Collaborative Deterministic-Probabilistic Learning for Real-World Spatiotemporal Dynamics

Zhi Sheng[1], Yuan Yuan[2], Yudi Zhang[3], Jingtao Ding[1], Yong Li[1]
[1]Center for Urban Science and Computation, Tsinghua University
[2]Courant Institute of Mathematical Sciences, New York University
[3]School of Architecture, Tsinghua University
y.y@nyu.edu,liyong07@tsinghua.edu.cn

Abstract

Probabilistic forecasting is crucial for high-stakes real-world spatiotemporal systems, such as climate dynamics, renewable energy integration, and urban mobility, where quantifying uncertainty is essential for reliable, risk-aware decision-making. While governed by underlying physical laws, these chaotic systems are dominated by high-frequency stochastic fluctuations and spatially heterogeneous variance. Standard deterministic methods effectively capture mean trends but fail to quantify intrinsic uncertainty, while pure generative AI often struggles with computational scalability and the risk of distorting temporal consistency. In this work, we propose **CoST**, a **Co**llaborative framework designed for diverse **S**patio**T**emporal scientific systems. Drawing on the Reynolds decomposition in fluid dynamics, CoST separates a system's evolution into a predictable "mean flow" and turbulent "residual fluctuations". We leverage a powerful deterministic backbone to capture governing physical trends, paired with a scale-aware diffusion model dedicated to learning the residual uncertainties while adapting to spatial heterogeneity. Extensive experiments across ten real-world datasets from climate, energy, communication, and urban systems show that CoST achieves 25% performance gains over state-of-the-art baselines, while reducing computational overhead by over 10×. This work offers a scalable, scientifically consistent solution for trustworthy forecasting in data-rich scientific domains. Code and datasets are available at https://github.com/tsinghua-fib-lab/CoST.

1 Introduction

Quantifying uncertainty in chaotic spatiotemporal systems, such as climate dynamics, renewable energy integration, and urban mobility, is critical for trustworthy scientific decision-making [14, 35, 48, 55, 59]. Unlike static data, these systems evolve with complex physical or social laws, where accurate forecasting is essential for risk management, such as disaster preparedness in climate science and grid stability in energy systems [6, 42, 58, 64]. While deterministic methods effectively capture mean trends by minimizing reconstruction errors (e.g., MSE) [38, 71, 73], they fail to quantify the intrinsic stochasticity of these systems. Conversely, probabilistic methods aim to learn the full predictive distribution [30, 47, 70], providing the uncertainty estimates necessary for high-stakes domains.

However, developing effective probabilistic models for scientific spatiotemporal data faces three distinct challenges. First, these systems exhibit coupled deterministic-stochastic dynamics, characterized by a superposition of primary physical laws (e.g., seasonal cycles) and intrinsic stochastic fluctuations [9, 70]. Second, they involve profound spatial heterogeneity in uncertainty intensity [24, 71]. For instance, the stochastic fluctuations in coastal wind farms or urban transit hubs are far more volatile than in inland stations or residential zones [24, 71]. Third, real-world scientific applications demand computational efficiency, forecasts must be generated rapidly to support real-time simulation and decision-making [43, 54]. Recently, diffusion models have emerged as a powerful tool for probabilistic forecasting due to their ability to model complex distributions [47, 54, 62]. Yet, applying vanilla diffusion models to scientific forecasting remains computationally expensive and often struggles to capture long-term temporal consistency, as the progressive noise corruption can distort underlying physical trends [50, 53, 70]. While recent works attempt to guide diffusion with temporal constraints [47, 62], they typically force a single model to learn both the stable physical laws and the high-frequency noise, leading to sub-optimal convergence and high inference costs.

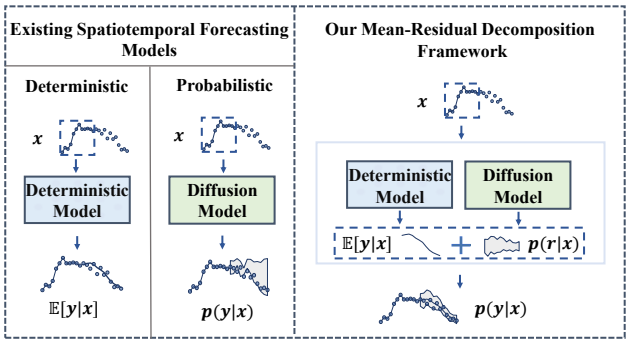

Figure 1: Comparison of existing models with our mean-residual decomposition framework.

To address these limitations, we draw inspiration from the Reynolds decomposition in fluid dynamics [44], which separates a flow field into a time-averaged mean component and a fluctuating turbulent component. We argue that spatiotemporal forecasting should follow a similar paradigm: a collaborative approach where a deterministic model captures the predictable "mean flow" (physical trends), and a probabilistic model focuses solely on the "residual fluctuations" (uncertainty). Our design offers two key advantages. First, it leverages the efficiency of established deterministic backbones to model the primary dynamics. Second, it allows the diffusion model to focus its generative capacity exclusively on the residuals, significantly reducing the complexity of the distribution it needs to learn.

Building on this insight, we propose **CoST**, a general framework that **Co**llaborates deterministic and diffusion models for diverse **S**patio**T**emporal systems. As illustrated in Figure 1, CoST first employs a deterministic backbone to estimate the conditional mean

$\mathbb{E}[y|x]$, capturing regular spatiotemporal patterns. Subsequently, a lightweight diffusion model learns the residual distribution $p(r|x)$, providing fine-grained uncertainty quantification. To handle the spatial heterogeneity of fluctuation magnitudes, we introduce a scale-aware diffusion mechanism that conditions the generation process on location-specific variance priors, allowing the model to adaptively infer residual intensities for different locations. We validate CoST across ten datasets in climate, energy, and urban systems. Crucially, in Sea Surface Temperature (SST) analysis, CoST correctly isolates high-variance regions associated with the thermocline gradients of ENSO, demonstrating its potential as a trustworthy tool for earth science discovery. In summary, our main contributions are as follows:

- We introduce a physics-inspired perspective for spatiotemporal modeling that mirrors Reynolds decomposition, integrating deterministic trend learning with probabilistic residual modeling in a collaborative framework.
- We propose **CoST**, a unified framework that employs a deterministic backbone for mean estimation and a scale-aware diffusion model for residual uncertainty. This design effectively handles spatial heterogeneity and simplifies the generative learning task.
- We conduct extensive evaluations on ten real-world datasets across climate science, energy systems, communication networks, and urban environments. Results demonstrate that CoST achieves an average improvement of 25% in probabilistic calibration and accuracy over state-of-the-art baselines while reducing computational overhead by over 10×, validating its potential for AI-accelerated scientific discovery.

2 Related Work

We provide definitions and related work on spatiotemporal deterministic forecasting and probabilistic forecasting in Appendix A.

Spatiotemporal Forecasting in Scientific Domains. Forecasting spatiotemporal dynamics is fundamental to disciplines ranging from climate modeling [22, 35, 55] to energy grid management [23, 59, 67]. Traditionally, these fields rely on process-based solvers, such as Numerical Weather Prediction (NWP), which simulate physical states via complex partial differential equations (PDEs) [4, 7, 29]. While physically rigorous, their high computational cost often limits its real-time application [4, 29]. To address these bottlenecks, data-driven surrogates like CNNs [28, 65] and Transformers [41, 63] have emerged as efficient alternatives [22, 23, 67]. However, these deterministic methods yield simple point estimates, failing to quantify the uncertainty essential for high-stakes risk assessment (e.g., extreme weather) [35, 48, 72]. Consequently, Diffusion Models have gained prominence for their superior ability to model complex, multi-modal distributions compared to GANs or VAEs [32, 35]. These models have achieved state-of-the-art performance across diverse scientific domains, including climate, energy, and human mobility simulation [14, 35, 48, 55, 59].

Diffusion-based spatiotemporal forecasting. While diffusion models excel at modeling complex distributions, their application to scientific spatiotemporal forecasting faces challenges in maintaining temporal coherence and physical consistency. Pure diffusion approaches often treat forecasting as conditional generation, which can lead to hallucinations that violate physical continuity or distort long-term trends [17, 57, 62, 70]. Although recent works have attempted to mitigate this by injecting temporal priors or redefining denoising transitions [47, 50, 53, 54], these methods typically burden a single model with the dual task of capturing both deterministic laws and stochastic fluctuations, resulting in computational inefficiency and suboptimal convergence. To address these limitations, a growing trend involves hybridizing deterministic backbones with diffusion models. In climate science, approaches like CorrDiff and CasCast use diffusion to refine or downscale coarse deterministic predictions [20, 39]. Similarly, in time series forecasting, methods such as TMDM and DiffCast condition the diffusion process on representations from deterministic Transformers [30, 69]. Distinguishing our framework from these approaches, CoST goes beyond simple refinement or coupled training. We introduce a generalized Mean-Residual Decomposition strategy that structurally decouples the task: a deterministic backbone captures stable physical trends (mean flow), while a diffusion model focuses exclusively on the residual distribution. Unlike prior domain-specific hybrids, CoST is a general framework that incorporates a scale-aware mechanism to explicitly handle spatial heterogeneity, making it broadly applicable across diverse scientific systems.

3 Preliminaries

We provide a summary of notations used in this paper in Appendix B.1 for clarity.

Spatiotemporal systems. Spatiotemporal systems underpin many domains such as climate science, energy, communication networks, and urban environments. The data recording spatiotemporal dynamics are typically represented as a tensor $\mathbf{x} \in \mathbb{R}^{T\times V\times C}$, where T, V, and C denote the temporal, spatial, and feature dimensions, respectively. Depending on the spatial structure, the data can be organized as grid-structured ($V = H \times W$) or graph-structured (where V represents the set of nodes). Given a historical context $\mathbf{x}^{co} = \mathbf{x}^{t-M+1:t}$ of length M, the goal is to predict future targets $\mathbf{x}^{ta} = \mathbf{x}^{t+1:t+P}$ over a horizon P using a model $\mathcal{F}$.

