why Parameter-Space Diffusion?

In the context of this paper, modelling uncertainty in the parameter space offers clear advantages over modelling in the data space. Let the true posterior distribution of the dynamic parameters be $p(\Lambda_t / C_t)$. In the context of regime switching, this distribution is typically multimodal; for example, it may be a mixture distribution comprising stable and highly volatile modes. If standard variational inference methods employ a Gaussian approximation $q_\varphi(\Lambda_t / C_t) = N(\mu, \sigma^2)$, they are subject to the mode-seeking behaviour of the reverse Kullback–Leibler divergence, which often causes the distribution to collapse into a single mode, or to cover the mean locations of multiple modes with excessive variance.

Observation 1 (Expressiveness). *A diffusion-based posterior $q_{diff}(\Lambda_t / C_t)$ could represent multi-modal and heavy-tailed distributions over spectral parameters more flexibly than a unimodal Gaussian q_N, because the reverse process can transform a simple prior into complex conditional densities through iterative denoising.*

By lifting the latent variable Λ to a diffusion process, we enabled the model to capture the full landscape of dynamical possibilities. Furthermore, since $\dim(\Lambda) \ll \dim(x_{t+1:t+H})$, the diffusion process was far more efficient to sample than generating the full output sequence pixel-by-pixel.

Problem setup and notation

We considered a multivariate time series $x_{1:T} \in \mathbb{R}^D$ and forecasting horizons $H \in \{3, 7, 14, 30, 60\}$. At time t, the model observed a window $x_{t-L+1:t}$ and predicted the distribution of $x_{t+1:t+H}$. We denoted the context features by $C_t = f_\psi(x_{t-L+1:t+0})$, where f_ψ was a deterministic encoder implemented by a selective state space backbone[Gu et al., 2023, Gu et al., 2022]. The goal was to learn a predictive distribution $p_\theta(x_{t+1:t+H} \mid x_{t-L+1:t})$ that was sharp yet calibrated, so that prediction intervals achieved nominal coverage while remaining informative. A standard approach in probabilistic deep forecasting was to predict parameters of a distribution over observations, such as Gaussian mean and variance, possibly with quantiles[Salinas et al., 2020, Wen et al., 2017]. Another approach was to sample latent variables and propagate them through a model. We followed the second direction, but placed uncertainty on spectral dynamics parameters rather than directly on output noise. This design better captures regime uncertainty, because multiple plausible dynamics can coexist in the posterior given the same recent context.

Selective state space backbone

We assumed a state space recurrence of the form

$$h_{t+1} = A(\Lambda_t)h_t + Bu_t, \quad y_t = Ch_t + Du_t \tag{1}$$

where h_t was the latent state, u_t was an input derived from x_t, and $A(\Lambda_t)$ was parameterized by a spectral vector $\Lambda_t \in \mathbb{C}^N$. In practice, Λ_t was represented by concatenated real and imaginary parts in $\mathbb{R}^{2N}$. Selective state space models used structured parameterizations and gating to enable efficient long-sequence modeling [Gu et al., 2023, Gu et al., 2022], and the backbone encoder f_ψ produced context C_t that summarized the recent window. The forecast head then rolled out the recurrence for H steps to produce predictive samples. We focused on uncertainty in Λ_t because it controlled stability and oscillatory behavior, and it could be interpreted as a compact representation of dynamical regimes. When markets switched regimes, effective damping and frequency content could change, and we wanted the posterior over Λ_t to reflect that change.

In a Gaussian variational baseline, we set

$$\Lambda_t = \mu_\varphi(C_t) + \sigma_\varphi(C_t) \odot \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I) \tag{2}$$

which was a unimodal approximation. This baseline was simple and often effective under mild non-stationarity, but it could not represent a bimodal posterior without inflating variance, which then reduced sharpness and could distort interval widths.

VDSA: parameter-space diffusion posterior

VDSA replaced the unimodal posterior with a conditional diffusion model over Λ_t. We defined a forward noising process over diffusion steps $k = 1, \ldots, K$:

$$q(\Lambda_t^{(k)} \mid \Lambda_t^{(k-1)}) = \mathcal{N}(\Lambda_t^{(k)}; \sqrt{1-\beta_k}\,\Lambda_t^{(k-1)}, \beta_k I) \tag{3}$$

with $\Lambda_t^{(0)}$ being a target spectral vector and $\Lambda_t^{(K)} \approx \mathcal{N}(0, I)$. The reverse process was parameterized by a denoiser ε_θ conditioned on C_t:

$$p_\theta(\Lambda_t^{(k-1)} \mid \Lambda_t^{(k)}, C_t) = \mathcal{N}(\Lambda_t^{(k-1)}; \mu_\theta(\Lambda_t^{(k)}, k, C_t), \sigma_k^2 I) \tag{4}$$

We trained in the standard noise-prediction form[Ho et al., 2020]:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,k,\varepsilon}\left[\left\| \varepsilon - \varepsilon_\theta(\Lambda_t^{(k)}, k, C_t) \right\|_2^2\right] \tag{5}$$

where $\Lambda_t^{(k)}$ was constructed by noising $\Lambda_t^{(0)}$ and $\varepsilon \sim \mathcal{N}(0, I)$.

The target $\Lambda_t^{(0)}$ could be treated as a learnable latent variable aligned with forecasting, or it could be derived from a spectral estimator module; in our implementation we treated it as an auxiliary latent predicted from C_t and optimized jointly. To remove ambiguity about this auxiliary latent, we used a spectral amortizer $a_{(1)}$ that mapped the context vector to an unconstrained spectral code, $\eta_t = a_{(1)}(C_t) \in \mathbb{R}^{2N}$. The stabilized clean target was then defined as $\Lambda_t^{(0)} = T_{stab}(\eta_t)$, where T_{stab} applied the stability transforms described below. During teacher training, $\Lambda_t^{(0)}$ was optimized jointly with the SSM backbone and diffusion denoiser rather than estimated from a separate offline procedure. The diffusion loss regularized the conditional density around these task-relevant spectra, while the forecasting loss prevented the spectral amortizer from learning arbitrary modes that were disconnected from predictive performance.

In the Gaussian spectral baseline, the same context encoder and stability transform were used, but the posterior was restricted to $\Lambda_t = T_{stab}(\mu_\varphi(C_t) + \sigma_\varphi(C_t) \odot \varepsilon)$, so the comparison isolated the effect of replacing a unimodal posterior with a

diffusion posterior. Given a sampled spectral vector $\hat{\Lambda}_t$, we ran the state space recurrence forward and produced a predictive sample $\hat{x}_{t+1:t+H}$. Repeating this procedure M times yielded a Monte Carlo approximation to the predictive distribution, which supported CRPS, pinball loss, and interval computation; we used $M = 50$ in our main experiments unless noted otherwise.

For a set of samples $\{\hat{x}_{t+1:t+H}^{(m)}\}_{m=1}^{M}$, the empirical CRPS for a scalar target x was computed as

$$\widehat{\mathrm{CRPS}}(F,x) = \frac{1}{M}\sum_{m=1}^{M}|\hat{x}^{(m)} - x| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{j=1}^{M}|\hat{x}^{(m)} - \hat{x}^{(j)}|. \tag{6}$$

The forecasting objective combined this sample-based CRPS with a small point-forecast term on the Monte Carlo mean,

$$L_{forecast} = \mathrm{CRPS}(F_\Theta, x_{t+1:t+H}) + \tau \left\| \bar{x}_{t+1:t+H} - x_{t+1:t+H} \right\|_1, \tag{7}$$

where τ was set to 0.1 in the main runs. This loss choice made calibration part of the training signal rather than a post-hoc evaluation metric. Uncertainty was injected at the dynamics level, so multiple trajectories could correspond to different plausible regime interpretations.

