CogDPM: Diffusion Probabilistic Models via Cognitive Predictive Coding

Kaiyuan Chen [*1] **Xingzhuo Guo** [*1] **Yu Zhang** [1] **Jianmin Wang** [1] **Mingsheng Long** [1]

Abstract

Predictive Coding (PC) is a theoretical framework in cognitive science suggesting that the human brain processes cognition through spatiotemporal prediction of the visual world. Existing studies have developed spatiotemporal prediction neural networks based on the PC theory, emulating its two core mechanisms: Correcting predictions from residuals and hierarchical learning. However, these models do not show the enhancement of prediction skills on real-world forecasting tasks and ignore the *Precision Weighting* mechanism of PC theory. The precision weighting mechanism posits that the brain allocates more attention to signals with lower precision, contributing to the cognitive ability of human brains. This work introduces the *Cognitive Diffusion Probabilistic Models* (CogDPM), which demonstrate the connection between diffusion probabilistic models and PC theory. CogDPM features a precision estimation method based on the hierarchical sampling capabilities of diffusion models and weight the guidance with precision weights estimated by the inherent property of diffusion models. We experimentally show that the precision weights effectively estimate the data predictability. We apply CogDPM to real-world prediction tasks using the United Kindom precipitation and ERA surface wind datasets. Our results demonstrate that CogDPM outperforms both existing domain-specific operational models and general deep prediction models by providing more proficient forecasting.

[*]Equal contribution [1]School of Software, BNRist, Tsinghua University. Kaiyuan Chen <cky21@mails.tsinghua.edu.cn>. Correspondence to: Mingsheng Long <mingsheng@tsinghua.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Predictive Coding (PC) is a theoretical construct in cognitive science, positing that the human brain cognizes the visual world through predictive mechanisms (Spratling, 2017; Hohwy, 2020). The PC theory elucidates that the brain hierarchically amends its perception of the environment by anticipating changes in the visual world. Researchers have developed computational models based on the PC theory to simulate the brain's predictive mechanisms (Keller & Mrsic-Flogel, 2018). Neuroscientists employ these models to empirically validate the efficacy of the PC theory and to find new characteristics. Precision weighting, a pivotal feature of the PC theory, suggests that the brain assigns more attention to signals with lower precision by using precision as a filter in weighting prediction errors.

With the advancement of deep learning, predictive learning has emerged as one of the principal learning methods (Rane et al., 2020; Bi et al., 2023). Neural networks are now capable of making effective predictions in video data (Shi et al., 2015; Wang et al., 2017; Ho et al., 2022c). Deep video prediction models have rich applications, such as weather forecasting (Ravuri et al., 2021; Zhang et al., 2023) and autonomous driving simulation (Wang et al., 2018; Wen et al., 2023).

Researchers design cognitively inspired video prediction models utilizing the PC theory. PredNet (Lotter et al., 2020), which employs multi-layer ConvLSTM (Shi et al., 2015) networks to predict the next frame in a video sequence, is responsible for predicting the residual between the outcomes of a network layer and the ground truth values. However, the predictive capability of PredNet does not show significant improvement over non-hierarchical video prediction models and has not been validated in real-world video prediction tasks. We posit that the hierarchical modeling mechanism in PredNet is not effectively implemented. PredNet directly targets low signal-to-noise ratio residuals as learning objectives, which complicates the learning process, and fails to extract fundamentally distinct features between layers. Additionally, PredNet lacks the capability to model precision, leading to uniform weighting in learning residuals across different regions. This results in redundant noise information becoming a supervisory signal and hinders the model's ability to learn from important information.

In this study, we propose PC-inspired *Cognitive Diffusion Probabilistic Models* (CogDPM), which align the main features of PC theory with Diffusion Probabilistic Models (DPMs), a specialized branch of deep generative models. The CogDPM framework innovatively abstracts the multi-step inference process characteristic of Diffusion Probabilistic Models into a hierarchically structured model, where each layer is responsible for processing signals at distinct spatiotemporal scales. This hierarchical approach allows for a progressive enhancement in the model's interpretation of sensory inputs, actively working to reduce prediction errors through iterative refinement. A key feature of the CogDPM framework is its ability to estimate spatiotemporal precision weights based on the variance of states in each hierarchical layer. This methodology plays a crucial role in optimizing the overall precision of predictions, and represents a novel advancement in predictability modeling.

We verify the effectiveness of precision weights as well as the predictions skills of CogDPM on real-world spatiotemporal forecasting tasks. To verify precision weights, we use synthetic motion datasets of both rigid body and fluid. Results show precision weights get higher salience on the hard-to-predict region. To validate the prediction capabilities of CogDPM, we apply CogDPM to real-world tasks including precipitation nowcasting (Shi et al., 2015; Ravuri et al., 2021) and high wind forecasting (Barbounis et al., 2006; Soman et al., 2010). We evaluate CogDPM through case studies focusing on extreme weather events and scientific numerical metrics. CogDPM outperforms operational domain-specific models FourCastNet (Pathak et al., 2022) and DGMR (Ravuri et al., 2021) as well as the general deep predictive models. We demonstrate that CogDPM has strong extreme event prediction capabilities and verify the effectiveness of precision estimations of CogDPM which provide useful information for weather-driven decision-making.

In summary, we identify the following advantages of CogDPM:

- CogDPM aligns diffusion probabilistic models with Predictive Coding theory, which inherently integrates hierarchy prediction error minimization with precision-weighting mechanics.
- CogDPM delivers skillful and distinct prediction results, particularly in scientific spatiotemporal forecasting, demonstrating a marked improvement in probabilistic forecasting metrics.
- CogDPM presents a novel method for predictability estimation, providing index of confidence modeling for probabilistic forecasting.

2. Related Work

Predictive Learning. Predictive learning is a subfield of machine learning that utilizes historical data to make predictions about future events or outcomes. As an important aspect of human cognition that plays a crucial role in our ability to perceive and understand the world, spatiotemporal predictive learning has triggered a substantial amount of research efforts, such as ConvLSTM (Shi et al., 2015), PredRNN (Wang et al., 2017), and ModeRNN (Yao et al., 2023). Recently, diffusion models (Ho et al., 2020) have been successfully applied in video generation (Ho et al., 2022a) so as to capture spatiotemporal correlations, showing a promising trend as a spatiotemporal predictive learning framework.

Predictive Coding. In neuroscience, predictive coding is a theory of brain function about how brains create predictions about the sensory input. Rao & Ballard translates the idea of predictive coding into a computational model based on extra-classical receptive-field effects, and shows the brain mechanism of trying to efficiently encode sensory data using prediction. Further research in neuroscience (Friston, 2009; Clark, 2013; Emberson et al., 2015; Spratling, 2017) presents different interpretations of predictive coding theory.

Predictive Coding Neural Networks. The development of deep learning has arisen plenty of deep predictive networks with cognition-inspired mechanisms. PredNet (Lotter et al., 2016) implements hierarchical predictive error with ConvLSTM for spatiotemporal prediction using principles of predictive coding. CPC (Oord et al., 2018; Henaff, 2020) and MemDPC (Han et al., 2020) incorporate contrastive learning in the latent space via a predictive-coding-based probabilistic loss. PCN (Wen et al., 2018; Han et al., 2018) proposes a bi-directional and recurrent network to learn hierarchical image features for recognition. Such models introduce the motivation of predictive coding in their task-specific manners. However, these works ignore precision weighting, a pivotal mechanism in PC theory. Besides, these works have not explored a proper PC-based framework of diffusion models.

3. Method

Spatiotemporal forecasting involves extracting patterns from a sequence of vector fields $\mathbf{c}^{-N_0:0}$ and providing future evolution $\mathbf{x}^{1:N}$. We give a brief introduction to the framework of predictive coding and propose our CogDPM for implementing Predictive Coding into spatiotemporal forecasting.

To avoid confusion, we use the superscript N to represent different moments allowed by time, and the subscript t to denote the ordinal number of the inference steps in the diffusion model.

3.1. CogDPM via Predictive Coding

Figure 1a presents a conceptual demonstration of a predictive coding (PC) system. Based on PC theory, we propose *Cognitive Diffusion Probabilistic Models (CogDPM)* for spatiotemporal forecasting based on multi-step denoising (Ho et al., 2020), which realizes the core mechanisms of hierarchical inference and prediction error minimization. Fig. 1b shows the framework of CogDPM, which takes past observations as input to forecast the evolution of future fields and estimate corresponding prediction error.

Hierarchical Inference. Predictive coding theory describes that the brain makes spatiotemporal predictions of the sensations through hierarchical inference with multilayer organized estimators (Walsh et al., 2020). While different layers of the PC system are responsible for processing features at different spatial scales, the hierarchical system gradually performs prediction error minimization and converges on a final consistent predictions (Wiese & Metzinger, 2017). CogDPM aligns the multi-step inference of DPM with the hierarchical inference of the PC system. In the inference phase of CogDPM, the forecast is gradually generated in the hidden states evolution process from $\mathbf{x}_T, \mathbf{x}_{T-1}, \dots$ to $\mathbf{x}_0$, where $\mathbf{x}_T$ is a Gaussian prior and $\mathbf{x}_0$ indicates the generated target distribution of forecast. CogDPM inherits the properties of DPM that the different inference steps have varying spatial and temporal scales of feature expression capabilities (Zheng et al., 2022). In the initial stages of inference, the model yields holistic and vague results. As it approaches the final steps, the model shifts its focus towards supplementing with detailed information, which is also aligned with the hierarchical property of the PC system. In each internal inference step, the guidance of the diffusion model plays a similar role with the error units of the PC system, taking observation sequence as input and strengthen the correlation between generated results and observations (Dhariwal & Nichol, 2021).