Conditional diffusion models. The diffusion-based forecasting includes a forward process and a reverse process. In the forward process, noise is added incrementally to the target data $\mathbf{x}_0^{ta}$, gradually transforming the data distribution into a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. At any diffusion step, the corrupted target data can be computed using the one-step forward equation:

$$\mathbf{x}_n^{ta} = \sqrt{\bar{\alpha}_n}\mathbf{x}_0^{ta} + \sqrt{1-\bar{\alpha}_n}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\bar{\alpha}_n = \prod_{i=1}^{n} \alpha_i$ and $\alpha_n = 1 - \beta_n$. In the reverse process, prediction begins by first sampling $\mathbf{x}_N^{ta}$ from the standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, followed by a denoising procedure through the following Markov process:

$$\begin{aligned} p_\theta(\mathbf{x}_{0:N}^{ta}) &:= p(\mathbf{x}_N^{ta}) \prod_{n=1}^{N} p_\theta(\mathbf{x}_{n-1}^{ta}|\mathbf{x}_n^{ta}, \mathbf{x}_0^{co}), \\ p_\theta(\mathbf{x}_{n-1}^{ta}|\mathbf{x}_n^{ta}) &:= \mathcal{N}(\mathbf{x}_{n-1}^{ta}; \mu_\theta(\mathbf{x}_n^{ta}, n|\mathbf{x}_0^{co}), \Sigma_\theta(\mathbf{x}_n^{ta}, n)), \\ \mu_\theta(\mathbf{x}_n^{ta}, n|\mathbf{x}_0^{co}) &= \frac{1}{\sqrt{\bar{\alpha}_n}}\left(\mathbf{x}_n^{ta} - \frac{\beta_n}{\sqrt{1-\bar{\alpha}_n}}\epsilon_\theta(\mathbf{x}_n^{ta}, n|\mathbf{x}_0^{co})\right) \end{aligned} \tag{2}$$

where the variance $\Sigma_\theta(\mathbf{x}_n^{ta}, n) = \frac{1-\bar{\alpha}_{n-1}}{1-\bar{\alpha}_n}\beta_n$, and $\epsilon_\theta(\mathbf{x}_n^{ta}, n|\mathbf{x}_0^{co})$ is predicted by the denoising network trained by the loss function below:

$$\mathcal{L}(\theta) = \mathbb{E}_{n,\mathbf{x}_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta(\mathbf{x}_n^{ta}, n|\mathbf{x}_0^{co})\right\|_2^2\right]. \tag{3}$$

Evaluations of probabilistic forecasting. We argue that probabilistic forecasting should be assessed from two key perspectives: *Data Distribution*—the predicted distribution should match the empirical distribution, and *Prediction Usability*—prediction intervals should achieve high coverage while remaining sharp. While metrics like CRPS, MAE and RMSE are widely used, they fail to assess: **(i)** the accuracy of quantile-wise coverage; **(ii)** whether the interval width reflects true uncertainty. To address this, we introduce Quantile Interval Coverage Error (QICE) [21] and Interval Score (IS) [19] as complementary metrics.

(i) QICE measures the mean absolute deviation between the empirical and expected proportions of ground-truth values falling into each of equal-sized quantile intervals. QICE evaluates how well the predicted distribution aligns with the expected coverage across quantiles, which is defined as follows:

$$\begin{aligned}\text{QICE} &:= \frac{1}{M_{\text{QIs}}}\sum_{m=1}^{M_{\text{QIs}}}\left|r_m - \frac{1}{M_{\text{QIs}}}\right|,\\ r_m &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{1}_{y_n \geq \hat{y}_n^{\text{low}_m}} \cdot \mathbb{1}_{y_n \leq \hat{y}_n^{\text{high}_m}},\end{aligned} \tag{4}$$

where $\hat{y}_n^{\text{low}_m}$ and $\hat{y}_n^{\text{high}_m}$ denote the bounds of the m-th quantile interval for y_n. Ideally, each QI should contain $1/M_{\text{QIs}}$ of the observations, yielding a QICE of 0. Lower QICE indicates better alignment between predicted and true distributions.

(ii) IS evaluates prediction interval (PI) quality by jointly accounting for sharpness and empirical coverage, and is defined as:

$$\begin{aligned}\text{IS} := \frac{1}{N}\sum_{n=1}^{N}\Bigg[&(u_n^{\alpha_{CI}} - l_n^{\alpha_{CI}}) + \frac{2}{\alpha_{CI}}(l_n^{\alpha_{CI}} - y_n)\mathbb{1}_{y_n < l_n^{\alpha_{CI}}}\\ &+ \frac{2}{\alpha_{CI}}(y_n - u_n^{\alpha_{CI}})\mathbb{1}_{y_n > u_n^{\alpha_{CI}}}\Bigg],\end{aligned} \tag{5}$$

where $u_n^{\alpha_{CI}}$ and $l_n^{\alpha_{CI}}$ are the upper and lower bounds of the central prediction interval for the n-th data point, derived from the corresponding predictive quantiles. A narrower interval improves the score, while missed coverage incurs a penalty scaled by α_{CI}. Lower IS indicates better performance.

4 Methodology

In this section, we propose CoST, a unified framework that combines the strengths of deterministic and diffusion models. Specifically, we first train a deterministic model to predict the conditional mean, capturing the regular spatiotemporal patterns. Then, guided by a customized fluctuation scale, we employ a scale-aware diffusion model to learn the residual distribution, enabling fine-grained uncertainty modeling. An overview of the CoST architecture is shown in Figure 2.

4.1 Theoretical Analysis of Mean-Residual Decomposition

Current diffusion-based probabilistic forecasting approaches typically employ a single diffusion model to capture the full distribution of data, incorporating both the regular spatiotemporal patterns and the random fluctuations. However, jointly modeling these components remains challenging [70]. Inspired by Mardani et al. [39] and the Reynolds decomposition in fluid dynamics [44], we propose to divide probabilistic forecasting into two parts: predicting the conditional mean and modeling the residual distribution. The spatiotemporal data $\mathbf{x}^{ta}$ can therefore be expressed as:

$$\mathbf{x}^{ta} = \underbrace{\mathbb{E}[\mathbf{x}^{ta}|\mathbf{x}^{co}]}_{:=\boldsymbol{\mu}(Deterministic)} + \underbrace{(\mathbf{x}^{ta} - \mathbb{E}[\mathbf{x}^{ta}|\mathbf{x}^{co}])}_{:=\mathbf{r}(Diffusion)}, \tag{6}$$

where $\boldsymbol{\mu}$ is the conditional mean representing the regular patterns, and $\mathbf{r}$ is the residual representing the random variations. If the deterministic model approximates the conditional mean accurately, the expected residual becomes negligible, i.e. $\mathbb{E}[\mathbf{r}|\mathbf{x}^{co}] \approx 0$, and we can obtain that $\text{var}(\mathbf{r}|\mathbf{x}^{co}) = \text{var}(\mathbf{x}^{ta}|\mathbf{x}^{co})$. Based on the law of total variance [5], we can express the variance of the target data and residuals as:

$$\begin{aligned}\text{var}(\mathbf{r}) &= \mathbb{E}[\text{var}(\mathbf{r}|\mathbf{x}^{co})] + \underbrace{\text{var}(\mathbb{E}[\mathbf{r}|\mathbf{x}^{co}])}_{=0},\\ \text{var}(\mathbf{x}^{ta}) &= \mathbb{E}[\text{var}(\mathbf{x}^{ta}|\mathbf{x}^{co})] + \underbrace{\text{var}(\mathbb{E}[\mathbf{x}^{ta}|\mathbf{x}^{co}])}_{\geq 0}.\end{aligned} \tag{7}$$

Due to $\text{var}(\mathbf{r}|\mathbf{x}^{co}) = \text{var}(\mathbf{x}^{ta}|\mathbf{x}^{co})$, we have $\text{var}(\mathbf{r}) \leq \text{var}(\mathbf{x}^{ta})$. Moreover, the highly dynamic nature of the spatiotemporal system results in a larger $\text{var}(\mathbb{E}[\mathbf{x}^{ta}|\mathbf{x}^{co}])$, which consequently makes $\text{var}(\mathbf{r})$ smaller compared to $\text{var}(\mathbf{x}^{ta})$. Our core idea is that if a deterministic model can accurately predict the conditional mean, that is, $\boldsymbol{\mu} \approx \mathbb{E}_\theta[\mathbf{x}^{ta}|\mathbf{x}]$, then the diffusion model can be dedicated solely to learning the simpler residual distribution. This design avoids the challenge diffusion models face in modeling complex spatiotemporal dynamics, while fully exploiting their strength in uncertainty estimation. By collaborating high-performing deterministic architectures and diffusion models, our method effectively captures regular dynamics and models uncertainty via residual learning.

4.2 Mean Prediction via Deterministic Model

To capture the conditional mean $\mathbb{E}_\theta[\mathbf{x}^{ta}|\mathbf{x}^{co}]$, our framework leverages existing high-performance deterministic architectures, which are designed to capture complex spatiotemporal dynamics efficiently. In our main experiments, we use the STID [52] model as the backbone for mean prediction, and also validate our framework with ConvLSTM [56], STNorm [16], and iTransformer [37] to ensure its generality (See Section 5.4). In the first stage of training, we pretrain the deterministic model for 50 epochs using historical conditional inputs $\mathbf{x}^{co}$ to output the mean estimate $\mathbb{E}_\theta[\mathbf{x}^{ta}|\mathbf{x}^{co}]$. The model is trained with the standard $\mathcal{L}_2$ loss:

$$\mathcal{L}_2 = \left\|\mathbb{E}_\theta[\mathbf{x}^{ta}|\mathbf{x}^{co}] - \mathbf{x}^{ta}\right\|_2^2. \tag{8}$$

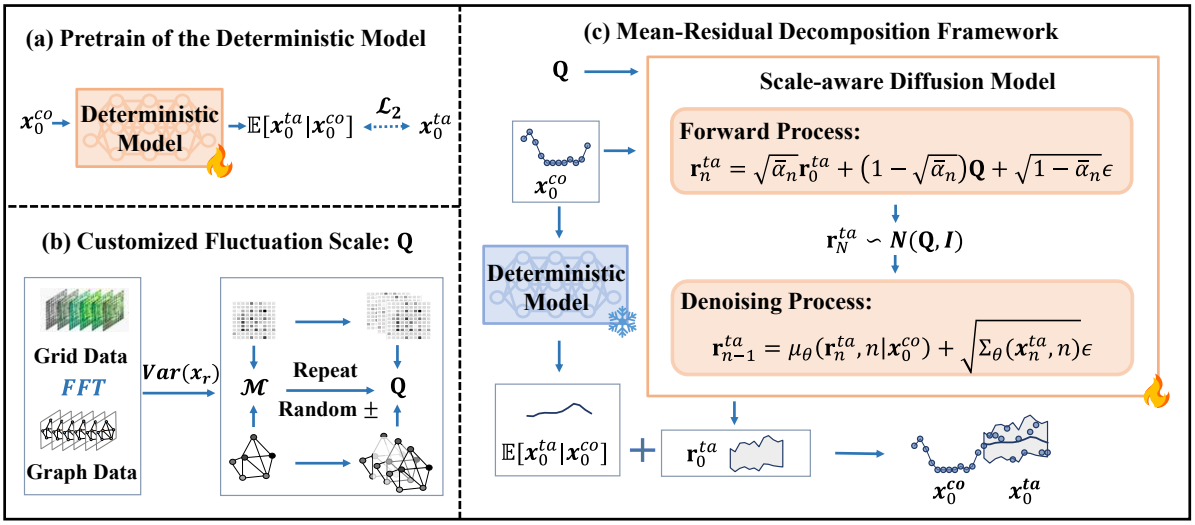

Figure 2: Overview of CoST: (a) Pretraining of the deterministic model; (b) Computation of the customized fluctuation scale; (c) Overall framework of the mean-residual decomposition.