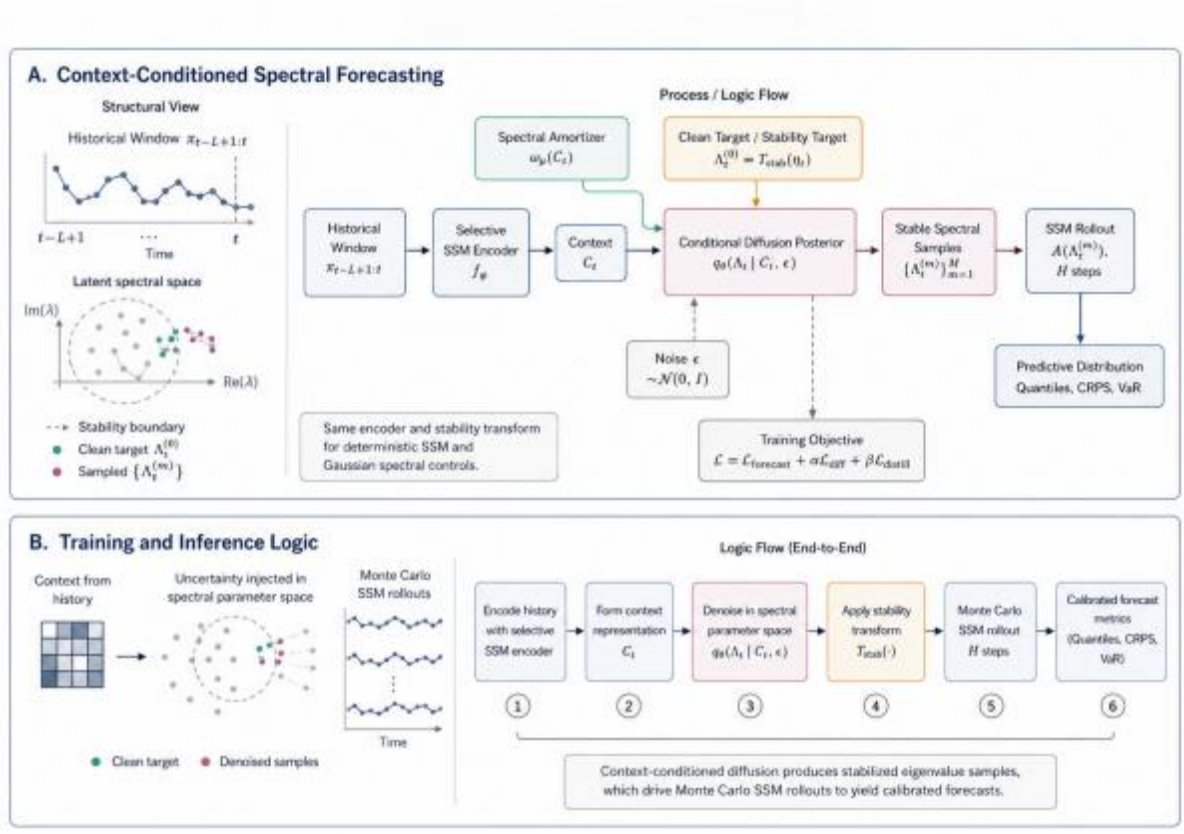


Figure 1. Context-conditioned spectral forecasting architecture and training/inference logic. (A) Structural view of VDSA. The historical window $x_{t-L+1:t}$ is encoded by a selective SSM into context C_t, while a spectral amortizer provides the clean stability target $\Lambda_t^{(0)} = T_{stab}(\eta_t)$. A conditional diffusion posterior maps Gaussian noise and context into stable spectral samples $\{\Lambda_t^{(m)}\}_{m=1}^{M}$, which drive Monte Carlo SSM rollouts and predictive metrics. (B) End-to-end logic showing that uncertainty is injected in spectral parameter space, followed by stability transformation, SSM rollout, and calibrated forecast evaluation. The deterministic SSM and Gaussian spectral controls use the same encoder and stability transform, so the comparison isolates the effect of replacing a unimodal posterior with parameter-space diffusion.

Few-step deterministic sampling

A practical concern is that sampling $\hat{\Lambda}_t$ with K diffusion steps can be slow. We adopted deterministic sub-sampling similar to denoising diffusion implicit models (DDIM)[Song et al., 2021] to reduce steps at inference. Let $\bar{\alpha}_k = \prod_{i=1}^{k}(1 - \beta_i)$. For a diffusion

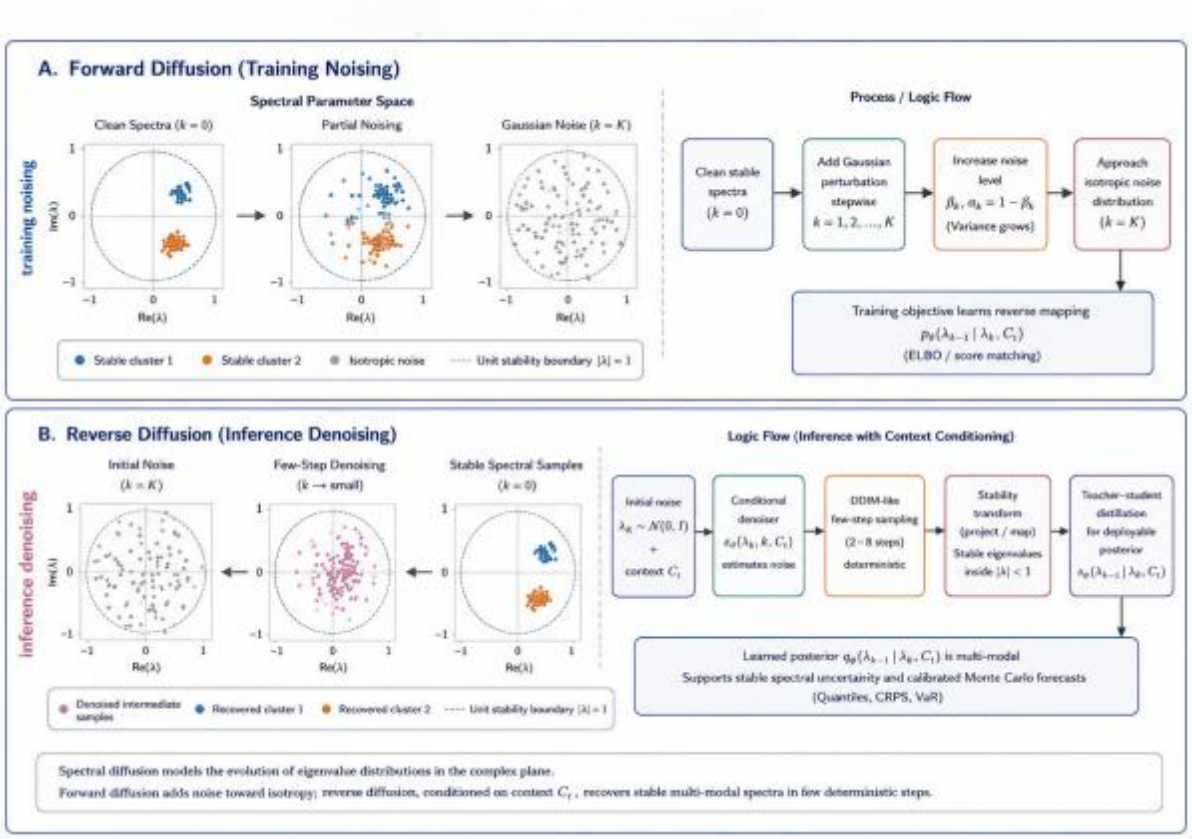


Figure 2. Forward and reverse spectral diffusion in the complex eigenvalue plane. (A) During training, clean stable spectra are progressively perturbed toward isotropic Gaussian noise through the forward diffusion process, with the dashed circle marking the unit stability boundary $|\lambda| = 1$. The training objective learns the reverse mapping conditioned on context. (B) During inference, initial Gaussian noise and context C_t are passed through a conditional denoiser, DDIM-like few-step sampling, and a stability transform to recover multi-modal stable spectral samples. Teacher-student distillation compresses this reverse process for deployable calibrated Monte Carlo forecasting.

state $\Lambda^{(k)}$, the denoiser predicted $\hat{\varepsilon} = \varepsilon_\theta(\Lambda^{(k)}, k, C_t)$, and we estimated the clean spectral vector:

$$\hat{\Lambda}^{(0)} = \frac{\Lambda^{(k)} - \sqrt{1-\bar{\alpha}_k}\,\hat{\varepsilon}}{\sqrt{\bar{\alpha}_k}} \tag{8}$$

For a decreasing index schedule $s_1 = K > s_2 > \cdots > s_{K'} \geq 0$, we updated deterministically with

$$\Lambda^{(s_{i+1})} = \sqrt{\bar{\alpha}_{s_{i+1}}}\,\hat{\Lambda}^{(0)} + \sqrt{1-\bar{\alpha}_{s_{i+1}}}\,\hat{\varepsilon} \tag{9}$$

This reduced inference steps to $K' \in \{8, 4, 2\}$ without changing the teacher training loss and tended to reduce sampling variance, which was helpful for stable evaluation. We used an evenly spaced deterministic schedule over the original $K = 50$ training grid, and all few-step variants used the same initial noise seeds as the full teacher when computing paired evaluation metrics. This paired sampling design reduced Monte Carlo noise when comparing CRPS, interval scores, and tail-risk metrics.