Prediction Error Minimization. Each layer in the PC system outputs two key components: predictions for future sensations and estimations of prediction errors (van Elk, 2021). This process is enabled by interactions between two functionally distinct neural sub-components in the layer: expectation units and error units (Walsh et al., 2020). The expectation unit updates expected sensory states from the previous level to the error units, without directly receiving sensory-driven signals as input. The error unit receives and analyzes the discrepancies between perceptual and expected sensory states to compute the error, which is then fed back to the expectation unit in the next layer. The goal of the information transfer between multiple layers is to minimize prediction errors, ultimately resulting in more accurate environmental perceptions. CogDPM couples a generative DPM G_θ with a perceptual DPM P_θ, where θ represents their sharing parameters. The previous state $\mathbf{x}_t$ is the sharing input of both models, while observations $\mathbf{c}$ can only be attached by the perceptual DPM. With the previous state as observation, the perceptual DPM acts as sensory stimuli and thus aligns with the bottom-up process in the PC system. The generative DPM, as a comparison, performs as the top-down prediction based on conceptual knowledge. Fig. 1c provides detailed schematic diagram of a single step in CogDPM. Given the outputs $G_\theta(\mathbf{x}_t)$ and $P_\theta(\mathbf{x}_t, \mathbf{c})$ separately for each step t, the guidance for predictive error minimization can be expressed by:

$$\text{Guidance}[\mathbf{x}_t] = P_\theta(\mathbf{x}_t, \mathbf{c}) - G_\theta(\mathbf{x}_t), \tag{1}$$

i.e., the difference between sensations and predictions.

3.2. Precision Weighting in CogDPM

Precision weighting stands as the pivotal mechanism for filtering information transmitted between adjacent layers. It posits that the brain expends more effort in comprehending imprecise information, recognizing that sensory input often contains a substantial proportion of redundant information, which does not necessitate repetitive processing (Hohwy, 2020). During each error minimization phase of the predictive coding (PC) approach, the error unit generates precision maps. These maps selectively filter the signal transmitted to the subsequent layer, assigning greater weight to signals characterized by higher imprecision.

Following precision weighting in PC theory, our goal is to design a modeling of imprecision for each denoising process of CogDPM. We therefore delve into the progressive denoising mechanism in the backward process of DPMs. In each denoising step for $\mathbf{x}_t$, the model predicts a noise towards the corresponding groundtruth $\mathbf{x}_0$ (Song et al., 2020). The model usually shifts $\mathbf{x}_t$ into $\mathbf{x}_{t-1}$ within a tiny step and recursively performs the process to get $\mathbf{x}_0$, but can either directly obtain $\mathbf{x}_0$ within a single larger step. If the direct predictions from step t and from step $t+1$ with generative DPM G_θ differ in a significant manner for a certain spatiotemporal region, the single step produces inconsistent signal from previous steps, indicating the imprecision of the generative model at such region of the current state. Hence, we use the fluctuation field of direct predictions $\mathbf{x}_0$ from $\{\mathbf{x}_t, \dots, \mathbf{x}_{t+k-1}\}$ to estimate such imprecision of state $\mathbf{x}_t$ for each coordinate, formulated by Eq. (2):

$$\mathrm{U}[\mathbf{x}_t] = \mathrm{Var}\left[\mathbb{E}_{G_\theta}\left[\mathbf{x}_0 \mid \mathbf{x}_t\right], \dots, \mathbb{E}_{G_\theta}\left[\mathbf{x}_0 \mid \mathbf{x}_{t+k-1}\right]\right], \tag{2}$$

where Var stands for the variance field along the denoising step, and k is the hyperparameter for window length. In this way, CogDPM provides a modeling of the inverse precision field for multiscale spatiotemporal coordinates in the inference steps. Since only the past observation is given in the

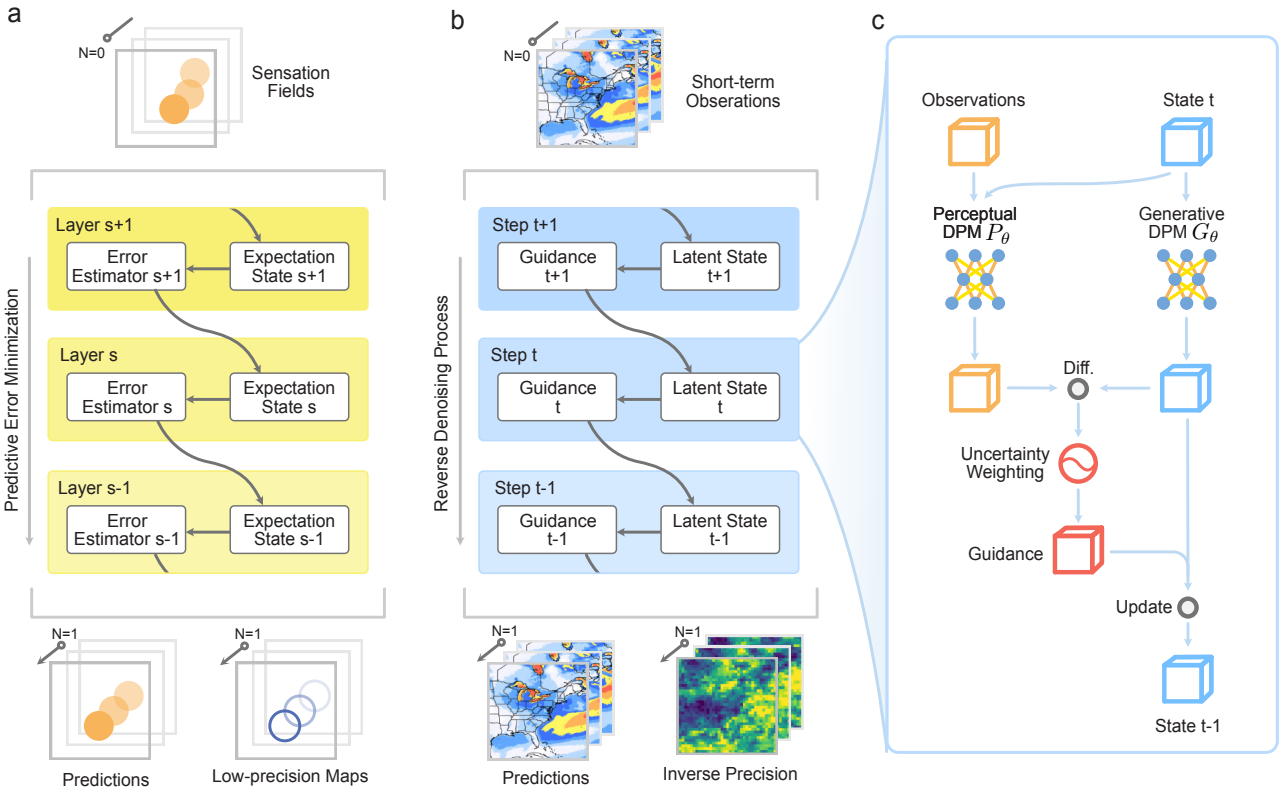

Figure 1. **a**, A general predictive coding framework. The system recognizes the sensation fields with hierarchy error units and expectation units and generates the predictions and precision maps during the process. **b**, Cognitive Diffusion Probabilistic Models (CogDPM) framework, providing predictions and precision weights with multi-step denoising process. **c**, Updates of latent states with precision-weighted predictive error.

forecasting tasks, this precision is a good substitution for the actual precision to weight the minimization. We implement precision weighting in the CogDPM framework, which can be formulated as Eq. (3),

$$\mathbf{x}_{t-1} = G_\theta(\mathbf{x}_t) + f(\mathrm{U}[\mathbf{x}_t]) \cdot \mathrm{Guidance}[\mathbf{x}_t], \tag{3}$$

where f is a parameter-free normalization function shown in Eq. (8). Precision weighting helps to control the balance between diversity and the alignments with the observation, with larger guidance increasing the alignments and decreasing the diversity or the quality of generations. Through this precision weighting mechanism, CogDPM strategically allocates greater guidance intensity to regions with lower predictability, thereby enhancing local precision in a focused manner.

Computational details. The framework of a standard DPM starts with $\mathbf{x}_0$ sampled from data distribution, and latent states $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ following the forward process along a Markov chain as Eq. (4).

$$q(\mathbf{x}_{t+1} \mid \mathbf{x}_t) = \mathcal{N}\left(\sqrt{\alpha_t}\mathbf{x}_t, \sqrt{1-\alpha_t}\mathbf{I}\right), \tag{4}$$

where $\{\alpha_t\}_{t=1,2,\ldots,T}$ are constant parameters. Each latent state is a corrupted estimation for the future inputs with the three-dimensional shape of $N \times H \times W$.