4.3 Residual Learning via Diffusion Model

The residual distribution of spatiotemporal data is not independently and identically distributed (i.i.d.) nor does it follow a fixed distribution, such as $\mathcal{N}(0, \sigma)$. Instead, it often exhibits complex spatiotemporal dependence and heterogeneity. We use the diffusion model to focus on learning the distribution of residual $\mathbf{r}^{ta} = \mathbf{x}^{ta} - \mathbb{E}_\theta[\mathbf{x}^{ta}|\mathbf{x}^{co}]$. Accordingly, the target data $\mathbf{x}^{ta}$ for diffusion models in Eqs. (1), (2), and (3) is replaced by $\mathbf{r}^{ta}$. We incorporate timestamp information as a condition in the denoising process and concatenate the context data $\mathbf{x}_0^{co}$ with noised residual $\mathbf{r}_n^{ta}$ as input to capture real-time fluctuations. Notably, no noise is added to $\mathbf{x}_0^{co}$ during diffusion training or inference. To model the spatial patterns of the residuals, we propose a scale-aware diffusion process to further distinguish the heterogeneity for different spatial units. In this section, we detail the calculation of $\mathbf{Q}$ and how it is integrated into the scale-aware diffusion process.

(i) Customized fluctuation scale. Specifically, we apply the Fast Fourier Transform (FFT) to spatiotemporal sequences in the training set to quantify fluctuation levels in different spatial units and use the custom scale $\mathbf{Q}$ as input to account for spatial heterogeneity in residual. We use FFT not just as a feature extractor, but to isolate the high-frequency 'turbulent' components that correspond to the residual term in Reynolds decomposition. Specifically, we first employ FFT to extract the fluctuation components for each spatial unit within the training set. The detailed steps are as follows:

$$
\begin{aligned}
&\mathbf{A}_\mathrm{k} = |\mathrm{FFT}(\mathbf{x})_\mathrm{k}|, \quad \phi_\mathrm{k} = \phi\left(\mathrm{FFT}(\mathbf{x})_\mathrm{k}\right), \quad \mathbf{A}_{\max} = \max_{\mathrm{k}\in\{1,\cdots,\lfloor \frac{L}{2} \rfloor+1\}} \mathbf{A}_\mathrm{k}, \\
&\mathcal{K} = \left\{ \mathrm{k} \in \left\{1, \cdots, \left\lfloor \frac{L}{2} \right\rfloor + 1\right\} : \mathbf{A}_\mathrm{k} < 0.1 \times \mathbf{A}_{\max} \right\}, \\
&\mathbf{x}_\mathbf{r}[i] = \sum_{\mathrm{k}\in\mathcal{K}} \mathbf{A}_\mathrm{k} \Big[\cos\left(2\pi \mathbf{f}_\mathrm{k} i + \phi_\mathrm{k}\right) + \cos\left(2\pi \bar{\mathbf{f}}_\mathrm{k} i + \bar{\phi}_\mathrm{k}\right) \Big],
\end{aligned}
\tag{9}
$$

where $\mathbf{A}_\mathrm{k}, \phi_\mathrm{k}$ reprent the amplitude and phase of the k−th frequency component. L is the temporal length of the training set. $\mathbf{A}_{\max}$ is the maximum amplitude among the components, obtained using the max operator. $\mathcal{K}$ represents the set of indices for the selected residual components. $\mathbf{f}_\mathrm{k}$ is the frequency of the k-th component. $\bar{\mathbf{f}}_\mathrm{k}, \bar{\phi}_\mathrm{k}$ represent the conjugate components. $\mathbf{x}_\mathbf{r}$ ref to the extracted residual component of the training set. We then compute the variance σ_v^2 of the residual sequence for each location v and expand it to match the shape as $\mathbf{r}_0^{ta} \in \mathbb{R}^{B\times V\times P}$, where B represents the batch size. And we can get the variance tensor $\mathcal{M}$:

$$
\mathcal{M}_{b,v,p} = \sigma_v^2, \forall b \in \{1, \cdots, B\}, \forall v \in \{1, \cdots, V\}, \forall p \in \{1, \cdots, P\}.
\tag{10}
$$

The residual fluctuations are bidirectional, encompassing both positive and negative variations, so we generate a random sign tensor $\mathbf{S} \in \mathbb{R}^{B\times V\times P}$ for $\mathcal{M}$, where each element $S_{b,v,p}$ of $\mathbf{S}$ is sampled from a Bernoulli distribution with $p = 0.5$. The customized fluctuation scale $\mathbf{Q}$ is computed as:

$$
\begin{aligned}
&\mathbf{Q}_{b,v,p} = S_{b,v,p} \times \mathcal{M}_{b,v,p}, \\
&\forall b \in \{1, \cdots, B\}, \forall v \in \{1, \cdots, V\}, \forall p \in \{1, \cdots, P\}.
\end{aligned}
\tag{11}
$$

Then $\mathbf{Q}$ is used as the input of the denoising network.

(ii) Scale-aware diffusion process. The vanilla diffusion models assume a shared prior distribution $\mathcal{N}(0, I)$ across all spatial locations, failing to capture spatial heterogeneity. To further model such differences, we adopt the technique proposed by Han et al. [21] to make the residual learning location-specific conditioned on $\mathbf{Q}$. Specifically, we redefine the noise distribution at the endpoint of the diffusion process as follows:

$$
p(\mathbf{r}_N^{ta}) = \mathcal{N}(\mathbf{Q}, I),
\tag{12}
$$

Accordingly, the Eq (1) in the forward process is rewritten as:

$$
\mathbf{r}_n^{ta} = \sqrt{\bar{\alpha}_n}\mathbf{r}_0^{ta} + (1 - \sqrt{\bar{\alpha}_n})\mathbf{Q} + \sqrt{1 - \bar{\alpha}_n}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I).
\tag{13}
$$

And in the denoising process, we sample $\mathbf{r}_N^{ta}$ from $\mathcal{N}(\mathbf{Q}, I)$, and denoise it use Eq (2), the computation of $\mu_\theta(\mathbf{r}_n^{ta}, n|\mathbf{x}_0^{co})$ in Eq (2) is modified as:

$$\mu_\theta(\mathbf{r}_n^{ta}, n|\mathbf{x}_0^{co}) = \frac{1}{\sqrt{\bar{\alpha}_n}}\left(\mathbf{r}_n^{ta} - \frac{\beta_n}{\sqrt{1-\bar{\alpha}_n}}\epsilon_\theta(\mathbf{r}_n^{ta}, n|\mathbf{x}_0^{co})\right) + (1 - \frac{1}{\sqrt{\bar{\alpha}_n}})\mathbf{Q}. \tag{14}$$

This modification allows the diffusion process to be conditioned on location-specific priors $\mathbf{Q}$, enhancing its ability to model spatial heterogeneity in uncertainty.

4.4 Training and Inference

We adopt a two-stage training procedure: first pretraining a deterministic model to predict the conditional mean, then training a diffusion model to capture the residual distribution (Appendix Algorithm 1). During inference, the deterministic model provides the mean prediction, while the diffusion model estimates residuals; their outputs are combined to form the final forecast (Appendix Algorithm 2).

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate our method on ten datasets spanning four domains, including climate (SST-CESM2 and SST-ERA5), energy (SolarPower), communication (MobileNJ and MobileSH), and urban systems (CrowdBJ, CrowdBM, TaxiBJ, BikeDC and Los-Speed), each featuring distinct spatiotemporal characteristics. Detailed information on the datasets can be found in Appendix D.1.

Baselines. We compare against nine representative state-of-the-art baselines commonly adopted in spatiotemporal modeling, including: **Gaussian Process (GP)**, **DeepState** [46], **D3VAE** [31], **DiffSTG** [62], **TimeGrad** [47], **CSDI** [57], **DYffusion** [50], **TMDM** [30], and **NPDiff** [54]. Detailed descriptions of each baseline are provided in Appendix D.2.

Metrics. We evaluate performance using two deterministic metrics (MAE, RMSE) and three probabilistic metrics (CRPS, QICE, IS). For QICE, we set $M_{\text{QIs}} = 10$ bins following its original design [21], which offers a balanced trade-off between granularity and stability. For IS, we choose a confidence level of 90% (i.e., $\alpha_{CI} = 0.1$) following common practice in spatiotemporal forecasting tasks [47, 57].

Experimental configuration. We define short-term forecasting as predicting the next 12 steps from the previous 12 observations [54, 62], and long-term forecasting as predicting the next 64 steps from the previous 64 [27, 71]. As temporal granularity varies across datasets, the actual durations differ. Full configurations are in Appendix D.3.

5.2 Overall Performance

Short-term forecasting. Table 1 reports probabilistic metrics for short-term forecasting, with additional results in Appendix Table 8. CoST demonstrates superior reliability, achieving the best QICE scores across all five datasets. This indicates exceptional distribution calibration, a critical attribute for trustworthy scientific modeling. On average, CoST yields improvements of 17.4% in CRPS, 46.6% in QICE, and 16.5% in IS over state-of-the-art baselines.

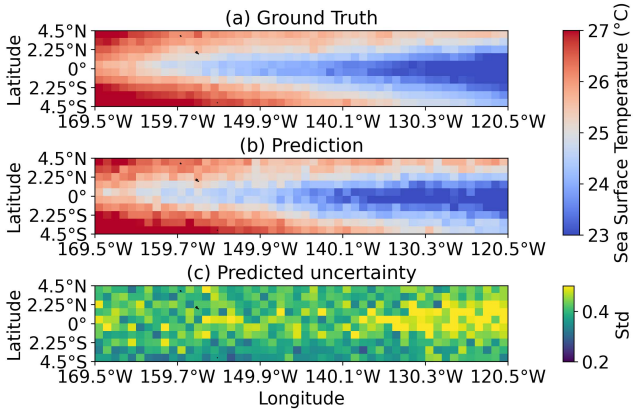

Figure 3: (a) and (b) show the ground-truth and predicted value of SST, and (c) displays the spatial distribution of forecasting uncertainty.

While specific models like DYffusion or NPDiff show competitive performance on individual metrics (e.g., CRPS in TaxiBJ or IS in SolarPower), CoST offers the most balanced and robust performance across diverse domains. Deterministic results (Table 2, Appendix Table 9) show reductions of 7% in MAE and 6.1% in RMSE, confirming that integration with a strong conditional mean estimator enhances regular pattern capture. In contrast, TMDM underperforms other baselines, mainly because its time-series–oriented design limits generalization to spatiotemporal data, and its focus on long sequences weakens short-term prediction, as evidenced in Appendix Tables 3 and 10. Further experiments on the ETTh1 and ETTh2 time-series datasets show notable performance gains (Appendix Table 11).