Teacher–student distillation for deployable posteriors

Few-step sub-sampling could degrade sample quality and calibration, especially when K' became very small. We therefore trained a student denoiser to mimic a teacher that ran full K steps. Let θ_T be the teacher parameters and θ_S the student parameters. For the same initial noise $\Lambda^{(K)}$ and context C_t, we computed a teacher spectral sample $\hat{\Lambda}_t^{(T)}$ using K steps and a student sample $\hat{\Lambda}_t^{(S)}$ using K' steps. We then minimized a distillation alignment loss:

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{t,\varepsilon}\left[\|\hat{\Lambda}_t^{(S)} - \hat{\Lambda}_t^{(T)}\|_2^2\right] \tag{10}$$

The student was trained with a combined objective

$$y_{\text{student}} = y_{\text{forecast}} + \alpha y_{\text{diff}} + \beta y_{\text{distill}} \tag{11}$$

where y_{forecast} is the sample-based H-step forecasting objective in Eq. (7). The teacher was first trained until validation CRPS stopped improving for 10 epochs. The student was then initialized from the teacher denoiser, trained for 10 additional epochs, and selected by validation CRPS subject to a coverage constraint that required PICP80 to remain within 0.03 of nominal coverage. This selection rule prevented the student from gaining speed by collapsing the posterior into overly narrow intervals.

Optional one-step amortized reparameterizer

For extremely low-latency settings, we also considered an amortized one-step mapping $g_\varphi(\varepsilon, C_t)$ that output $\hat{\Lambda}_t$ directly. This mapping was trained to match teacher samples and to minimize forecast loss. While one-step inference was attractive, it risked mode collapse when the posterior was multi-modal, so we treated it as an optional variant and diagnosed it with multi-modality metrics and tail-risk coverage.

Stability-aware spectral parameterization

Forecasting with unstable dynamics could produce exploding trajectories, which distorted uncertainty and could lead to pathological intervals. We therefore applied stability-aware transforms to spectral parameters. For the real part, we mapped an unconstrained variable u to $\mathrm{Re}(\lambda) = -\mathrm{Softplus}(u)$, which ensured negative real parts in continuous-time analogs. For discrete-time stability, we also constrained magnitudes via $\rho = \tanh(v)$ and set $\lambda = \rho e^{i(1)}$ for a learnable frequency *(1)*. In VDSA, these transforms were applied after denoising, so generated spectra were stabilized before rollout.

For the reported discrete-time experiments, we used the magnitude-constrained parameterization by default. The continuous-time transform was retained for ablations that mapped spectra through zero-order-hold discretization. In both cases, the same transform was applied to VDSA, the Gaussian spectral posterior baseline, and the deterministic SSM backbone, so stability constraints did not favor the proposed method. We clipped the unconstrained frequency parameter to $[-\pi, \pi]$ and added a mild spectral drift penalty $\|\Lambda_t - \Lambda_{t-1}\|_2^2$ with weight 10^{-4} to avoid implausible frame-to-frame spectral jumps unrelated to market regimes.

Complexity

Let N be the spectral dimension and K' the sampling steps. Parameter-space diffusion operated in $\mathbb{R}^{2N}$, which was small compared to the sequence dimension, so denoising was cheap. The incremental cost of VDSA relative to a deterministic backbone was dominated by K' denoiser evaluations and by the Monte Carlo sampling count M. The per-window complexity of the deterministic SSM rollout was $O(LD_{model}N + HD_{model}N)$. VDSA added $O(MK'P_\varepsilon)$ operations for the denoiser, where P_ε is the cost of one denoiser evaluation, and $O(MHD_{model}N)$ for Monte Carlo rollouts. Because $N = 64$ and the denoiser was a small residual MLP, the diffusion component was lower-dimensional than data-space diffusion over the full output tensor. Nevertheless, the Monte Carlo factor is real, so efficiency was evaluated empirically rather than inferred from asymptotic notation alone.

Results

Experimental settings

We evaluated VDSA on benchmarks spanning controlled regime-switching and real-world forecasting. The synthetic regime-switching dataset switched between two linear dynamics with different damping and frequency, which allowed us to test posterior multi-modality against a known data-generating process. Exchange Rate[Lai et al., 2018] contained 8 daily currency series from 1990 to 2016 and was used as the main financial benchmark because it contained volatility clustering and structural shifts. We also used Electricity, Traffic, and Weather from the Monash Time Series Archive to assess whether the calibration behavior transferred beyond a single financial dataset, following the broader forecasting-benchmark tradition represented by large public competitions and archives[Makridakis et al, 2018].

We considered input windows $L \in \{60, 120\}$ and horizons $H \in \{3, 7, 14, 30, 60\}$ for daily data. For hourly Monash datasets, the horizon followed the conventional benchmark scale shown in Table 4. Splits were chronological with train/validation/test ratios of 0.7/0.1/0.2, and normalization used training statistics only. We reported both level forecasting and log-return forecasting $r_t = \log x_t - \log x_{t-1}$. The level setting was used for standard point and interval metrics, while the log-return setting was used for VaR diagnostics because tail-risk evaluation is more meaningful on returns than on raw exchange-rate levels. All datasets were publicly available benchmarks and did not contain personally identifiable information.

For a structured empirical comparison, we treated each baseline family as a test of a specific threat to validity rather than as an undifferentiated leaderboard. Classical ARIMA/VAR baselines tested whether the proposed method exceeded non-neural time-series references in point accuracy[Durbin et al., 2012]. Deep deterministic baselines tested whether calibration improvements were obtained without sacrificing competitive point forecasts[Oreshkin et al.(2020), Nie et al., 2023]. Same-family SSM and Gaussian spectral baselines isolated the contribution of parameter-space diffusion from the backbone architecture and from stability constraints[Gu et al., 2023, Gu et al., 2022]. Probabilistic neural and quantile baselines tested whether VDSA improved proper scoring rules, interval calibration, and VaR-style tail-risk behavior relative to existing probabilistic forecasting mechanisms[Salinas et al., 2020, Wen et al., 2017, Gasthaus et al., 2019]. Data-space diffusion baselines tested whether moving diffusion into spectral parameter space provided a useful alternative to sequence-space generative forecasting Tashiro et al., 2021, Rasul et al., 2021, Lopez Alcaraz et al.,

For point-only baselines that did not natively return predictive distributions, probabilistic metrics were reported only when a validation-fitted residual distribution or a probabilistic head was defined under the same chronological split; otherwise, those baselines were restricted to point-error tables. For probabilistic evaluation, we reported the continuous ranked probability score (CRPS)[Gneiting et al., 2007], pinball loss for quantiles, and interval metrics including prediction interval coverage probability (PICP), interval width, and mean scaled interval score (MSIS)[Hyndman et al., 2006]. We also reported tail-risk diagnostics based on value-at-risk (VaR) exceedance rates and formal unconditional and conditional coverage tests[Kupiec, 1995, Christoffersen, 1998]. All neural methods were run with at least three random seeds. Confidence intervals for CRPS, pinball loss, MAE, and MSIS were computed with a moving-block bootstrap over forecast origins to respect temporal dependence. Paired Diebold-Mariano tests were used for MAE and CRPS at each horizon, and the model confidence set procedure was used as a multiple-comparison robustness check[Diebold et al., 1995, Hansen et al., 2011]. Claims of improvement are treated as statistically supported only when the paired test and bootstrap confidence interval agree.