In each step of the backward process, we update the latent state with the denoising network ϵ_θ. We denote the sensation input as $\mathbf{c}$, which has a shape of $N_0 \times H \times W$. The perceptual model P_θ and generative model G_θ can be preformed separately as Eq. (5) and (6).

$$P_\theta(\mathbf{x}_t, \mathbf{c}) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, \mathbf{c})\right), \tag{5}$$

$$G_\theta(\mathbf{x}_t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, \varnothing)\right), \tag{6}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and ϵ_θ is the denoising network of the DPM. CogDPM provides inverse precision estimation with Eq. (2), and $\mathbb{E}_{G_\theta}[\mathbf{x}_0 \mid \mathbf{x}_t]$ can be computed as Eq. (7):

$$\mathbb{E}_{G_\theta}[\mathbf{x}_0 \mid \mathbf{x}_t] = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, \varnothing)\right). \tag{7}$$

For implementation, we push $G_\theta(\mathbf{x}_t)$ into the estimation queue with a maximal queue length of k, and estimate the precision with Eq. (2). Thus, we can merge $G_\theta(\mathbf{x}_t)$ and $P_\theta(\mathbf{x}_t, \mathbf{c})$ with respect to the control of precision with Eq. (3). Considering numerical stability, we normalize the

inverse precision field in $\mathrm{U}(\mathbf{x}_t)$ and clip the value in a fixed range. The formulation of f is following:

$$f(\mathbf{w}) = \lambda \cdot \mathrm{clip}\left(\frac{\mathbf{w} - \bar{\mathbf{w}}}{\sigma(\mathbf{w})}, 0, 1\right) + 1, \tag{8}$$

where $\bar{\mathbf{w}}$ and $\sigma(\mathbf{w})$ are the mean and standard error of $\mathbf{w}$, λ is a constant that controls the guidance strength. Finally, we merge $G_\theta(\mathbf{x}_t)$ and $P_\theta(\mathbf{x}_t, \mathbf{c})$ with the guidance weight by inverse precision as Eq. (3). The pseudo code of the inference process of CogDPM framework is shown in Algorithm 1.

Objective function. CogDPM follows the training schema in diffusion probabilistic model (Ho et al., 2020) that predicts the noise from the corrupted inputs. We denote the loss term as $\mathcal{L}(\theta)$. The denoising U-Net ϵ_θ has parameters θ, and takes the corrupted future observations $\mathbf{x}_s$, contexts $\mathbf{c}$ and the scalar diffusion step s as input. We adopt the L1 loss to minimize the error between injected noise and the prediction of the denoising U-Nets.

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\epsilon,\mathbf{c}}\left[\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \mathbf{c}, t\right)\|_1\right] \tag{9}$$

To jointly train the conditional and unconditional models, $\mathbf{c}$ is replaced by $Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ with 10% probability.

Algorithm 1 Inference Process of CogDPM framework

Input: Context input $\mathbf{c}$, denosing model ϵ_θ, maximul queue length L
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}_\mathbf{x}, \mathbf{I}_\mathbf{x})$
Define free estimation queue Q^{free}
for $t = T$ **to** 1 **do**
  $\epsilon_\mathbf{c} \sim \mathcal{N}(\mathbf{0}_\mathbf{c}, \mathbf{I}_\mathbf{c})$
  $\epsilon_t^{\text{cond}} = \epsilon_\theta(\hat{\mathbf{x}}_t, \mathbf{c})$ {Network output with condition $\mathbf{c}$.}
  $\epsilon_t^{\text{free}} = \epsilon_\theta(\hat{\mathbf{x}}_t, \epsilon_\mathbf{c})$ {Network output without condition.}
  $P_\theta(\mathbf{x}_t, \mathbf{c}) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x_t} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t^{\text{cond}})$
  $G_\theta(\mathbf{x}_t) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x_t} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t^{\text{free}})$
  $\hat{\mathbf{x}}_{t\to 0} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t^{\text{free}}\right)$ {Estimate $\mathbf{x}_0$ with $\mathbf{x}_t$.}
  Push $\hat{\mathbf{x}}_{t\to 0}$ into Q^{free}
  if Length of Q^{free} exceeds L **then**
    Drop last term from Q^{free}
  end if
  Get inverse precision estimation $\mathbf{w} = f(\mathrm{Var}(Q^{\text{free}}))$
  $\mathbf{x}_{t-1} = G_\theta(\mathbf{x}_t) + \mathbf{w} \cdot (P_\theta(\mathbf{x}_t, \mathbf{c}) - G_\theta(\mathbf{x}_t))$ {Prediction error minimization with precision weighting.}
end for
Output: $\mathbf{x}_0$

4. Experiments

We demonstrate that by incorporating the novel design inspired by the cognitive predictive process, CogDPM can deliver more skillful and improved results in tasks of scientific spatiotemporal field prediction.

4.1. Synthesis Data Experiments

In this section, we compare the predictive performance of CogDPM with other mainstream deep predictive networks and investigate the interpretability of Precision weighting within the CogDPM framework in the context of spatiotemporal prediction. We expect high correlation between the precision estimation and the predictability of CogDPM. The inverse precision estimator should allocate more attention to the region with higher prediction difficulty.

Benchmarks. We conduct experiments on the MovingMNIST dataset (Wu et al., 2021), which simulates the motion of rigid bodies, and the Turbulence flow dataset, which models fluid dynamics. The Moving MNIST dataset is generated with the same method as (Wu et al., 2021). We create sequences with 20 frames, and each frame contains three handwriting digits. The motion of digits consists of transition, reflection, and rotation. Models predict the next 16 frames with 4 continuous context frames. The turbulent flow dataset is proposed by (Rui et al., 2020). We follow the same dataset parameters as Rui et al. and generate a sequence with 15 frames and 64 x 64 grids on each frame. Four frames are taken to predict the next 11 frames.

We have selected a diverse array of deep spatiotemporal forecasting models as baselines for our study. These include the Transformer-based spatiotemporal forecasting model FourCastNet (Pathak et al., 2022) , RNN-type networks such as MotionRNN (Wu et al., 2021) and PredRNN-v2 (Wang et al., 2022), the physics-inspired predictive model PhyDNet (Guen & Thome, 2020), and a predictive DPM model that employs naive Classifier-free Guidance (Ho & Salimans, 2021) and utilizes the same network architecture as CogDPM.

For the evaluation metrics, we have chosen the Neighborhood-based CRPS (Continuous Ranked Probability Score), CSI (Critical Success Index), and FSS (Fractional Skill Score), which are commonly used in scientific forecasting tasks. The CRPS metric emphasizes the ensemble forecasting capabilities of the model, with lower values indicating better predictive performance. On the other hand, the CSI and FSS metrics focus on assessing the accuracy of the model's predictions in peak regions, with higher values denoting stronger predictive capabilities. The implementation details of these metrics are provided in the appendix D, and we will continue to employ them in subsequent experiments on real-world datasets.

Methods / Metrics	MovingMNIST				Turbulence			
	CRPS ↓		CSI ↑ (w5)	FSS ↑ (w5)	CRPS ↓		CSI ↑ (w5)	FSS ↑ (w5)
	(w8, avg)	(w8, max)			(w8, avg)	(w8, max)		
FourCastNet	0.0619	0.2288	0.1915	0.3261	0.0098	0.0119	0.3761	0.6558
MotionRNN	0.0377	0.1232	0.4859	0.6758	0.0037	0.0046	0.7235	0.9354
PhyDNet	0.0325	0.0983	0.6161	0.7969	0.0079	0.009	0.5456	0.8254
PredRNN-v2	**0.027**	0.0774	0.688	0.8471	0.0033	0.0042	0.7529	0.9507
DPM	0.0323	0.082	0.6959	0.822	0.0023	0.0096	0.6725	0.9668
CogDPM (ours)	**0.027**	**0.0697**	**0.7365**	**0.8588**	**0.0023**	**0.0034**	**0.7962**	**0.9722**

Table 1. Numerical Evaluation of Prediction Skills on MovingMNIST and Turbulence Datasets

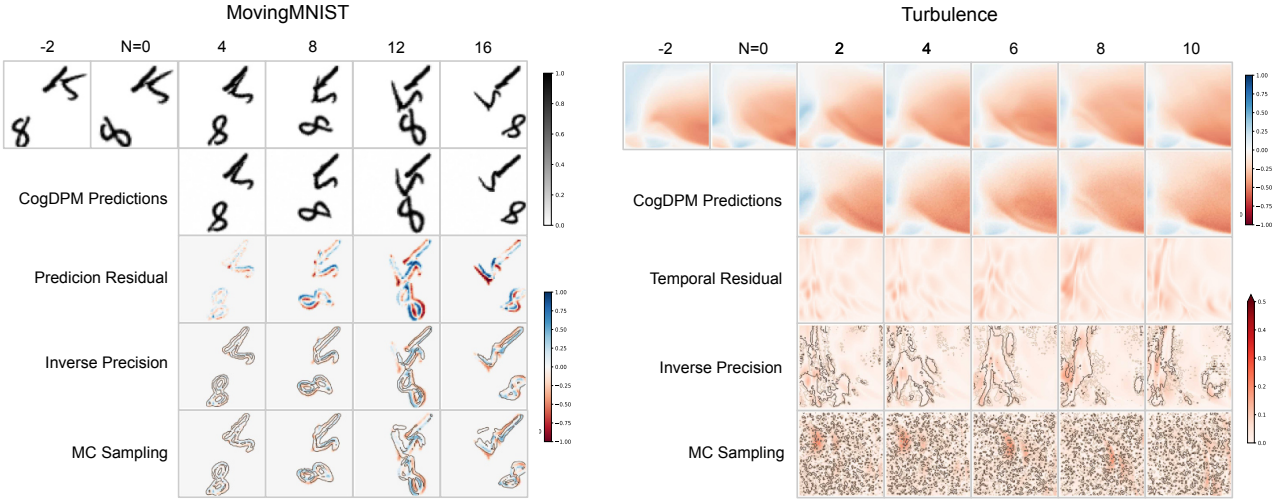

Figure 2. Predictions and inverse precision of CogDPM on rigid-body MovingMNIST dataset *(left)* and Turbulence flow dataset *(right)*.

Numerical Results Table 4 presents the numerical evaluation results for two datasets. Here, w denotes the window size employed in the Neighborhood-based assessment method, while avg and max represent the average and maximum values obtained from this method, respectively. The CogDPM model demonstrates consistent improvements over the baseline models in terms of the CRPS, which measures the average ensemble forecasting capability, as well as the CSI and FSS indicators, which assess the accuracy of the model's predictions in the peak regions. Additionally, when compared to the DPM model based on naive Classifier-free Guidance, CogDPM exhibits superior performance. This underscores the beneficial impact of introducing the Precision Weighting mechanism on enhancing the model's predictive efficacy.