Long-term forecasting. As shown in Table 3, CoST demonstrates superior performance in long-term probabilistic forecasting. Most notably, it achieves a remarkable 70.4% average improvement in QICE, indicating that our model provides significantly better-calibrated uncertainty estimates compared to state-of-the-art baselines. In terms of probabilistic accuracy (CRPS), CoST consistently matches or outperforms the strongest competitor, CSDI, achieving an average improvement of roughly 15.0% (e.g., on Los-Speed and Climate). Despite relying on a lightweight MLP architecture, CoST effectively captures long-range dependencies, often surpassing the Transformer-based CSDI. Furthermore, as detailed in Section 5.7, this simpler design translates to significantly better training efficiency and inference speed. Regarding deterministic accuracy (Table 10), CoST remains highly competitive. It achieves the lowest MAE and RMSE on three out of five datasets (MobileSH, CrowdBJ, and Los-Speed). Even in cases where specialized baselines like DYffusion (SST) or TimeGrad (CrowdBM) excel, CoST ranks as the second-best, demonstrating that our framework enhances uncertainty quantification without compromising the accuracy of the mean prediction.

5.3 Case study of SST forecasting.

To assess our model's ability to quantify uncertainty under complex climate dynamics, we evaluate its performance in a key region for ENSO-related Sea Surface Temperature (SST) forecasting.

Table 1: Short-term forecasting results in terms of CRPS, QICE, and IS. Bold indicates the best performance, while underlining denotes the second-best. DYffusion is limited to grid-format data, and "-" denotes results that are not applicable.

Model	Climate			MobileSH			TaxiBJ			SolarPower			CrowdBJ		
	CRPS	QICE	IS	CRPS	QICE	IS	CRPS	QICE	IS	CRPS	QICE	IS	CRPS	QICE	IS
GP	0.083	0.158	9.98	0.495	0.120	6.90	0.217	0.137	258.9	0.732	0.222	769.0	0.601	0.152	17.1
DeepState	0.027	0.010	5.05	0.441	0.043	0.651	0.384	0.050	470.23	0.654	0.097	656.8	0.630	0.054	34.7
D3VAE	0.053	0.071	15.8	0.856	0.105	1.73	0.433	0.160	985.7	0.475	0.083	731.1	0.668	0.099	53.6
DiffSTG	0.026	0.068	7.42	0.303	0.078	0.526	0.299	0.074	416.5	0.213	0.068	240.6	0.436	0.089	32.1
TimeGrad	0.042	0.147	16.0	0.489	0.143	0.759	0.170	0.102	213.2	1.000	0.128	781.7	0.385	0.113	48.6
CSDI	0.027	0.019	5.18	0.200	0.052	0.295	0.122	0.048	121.8	0.267	0.050	221.6	0.306	0.028	16.4
TMDM	0.198	0.127	17.4	1.81	0.126	14.1	0.493	0.113	961.0	0.845	0.124	992.7	1.48	0.127	77.4
NPDiff	0.022	0.031	4.24	0.201	0.106	0.627	0.222	0.112	474.1	0.209	0.020	**175.3**	0.287	0.120	34.5
DYffusion	**0.020**	0.123	12.4	0.230	0.096	0.573	**0.084**	0.054	99.5	-	-	-	-	-	-
CoST	0.021	**0.009**	**4.04**	**0.147**	**0.014**	**0.215**	0.100	**0.023**	**95.3**	**0.208**	**0.019**	192.1	**0.215**	**0.014**	**11.5**

Table 2: Short-term forecasting results in terms of MAE and RMSE. Bold indicates the best performance, while underlining denotes the second-best. DYffusion is limited to grid-format data, and "-" denotes results that are not applicable.

Model	Climate		MobileSH		TaxiBJ		SolarPower		CrowdBJ	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
GP	1.51	1.64	0.359	0.611	28.0	30.5	43.9	97.1	3.37	4.70
DeepState	0.960	1.21	0.100	0.135	55.3	77.0	41.7	78.5	5.64	8.04
D3VAE	1.75	2.31	0.186	0.373	49.3	84.8	60.1	122.8	5.16	10.1
DiffSTG	0.90	1.13	0.066	0.103	41.8	69.4	31.1	63.8	3.68	6.63
TimeGrad	1.31	1.48	0.047	0.053	29.1	34.1	39.3	94.8	4.37	5.43
CSDI	0.94	1.20	0.044	0.075	18.2	31.6	38.8	69.6	2.71	5.51
TMDM	4.29	5.38	0.526	0.660	74.5	96.7	65.2	137.9	12.8	18.5
NPDiff	0.79	1.07	0.037	0.057	26.7	52.2	32.1	53.6	2.05	3.27
DYffusion	0.86	1.07	0.050	0.072	**12.3**	**18.0**	-	-	-	-
CoST	**0.74**	**0.96**	**0.033**	**0.051**	15.1	25.6	**29.7**	**51.9**	**1.92**	**3.04**

Table 3: Long-term forecasting results in terms of CRPS, QICE, and IS. Bold indicates the best performance, while underlining denotes the second-best. DYffusion is limited to grid-format data, and "-" denotes results that are not applicable.

Model	MobileSH			Climate			CrowdBJ			CrowdBM			Los-Speed		
	CRPS	QICE	IS	CRPS	QICE	IS	CRPS	QICE	IS	CRPS	QICE	IS	CRPS	QICE	IS
GP	0.537	0.112	7.13	0.086	0.146	9.18	0.660	0.143	19.4	0.622	0.153	76.5	0.910	0.149	112.6
DeepState	0.707	0.066	0.924	0.031	0.018	6.09	0.925	0.073	42.4	1.02	0.080	123.2	0.133	0.090	68.5
D3VAE	0.798	0.129	1.830	0.075	0.083	24.0	0.710	0.109	63.9	0.674	0.108	152.3	0.138	0.101	113.2
DiffSTG	0.374	0.107	0.923	0.027	0.077	7.90	0.370	0.094	31.3	0.400	0.073	67.1	0.124	0.080	104.6
TimeGrad	0.245	0.075	0.408	0.041	0.101	14.2	0.371	0.073	32.4	0.237	0.049	33.9	0.192	0.081	98.8
CSDI	0.158	0.045	**0.216**	0.036	0.073	6.80	0.229	0.038	12.0	0.235	0.052	33.7	0.134	0.090	**59.2**
TMDM	0.799	0.127	16.1	0.093	0.115	7.36	0.751	0.127	77.5	0.346	0.125	187.7	0.904	0.121	837.0
NPDiff	0.204	0.102	0.611	0.109	0.115	41.3	0.288	0.114	33.6	0.331	0.111	90.8	1.366	0.126	950.4
DYffusion	0.308	0.086	0.550	0.030	0.147	15.2	-	-	-	-	-	-	-	-	-
CoST	**0.158**	**0.016**	0.218	**0.024**	**0.011**	**4.87**	**0.217**	**0.011**	**11.5**	**0.235**	**0.009**	**31.2**	**0.089**	**0.040**	64.6

As shown in Figure 3, our model produces high-fidelity SST forecasts that closely match ground truth across both warm pool and cold tongue regions. In addition to accurate mean predictions, it provides well-calibrated uncertainty estimates, revealing elevated variance in the central equatorial Pacific, especially near 0° latitude and 140°–130°W, where sharp thermocline gradients and nonlinear feedbacks make forecasting particularly challenging. These high-uncertainty areas align with known regions of model divergence in

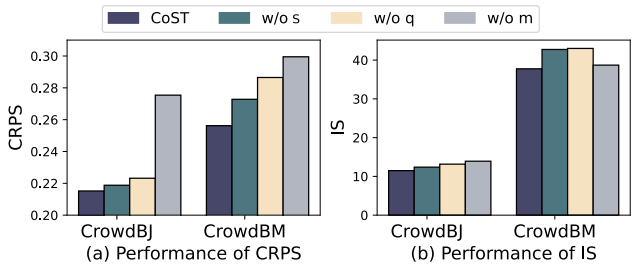

Figure 4: Ablation study on the CrowdBJ and CrowdBM comparing variants in terms of (a) CRPS and (b) IS.

climate science [8, 9, 40], demonstrating that our method delivers both accurate predictions and geophysically consistent uncertainty estimates.

5.4 Framework generalization.

To demonstrate the generality of CoST, we instantiate it with four representative spatiotemporal forecasting models: STID [52], STNorm [16], ConvLSTM [56], and iTransformer [37]. These models cover a diverse set of deep learning architectures, including CNNs, RNNs, MLPs, and Transformers. As shown in Table 4, CoST consistently enhances the performance of these backbones by effectively integrating deterministic and probabilistic modeling. Compared to using a single diffusion model, CoST yields more accurate predictions and better-calibrated uncertainty estimates, validating the framework's broad applicability and effectiveness.

5.5 Ablation Study

We perform an ablation study to assess the contribution of each proposed module. Specifically, we construct three model variants by progressively removing key components: **(w/o s)** removes the scale-aware diffusion process; **(w/o q)** excludes the customized fluctuation scale as a prior; **(w/o m)** removes the conditional mean predictor, relying solely on the diffusion model. Experiments on two datasets (Figure 4) show that the deterministic predictor notably improves performance by capturing regular spatiotemporal patterns, while also reducing the diffusion model's complexity. Adding the customized fluctuation scale further enhances accuracy, indicating its utility in providing valuable fluctuation information across different spatial units. And the scale-aware diffusion process enables the diffusion model to better utilize this condition.

5.6 Qualitative Analysis

Analysis of distribution alignment. As shown in Figure 5, the ground truth exhibits clear spatiotemporal multi-modality. In Figure 5(a), three peaks likely correspond to different time points or varying states at the same time. CoST accurately captures all three peaks, while CSDI only fits two, showing CoST's superior multi-modal modeling. In Figure 5(b), both models capture two peaks, but CoST aligns better with the peak spacing in the true distribution, reflecting stronger temporal sensitivity. These strengths arise from CoST's hybrid design: the diffusion component models residual uncertainty to capture multi-modal traits, while the deterministic

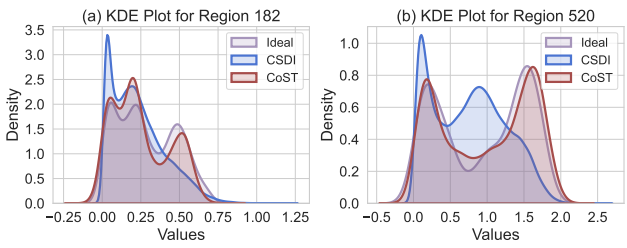

Figure 5: KDE plots of the MobileSH dataset for different regions: (a) Region 182, (b) Region 520.