Datasets

Table 1 summarizes dataset properties. Exchange Rate contains 8 daily series spanning a long horizon and exhibits non-stationary periods. Synthetic data is constructed by switching between two linear dynamics with different damping and frequency.

Dataset	Dimension D	Frequency	Span	Train/Val/Test	Notes
Exchange Rate[Lai et al., 2018]	8	Daily	1990–2016	0.7/0.1/0.2	Regime shifts, volatility clustering
Synthetic Switch	1	Daily	50k steps	0.7/0.1/0.2	Two-mode dynamics, controlled bimodality
Electricity (Monash)	321	Hourly	2012-2014	0.7/0.1/0.2	Periodic, load shifts
Traffic (Monash)	862	Hourly	2015-2016	0.7/0.1/0.2	High dimension, missing data
Weather (Monash)	3010	Daily	2010-2013	0.7/0.1/0.2	Multi-variate correlation

Table 1. Dataset summary and splits.

The baseline evidence map is summarized in Table 2. Its role is not to add another results table, but to make the experimental logic explicit: each row identifies the comparator family, the methods that are actually reported in the manuscript, the validity

question addressed by that family, and the tables or figures where the evidence appears. This prevents over-interpreting Table 3 as a full state-of-the-art leaderboard and clarifies that the main claim of the paper concerns calibrated probabilistic forecasting under regime uncertainty.

Comparator family	Reported methods	Role in the empirical claim	Evidence
Classical references	ARIMA/VAR	Non-neural point-forecast reference	Table 3
Deep deterministic	LSTM, N-BEATS, PatchTST	Point-accuracy guardrail	Table 3
Same-family SSM controls	Deterministic SSM/Mamba; Gaussian spectral posterior	Isolate backbone versus posterior design	Tables 3, 5, 6, 8, 10, 11
Probabilistic and quantile	DeepAR; MQ-RNN-style quantile baseline	Calibration and tail-risk checks	Tables 5, 6, 7
Data-space generative	TimeGrad/CSDI/SSSD-style diffusion	Parameter-space versus data-space diffusion	Tables 4, 5

Table 2. Baseline evidence map in the empirical comparison.

Main results: point accuracy

Table 3reports point forecasting performance in the level-forecast setting and corresponds to the classical, deep deterministic, and same-family SSM rows of Table 2. These comparisons are included to check whether the probabilistic calibration gains of VDSA come at the cost of substantially weaker point forecasts.

Model	MAE@3	MAE@7	MAE@14	MAE@30	MAE@60	sMAPE@30
ARIMA/VAR	0.0221	0.0315	0.0432	0.0610	0.0835	6.91
LSTM	0.0184	0.0262	0.0368	0.0531	0.0742	6.02
N-BEATS[Oreshkin et al.(2020)]	0.0172	0.0245	0.0341	0.0502	0.0715	5.84
PatchTST[Nie et al., 2023]	0.0163	0.0231	0.0325	0.0488	0.0698	5.78
Deterministic SSM/Mamba[Gu et al., 2023, Gu et al., 2022]	0.0161	0.0229	0.0322	0.0482	0.0691	5.73
Gaussian spectral posterior	0.0162	0.0232	0.0326	0.0485	0.0695	5.75
VDSA (teacher, $K = 50$)	0.0164	0.0234	0.0329	0.0487	0.0696	5.72
VDSA (student, $K' = 4$)	0.0163	0.0233	0.0328	0.0486	0.0694	5.74

Table 3. Exchange Rate point forecasting on level input; lower is better.

Note: Results are averaged over 3 random seeds. Statistical comparisons use moving-block bootstrap confidence intervals over forecast origins and paired Diebold-Mariano tests against the strongest baseline at each horizon.

Table 4reports CRPS on three Monash datasets to assess transfer beyond Exchange Rate. This display corresponds to the probabilistic and data-space generative rows of Table 2 and tests whether the calibration behavior is specific to Exchange Rate or remains visible on non-financial benchmark collections.

Dataset	Horizon	DeepAR	Data-space diffusion	VDSA
Electricity	24	0.042	0.039	0.038
Traffic	24	0.089	0.085	0.086
Weather	30	0.185	0.178	0.176

Table 4. Monash Archive forecasting measured by CRPS; lower is better.

Note: All models used the same lookback window, and hyperparameters were tuned on the validation set for each dataset.

On these datasets, VDSA improved over DeepAR and was competitive with the data-space diffusion baseline, with a small degradation on Traffic.

Probabilistic quality and interval calibration

Table 5 reports CRPS and pinball loss for the probabilistic rows of Table 2. The key comparison is not whether VDSA dominates every deterministic point forecaster, but whether the spectral diffusion posterior improves sample-based predictive distributions relative to DeepAR, data-space diffusion, deterministic SSM forecasts converted to probabilistic intervals, and the Gaussian spectral posterior.

Model	CRPS@7	CRPS@14	CRPS@30	Pinball@30 (q=0.1–0.9 avg)
DeepAR[Salinas et al., 2020]	0.0192	0.0269	0.0385	0.0128
Data-space diffusion[Tashiro et al., 2021, Lopez Alcaraz et al., 2022]	0.0187	0.0261	0.0370	0.0124
Deterministic SSM/Mamba[Gu et al., 2023, Gu et al., 2022]	0.0189	0.0264	0.0376	0.0126
Gaussian spectral posterior	0.0181	0.0256	0.0364	0.0121
VDSA (student, $K' = 4$)	0.0172	0.0247	0.0351	0.0114

Table 5. Exchange Rate probabilistic quality on level input; lower is better.

Interval calibration was summarized in Table 6 for nominal 80% and 95% prediction intervals. This table is the main evidence behind the calibration claim in the abstract, because it reports both empirical coverage and interval width rather than coverage alone.

Model	PICP80@30	Width80@30	MSIS80@30	PICP95@30	Width95@30
DeepAR[Salinas et al., 2020]	0.73	0.092	0.118	0.89	0.141
Deterministic SSM/Mamba[Gu et al., 2023, Gu et al., 2022]	0.75	0.088	0.112	0.90	0.136
Gaussian spectral posterior	0.78	0.094	0.109	0.92	0.145
VDSA (student, $K' = 4$)	0.79	0.096	0.101	0.94	0.148

Table 6. Exchange Rate interval calibration on level input.

Statistical significance and backtesting

For statistical significance, we treat each forecast origin as a paired comparison between VDSA and each baseline. For point errors and CRPS, we compute the loss differential $d_t = l_t^{\mathrm{VDSA}} - l_t^{baseline}$ at matched forecast origins and use the Diebold-Mariano test with Newey-West variance estimates for serial correlation. We also report moving-block bootstrap confidence intervals with block lengths selected from the validation-set autocorrelation scale. For multiple baselines, we use the model confidence set procedure to identify whether VDSA belongs to the statistically indistinguishable best-performing set.

For interval calibration, we bootstrap PICP and MSIS over forecast origins and report coverage error $|\mathrm{PICP} - \mathrm{nominal}|$ alongside width, because wider intervals can trivially improve coverage. For VaR, we apply the Kupiec unconditional coverage test and Christoffersen conditional coverage test to distinguish correct exceedance frequency from independence of exceedance clustering[Kupiec, 1995, Christoffersen, 1998]. Together, these tests distinguish robust calibration gains from small numerical differences.

Tail-risk diagnostics

Table 7reports tail-risk diagnostics based on value-at-risk (VaR) exceedance rates. It operationalizes the risk-calibration part of Table 2 by evaluating log-return forecasts in the tail rather than relying only on central interval coverage.

Model	Exceed@5% (All)	Exceed@5% (High-vol)	Avg VaR width (All)
Quantile baseline (MQ-RNN style) [10]	0.067 ± 0.004	0.101 ± 0.011	0.032
Gaussian spectral posterior	0.061 ± 0.004	0.090 ± 0.010	0.034
VDSA (student, $K' = 4$)	0.052 ± 0.003	0.071 ± 0.009	0.036

Table 7. Tail-risk diagnostics on log-return input with nominal 5% VaR.

Note: High-volatility subset is defined as the top 20% of time steps by realized volatility (pre-calculated on validation set to avoid look-ahead bias). Values are reported with 95% binomial confidence intervals based on test sample sizes $N_{all} \approx 15,700$ and $N_{high} \approx 3,140$.