Interpretability of precision weights. Figure 2 presents the outcomes of the CogDPM model. The initial two rows delineate the ground truth images alongside the corresponding prediction results generated by CogDPM. The third row illustrates the prediction residuals, representing the discrepancies between the actual and predicted data as depicted in the preceding rows. The fourth row features images that overlay the inverse precision map, highlighting the top 20% of values with a black contour line, against a backdrop of the residual map. The fifth row shows the precision map estimated by Monte Carlo sampling which estimate the prediction confidence with the variation among multiple independent predictions with difference noise prior (Zhang, 2021).

CogDPM provides reasonable predictions in both datasets. In the prediction of rigid body motion, the estimated Inverse Precision effectively encompasses the Precision Residuals, which are primarily located at the edges of objects. The edges of objects present a greater challenge for prediction compared to blank areas or the interior of objects. This outcome aligns with our expectations for the estimation of the precision map. Precision estimated with MC sampling works similarly but provide more false positive region in frame 12 and 14.

In the prediction of fluid motion, regions with large temporal residuals exhibit higher accelerations, indicating increased predictive difficulty. The estimated Inverse Precision in-

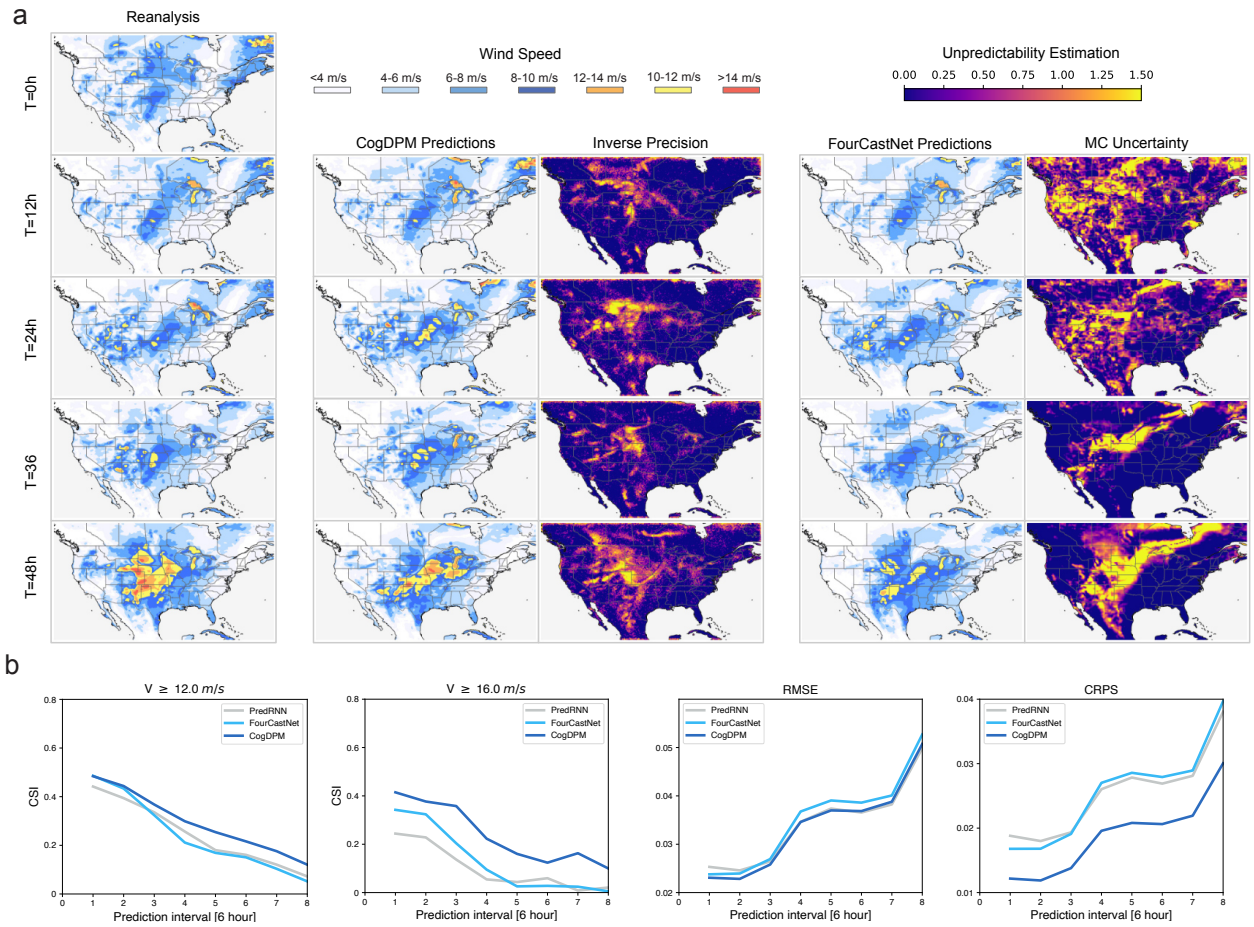

Figure 3. **Experiments on high wind forecasting. a**, a Case study of the ERA5 wind forecast from 2017-03-04 18:00. High wind and tornadoes attacked the Mideast USA at 2017-03-06 18:00(T=48h) (Twin Cities, 2017). CogDPM provides alarming forecasts, covering states with the most severe weather reports, Iowa and Missouri. CogDPM precision indicate the credibility of the predictions, helping forecasters to identify the missing and false positive regions. **b**, Numerical scores on ERA5 wind dataset from 2017-01-01 to 2019-12-31. We report CSI with 12 m/s (first) and 16 m/s (second) threshold, RMSE (third), and CRPS across four ensembles (fourth).

deed covers the Temporal Residuals well, meeting our expectations. We observe that in both fluid and rigid body motion prediction tasks, the Precision weights of CogDPM exhibit varying styles, yet consistently depict the model's confidence on current case. On comparison, MC sampling method almost fails in this case due to the over-confidence of the prediction result. Difference among multiple predictions have no significant signals but random noise. While, the CogDPM is not effected because its precision describe the continuous enhancing process of model's confidence during the hierarchy inference.

4.2. Surface Wind Forecasting Experiments

Benchmarks. We first evaluate our model by applying it to the task of surface wind forecasting, using the ERA5 reanalysis dataset (Hersbach et al., 2023). Accurate wind field forecasting is crucial for various applications in energy and weather domains. Ensemble forecasting is a key technique to provide more useful information for the forecasters, which provides multiple predictions and the confidence of its predictions. We show that CogDPM not only provide better ensemble forecasts results, but also estimate the prediction confidence with its precision weights.

We choose real-world operational metrics for evaluation. In the meteorology domain, forecasters focus on evaluating the risk of high wind and confirming the time for extreme weather issue warnings. On this purpose, we use Critical Success Index (CSI) to measure the consistency between heavy wind regions in forecasts and ground truths. In the energy domain, accurate wind field forecasting supports the prediction of wind power, which is essential for the fluctuation control of clean energy generation (Marugán et al., 2018). Absolute wind speed is the dominant factor that affect the power production of the wind turbine (Porté-Agel et al., 2013); thus, we consider pixel-wise Root Mean Square Error (RMSE) and Radially Continuous ranked prob-

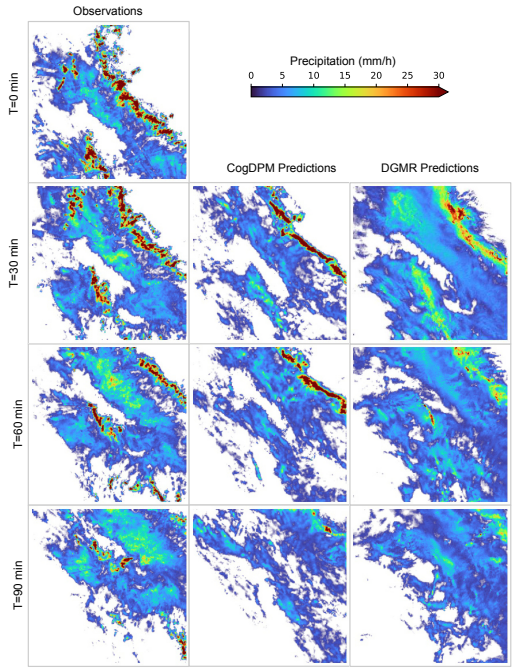

Figure 4. **Experiments on precipitation nowcasting.** Case study on an extreme precipitation event starting on 2019-07-24 at 03:15 in the UK timezone, CogDPM successfully predicts movement and intensity variation of the squall front, while DGMR produces results with early dissipation.