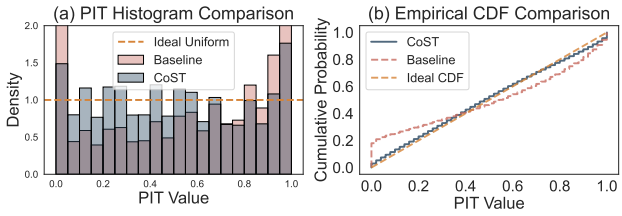

Figure 6: PIT analysis on the MobileSH dataset: (a) PIT histogram and (b) PIT empirical CDF.

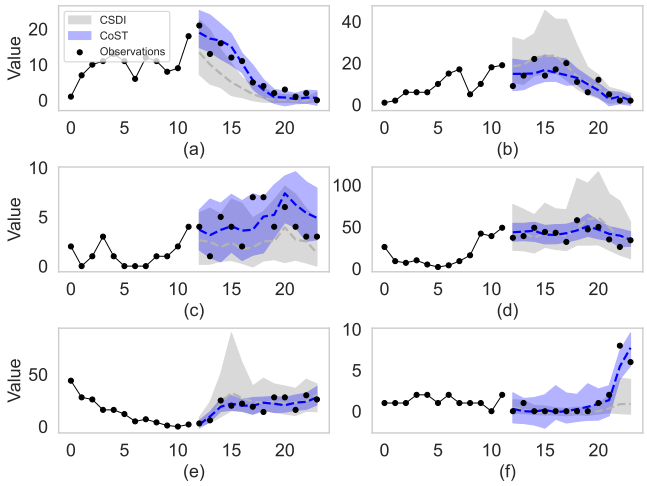

Figure 7: Visualizations of predictive uncertainty for both CSDI and CoST on the CrowdBJ dataset. The shaded regions represent the 90% confidence interval. The dashed lines denote the median of the predicted values for each model.

backbone learns regular trends. Additionally, we present the PIT (Probability Integral Transform) histogram in Figure 6 (a) and the PIT empirical cumulative distribution function (CDF) in Figure 6 (b) to visually reflect the alignment of the full distribution. Ideally, the true values' quantiles in the predictive distribution should follow a uniform distribution, corresponding to the dashed line in Figure 6 (a). In the case of perfect calibration, the PIT CDF should closely resemble the yellow diagonal line. Clearly, our model outperforms CSDI.

Table 4: Performance of different deterministic backbone models within the CoST framework. "Diffusion (w/o m)" denotes the results obtained using a single diffusion model.

Model	Climate					MobileNJ					BikeDC				
	MAE	RMSE	CRPS	QICE	IS	MAE	RMSE	CRPS	QICE	IS	MAE	RMSE	CRPS	QICE	IS
Diffusion (w/o m)	1.070	1.361	0.030	0.030	6.58	0.195	0.6711	0.159	0.036	1.364	2.387	10.79	1.090	0.059	12.6
+iTransformer	0.818	1.088	0.023	0.018	4.83	0.122	0.207	0.123	0.021	0.815	0.526	2.23	0.454	0.035	3.82
Reduction	23.6%	20.1%	23.3%	40.0%	26.6%	37.4%	69.2%	22.6%	41.7%	40.2%	78.0%	79.3%	58.3%	40.7%	69.7%
+ ConvLSTM	0.889	1.151	0.027	0.024	5.54	0.137	0.231	0.120	0.025	0.913	0.454	2.01	0.443	0.037	6.07
Reduction	16.9%	15.4%	10.0%	20.0%	15.8%	29.7%	65.6%	24.5%	30.6%	33.1%	81.0%	81.4%	59.4%	37.3%	51.8%
+STNorm	0.819	1.066	0.023	0.007	4.52	0.144	0.276	0.123	0.016	0.825	0.600	2.71	0.500	0.029	3.74
Reduction	23.5%	21.7%	23.3%	76.7%	31.3%	26.2%	58.9%	22.6%	55.6%	39.5%	74.9%	74.9%	54.1%	50.8%	70.3%

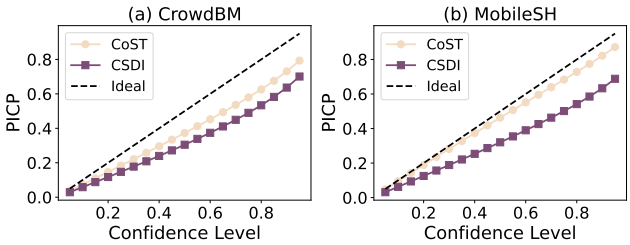

Figure 8: PICP comparison between our model and CSDI on CrowdBM and MobileSH.

Table 5: Comparison of training and inference time on the MobileSH dataset.

Model	Train Time	Inference Time
D3VAE	3min 27s	2min 15s
DiffSTG	24min 16s	18min 38s
TimeGrad	5min	2min
CSDI	48min 40s	38min 49s
DyDiffusion	33h	3h
CoST	2min	50s

Analysis of prediction quality. To intuitively demonstrate the effectiveness of our predictions, we visualize results on the CrowdBJ dataset in Figure 7 (More Cases shown in Appendix Figure 10), comparing our model with the best baseline, CSDI. As shown in Figures 7 (a, c, f), our model, aided by a deterministic backbone, better captures regular spatiotemporal patterns. Meanwhile, the diffusion module enhances uncertainty modeling by focusing on residuals, as reflected in Figures 7 (b, d, e). Beyond sample-level comparison, we evaluate prediction interval calibration via dynamic quantile error curves on CrowdBM and MobileSH (Figure 8). For each confidence level α, we compute the corresponding quantile interval and its Prediction Interval Coverage Probability (PICP). Closer alignment with the diagonal (black dashed line) indicates better calibration. Our model consistently outperforms CSDI in this regard.

5.7 Computational Cost

We benchmark training and inference time (including 50 sampling iterations and pretraining for our mean predictor) on the MobileSH dataset. As shown in Table 5, CoST is markedly more efficient than existing probabilistic models. The efficiency comes from modeling only the residual distribution with a lightweight denoising network. In contrast, other baselines incur high computational costs by modeling the full data distribution with complex networks, while our approach is well-suited for time-sensitive applications such as mobile traffic prediction.

6 Conclusion

In this work, we present CoST, a general framework for trustworthy forecasting in chaotic spatiotemporal systems. Inspired by the Reynolds decomposition from fluid dynamics, CoST decouples system evolution into predictable physical trends and stochastic residual fluctuations. This structure bridges deterministic modeling with diffusion-based generative learning, ensuring that uncertainty quantification remains physically grounded rather than hallucinatory. Extensive evaluations across climate science, renewable energy, and urban dynamics show that CoST not only achieves state-of-the-art accuracy but also delivers practical scientific utility. By effectively modeling the superposition of physical laws and intrinsic noise, CoST provides a scalable and interpretable pathway for AI-driven discovery in data-rich scientific domains.

7 Limitations and Ethical Considerations

This work utilizes multiple datasets involving human activity. All mobility data are anonymized and aggregated at the grid level, ensuring no individual trajectories can be reconstructed, complying with standard privacy protocols. The SolarPower data is proprietary; we have permission to use it for research but cannot release it publicly. Our approach also has limitations. First, the effectiveness of the mean–residual decomposition depends on the deterministic backbone; if the mean predictor fails under extreme or unprecedented events, the residual diffusion model cannot fully compensate. Second, although inspired by Reynolds decomposition, the current framework does not explicitly enforce physical conservation laws, which we leave to future physics-constrained extensions.

GenAI Disclosure

In compliance with the stated policy, we report the use of a large language model (LLM) as a general-purpose assistance tool in the writing of this paper. Its application was confined to copy-editing and language polishing, such as correcting grammar and syntax. The LLM played no role in the research ideation or the generation of the substantive content of this work.

References

[1] Lei Bai, Lina Yao, Salil Kanhere, Xianzhi Wang, Quan Sheng, et al. 2019. Stg2seq: Spatial-temporal graph to sequence model for multi-step passenger demand forecasting. *arXiv preprint arXiv:1905.10069* (2019).

[2] Lei Bai, Lina Yao, Salil S Kanhere, Zheng Yang, Jing Chu, and Xianzhi Wang. 2019. Passenger demand forecasting with multi-task convolutional recurrent neural networks. In *Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, PAKDD 2019, Macau, China, April 14-17, 2019, Proceedings, Part II 23*. Springer, 29–42.

[3] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. 2020. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems* 33 (2020), 17804–17815.

[4] Peter Bauer, Alan Thorpe, and Gilbert Brunet. 2015. The quiet revolution of numerical weather prediction. *Nature* 525, 7567 (2015), 47–55.

[5] Dimitri Bertsekas and John N Tsitsiklis. 2008. *Introduction to probability*. Vol. 1. Athena Scientific.

[6] Oussama Boussif, Ghait Boukachab, Dan Assouline, Stefano Massaroli, Tianle Yuan, Loubna Benabbou, and Yoshua Bengio. 2023. Improving* day-ahead* solar irradiance time series forecasting by leveraging spatio-temporal context. *Advances in Neural Information Processing Systems* 36 (2023), 2342–2367.

[7] Robert Joseph Broderick, Jimmy Edward Quiroz, Matthew J Reno, Abraham Ellis, Jeff Smith, and Roger Dugan. 2012. *Time series power flow analysis for distribution connected PV generation*. Technical Report. Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States).

[8] Mark A Cane. 2005. The evolution of El Niño, past and future. *Earth and Planetary Science Letters* 230, 3-4 (2005), 227–240.

[9] Dan Cao, Jiahua Zhang, Lan Xun, Shanshan Yang, Jingwen Wang, and Fengmei Yao. 2021. Spatiotemporal variations of global terrestrial vegetation climate potential productivity under climate change. *Science of The Total Environment* 770 (2021), 145320.

[10] Haoye Chai, Tao Jiang, and Li Yu. 2024. Diffusion Model-based Mobile Traffic Generation with Open Data for Network Planning and Optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4828–4838.

[11] Changlu Chen, Yanbin Liu, Ling Chen, and Chengqi Zhang. 2022. Bidirectional spatial-temporal adaptive transformer for urban traffic flow forecasting. *IEEE Transactions on Neural Networks and Learning Systems* 34, 10 (2022), 6913–6925.

[12] Jiayuan Chen, Shuo Zhang, Xiaofei Chen, Qiao Jiang, Hejiao Huang, and Chonglin Gu. 2021. Learning traffic as videos: a spatio-temporal VAE approach for traffic data imputation. In *International Conference on Artificial Neural Networks*. Springer, 615–627.

[13] Weihuang Chen, Fangfang Wang, and Hongbin Sun. 2021. S2tnet: Spatio-temporal transformer networks for trajectory prediction in autonomous driving. In *Asian conference on machine learning*. PMLR, 454–469.

[14] Chen Chu, Hengcai Zhang, Peixiao Wang, and Feng Lu. 2024. Simulating human mobility with a trajectory generation framework based on diffusion model. *International Journal of Geographical Information Science* 38, 5 (2024), 847–878.

[15] Miguel Ángel De Miguel, José María Armingol, and Fernando Garcia. 2022. Vehicles trajectory prediction using recurrent VAE network. *IEEE Access* 10 (2022), 32742–32749.