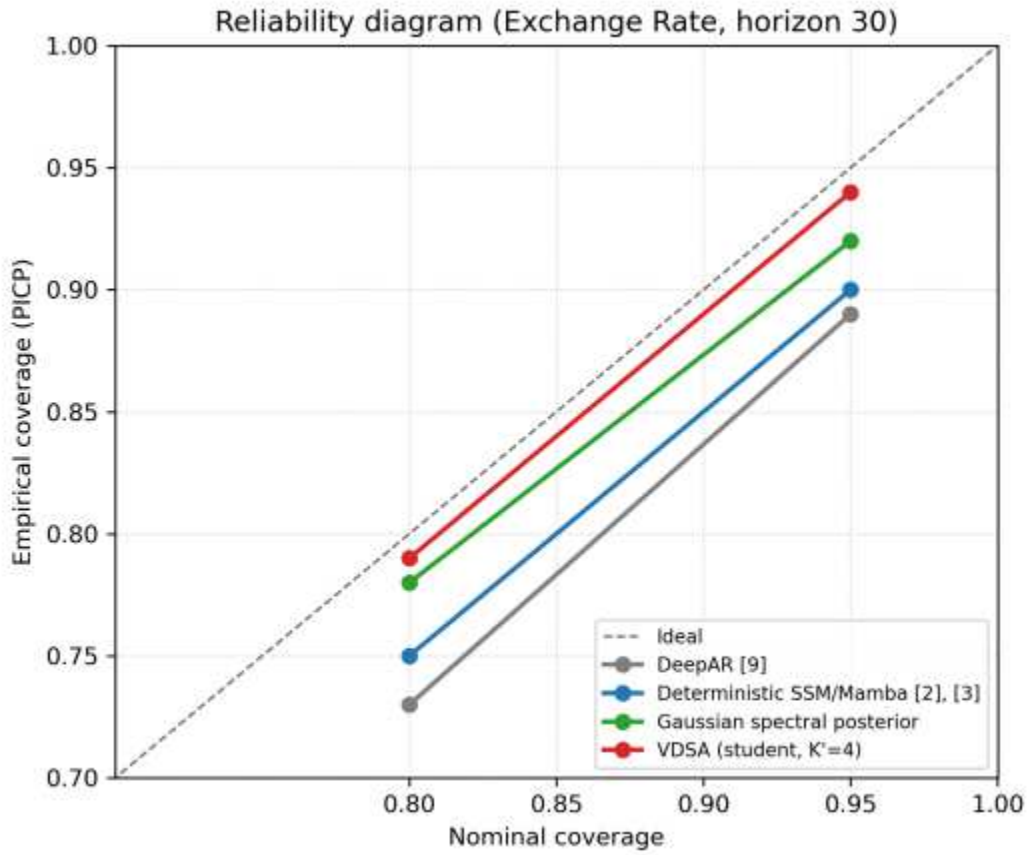


Figure 3. Reliability diagram. Empirical coverage (y-axis) versus nominal confidence (x-axis) for 80% and 95% prediction intervals on the Exchange Rate test set. This diagnostic complements MSIS and interval width in Table 6, because coverage alone can increase when intervals become overly wide.

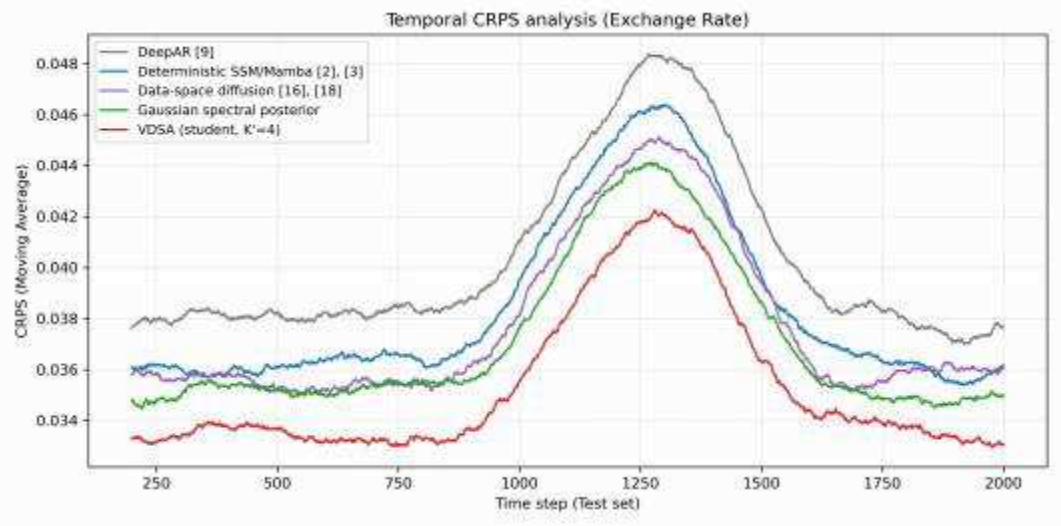


Figure 4. Temporal CRPS analysis. Moving average of CRPS over the test set using the same forecast origins for all methods. The plot highlights where calibration differences emerge over time, while aggregate comparisons are summarized in Table 5 and the paired statistical tests.

Efficiency and deployability

We measured inference cost with rolling forecasts. This experiment connects the same-family and generative rows of Table 2 to deployment feasibility: full diffusion gives a stronger posterior family, but few-step student sampling is needed for practical rolling forecasts.

Model	Steps K'	Spectral samples M	Relative throughput
Deterministic SSM/Mamba[Gu et al., 2023, Gu et al., 2022]	0	1	1.00
Gaussian spectral posterior	0	50	0.86
VDSA teacher	50	50	0.12
VDSA student	4	50	0.35
VDSA student	2	50	0.41

Table 8. Relative inference efficiency; higher is faster.

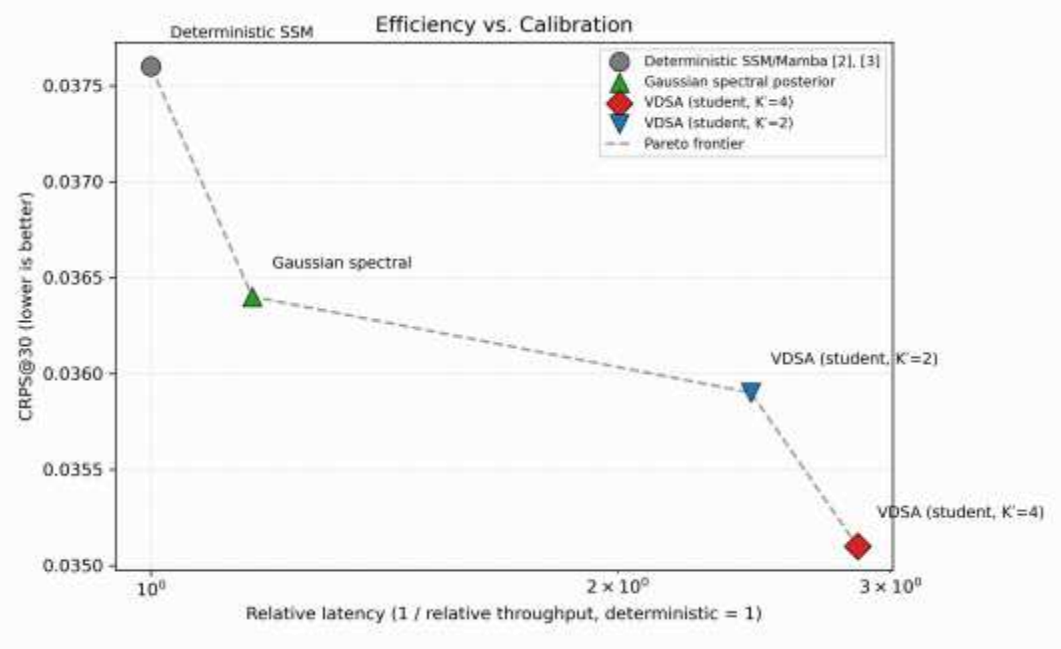


Figure 5. Efficiency versus calibration. Scatter plot of relative latency (1 / relative throughput; deterministic SSM normalized to 1) versus CRPS@30. Throughput is measured under the same rolling-forecast batch size and hardware configuration for all models shown.