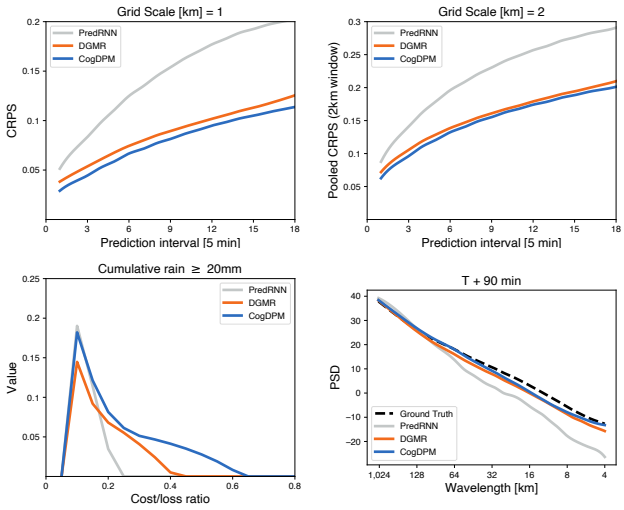

Figure 5. **Experiments on precipitation nowcasting.** Numerical verification scores on sampled the United Kingdom precipitation dataset in 2019. CRPS is computed with four ensembles for spatial pooling size 1km x 1km (left top) and 2 km x 2 km (right top); Economic value with 20 mm/h accumulative rain threshold (left bottom); Radially averaged power spectral density on predictions at 90 minutes (right bottom). CogDPM surpasses the operational forecast model DGMR in ensemble forecasting precision and forecast skillfulness.

ability score (CRPS) on wind speed for the evaluation of this scenario (Barbounis et al., 2006). Applendix D shows detailed implemation of these metrics.

Results. We use the ERA5 reanalysis surface wind data and crop patches centered in the US spanning from 1979 to 2021. We evaluate predictions for the next 48 hours with 6-hour intervals using the observations in past 24 hours. We compare the proposed method with FourCastNet (Pathak et al., 2022), a domain-specialized network for reanalysis field forecasting, and predictive recurrent networks for deterministic video prediction. FourCastNet provides ensemble forecasts based on the Gaussian disturbance on the initial states following (Evensen, 2003).

Figure 3a shows studies on a case starting from 2017-03-04 18:00. The results from FourCastNet indicate a failure to accurately forecast the growing high wind region, and the high wind region is underestimated in the 48-hour forecast. In contrast, results from CogDPM not only locate the high wind region more accurately, but also provide intensity estimates much closer to the ground truth, supporting the need for 48-hour-ahead precautions. CogDPM are capable of providing alarming forecasts around 2017-03-06 18:00, when high wind and tornadoes attacked the Mideast USA[1].

We also visualize the inverse precision fields corresponding to the forecasts, since confidence estimation provide key information for decision-making. In the forecast for the first 24 hours, the uncertainty fields given by FourCastNet are relatively dispersed and not closely related to the evolution of the associated wind field. In the next time period to the 48 hours, FourCastNet produces unreasonable estimates for the windless area in the upper right corner. The inverse precision fields given by CogDPM had much closer correlations to the weather process. In the 48-hour forecast, CogDPM underestimated the forecast intensity in Wyoming and Colorado, but allocated lower precision on that region.

Figure 3b shows that CogDPM outperforms baseline methods on CSI, particularly for heavier wind thresholds. For the measurement of RMSE, we take the mean across eight ensemble forecasts for all methods. Although DPMs are not directly optimized by the Mean Squared Error (MSE) loss, the mean ensemble results are competitive with predictive models trained with MSE losses. The CogDPM exhibits a lower CRPS across all prediction times, indicating its ability to effectively generate ensemble forecasts.

Our results demonstrate that CogDPM is capable of making predictions under severe conditions, supported by the

[1]Summary of March 06 2017 Severe Weather Outbreak - Earliest Known Tornado in Minnesota's History, https://www.weather.gov/mpx/SevereWeather_06March2017

probabilistic forecast ability of the PEM process, while deterministic models avoid predicting severe cases to reduce mistake-making risk.

4.3. Precipitation Nowcasting Experiments

Benchmarks. We evaluate our model on the precipitation nowcasting task using the United Kingdom precipitation dataset (Ravuri et al., 2021). Precipitation nowcasting aims to predict high-resolution precipitation fields up to two hours ahead, which provides socioeconomic value on weather-dependent decision-making (Ravuri et al., 2021). Precipitation data is extremely unbalanced on spatiotemporal scales, demanding nowcasting models to focus on vital parts of the field. Fig. 4a shows a case study selected by the chief meteorologist from MetOffice (Ravuri et al., 2021), which involves a squall line sweeping across the United Kingdom. We choose DGMR as a strong baseline on skillful nowcasting (Ravuri et al., 2021), which is data-driven method that forecast precipitation with a generative adversarial network. DGMR is also the operational method deployed by Met Office of the United Kingdom.

Results. In Figure 4, our results accurately forecast both the trajectory and intensity fluctuations of the squall line, as depicted by the red precipitation line in the top right segment. CogDPM's forecasts consistently show the squall line progressing over 30 and 60 minutes, followed by dissipation at the 90-minute mark, mirroring actual events. Conversely, predictions from DMGR indicate a rapid dissipation of the squall line within 30 minutes, and significantly weaker outcomes are projected for the 60-minute mark. We posit that the suboptimal performance of the DGMR model is attributable to the simultaneous use of generative loss and pixel-wise alignment loss functions during its training phase, which leads to unstable training process and still keeps the drawback of dissipation of deterministic alignments. While the generative loss alone is capable of simulating realistic meteorological processes, it falls short in accurately predicting the extent of precipitation and is abandoned in DGMR. On the contrary, CogDPM does not require additional deterministic alignment during training but enhances precision with precision-weighted guidance during inference steps. We present additional case studies in Appendix F.

We further explore the numerical evaluations in Fig 5 with metrics on different forecast properties focusing on the accuracy, reality and diversity. Radially Continuous ranked probability score (CRPS) measures the alignment between probabilistic forecast and the ground truth. We also report the spatially aggregated CRPS (Ravuri et al., 2021) to test prediction performance across different spatial scales. Details of these metrics can be found in Extended Data. The first row in Fig 4 shows CogDPM consistently outperforms baseline models for the whole time period. We adopt the decision-analytic model to evaluate the Economic value of ensemble predictions (Ravuri et al., 2021). Curves in Figure 5 with greater under-curve area provide better economic value, and CogDPM outperforms baseline models in this regard. Radially averaged power spectral density (PSD) evaluates the variations of spectral characteristics on different spatial scale. CogDPM achieves the minimal gap with ground truth characteristics.

The superior performance metrics of CogDPM stem from its diffusion models' ability to emulate the hierarchical inference of predictive coding, resulting in smaller prediction errors compared to single-step forecasting models. Furthermore, the integration of precision weighting allows the model to dynamically assess the precision of inputs and adjust the intensity of conditional control accordingly. This targeted approach effectively reduces errors in areas that are challenging to predict, thereby enhancing the accuracy of the model in delineating boundaries and extreme regions.

5. Discussion

CogDPM is related to classifier-free diffusion models (Ho & Salimans, 2021), which enhance the class guidance with a conditional DPM and an unconditional DPM. CogDPM framework builds the connection between classifier-free diffusion models and predictive coding. We also introduce the precision estimation method with the reverse diffusion process and use precision to control the guidance strength in spatiotemporal scales. We adopt the ablation study to show the enhancement in prediction skills of the CogDPM framework compared with the vanilla CFG method in appendix E.

Active inference (Parr et al., 2019) is also a widely discussed theory of the predictive coding framework, which states that cognition system actively interact with the environment to minimize the prediction error. Active inference is omitted in this work. We take a computational predictive coding model with both active inference and precision weighting as the future work.

6. Conclusion

We propose CogDPM, a novel spatiotemporal forecasting framework based on diffusion probabilistic models. CogDPM shares main properties with predictive coding and is adapted for field prediction tasks. The multi-step reverse diffusion process models the hierarchy of predictive error minimization. The precision of a latent expectation can be estimated from the variance of states in the neighboring levels. The CogDPM framework has demonstrated its ability to provide skillful spatiotemporal predictions in precipitation nowcasting and wind forecasting. Case studies and numeric evaluations demonstrate that CogDPM provides competitive forecasting skills.

Acknowledgement

This work was supported by the National Key Research and Development Plan (2021YFC3000905), the National Natural Science Foundation of China (U2342217 and 62022050), the BNRist Innovation Fund (BNR2024RC01010), and the National Engineering Research Center for Big Data Software.

Impact Statement

This paper presents work whose goal is to advance the deep learning research for a PC-based spatiotemporal forecasting framework. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., and Babenko, A. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2021.

Barbounis, T. G., Theocharis, J. B., Alexiadis, M. C., and Dokopoulos, P. S. Long-term wind speed and power forecasting using local recurrent neural network models. *IEEE Transactions on Energy Conversion*, 21(1):273–284, 2006.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv:2311.15127*, 2023a.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575, June 2023b.

Clark, A. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings of the 9-th International Conference on Representation Learning*, 2021.

Emberson, L. L., Richards, J. E., and Aslin, R. N. Top-down modulation in the infant brain: Learning-induced expectations rapidly affect the sensory cortex at 6 months. *Proceedings of the National Academy of Sciences*, 112 (31):9585–9590, 2015.

Evensen, G. The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53:343–367, 2003.

Friston, K. The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, 13(7):293–301, 2009.

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

Guen, V. L. and Thome, N. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11474–11484, 2020.

Han, K., Wen, H., Zhang, Y., Fu, D., Culurciello, E., and Liu, Z. Deep predictive coding network with local recurrent processing for object recognition. *Advances in neural information processing systems*, 31, 2018.

Han, T., Xie, W., and Zisserman, A. Memory-augmented dense predictive coding for video representation learning. In *European conference on computer vision*, pp. 312–329. Springer, 2020.

Harris, D., Foufoula-Georgiou, E., Droegemeier, K. K., and Levit, J. J. Multiscale statistical properties of a high-resolution precipitation forecast. *Journal of Hydrometeorology*, 2(4):406–418, 2001.

Harvey, W., Naderiparizi, S., Masrani, V., Weilbach, C., and Wood, F. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35: 27953–27965, 2022.

Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pp. 4182–4192. PMLR, 2020.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N. Era5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2023.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022a.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation, 2022b.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *arXiv:2204.03458*, 2022c.