[16] Jinliang Deng, Xiusi Chen, Renhe Jiang, Xuan Song, and Ivor W Tsang. 2021. St-norm: Spatial and temporal normalization for multi-variate time series forecasting. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 269–278.

[17] Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Bernie Wang. 2023. Prediff: Precipitation nowcasting with latent diffusion models. *Advances in Neural Information Processing Systems* 36 (2023), 78621–78656.

[18] Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. 2019. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3656–3663.

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

[20] Junchao Gong, Lei Bai, Peng Ye, Wanghan Xu, Na Liu, Jianhua Dai, Xiaokang Yang, and Wanli Ouyang. 2024. CasCast: skillful high-resolution precipitation nowcasting via cascaded modelling. In *Proceedings of the 41st International Conference on Machine Learning*. 15809–15822.

[21] Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. 2022. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems* 35 (2022), 18100–18115.

[22] Yan Han, Lihua Mi, Lian Shen, CS Cai, Yuchen Liu, Kai Li, and Guoji Xu. 2022. A short-term wind speed prediction method utilizing novel hybrid deep learning algorithms to correct numerical weather forecasting. *Applied Energy* 312 (2022), 118777.

[23] Xinyue Hu, Haoji Hu, Saurabh Verma, and Zhi-Li Zhang. 2020. Physics-guided deep neural networks for power flow analysis. *IEEE Transactions on Power Systems* 36, 3 (2020), 2082–2092.

[24] Zhanhong Jiang, Chao Liu, Adedotun Akintayo, Gregor P Henze, and Soumik Sarkar. 2017. Energy prediction using spatiotemporal pattern networks. *Applied Energy* 206 (2017), 1022–1039.

[25] Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Shao, Jincai Huang, Junbo Zhang, and Yu Zheng. 2023. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[26] Junchen Jin, Dingding Rong, Tong Zhang, Qingyuan Ji, Haifeng Guo, Yisheng Lv, Xiaoliang Ma, and Fei-Yue Wang. 2022. A GAN-based short-term link traffic prediction approach for urban road networks under a parallel learning framework. *IEEE Transactions on Intelligent Transportation Systems* 23, 9 (2022), 16185–16196.

[27] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728* (2023).

[28] Shahab Kareem, Zhala Jameel Hamad, and Shavan Askar. 2021. An evaluation of CNN and ANN in prediction weather forecasting: A review. *Sustainable Engineering and Innovation* 3, 2 (2021), 148.

[29] Tiruvalam Natarajan Krishnamurti and Lahouari Bounoua. 2018. *An introduction to numerical weather prediction techniques*. CRC press.

[30] Yuxin Li, Wenchao Chen, Xinyue Hu, Bo Chen, Mingyuan Zhou, et al. 2024. Transformer-modulated diffusion models for probabilistic multivariate time series forecasting. In *The Twelfth International Conference on Learning Representations*.

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

[32] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926* (2017).

[33] Lequan Lin, Zhengkun Li, Ruikun Li, Xuliang Li, and Junbin Gao. 2024. Diffusion models for time-series applications: a survey. *Frontiers of Information Technology & Electronic Engineering* 25, 1 (2024), 19–41.

[34] Zhihui Lin, Maomao Li, Zhuobin Zheng, Yangyang Cheng, and Chun Yuan. 2020. Self-attention convlstm for spatiotemporal prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 11531–11538.

[35] Fenghua Ling, Zeyu Lu, Jing-Jia Luo, Lei Bai, Swadhin K Behera, Dachao Jin, Baoxiang Pan, Huidong Jiang, and Toshio Yamagata. 2024. Diffusion model-based probabilistic downscaling for 180-year East Asian climate reconstruction. *npj Climate and Atmospheric Science* 7, 1 (2024), 131.

[36] Lingbo Liu, Ruimao Zhang, Jiefeng Peng, Guanbin Li, Bowen Du, and Liang Lin. 2018. Attentive crowd flow machines. In *Proceedings of the 26th ACM international conference on Multimedia*. 1553–1561.

[37] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. [n. d.]. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*.

[38] Ziqing Ma, Wenwei Wang, Tian Zhou, Chao Chen, Bingqing Peng, Liang Sun, and Rong Jin. 2024. FusionSF: Fuse Heterogeneous Modalities in a Vector Quantized Framework for Robust Solar Power Forecasting. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5532–5543.

[39] Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, et al. 2023. Residual corrective diffusion modeling for km-scale atmospheric downscaling, 2024. *URL https://arxiv. org/abs/2309.15214* (2023).

[40] Michael J McPhaden, Stephen E Zebiak, and Michael H Glantz. 2006. ENSO as an integrating concept in earth science. *science* 314, 5806 (2006), 1740–1745.

[41] Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Romit Maulik, Rao Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. 2024. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. *Advances in Neural Information Processing Systems* 37 (2024), 68740–68771.

[42] TN Palmer. 2012. Towards the probabilistic Earth-system simulator: A vision for the future of climate and weather prediction. *Quarterly Journal of the Royal Meteorological Society* 138, 665 (2012), 841–861.

[43] Tim Palmer. 2014. Climate forecasting: Build high-resolution global climate models. *Nature* 515, 7527 (2014), 338–339.

[44] Stephen B Pope. 2001. Turbulent flows. *Measurement Science and Technology* 12, 11 (2001), 2020–2021.

[45] Yanjun Qin, Haiyong Luo, Fang Zhao, Yuchen Fang, Xiaoming Tao, and Chenxing Wang. 2023. Spatio-temporal hierarchical MLP network for traffic forecasting. *Information Sciences* 632 (2023), 543–554.

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

[47] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. 2021. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*. PMLR, 8857–8868.

[48] RN Ravikumar and S Aarthi. 2026. Forecasting Wind and Solar Energy With GANs and Diffusion Models: A New Frontier in Renewable Energy Prediction. In *Enhancing Renewable Energy Systems With Generative AI*. IGI Global Scientific Publishing, 31–58.

[49] Stephen Roberts, Michael Osborne, Mark Ebden, Steven Reece, Neale Gibson, and Suzanne Aigrain. 2013. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371, 1984 (2013), 20110550.

[50] Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. 2023. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. *Advances in neural information processing systems* 36 (2023), 45259–45287.

[51] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting* 36, 3 (2020), 1181–1191.

[52] Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. 2022. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4454–4458.

[53] Lifeng Shen and James Kwok. 2023. Non-autoregressive conditional diffusion models for time series prediction. In *International Conference on Machine Learning*. PMLR, 31016–31029.

[54] Zhi Sheng, Yuan Yuan, Jingtao Ding, and Yong Li. 2025. Unveiling the Power of Noise Priors: Enhancing Diffusion Models for Mobile Traffic Prediction. *arXiv preprint arXiv:2501.13794* (2025).

[55] Jimeng Shi, Bowen Jin, Jiawei Han, Sundararaman Gopalakrishnan, and Giri Narasimhan. 2025. CoDiCast: Conditional Diffusion Model for Global Weather Forecasting with Uncertainty Quantification. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*. 9853–9861.

[56] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems* 28 (2015).

[57] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems* 34 (2021), 24804–24816.

[58] Lucas R Vargas Zeppetello, Adrian E Raftery, and David S Battisti. 2022. Probabilistic projections of increased heat stress driven by climate change. *Communications Earth & Environment* 3, 1 (2022), 183.

[59] Junfei Wang, Darshana Upadhyay, Marzia Zaman, and Pirathayini Srikantha. 2025. Synthetic Power Flow Data Generation Using Physics-Informed Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2504.17210* (2025).

[60] Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and S Yu Philip. 2018. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *International conference on machine learning*. PMLR, 5123–5132.

[61] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. 2017. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems* 30 (2017).

[62] Haomin Wen, Youfang Lin, Yutong Xia, Huaiyu Wan, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2023. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*. 1–12.

[63] Ruowu Wu, Yandan Liang, Lianlei Lin, and Zongwei Zhang. 2024. Spatiotemporal Multivariate Weather Prediction Network Based on CNN-Transformer. *Sensors (Basel, Switzerland)* 24, 23 (2024), 7837.

[64] Peng Xie, Tianrui Li, Jia Liu, Shengdong Du, Xin Yang, and Junbo Zhang. 2020. Urban flow prediction from spatiotemporal data using machine learning: A survey. *Information Fusion* 59 (2020), 1–12.

[65] Hongwei Yang, Jie Yan, Yongqian Liu, and Zongpeng Song. 2022. Statistical downscaling of numerical weather prediction based on convolutional neural networks. *Global Energy Interconnection* 5, 2 (2022), 217–225.

[66] Yiyuan Yang, Ming Jin, Haomin Wen, Chaoli Zhang, Yuxuan Liang, Lintao Ma, Yi Wang, Chenghao Liu, Bin Yang, Zenglin Xu, et al. 2024. A survey on diffusion models for time series and spatio-temporal data. *arXiv preprint arXiv:2404.18886* (2024).

[67] Yan Yang, Zhifang Yang, Juan Yu, Baosen Zhang, Youqiang Zhang, and Hongxin Yu. 2019. Fast calculation of probabilistic power flow: A model-based deep learning approach. *IEEE Transactions on Smart Grid* 11, 3 (2019), 2235–2244.

[68] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. 2020. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 507–523.

[69] Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. 2024. Diffcast: A unified framework via residual diffusion for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27758–27767.

[70] Xinyu Yuan and Yan Qiao. 2024. Diffusion-ts: Interpretable diffusion for general time series generation. *arXiv preprint arXiv:2403.01742* (2024).

[71] Yuan Yuan, Jingtao Ding, Jie Feng, Depeng Jin, and Yong Li. 2024. Unist: A prompt-empowered universal model for urban spatio-temporal prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4095–4106.

[72] Yuan Yuan, Chonghua Han, Jingtao Ding, Depeng Jin, and Yong Li. 2024. Urbandit: A foundation model for open-world urban spatio-temporal learning. *arXiv preprint arXiv:2411.12164* (2024).

[73] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[74] Liang Zhang, Jianqing Wu, Jun Shen, Ming Chen, Rui Wang, Xinliang Zhou, Cankun Xu, Quankai Yao, and Qiang Wu. 2021. SATP-GAN: Self-attention based generative adversarial network for traffic flow prediction. *Transportmetrica B: Transport Dynamics* 9, 1 (2021), 552–568.

[75] Zijian Zhang, Ze Huang, Zhiwei Hu, Xiangyu Zhao, Wanyu Wang, Zitao Liu, Junbo Zhang, S Joe Qin, and Hongwei Zhao. 2023. MLPST: MLP is All You Need for Spatio-Temporal Prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 3381–3390.

A Related Work

A.1 Spatiotemporal deterministic forecasting.