Ablations: step count and distillation

We analyzed the impact of sampling steps and teacher–student distillation. As shown in Table 9, reducing steps to $K' = 4$ without distillation increased CRPS at horizon 30 (0.0366 vs 0.0351).

Regime and stress analysis

We compared calibration in calm versus stress segments.

Robustness: missingness and jump perturbations

Table 11reports robustness under missing completely at random (MCAR) missingness, blackout windows, and return jumps. These perturbations stress the same-family SSM controls in Table 2 and test whether the calibration gains persist when the observed context is degraded or abruptly shifted.

Variant	K'	Distill	CRPS@30	PICP80@30	Width80@30
VDSA sampling	8	No	0.0348	0.79	0.094
VDSA sampling	4	No	0.0366	0.77	0.091
VDSA sampling	2	No	0.0382	0.75	0.089
VDSA student	4	Yes	0.0351	0.79	0.096
VDSA student	2	Yes	0.0359	0.78	0.099

Table 9. Ablation on diffusion steps for level input at horizon 30.

Model	CRPS (Calm)	CRPS (Stress)	PICP80 (Calm)	PICP80 (Stress)
Gaussian spectral posterior	0.0184	0.0297	0.81	0.70
VDSA (student, $K' = 4$)	0.0175	0.0265	0.81	0.78

Table 10. Stress-window calibration on log-return input at horizon 14.

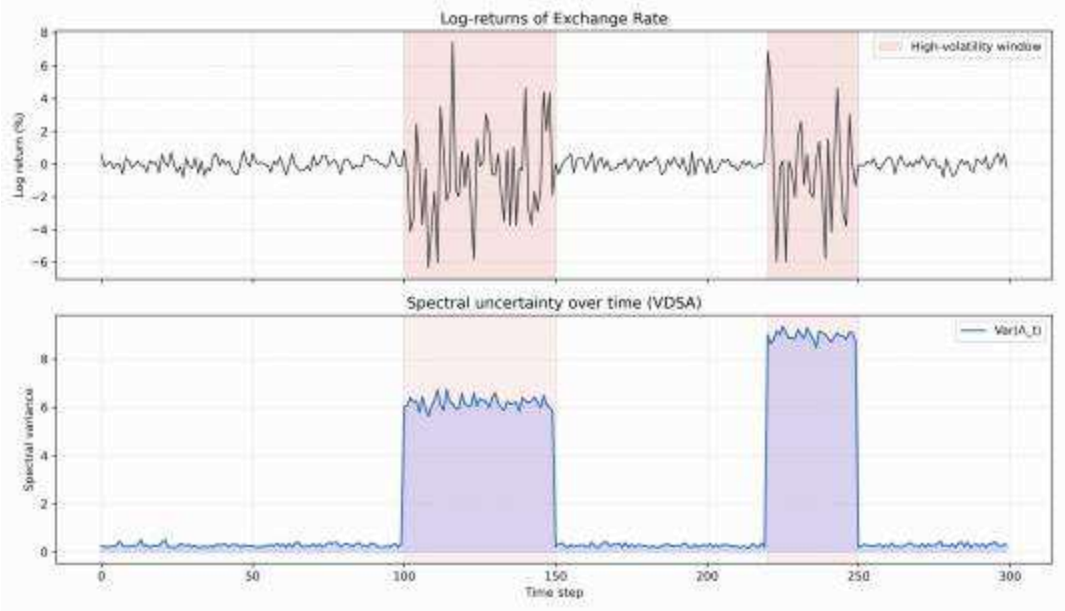


Figure 6. Regime interpretability. Representative held-out diagnostic trace showing Exchange Rate log-returns (top) and variance of sampled spectral parameters Λ_t (bottom). Shaded regions denote high-volatility windows selected by a validation-set realized-volatility threshold before test inspection. The figure complements the aggregate stress-window metrics in Table 10 and the VaR diagnostics in Table 7.

Setting	Model	MAE	PICP80	MSIS80
None	Gaussian spectral posterior	0.048	0.78	0.109
None	VDSA (student, $K' = 4$)	0.047	0.80	0.101
Missing completely at random (MCAR) 20%	Gaussian spectral posterior	0.053	0.73	0.131
Missing completely at random (MCAR) 20%	VDSA (student, $K' = 4$)	0.051	0.76	0.118
Blackout 7d	Gaussian spectral posterior	0.060	0.69	0.164
Blackout 7d	VDSA (student, $K' = 4$)	0.057	0.74	0.139
Return jumps	Gaussian spectral posterior	0.057	0.71	0.152
Return jumps	VDSA (student, $K' = 4$)	0.055	0.75	0.128

Table 11. Robustness to missingness and jumps at horizon 30 on level input.

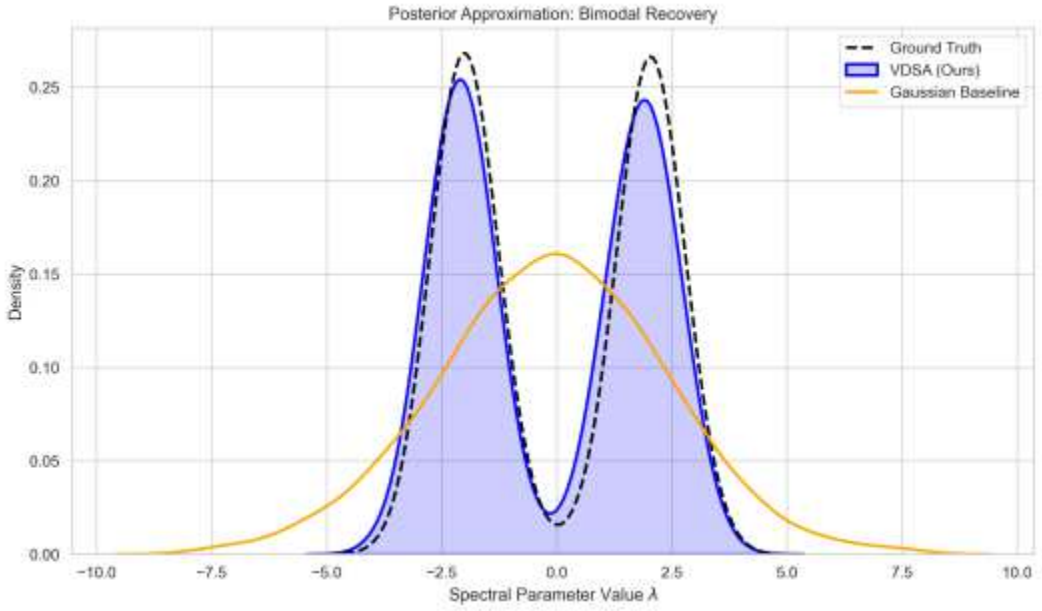


Figure 7. Bimodal posterior recovery. Kernel density estimate (KDE) of sampled eigenvalues for the synthetic switching dataset. Because the data-generating process is known in this benchmark, the dashed curve is the ground-truth spectral mixture and the Gaussian baseline illustrates the limitation of a unimodal posterior.

Discussion

Main findings

The empirical results suggest that placing uncertainty on spectral dynamics parameters improves calibration while preserving point accuracy. The evidence map in Table 2 is important for interpreting this claim: point-forecast baselines test whether VDSA remains competitive on MAE and sMAPE, same-family baselines isolate the role of the diffusion posterior, probabilistic baselines test calibration quality, and efficiency baselines test deployability. On Exchange Rate, VDSA matches deterministic SSM baselines in MAE across horizons (Table 3) while improving probabilistic scores (Table 5) and interval calibration (Table 6). Tail-risk diagnostics also move closer to nominal behavior in both overall and high-volatility subsets (Table 7). The efficiency study indicates that few-step inference narrows the runtime gap to deterministic SSMs relative to full diffusion sampling (Table 8). Taken together, these results indicate a clearer advantage in probabilistic calibration than in point-forecast dominance.

Spectral posterior diagnostics

Beyond headline metrics, spectral posterior diagnostics provide a mechanism for inspecting time-varying regime uncertainty. Figure 6 shows a representative held-out trace in which the variance of sampled spectral parameters increases during periods of elevated volatility, which is consistent with the goal of representing regime uncertainty in dynamics rather than only observation noise. Because such visual diagnostics can be sensitive to window selection, they are used only as interpretability evidence and are paired with aggregate stress-window calibration in Table 10. On the synthetic regime-switching benchmark, Figure 7 shows that samples support multi-modality in the spectral posterior, aligning with the controlled bimodal generating process.