Hohwy, J. New directions in predictive processing. *Mind & Language*, 35(2):209–223, 2020.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Höppe, T., Mehrjou, A., Bauer, S., Nielsen, D., and Dittadi, A. Diffusion models for video prediction and infilling. *Transactions on Machine Learning Research*, 2022.

Jolliffe, I. T. and Stephenson, D. B. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons, 2012.

Keller, G. B. and Mrsic-Flogel, T. D. Predictive processing: a canonical cortical computation. *Neuron*, 100(2):424–435, 2018.

Lotter, W., Kreiman, G., and Cox, D. Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations*, 2016.

Lotter, W., Kreiman, G., and Cox, D. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature machine intelligence*, 2(4):210–219, 2020.

Marugán, A. P., Márquez, F. P. G., Perez, J. M. P., and Ruiz-Hernández, D. A survey of artificial neural network in wind energy systems. *Applied energy*, 228:1822–1836, 2018.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

Parr, T., Corcoran, A. W., Friston, K. J., and Hohwy, J. Perceptual awareness and active inference. *Neuroscience of consciousness*, 2019(1):niz012, 2019.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv:2202.11214*, 2022.

Porté-Agel, F., Wu, Y.-T., and Chen, C.-H. A numerical study of the effects of wind direction on turbine wakes and power losses in a large wind farm. *Energies*, 6(10): 5297–5313, 2013.

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., and Foresti, L. Pysteps: an open-source python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, 12(10):4185–4219, 2019.

Rane, R. P., Szügyi, E., Saxena, V., Ofner, A., and Stober, S. Prednet and predictive coding: A critical review. In *Proceedings of the 2020 international conference on multimedia retrieval*, pp. 233–241, 2020.

Rao, R. P. and Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

Rasul, K., Seward, C., Schuster, I., and Vollgraf, R. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, pp. 8857–8868. PMLR, 2021.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

Roberts, N. M. and Lean, H. W. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136(1): 78–97, 2008.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

Rui, W., Karthik, K., Mustafa, Albert, A., and Yu, R. Towards physics-informed deep learning for turbulent flow prediction. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020.

Schaefer, J. T. The critical success index as an indicator of warning skill. *Weather and Forecasting*, 5(4):570–575, 1990.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

Sinclair, S. and Pegram, G. Empirical mode decomposition in 2-d space and time: a tool for space-time rainfall analysis and nowcasting. *Hydrology and Earth System Sciences*, 9(3):127–137, 2005.

Soman, S. S., Zareipour, H., Malik, O., and Mandal, P. A review of wind power and wind speed forecasting methods with different time horizons. In *North American power symposium 2010*, pp. 1–8. IEEE, 2010.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Spratling, M. A review of predictive coding algorithms. *Brain and Cognition*, 112:92–97, 2017. ISSN 0278-2626.

Twin Cities, M. W. F. O. Summary of march 06 2017 severe weather outbreak - earliest known tornado in minnesota's history, 2017. URL https://www.weather.gov/mpx/SevereWeather_06March2017.

van Elk, M. A predictive processing framework of tool use. *Cortex*, 139:211–221, 2021. ISSN 0010-9452. doi: https://doi.org/10.1016/j.cortex.2021.03.014.

Voleti, V., Jolicoeur-Martineau, A., and Pal, C. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, 35:23371–23385, 2022.

Walsh, K. S., McGovern, D. P., Clark, A., and O'Connell, R. G. Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the new York Academy of Sciences*, 1464(1):242–268, 2020.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. Video-to-video synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 1152–1164, 2018.

Wang, Y., Long, M., Wang, J., Gao, Z., and Yu, P. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *Advances in Neural Information Processing Systems*, pp. 879–888, 2017.

Wang, Y., Wu, H., Zhang, J., Gao, Z., Wang, J., Yu, P. S., and Long, M. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Wen, H., Han, K., Shi, J., Zhang, Y., Culurciello, E., and Liu, Z. Deep predictive coding network for object recognition. In *International conference on machine learning*, pp. 5266–5275. PMLR, 2018.

Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., and Zhang, X. Panacea: Panoramic and controllable video generation for autonomous driving. *arXiv:2311.16813*, 2023.

Wiese, W. and Metzinger, T. K. Vanilla pp for philosophers: A primer on predictive processing. In Metzinger, T. K. and Wiese, W. (eds.), *Philosophy and Predictive Processing*, chapter 1. MIND Group, Frankfurt am Main, 2017. ISBN 9783958573024. doi: 10.15502/9783958573024.

Wu, H., Yao, Z., Wang, J., and Long, M. Motionrnn: A flexible model for video prediction with spacetime-varying motions, 2021.

Xu, M., Yu, L., Song, Y., Shi, C., Ermon, S., and Tang, J. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2021.

Yang, R., Srivastava, P., and Mandt, S. Diffusion probabilistic modeling for video generation. *Entropy*, 25(10):1469, 2023.

Yao, Z., Wang, Y., Wu, H., Wang, J., and Long, M. Modernn: Harnessing spatiotemporal mode collapse in unsupervised predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Zhang, J. Modern monte carlo methods for efficient uncertainty quantification and propagation: A survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13 (5):e1539, 2021.

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., and Wang, J. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619(7970):526–532, 2023.

Zheng, G., Li, S., Wang, H., Yao, T., Chen, Y., Ding, S., and Li, X. Entropy-driven sampling and training scheme for conditional diffusion generation. In *European Conference on Computer Vision*, pp. 754–769. Springer, 2022.

A. Preliminary

Diffusion Models. Diffusion models are the state-of-the-art deep generative models on image synthesis(Dhariwal & Nichol, 2021; Song & Ermon, 2019; Ho et al., 2020), and have been explored widely in numerous tasks, such as computer vision(Baranchuk et al., 2021), time series modeling (Rasul et al., 2021) and molecular graph modeling (Hoogeboom et al., 2022; Xu et al., 2021). Diffusion probabilistic models (DPMs), a major paradigm in diffusion models, handly construct the forward process $q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)$ by progressively injecting noise to a data distribution $q(\mathbf{x}_0)$, and generate samples with a denoising backward process. Formally, we define a Markov forward process q with latent variables $\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_T$, which follows Eq. (10) and Eq. (11):

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}), \tag{10}$$

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t \mid \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}), \tag{11}$$

where $\beta_t \in (0,1),\ t = 1, \dots, T$ schedule the forward process. In this work, we select cosine β scheduling (Nichol & Dhariwal, 2021). Eq. (12) allows to directly sample an arbitrary latent variable conditioned on the input $\mathbf{x}_0$. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, we formulate the marginal distribution as:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t \mid \sqrt{\bar{\alpha}_t}\mathbf{x_0}, (1-\bar{\alpha}_t)\mathbf{I}). \tag{12}$$

We define the reverse process for Eq. (10) and Eq. (11) as $p_\theta(\mathbf{x}_{0:T})$, with initial state $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T \mid \mathbf{0}, \mathbf{I})$ and parameterized marginal distributions:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t), \tag{13}$$

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1} \mid \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)). \tag{14}$$

Diffusion models transfer the goal of generating target distribution into minimizing distance between forward and backward processes, formulated in Eq. (16). Eq. (17) shows that the alignment between two processes can be factorized into the alignment between marginal conditional distributions.

$$\min_{\{\mu_t, \Sigma_t^2\}_{t=1}^T} L_{\text{vb}}(q, p_\theta) \tag{15}$$

$$\Leftrightarrow \min_{\{\mu_t, \Sigma_t^2\}_{t=1}^T} D_{\text{KL}}(q(\mathbf{x}_{0:T}) \| p_\theta(\mathbf{x}_{0:T})) \tag{16}$$

$$\Leftrightarrow \min_{\{\mu_t, \Sigma_t^2\}_{t=1}^T} \sum_{t=1}^{T} D_{\text{KL}}(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)). \tag{17}$$

Ho et al. (Ho et al., 2020) adopt a denoising network $\epsilon_\theta(\mathbf{x}_t, t)$ to parameterize $\mu_\theta(\mathbf{x}_t, t)$, and simplify the above loss as Eq. (18):

$$\mathrm{E}_{t \sim \mathcal{U}(1,T), \mathbf{x_0} \sim q(\mathbf{x}_0) \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2. \tag{18}$$

We train the diffusion models with Eq. (18), and CogDPM inference results with the same methodology as Eq. (14).

Video generation with Diffusion Models. Considering the success in image synthesis, diffusion models can be naturally applied to video generation and prediction tasks. Unlike static images, video generation faces two main problems: 1. considerable computation consumption, which also cause slow inference speed; 2. inconsistency between adjacent frames. For the first problem, previous works focus on enhancing the computation efficiency of the back-bone U-Net in diffusion models (Ho et al., 2022c; Voleti et al., 2022). Video-Image joint training helps accelerating the optimization progress (Ho et al., 2022c).

Yang et al. (2023) combine a deterministic prediction model with a residual prediction diffusion model. The diffusion model learn to generate the stochastic error between the deterministic prediction and the ground truth video, and the deterministic model ensure the continuity between adjacent frames. Ho et al. (2022c) introduce a gradient based conditional sampling method to improve temporally coherency, and also adapt cascaded architectures from image-based to video-based diffusion models(Ho et al., 2022a), leading to enhanced video quality.

To maintain temporal consistency across extended video sequences, researchers also design auto-regressive conditioning procedures (Voleti et al., 2022; Harvey et al., 2022; Höppe et al., 2022; Yang et al., 2023). These models recursively utilize preceding outputs as inputs to sequentially generate subsequent frames. Voleti et al. and Höppe et al. use masked sequences to train the model(Voleti et al., 2022; Höppe et al., 2022), while Yang et al. treat the process from a purely probabilistic view(Yang et al., 2023).