Deterministic forecasting of spatiotemporal systems focuses on point estimation. These models are typically trained with loss functions like MSE or MAE to learn the conditional mean $\mathbb{E}[y|x]$, capturing regular patterns. Common deep learning architectures include MLP-based [45, 52, 75], CNN-based [32, 36, 73], and RNN-based [2, 34, 60, 61] models, valued for their efficiency. GNN-based methods [1, 3, 18, 25] capture spatial dependencies in graph-based data, while Transformer-based models [6, 11, 13, 38, 68] are effective at modeling complex temporal dynamics.

A.2 Spatiotemporal probabilistic forecasting.

The core of probabilistic forecasting lies in modeling uncertainty, aiming to capture the full data distribution [57, 66]. This is particularly suited for modeling the stochastic nature of spatiotemporal systems. While early methods include classical Bayesian approaches like Gaussian Processes (GP) [49] and influential deep learning models such as DeepAR [51] and DeepStateSpace [46], recent advances have explored generative models such as GANs [26, 74], VAEs [12, 15], and diffusion models [10, 33]. Diffusion models, in particular, have gained traction for their ability to model complex distributions with stable training, yielding strong performance in spatiotemporal forecasting [47, 50, 53, 54].

B Background

B.1 Glossary

We summarize all notations and symbols used throughout the paper in Table 6.

B.2 Spatiotemporal Data

Spatiotemporal data typically come in two forms: (i) **Grid-based data**, where the spatial dimension V can be expressed in a two-dimensional form as $H \times W$, with H and W denoting height and width, respectively. (ii) **Graph-based data**, where V denotes the number of nodes in a spatial graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, defined by its set of nodes $\mathcal{V}$, the set of edges $\mathcal{E}$ and the adjacency matrix $\mathcal{A}$. Its elements a_{ij} show if there's an edge between node i and j in $\mathcal{V}$, $a_{ij} = 1$ when there's an edge and $a_{ij} = 0$ otherwise.

C Methodology

C.1 Algorithm

The training and inference procedures of CoST are summarized in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 Training

Stage 1: Pretraining of Deterministic Model $\mathbb{E}_\theta$
repeat
  Estimate the conditional mean $\mathbb{E}_\theta[\mathbf{x}_0^{ta}|\mathbf{x}_0^{co}]$.
  Update $\mathbb{E}_\theta$ using the following loss function:

$$\mathcal{L}_2 = \left\|\mathbb{E}_\theta[\mathbf{x}_0^{ta}|\mathbf{x}_0^{co}] - \mathbf{x}_0^{ta}\right\|_2^2$$

until The model has converged.
Stage 2: Training of Diffusion Model ϵ_θ
repeat
  Initialize $n \sim \text{Uniform}(1, \dots, N)$ and $\epsilon \sim \mathcal{N}(0, I)$.
  Calculate the target $\mathbf{r}_0^{ta} = \mathbf{x}_0^{ta} - \mathbb{E}_\theta[\mathbf{x}_0^{ta}|\mathbf{x}_0^{co}]$.
  Calculate noisy targets $\mathbf{r}_n^{ta}$ using Eq. (13).
  Update ϵ_θ using the following loss function:

$$\mathcal{L}(\theta) = \left\|\epsilon - \epsilon_\theta(\mathbf{r}_n^{ta}, n|\mathbf{x}_0^{co})\right\|_2^2$$

until The model has converged.

Algorithm 2 Inference

Input: Context data $\mathbf{x}_0^{co}$, customized fluctuation scale $\mathbf{Q}$, trained diffusion model ϵ_θ, trained deterministic model $\mathbb{E}_\theta$
Output: Target data $\mathbf{x}_0^{ta}$
Estimate the conditional mean $\mathbb{E}_\theta[\mathbf{x}_0^{ta}|\mathbf{x}_0^{co}]$
Sample $\mathbf{r}_N^{ta}$ from $\epsilon \sim \mathcal{N}(\mathbf{Q}, I)$
for $n = N$ to 1 **do**
  Estimate the noise $\epsilon_\theta(\mathbf{r}_n^{ta}, n|\mathbf{x}_0^{co})$
  Calculate the $\mu_\theta(\mathbf{r}_n^{ta}, n|\mathbf{x}_0^{co})$ using Eq. (14)
  Sample $\mathbf{r}_{n-1}^{ta}$ using Eq. (2)
end for
Return: $\mathbf{x}_0^{ta} = \mathbb{E}_\theta[\mathbf{x}_0^{ta}|\mathbf{x}_0^{co}] + \mathbf{r}_0^{ta}$

D Experiments

D.1 Datasets

In our experiments, we evaluate the proposed method on ten real-world datasets across four domains: **climate**, **energy**, **communication systems**, and **urban systems**. For climate forecasting, we train our models on the simulated SST-CESM2 dataset and evaluate them on the observational SST-ERA5 dataset, using the first 30 years for validation and the remaining years for testing. The remaining datasets are partitioned into training, validation, and test sets with a 6:2:2 ratio, and all datasets are standardized during training.Table 7 provides a summary of the datasets. The details are as follows:

- **Climate.** We utilize two datasets for sea surface temperature (SST) prediction in the Niño 3.4 region (5°S–5°N, 170°W–120°W), which is widely used for monitoring El Niño events: (i) SST-CESM2, simulated SST data from the CESM2-FV2 model of the CMIP6 project, covering the period from 1850 to 2014, with a spatial resolution of $1^\circ \times 1^\circ$. (ii) SST-ERA5: reanalysis data from ERA5, containing SST and 10-meter wind speed (U10/V10) variables from 1940 to 2025, with an original spatial resolution of approximately $0.25^\circ \times 0.25^\circ$. All data are regridded to a $1^\circ \times 1^\circ$ resolution for consistency. The CESM2 data are used for training, while the first 30 years of ERA5 are used for validation and the remaining years for testing.
- **Energy.** This dataset contains real-time meteorological measurements and photovoltaic (PV) power output collected from a PV power station in China, spanning from March 1st to December 31st, 2024. The features include: total active power output of the PV grid-connection point (P), ambient temperature, back panel temperature, dew point, relative humidity, atmospheric pressure, global horizontal irradiance (GHI), diffuse and direct radiation, wind direction and wind speed. Our forecasting task focuses on

Table 6: Glossary of notations and symbols used in this paper.

Symbol	Used for
$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$	Graph structure where $\mathcal{V}$ is the node set, $\mathcal{E}$ is the edge set, and $\mathbf{A}$ is the adjacency matrix.
$\mathbf{x} \in \mathbb{R}^{T \times V \times C}$	Spatiotemporal data.
T	The length of spatiotemporal series.
V	The number of spatial units.
C	The number of feature dimensions.
B	Batch size.
P	Prediction horizon.
M	Historical horizon.
N	The number of diffusion steps.
H	Height of the grid-based data.
W	Width of the grid-based data.
$\mathbf{Q}$	Customized fluctuation scale.
$\mathcal{M}$	The variance tensor.
$\mathbf{S}$	The random sign tensor.
$\{\cdot\}^{\mathrm{co}}$	Historical (conditional) term.
$\{\cdot\}^{\mathrm{ta}}$	Predicted (target) term.
$\{\cdot\}_{\mathrm{n}}$	Noisy data at n-th diffusion step.
μ	Mean.
$\mathbf{r}$	Residual.
ϵ	Gaussian noise.
$\mathcal{K}$	K The set of indices for the selected FFT components.
$\{\beta_n\}_{n=1}^{N}$	The noise schedule in the diffusion process.
$\alpha_n, \bar{\alpha}_n$	$\alpha_n = 1 - \beta_n$, $\bar{\alpha}_n = \prod_{i=1}^{n} \alpha_i$.
$\epsilon_\theta(\cdot)$	The denoising network with parameter θ.
α_{CI}	Significance level for the prediction interval.
$\mathbb{1}_{(\cdot)}$	Indicator function, which takes the value 1 when a certain condition is true, and 0 when the condition is false.

Table 7: The basic information of spatio-temporal data.

Dataset	Location	Type	Temporal Period	Spatial partition	Interval
SST-CESM2	Global (Niño 3.4)	Simulated SST	1850-2014	$1^\circ \times 1^\circ$	Monthly
SST-ERA5	Global (Niño 3.4)	Reanalysis SST / U10 / V10	1940-2025	$0.25^\circ \times 0.25^\circ$	Monthly
SolarPower	China (a PV station)	GHI / Weather / PV power	2024/03/01 - 2024/12/31	Station-level	15 min
TaxiBJ	Beijing	Taxi flow	2014/03/01 - 2014/06/30	32×32	Half an hour
BikeDC	Washington, D.C.	Bike flow	2010/09/20 - 2010/10/20	20×20	Half an hour
MobileSH	Shanghai	Mobile traffic	2014/08/01 - 2014/08/21	32×28	One hour
MobileNJ	Nanjing	Mobile traffic	2021/02/02 - 2021/02/22	20×28	One hour
CrowdBJ	Beijing	Crowd flow	2018/01/01 - 2018/01/31	1010	One hour
CrowdBM	Baltimore	Crowd flow	2019/01/01 - 2019/05/31	403	One hour
Los-Speed	Los Angeles	Traffic speed	2012/03/01 - 2012/03/07	207	5 min

GHI, which is the key variable for solar power prediction. Due to data privacy restrictions, the raw dataset cannot be publicly released.

- **Communication Systems.** Mobile communication traffic datasets are collected from two major cities in Shanghai and Nanjing, capturing the spatiotemporal dynamics of network usage patterns.
- **Urban Systems.** We adopt five widely used public datasets representing various urban sensing signals: (i) CrowdBJ and CrowdBM, crowd flow data from Beijing and Baltimore, respectively. (ii) TaxiBJ, taxi trajectory-based traffic flow data from Beijing. (iii) BikeDC, bike-sharing demand data from Washington D.C. (iv) Los-Speed, traffic speed data from the Los Angeles road network. These datasets have been extensively used in spatiotemporal forecasting research and provide diverse signals for evaluating model generality across cities and domains.

D.2 Baselines

We provide a brief description of the baselines used in our experiments:

- **GP (Gaussian Processes):**A non-parametric time series forecasting method that models data as a Gaussian process, offering uncertainty estimates and effective modeling of non-linear relationships.
- **DeepState [46]:** A deep learning framework for time series forecasting that integrates state space models with neural networks, enabling efficient probabilistic predictions by learning latent states and observation processes.
- **D3VAE [31]:** Aims at short-period and noisy time series forecasting. It combines generative modeling with a bidirectional variational auto-encoder, integrating diffusion, denoising, and disentanglement.
- **DiffSTG [62]:** First applies diffusion models to spatiotemporal graph forecasting. By combining STGNNs and diffusion models, it reduces prediction errors and improves uncertainty modeling.
- **TimeGrad [47]:** An autoregressive model based on diffusion models. It conducts probabilistic forecasting for multivariate time series and performs well on real-world datasets.
- **CSDI [57]:**Utilizes score-based diffusion models for time series imputation. It can leverage the correlations of observed values and also shows remarkable results on prediction tasks.
- **DYffusion [50]:** A training method for diffusion models in probabilistic spatiotemporal forecasting. It combines data temporal dynamics with diffusion steps and performs well in complex dynamics forecasting.
- **TMDM [30]:** TMDM integrates transformers with diffusion models for probabilistic time series forecasting, using transformer-based prior knowledge to enable accurate distribution forecasting and uncertainty estimation.
- **NPDiff [54]:** A general noise prior framework for mobile traffic prediction. It uses the data dynamics to calculate noise priors for the denoising process and achieve effective performance.