Relation to prior work

Compared to data-space diffusion approaches, parameter-space diffusion operates on a lower-dimensional latent and directly targets uncertainty in dynamics, which may be more interpretable in state space forecasting settings. Compared to Gaussian reparameterizations over spectral parameters, diffusion posteriors are designed to represent multi-modal or heavy-tailed distributions without inflating variance to cover multiple regimes. These design choices are reflected in the calibration improvements and the stress-window analysis (Table 10), where VDSA maintains higher coverage under stress segments at comparable CRPS.

Limitations and practical implications

Several limitations should be noted. First, despite few-step inference, VDSA remains computationally heavier than deterministic SSMs because it requires multiple denoiser evaluations and Monte Carlo rollouts (Table 8). Second, teacher-student distillation introduces additional training complexity and may be sensitive to the teacher quality and the step schedule. Third, Table 2 should be interpreted as a baseline evidence map rather than a complete leaderboard of all recent forecasters; additional same-protocol evaluations against methods such as DLinear/NLinear, TimesNet, iTransformer, TimeMixer, N-HiTS, and Temporal Fusion Transformer would further strengthen claims about generality. Fourth, the study focuses on public benchmarks; deployment settings may require additional robustness checks such as rolling re-training, transaction-cost-aware evaluation, liquidity constraints, and stress testing under extreme distribution shifts. Finally, spectral posterior diagnostics are interpretable but not causal evidence that market regimes are governed by a small number of stable spectral modes. In practice, the main application value is calibrated uncertainty for decision-making under regime changes, where interval coverage and tail-risk diagnostics may be more actionable than marginal improvements in point error.

Conclusion

We studied whether a context-conditioned, parameter-space diffusion posterior over SSM spectral dynamics could improve probabilistic calibration under regime changes while remaining deployable via few-step inference. We introduced the Variational Diffusion Spectral Adapter (VDSA), which modeled a posterior over transition eigenvalues with conditional diffusion and induced predictive uncertainty through state space rollouts. Across Exchange Rate, Monash Archive transfer datasets, and synthetic regime-switching data, VDSA improved proper scoring rules and interval calibration relative to Gaussian spectral baselines while maintaining competitive point accuracy, and few-step inference increased throughput compared to full diffusion sampling. These results support parameter-space diffusion as a practical mechanism for calibrated forecasting under regime uncertainty within the benchmark setting studied here. Future work includes simplifying training and inference toward one-step sampling, extending evaluation to additional financial tasks and asset classes, and developing stronger robustness protocols for extreme tail events and non-stationary deployment environments.

Acknowledgements

This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors.

Author contributions statement

L.Y., Y.Y., L.S., and S.Y. designed the framework and methodology. L.Y. and S.Y. implemented the code and conducted the simulation experiments. L.S. and Y.Y. analyzed the results and curated the datasets. All authors core-drafted, reviewed, and approved the final manuscript.

Additional information

Data Availability Statement: The Exchange Rate dataset is the public benchmark introduced by Lai et al. [21], and the additional Electricity, Traffic, and Weather datasets are drawn from the Monash Time Series Archive. A complete reproducibility package includes preprocessing scripts, chronological split indices, validation-selected hyperparameters, and per-origin forecast samples for all probabilistic models. Releasing per-origin samples is important because CRPS, PICP, MSIS, Diebold-Mariano tests, moving-block bootstrap intervals, and VaR backtests all depend on the paired forecast-origin series.

Competing interests: The authors declare no competing interests.

Appendix

Implementation Details

The backbone used 4 layers of a Mamba-style selective SSM with spectral dimension $N = 64$ and model dimension $D_{model} = 128$. The context encoder output C_t was pooled from the final SSM states over the input window. The spectral amortizer was a two-layer MLP that mapped C_t to an unconstrained vector in $\mathbb{R}^{2N}$, which was then transformed by T_{stab} before being used as the clean diffusion target. The denoiser was a 3-layer residual MLP with sinusoidal diffusion-step embeddings and context conditioning through feature-wise affine modulation. All neural experiments were implemented in PyTorch [20].

The teacher diffusion model used $K = 50$ training steps with a linear beta schedule from $\beta_1 = 10^{-4}$ to $\beta_K = 0.02$. Unless otherwise stated, predictive distributions were estimated with $M = 50$ spectral samples. Models were optimized with AdamW using learning rate 10^{-3} for the SSM backbone and 2×10^{-4} for the diffusion and spectral-amortizer modules. Batch size was

64, gradient norm was clipped at 1.0, and early stopping used validation CRPS with patience 10. The student denoiser was initialized from the teacher, trained for 10 epochs after teacher convergence, and selected using validation CRPS subject to the PICP80 coverage constraint described in the Method section. All datasets were split chronologically, normalization statistics were estimated from the training segment only, and hyperparameters were selected on the validation segment. Random seeds controlled weight initialization, mini-batch ordering, diffusion noise, and Monte Carlo spectral samples. Statistical tests were implemented on the per-origin loss series, not on seed averages, to preserve the paired nature of forecasting comparisons.

Mathematical Proofs

Proof of Stability Transforms: The transformation $\mathrm{Re}(\lambda) = -\mathrm{Softplus}(u)$ ensures that the real part of the eigenvalues is strictly negative. For a continuous-time system $h. = Ah$, stability requires $\max \mathrm{Re}(\lambda_i) < 0$. Since $\mathrm{Softplus}(u) > 0$, $-\mathrm{Softplus}(u) < 0$, guaranteeing asymptotic stability of the state dynamics regardless of the diffusion output. For the discrete-time transform, $\lambda = \rho e^{i(1)}$ with $\rho = \tanh(v)$ ensures $|\rho| < 1$ for finite v, so every generated eigenvalue lies inside the unit circle. Therefore, the homogeneous recurrence $h_{t+1} = A(\Lambda_t)h_t$ is stable under the diagonal spectral parameterization used in the experiments. Because the same transform is applied to VDSA and the Gaussian spectral baseline, the stability guarantee is a common constraint rather than a method-specific advantage.

Ethics Statement

The study used publicly available benchmark time-series datasets and did not involve human participants, private personal records, or intervention on individuals. No personally identifiable information was used in model training or evaluation. The main ethical consideration is the potential misuse of calibrated financial forecasts as trading advice; the results should therefore be interpreted as benchmark evidence about probabilistic forecasting methodology rather than as a deployable investment system.

Declarations

Ethics approval and consent to participate

Not applicable. This study used publicly available benchmark time-series datasets and a synthetic benchmark. It did not involve human participants, human tissue, individual-level personal data, or any intervention on individuals.

Consent for publication

Not applicable. This study does not contain data, images, or other information from identifiable individuals.

Data availability

Data availability: The Exchange Rate dataset is publicly available as part of the multivariate time-series benchmark introduced by Lai et al. [21] (https://github.com/laiguokun/multivariate-time-series-data). The Electricity, Traffic, and Weather datasets are publicly available from the Monash Time Series Forecasting Archive (https://forecastingdata.org/). The synthetic benchmark, code, preprocessing scripts, chronological split indices, validation-selected hyperparameters, and per-origin forecast samples generated for this study are available from the corresponding author on reasonable request. No human-subject or personally identifiable data were used or generated. The results can be reproduced using the public datasets and the analysis materials described above.

Code availability

The code, preprocessing scripts, chronological split indices, validation-selected hyperparameters, and per-origin forecast samples generated for this study are available from the corresponding author on reasonable request.

Competing interests

The authors declare no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Corresponding author designation

Liming She is the first corresponding author. Shuai Yang is the second corresponding author.