Blattmann et al. (2023b) extend 2D Latent Diffusion Models (LDMs) to 3D versions by inserting temporal layers between each blocks, which called spatial layers, in original U-Nets. Spatial layers concentrate on synthesizing individual frames in the video, while temporal layers are dedicated to the alignment between different frames. Leveraging pre-trained 2D LDMs, 3D LDMS can generate long videos without losing much image quality. Stable Video Diffusion is a large-scale implementation of 3D LDMs (Blattmann et al., 2023a), which exhibits its outstanding performance in video sythesis.

Previous works prove that DPMs can generate realistic and coherent videos with considerable diversity, and corresponding experiments mainly focus on general videos photoed by RGB cameras. Spatiotemporal forecasting is another brand of video prediction, which focus on scientific applications. In these tasks, beyond frame consistency, forecast accuracy, diversity and skillfulness are the mainly concerned metrics. In this work, we explore the field evolution forecasting with diffusion models, and enhance model forecasting value on real-world applications.

B. Implementation Details

Architecture design. CogDPM adopts the U-Net (Ronneberger et al., 2015; Ho et al., 2022c) backbone coupled with a vision transformer (ViT) (Dosovitskiy et al., 2021) encoder.

Inputs of the ViT encoder contain three parts: 1) patches of context cubes, 2) cube positional embedding, and 3) diffusion step embedding. The inputs are summed together after a linear projection and then fed into the ViT encoder. We adopt a random mask among the ViT inputs for better efficiency.

The architecture of the U-Net consists of a down-sampling tower, mid-blocks, an up-sampling tower, and short-cut connections. The down-sampling tower progressively reduces the spatial dimensions of the input while increasing the number of channels. The mid-blocks maintain the shape of the intermediate representations. The up-sampling tower reverses the operations of the down-sampling tower, and the short-cut connections provide direct connections between corresponding down-sampling and up-sampling blocks.

A single U-Net block employs separate neural modules for spatial and temporal modeling, including a spatial residual block with convolution neural networks and a temporal block with axial attention. We use cross-attention layers to merge the representations from the context encoder into the U-Net blocks. We repeat these blocks several times, followed by a bilinear interpolation operation that doubles or halves the spatial shapes.

We further introduce the cascade diffusion pipeline to accelerate sampling speed for the participation nowcasting task (Ho et al., 2022b). We parallel couple a low resolution CogDPM model with an additional super-resolution CogDPM, and allocate fewer inference steps on the super-resoluation model to decrease the total inference time compared to a single high-resolution model.

Evaluation. For the precipitation nowcasting, we follow the metrics used in (Ravuri et al., 2021), including CRPS, window pooled CRPS, economic values and PSD. The computational details are listed in the supplementary materials.

We uniformly sample 10,000 cases from the 512 km $\times$ 512 km cropped test dataset provided by (Ravuri et al., 2021). We crop the central 256 km $\times$ 256 km as model inputs to keep the same input shape with the pretrained DGMR. We take quantitative evaluations of the central 128 km $\times$ 128 km to avoid the boundary effects, as the similar method in (Ravuri et al., 2021).

For surface wind forecasting, we report CSI for the high wind precaution task and report CRPS and RMSE for the electric

power prediction task.

We also evaluate on the central 64×144 grids to concentrate on the land measurement while also eliminating the boundary effect from partial observation.

Training. We train a two-stage cascade diffusion model for the United Kingdom precipitation dataset and one-stage diffusion models for the others.

The one-stage models are trained for 1×10^6 steps. The learning rate is 2×10^{-5}, using Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We randomly replace conditions as i.i.d. standard Gaussian noise with a probability of 10%, following the classifier-free diffusion models (Ho & Salimans, 2021). On the ERA5 dataset, we train the model on 2 GPU cores (NVIDIA A100) for two weeks using a batch size of 16 per training step. On the turbulence flow and Moving MNIST dataset, we train the model on 1 GPU core (NVIDIA A100) for one week using a batch size of 36 per training step.

The two-stage model cascades a low-resolution DPM and a super-resolution DPM. The low-resolution DPM has the same training recipe as a model for Moving MNIST. The super-resolution DPM is trained for 5×10^6 steps, using 4 GPU cores (NVIDIA A100) for two weeks using a batch size of 8 per training step.

C. Datasets

In this study, we conduct experiments on synthesis datasets for interpreting precision estimations and on real-world datasets for evaluating prediction skills.

For interpreting precision estimations, we adopt The Moving MNIST dataset which is generated with the same method as (Wu et al., 2021). We create sequences with 20 frames, and each frame contains three handwriting digits. The motion of digits consists of transition, reflection, and rotation. We use four initial frames to predict the movements of the digits in the following 16 frames, and each frame has 64×64 grids. We generate 100,000 sequences for training, 1,000 for validation and 1,000 for testing. The turbulent flow dataset is proposed by (Rui et al., 2020). We follow the same dataset parameters as (Rui et al., 2020) and generate a sequence with 15 frames and 64 x 64 grids on each frame. Four frames are taken to predict the next 11 frames. We generate 20,000 sequences for training, 1,000 for validation, and 1,000 for testing.

For testing the wind field forecasting, we use the ERA5 dataset, a high-resolution global atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Hersbach et al., 2023). We select the region covering the United States, the region from longitude 130 degrees West to 60 degrees West, and latitude 20 degrees North to 56 degrees North. The dataset has the time scale from 1959 to 2019, and the 6-hour time interval. We use the 24-hour surface wind and zonal wind speed to predict the next 48 hours. We use the data from 1959-01-01 to 2013-12-31 for training, 2014-01-01 to 2016-12-31 for validation, and 2017-01-01 to 2019-12-31 for testing.

For evaluating the skill of precipitation nowcasting of CogDPM, we adopt the United Kingdom precipitation dataset, which contains radar composites from the Met Office RadarNet4 network from 2017 to 2019, and the experiment settings used in (Ravuri et al., 2021). The dataset is patched into $24 \times 256 \times 256$ composites with a time interval of 5 minutes and 1 km x 1 km spatial grids. The model forecasts precipitation fields of 90 minutes with 20-minute observations. We note that the UK precipitation dataset is a large high resolution spatiotemporal forecasting dataset which takes about 1 TB saved with TF Records. We follow the dataset splits and importance sampling techniques described in (Ravuri et al., 2021).

We list the detailed information of these datasets in the table C.

Table 2. Overview of Datasets for Different Tasks

Feature	MovingMNIST	Turbulence	US Surface wind	UK Precipitation
Image shape	(64, 64)	(64, 64)	(144, 280)	(256, 256)
Sequence length	20	15	12	22
Channel	1	2	3	1
Size of training set	100,000	100,000	73,000	5,788,800
Size of validation set	1,000	1,000	4,380	1,000
Size of test set	1,000	1,000	4,380	10,000

D. Verification Metrics

We outline the five standard evaluation metrics used in this article.

Critical Success Index (CSI) (Schaefer, 1990) quantifies the accuracy of binary predictive decisions, determining whether the target intensity value exceeds a specific threshold. To calculate this metric, we sum the hits, misses, and false alarms across all grid points and compute their ratio. CSI evaluates precision and recall simultaneously and is widely used to assess high wind forecasting. It is worth noting that the CSI metric counts the hit, miss, and false alarm over the entire test set.

CSI-Neighborhood (Jolliffe & Stephenson, 2012) is the CSI metric computed based on the neighborhood. Neighborhood methods evaluate the forecasts within a spatial window surrounding each grid, which can assess the 'closeness' of the forecasts. This metric is particularly suitable for verifying high-resolution forecasts.

Radially Continuous ranked probability score (CRPS) (Gneiting & Raftery, 2007) measures the alignment between probabilistic forecasts and ground truth data. It is widely used in the evaluation of weather forecasting models. The CRPS takes into account the entire distribution of forecasting probabilities, making it a more comprehensive evaluation metric compared to traditional point forecast metrics. We also report the neighborhood CRPS (Ravuri et al., 2021), also following (Jolliffe & Stephenson, 2012)

Fractional Skill Score (FSS) as described by Roberts and Lean (2008) (Roberts & Lean, 2008) represents another neighborhood-based metric, delineated by target thresholds. For each grid cell within a piece of test data, the proportion of surrounding cells exceeding a defined threshold within a spatial window is calculated. Subsequently, the summation of these proportions—predicted versus observed—across all grid cells is termed the Fractional Brier Score (FBS). The FSS is derived from the normalized FBS, relying on a threshold to determine local value occurrences and employs the Fractional Brier Score (FBS) to contrast predicted and observed value frequencies. Unlike the CSI-Neighborhood, which solely focuses on the accuracy of hit predictions, the FSS also facilitates the comparison of the rate of grid cells exceeding the threshold within a spatial window.

Power Spectral Density (PSD) (Harris et al., 2001; Sinclair & Pegram, 2005) measures the power distribution over each spatial frequency, comparing the intensity variability of forecasts to that of the observations. We use the PSD implementation from the PySTEPS package (Pulkkinen et al., 2019). Forecasts that have minor differences with observations are preferred.

Economic Value (Ravuri et al., 2021) evaluates the outcome of forecasts with a cost-loss ratio decision model. It shows the relative loss decrease with a forecasting-based precaution policy for a particular cost-loss ratio. For fair comparison, we follow the implementation from DGMR (Ravuri et al., 2021).