D.3 Experimental Configuration

In our experiment, for our model, we set the training maximum epoch for both the deterministic model and the diffusion model to 50, with early stopping based on a patience of 5 for both models. For the diffusion model, we set the validation set sampling number to 3, and the average metric computed over these samples is used as the criterion for early stopping. For the baseline models, we set the maximum training epoch to 100 and the early stopping patience also to 5. We set the number of samples to 50 for computing the experimental results presented in the paper. For the denoising network architecture, we adopt a lightweight variant of the MLP-based STID [52]. Specifically, we set the number of encoder layers to 8 and the embedding dimension to 128. The diffusion model employs a maximum of 50 diffusion steps, using a linear noise schedule with $\beta_1 = 0.0001$ and $\beta_N = 0.5$. During training, we set the initial learning rate to 0.001, and after 20 epochs, we adjust it to 4e-4. We use the Adam optimizer with a weight decay of 1e-6. All experiments are conducted with fixed random seeds. Models with lower GPU memory demands are run on NVIDIA TITAN Xp (12GB GDDR5X) and NVIDIA GeForce RTX 4090 (24GB GDDR6X) GPUs under a Linux environment. For the DYffusion [50] baseline, which requires substantially more resources, training is performed on NVIDIA A100 (80GB HBM2e) and A800 (40GB HBM2e).

D.4 Geographic Extent of the ENSO Region

To provide geographic context for the SST case study presented in Section 3, Figure 9 illustrates the global location and spatial extent of the selected region. The red box highlights the area from 4.5°S to 4.5°N and 169.5°W to 120.5°W in the central-to-eastern equatorial Pacific, a region known for strong ocean-atmosphere coupling and ENSO-related variability.

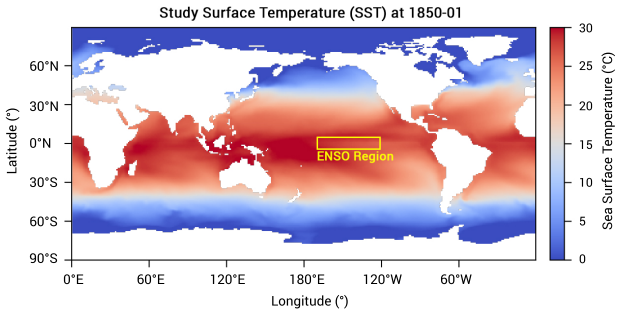

Figure 9: Global map indicating the spatial extent of ENSO region (highlighted in yellow). The region spans from 4.5°S to 4.5°N and 169.5°W to 120.5°W in the equatorial Pacific.

D.5 Additional Experimental Results

Table 8: Short-term forecasting results in terms of CRPS, QICE, and IS. Bold indicates the best performance, while underlining denotes the second-best. DYffusion is limited to grid-format data, and "-" denotes results that are not applicable.

Model	BikeDC			MobileNJ			CrowdBM			Los-Speed		
	CRPS	QICE	IS	CRPS	QICE	IS	CRPS	QICE	IS	CRPS	QICE	IS
GP	0.494	0.120	6.69	0.435	0.129	4.76	0.620	0.159	66.8	0.789	0.1579	61.7
DeepState	0.728	0.084	15.5	0.518	0.065	4.40	0.689	0.057	97.0	0.086	0.040	40.9
D3VAE	0.785	0.157	8.77	0.565	0.096	6.03	0.593	0.110	136.4	0.119	0.089	90.5
DiffSTG	0.692	0.157	8.08	0.291	0.071	3.11	0.453	0.047	68.5	0.078	0.045	50.9
TimeGrad	0.469	0.130	5.65	0.432	0.162	5.87	**0.240**	0.085	46.9	**0.031**	0.098	**20.8**
CSDI	0.529	0.057	4.79	0.111	0.039	0.80	0.390	0.054	61.1	0.059	0.026	30.8
TMDM	2.32	0.125	29.6	1.49	0.126	87.5	3.46	0.124	217.3	0.897	0.126	83.4
NPDiff	0.442	0.066	7.11	0.128	0.133	2.22	0.331	0.119	91.2	0.057	0.023	30.5
DYffusion	0.573	0.079	6.46	0.196	0.080	1.80	-	-	-	-	-	-
CoST	**0.419**	**0.028**	**3.45**	**0.089**	**0.032**	**0.66**	0.256	**0.027**	**37.8**	0.056	**0.023**	31.9

Table 9: Short-term forecasting results in terms of MAE and RMSE. Bold indicates the best performance, while underlining denotes the second-best. DYffusion is limited to grid-format data, and "-" denotes results that are not applicable.

Model	BikeDC		MobileNJ		CrowdBM		Los-Speed	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
GP	0.941	1.74	0.257	0.682	6.35	17.7	6.60	11.0
DeepState	1.98	3.81	0.582	0.827	13.9	23.2	6.50	9.23
D3VAE	0.871	3.59	0.580	1.135	11.0	24.7	8.28	11.9
DiffSTG	0.770	4.02	0.317	0.649	8.88	21.3	5.38	9.75
TimeGrad	0.843	**1.07**	0.340	0.357	10.1	12.4	**2.33**	**3.00**
CSDI	0.592	3.10	0.129	0.237	7.31	19.3	4.53	8.07
TMDM	2.44	4.11	3.27	4.10	72.9	94.8	9.42	13.9
NPDiff	**0.435**	1.90	0.123	0.175	5.42	13.7	4.07	7.64
DYffusion	0.480	1.37	0.222	0.357	-	-	-	-
CoST	0.492	1.76	**0.102**	**0.172**	**5.04**	**12.1**	4.05	7.30

Table 10: Long-term forecasting results in terms of MAE and RMSE. Bold indicates the best performance, while underlining denotes the second-best. DYffusion is limited to grid-format data, and "-" denotes results that are not applicable.

Model	MobileSH		SST		CrowdBJ		CrowdBM		Los-Speed	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
GP	0.399	0.627	1.52	1.65	2.52	3.75	7.37	20.7	6.69	11.2
DeepState	0.160	0.199	1.13	1.40	8.53	11.0	21.5	31.9	10.1	14.2
D3VAE	0.207	0.392	2.39	3.13	5.63	11.4	12.4	28.2	9.43	13.3
DiffSTG	0.078	0.125	0.94	1.19	3.04	6.37	7.59	18.8	7.77	14.2
TimeGrad	0.058	0.072	1.30	1.64	3.48	4.83	5.25	**7.40**	18.2	22.3
CSDI	**0.035**	0.057	1.31	1.63	1.99	3.64	**4.64**	12.4	11.3	15.0
TMDM	0.519	6.50	1.55	1.73	3.54	8.32	15.2	29.0	34.2	43.1
NPDiff	0.037	0.057	1.91	2.82	2.06	3.28	5.44	13.8	46.0	58.3
DYffusion	0.047	0.066	**0.85**	**1.06**	-	-	-	-	-	-
CoST	0.035	**0.053**	0.86	1.13	**1.92**	**3.05**	4.74	11.2	**5.94**	**10.8**

Table 11: Long-term forecasting performance comparison of TMDM on ETTh1 and ETTh2 Datasets.

Model	EETh1			EETh2		
	CRPS	QICE	IS	CRPS	QICE	IS
TMDM	0.395	0.041	4.8	0.196	0.018	2.2
CoST	0.311	0.007	1.6	0.109	0.007	0.78

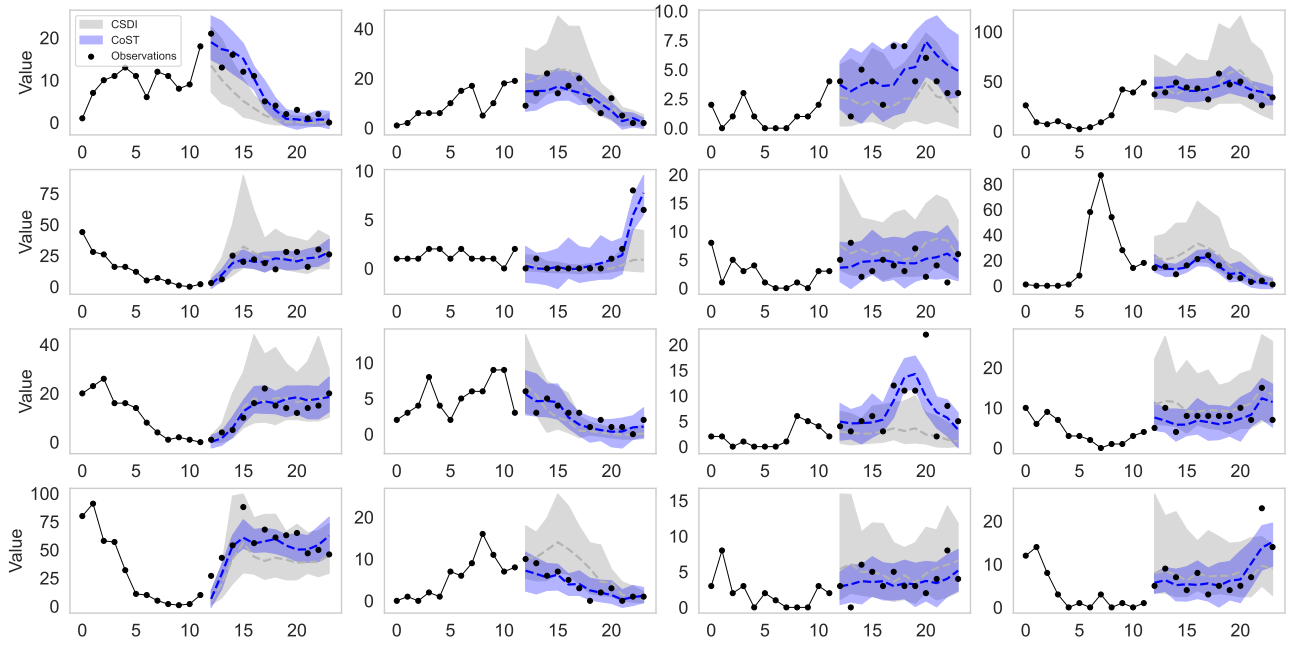

Figure 10: Visualizations of predictive uncertainty for both CSDI and CoST on the CrowdBJ dataset. The shaded regions represent the 90% confidence interval. The dashed lines denote the median of the predicted values for each model.