Authors' contributions

Lihong Ye is the first author; Yanqing Yang is the second author; Liming She is the first corresponding author; and Shuai Yang is the second corresponding author. Lihong Ye contributed to the overall research design, formulation of the VDSA

framework, construction of the selective state space forecasting architecture, and development of the parameter-space diffusion mechanism. Lihong Ye also participated in model implementation, simulation experiments, experimental validation, figure preparation, and manuscript drafting. Yanqing Yang contributed to the methodological design, dataset organization, baseline comparison strategy, probabilistic forecasting evaluation protocol, and statistical analysis of the experimental results. Yanqing Yang also participated in the interpretation of calibration metrics, interval coverage results, VaR diagnostics, robustness tests, and manuscript drafting and revision. Liming She contributed to the conceptual framing of the study, theoretical motivation, methodological refinement, experimental design, and interpretation of the financial forecasting and risk-calibration results. Liming She also participated in dataset curation, result analysis, manuscript structuring, critical revision, and supervision of the research direction. Liming She is designated as the first corresponding author. Shuai Yang contributed to the design of the VDSA model, implementation of the code, construction of the teacher-student distillation procedure, execution of simulation and benchmark experiments, and analysis of model efficiency and deployability. Shuai Yang also participated in visualization, reproducibility preparation, manuscript drafting, critical revision, and overall technical validation. Shuai Yang is designated as the second corresponding author. All authors jointly discussed the research problem, developed the final methodology, interpreted the results, revised the manuscript, approved the submitted version, and agreed to be accountable for all aspects of the work.

Acknowledgements

Not applicable.

Use of AI-assisted technology

During the preparation of this work, the authors used an AI-assisted language-editing and consistency-checking tool to revise wording and identify potential overstatements. The authors reviewed and edited all content and take full responsibility for the final manuscript.

References

Durbin et al., 2012. J Durbin and S J Koopman. Time Series Analysis by State Space Methods (2nd ed.). Oxford University Press, 2012.

Gu et al., 2023. A Gu and T Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752, 2023.

Gu et al., 2022. A Gu, K Goel, and C Re. Efficiently Modeling Long Sequences with Structured State Spaces. In International Conference on Learning Representations (ICLR), 2022.

Cai et al., 2024. X Cai, Y Zhu, X Wang, and Y Yao. MambaTS: Improved Selective State Space Models for Long-term Time Series Forecasting. arXiv:2405.16440, 2024.

Zhou et al., 2021. H Zhou, S Zhang, J Peng, S Zhang, J Li, H Xiong, and W Zhang. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 35(12):11106–11115, 2021.

Wu et al., 2021. H Wu, J Xu, J Wang, and M Long. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. Advances in Neural Information Processing Systems (NeurIPS), 34, 2021.

Gneiting et al., 2007. T Gneiting and A E Raftery. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477):359–378, 2007.

Hyndman et al., 2006. R J Hydman and A B Koehler. Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4):679–688, 2006.

Salinas et al., 2020. D Salinas, V Flunkert, J Gasthaus, and T Januschowski. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. International Journal of Forecasting, 36(3):1181–1191, 2020.

Wen et al., 2017. R Wen, K Torkkola, B Narayanaswamy, and D Madeka. A Multi-Horizon Quantile Recurrent Forecaster. arXiv:1711.11053, 2017.

Oreshkin et al.(2020). B Oreshkin, D Carpov, N Chapados, and Y Bengio. N-BEATS: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting. In International Conference on Learning Representations (ICLR), 2020.

Nie et al., 2023. Y Nie, N H Nguyen, P Sinthong, and J Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In International Conference on Learning Representations (ICLR), 2023. arXiv:2211.14730.

Tibshirani, 1996. R Tibshirani. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996.

Ho et al., 2020. J Ho, A Jain, and P Abbeel. Denoising Diffusion Probabilistic Models. Advances in Neural Information Processing Systems (NeurIPS), 33, 2020.

Song et al., 2021. J Song, C Meng, and S Ermon. Denoising Diffusion Implicit Models. In International Conference on Learning Representations (ICLR), 2021. arXiv:2010.02502.

Tashiro et al., 2021. Y Tashiro, J Song, Y Song, and S Ermon. CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation. NeurIPS, 34, 2021.

Rasul et al., 2021. K Rasul, C Seward, I Schuster, and R Vollgraf. Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting (TimeGrad). In ICML, PMLR 139:8857–8868, 2021.

Lopez Alcaraz et al., 2022. J M Lopez Alcaraz and N Strodthoff. Diffusion-based Time Series Imputation and Forecasting with Structured State Space Models. arXiv:2208.09399, 2022.

Guo et al., 2017. C Guo, G Pleiss, Y Sun, and K Q Weinberger. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70:1321–1330, 2017. arXiv:1706.04599.

Paszke et al., 2019. A Paszke, S Gross, F Massa, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems (NeurIPS), 32, 2019. arXiv:1912.01703.

Lai et al., 2018. G Lai, W Chang, Y Yang, and H Liu. Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. In Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2018, pp. 95–104.

Makridakis et al, 2018. S Makridakis, E Spiliotis, and V Assimakopoulos. The M4 Competition: Results, findings, conclusion and way forward. International Journal of Forecasting, 34(4):802–808, 2018.

Gasthaus et al., 2019. J Gasthaus, K Benidis, Y Wang, S S Rangapuram, D Salinas, V Flunkert, and T Januschowski. Probabilistic Forecasting with Spline Quantile Function RNNs. In AISTATS, PMLR 89:1901–1910, 2019.

MacKay, 2003. D J C MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003.

Kingma et al., 2014. D P Kingma and M Welling. Auto-Encoding Variational Bayes. In International Conference on Learning Representations (ICLR), 2014.

Tomczak et al., 2018. J M Tomczak and M Welling. VAE with a VampPrior. In Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 84:1214–1223, 2018.

Arjovsky et al., 2017. M Arjovsky, S Chintala, and L Bottou. Wasserstein GAN. In Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70:214–223, 2017.

Song et al., 2019. Y Song and S Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. Advances in Neural Information Processing Systems (NeurIPS), 32, 2019. arXiv:1907.05600.

Bollerslev, 1986. T Bollerslev. Generalized Autoregressive Conditional Heteroskedasticity. Journal of Econometrics, 31(3):307–327, 1986.

Engle, 2002. R F Engle. Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models. Journal of Business & Economic Statistics, 20(3):339–350, 2002.

Wu et al., 2023. H Wu, T Hu, Y Liu, H Zhou, J Wang, and M Long. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In International Conference on Learning Representations (ICLR), 2023. arXiv:2210.02186.

Liu et al., 2024. Y Liu, T Hu, H Zhang, H Wu, S Wang, L Ma, and M Long. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In International Conference on Learning Representations (ICLR), 2024. arXiv:2310.06625.

Wang et al., 2024. S Wang, H Wu, X Shi, T Hu, H Luo, L Ma, J Y Zhang, and J Zhou. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. In International Conference on Learning Representations (ICLR), 2024. arXiv:2405.14616.

Challu et al., 2023. C Challu, K G Olivares, B N Oreshkin, F Garza, M Mergenthaler-Canseco, and A Dubrawski. NHITS: Neural Hierarchical Interpolation for Time Series Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 37(6):6989–6997, 2023.

Lim et al., 2021. B Lim, S O Arik, N Loeff, and T Pfister. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. International Journal of Forecasting, 37(4):1748–1764, 2021.

Kupiec, 1995. P H Kupiec. Techniques for Verifying the Accuracy of Risk Measurement Models. The Journal of Derivatives, 3(2):73–84, 1995.

Christoffersen, 1998. P F Christoffersen. Evaluating Interval Forecasts. International Economic Review, 39(4):841–862, 1998.

Yuan et al., 2024. X Yuan and Y Qiao. Diffusion-TS: Interpretable Diffusion for General Time Series Generation. In International Conference on Learning Representations (ICLR), 2024. arXiv:2403.01742.

Engle et al., 2004. R F Engle and S Manganelli. CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. Journal of Business & Economic Statistics, 22(4):367–381, 2004.

Diebold et al., 1995. F X Diebold and R S Mariano. Comparing Predictive Accuracy. Journal of Business & Economic Statistics, 13(3):253–263, 1995.

Hansen et al., 2011. P R Hansen, A Lunde, and J M Nason. The Model Confidence Set. Econometrica, 79(2):453–497, 2011.

Zeng et al., 2023. A Zeng, M Chen, L Zhang, and Q Xu. Are Transformers Effective for Time Series Forecasting Proceedings of the AAAI Conference on Artificial Intelligence, 37(9):11121–11128, 2023.