E. Ablation Study

We demonstrate ablation experiments on precision weighting mechanism and network design. We compare the video diffusion model (VDM) (Ho et al., 2022c) with numerical evaluation. VDM shares the same training schema and U-Net architecture with CogDPM, and maintains a constant Classifier-free guidance during inference. Figure 6 reports the performance of MAE, RMSE, CRPS, and CSI metrics with different wind speed thresholds on the ERA5 surface wind dataset. The results show that CogDPM has significant enhancements in MAE, RMSE, and CRPS metrics, and achieves comparable or better results on CSI indices with different wind speed thresholds. This demonstrates that precision-weighted guidance can increase the precision of CogDPM in high-wind regions and provide more diverse forecasts, thereby improving the skillfulness of the forecast results.

F. Additional Case Studies

In Figure 7, we present additional case studies on ERA5 surface wind prediction. The inverse precision field of CogDPM indicate an informative unpredictable region.

In Figure 8, we present additional case studies on the United Kingdom precipitation dataset. The precision field of CogDPM effectively described the boundaries of the precipitation range.

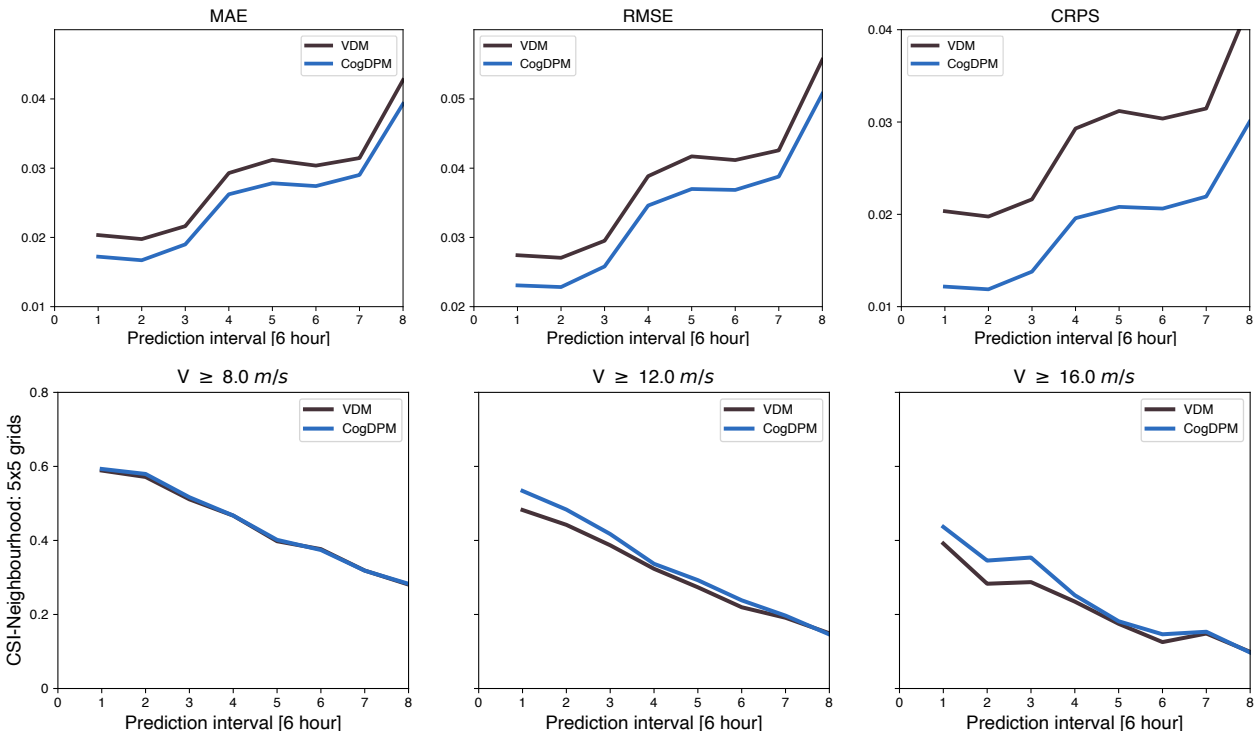

Figure 6. **Numerical comparison between CogDPM and video diffusion models (VDM) on the ERA5 wind forecast task.** The first row shows CSI metrics with thresholds of 8.0 m/s, 12.0 m/s and 16.0 m/s. The second row shows MAE, RMSE and CRPS relatively.

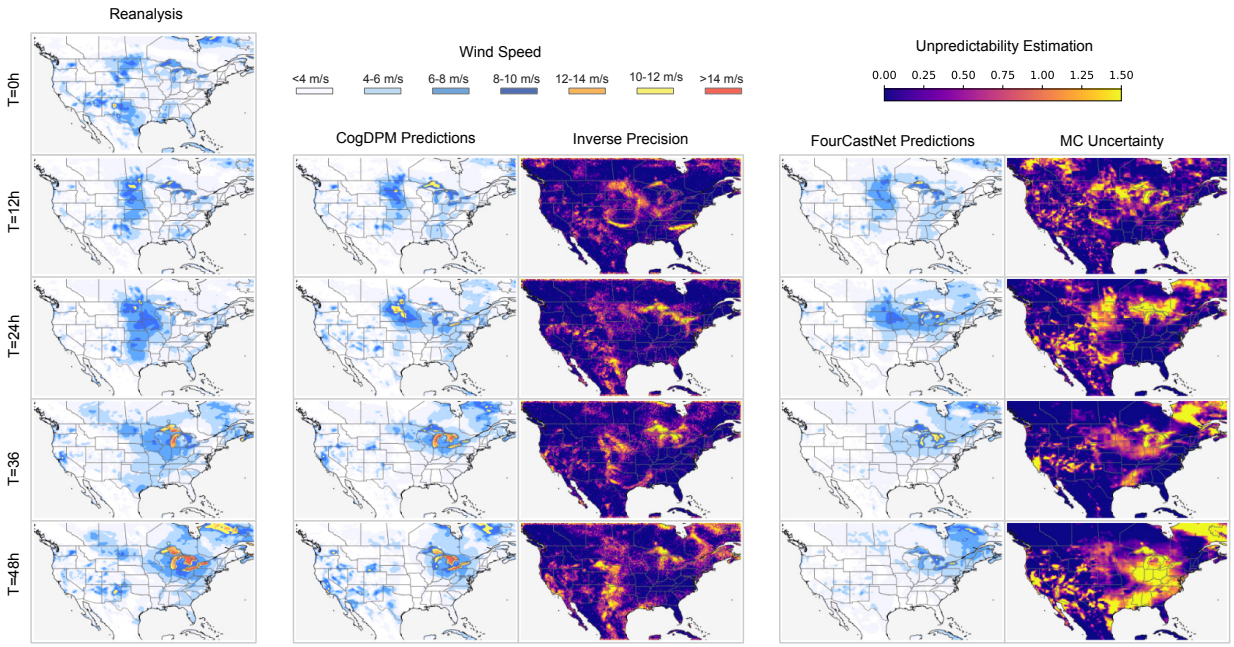

Figure 7. **Additional case studies on ERA5 surface wind prediction. a** Case study of the ERA5 wind forecast from 2017-01-02 18:00. The CogDPM successfully prediction the high wind region moving from mideast USA to the Great Lakes. FourCastNet overestimate the moving spatial scale and underestimate its intensity for T=36h and T=48h. The inverse precision field of CogDPM indicate an unpredictable region for mideast USA at T=48h where the prediction neglect. FourCastNet uncertainty focus on the east-south USA at T=48, but is irrelevant with the truth field. Maps produced by the Cartopy package.

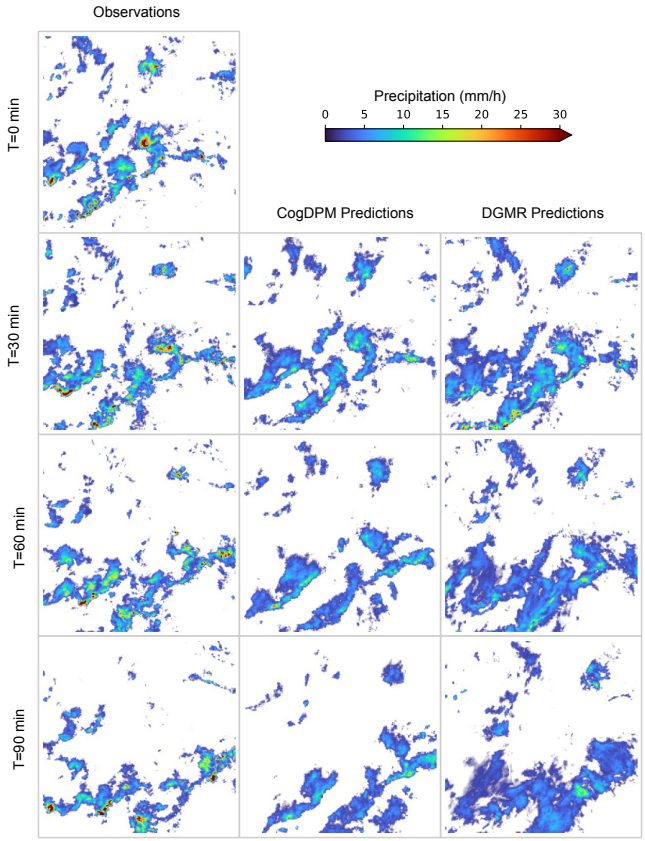

Figure 8. **Additional case studies on the United Kingdom precipitation dataset.** In this case, the forecast results of CogDPM maintained the bifurcated structure of the two squall lines and their precipitation range, with predicted locations closely matching actual observations. On the other hand, the results of DGMR showed the two squall lines merging at 60 minutes, and the 90-minute forecast significantly misreported the precipitation range. The precision field of CogDPM effectively described the boundaries of the precipitation range, while the precision field of DGMR duplicated the predicted precipitation intensity information.