Improving Tropical Cyclone Forecasting With Video Diffusion Models

Zhibo Ren
Department of Computing
Imperial College London
`zhibo.ren23@imperial.ac.uk`

Pritthijit Nath
Department of Applied Math and Theoretical Physics
University of Cambridge
`pn341@cam.ac.uk`

Pancham Shukla
Department of Computing
Imperial College London
`panchamkumar.shukla@imperial.ac.uk`

Abstract

Tropical cyclone (TC) forecasting is crucial for disaster preparedness and mitigation. While recent deep learning approaches have shown promise, existing methods often treat TC evolution as a series of independent frame-to-frame predictions, limiting their ability to capture long-term dynamics. We present a novel application of video diffusion models for TC forecasting that explicitly models temporal dependencies through additional temporal layers. Our approach enables the model to generate multiple frames simultaneously, better capturing cyclone evolution patterns. We introduce a two-stage training strategy that significantly improves individual-frame quality and performance in low-data regimes. Experimental results show our method outperforms the previous approach of Nath et al. by 19.3% in MAE, 16.2% in PSNR, and 36.1% in SSIM. Most notably, we extend the reliable forecasting horizon from 36 to 50 hours. Through comprehensive evaluation using both traditional metrics and Fréchet Video Distance (FVD), we demonstrate that our approach produces more temporally coherent forecasts while maintaining competitive single-frame quality. Code accessible at `https://github.com/Ren-creater/forecast-video-diffmodels`.

1 Introduction

Climate change is a pressing global issue causing unprecedented changes in the Earth's climate system, with tropical cyclones (TCs) receiving particular attention due to their increasing intensity and devastating impacts [1]. These extreme weather events pose a growing threat to global communities, causing extensive human suffering and economic losses [2]. Recent studies indicate that warming oceans (due to climate change) are leading to more powerful storms [3], making accurate TC forecasting increasingly critical for effective emergency preparedness and disaster response.

Traditional numerical weather prediction (NWP) methods [4], which rely on solving complex physical equations, have been the foundation of TC forecasting for decades. While these methods are rooted in well-established meteorological principles, they are computationally demanding and can encounter difficulties in capturing the highly non-linear and chaotic nature of atmospheric dynamics [5]. Such limitations, coupled with the increasing demand for more precise and timely predictions whilst being computationally inexpensive, have spurred interest in alternative or complementary approaches.

Machine learning (ML) has emerged as a promising tool for addressing these challenges, offering the potential to enhance or even replace traditional NWP methods. ML models can leverage vast amounts of observational and simulation data to identify intricate patterns and provide rapid inference [6]. Among these, deep learning (DL) techniques have been particularly effective, demonstrating success in a variety of TC-related tasks, including cyclone detection [7], track prediction [8], and precipitation forecasting [9]. Recent work by Nath et al. [10] has further highlighted the potential of diffusion

models in TC forecasting. However, their approach generates predictions on a frame-by-frame basis, which may limit its ability to capture critical temporal dependencies that span multiple time steps.

Recent advancements in video diffusion models have opened new possibilities for tackling such limitations. These models, which have achieved remarkable performance in areas such as video synthesis and editing [11; 12], explicitly account for temporal dynamics through specialised temporal layers. By generating multiple frames simultaneously, video diffusion models ensure better temporal coherence, a feature that is particularly crucial for weather forecasting applications where the evolution of atmospheric systems over time must be accurately represented.

Despite the recent progress, the challenge of TC forecasting still remains formidable. Skilful prediction requires models that can simultaneously capture spatial and temporal dependencies while ensuring stability and accuracy over longer prediction horizons. Although existing approaches have made considerable progress in short-term forecasting, issues such as temporal inconsistency and forecast degradation over time persist, leaving room for further innovation. Addressing these challenges is essential for pushing the boundaries of TC forecasting and delivering solutions that are both robust and practical.

To advance progress in this domain, our work makes the following contributions:

1. We present a novel application of video diffusion models for TC forecasting, enabling simultaneous generation of multiple frames to better capture temporal dynamics.
2. We introduce a two-stage training strategy that significantly improves model training stability and individual frame-quality, especially in low-data regimes.
3. We demonstrate superior long-horizon forecasting capability, extending reliable predictions from 36 (as demonstrated by Nath et al.) to 50 hours.
4. We establish Fréchet Video Distance (FVD) [13] as a more suitable metric for evaluating TC forecasts, providing better assessment of temporal coherence.

2 METHOD

Our approach builds upon the work of Nath et al. [10], extending their diffusion-based framework to incorporate temporal dynamics through video generation. The key insight is that treating TC evolution as a continuous process, rather than as a series of independent frame predictions, allows more effective modelling of storm dynamics.

2.1 DATA PROCESSING

We use the same dataset as Nath et al. [10], reorganising it into sequences of 10 consecutive frames for both infrared (IR) satellite images and corresponding ERA5 data. This results in 1,092 video sequences for training and 335 for testing. To handle corrupted data points, `NaN` values are replaced with zeros, and a consistent mask is applied during generation to ensure they do not influence the output.

2.2 VIDEO DIFFUSION MODEL

We adopt a 3D UNet architecture (illustrated in Figure A.1) following a design similar to that proposed by Ho et al. [11]. The model utilises temporal convolutions and attention mechanisms to generate sequences of 64x64 IR 10.8μm satellite imagery, conditioned on both the initial IR frame and corresponding ERA5 meteorological data. In particular, the model generates 10 forecast frames simultaneously, enabling it to better capture the temporal evolution of cyclonic systems. This architecture incorporates both classifier-free guidance [14] and dynamic thresholding for maintaining output quality within the normalised range. Detailed hyperparameter configurations used in model training are provided in Appendix B.1.

2.3 TWO-STAGE TRAINING

We introduce a two-stage training strategy crucial for model performance. In the low-data regime, we first conduct a single-frame stage of 200 epochs, focusing on individual frame prediction to

establish spatial understanding. Then, in the multi-frame stage, we train on 10-frame sequences for an additional 200 epochs to learn temporal dynamics. For the full dataset, we perform a single-frame stage of 100 epochs, followed by a multi-frame stage of 300 epochs to better capture temporal dynamics.

This curriculum learning approach significantly improves training stability and model performance. Notably, in the low-data regime experiments (see Section 3.1), while training without the first stage achieves similar FVD scores (402.98 v/s 402.03 w/ the first stage), the two-stage approach achieves superior single-frame quality with FID of 0.49 compared to 1.26 for direct training without the first stage. This gain in performance demonstrates that the initial single-frame training stage is essential for preserving frame-level accuracy even as we optimise for temporal coherence.

In the full dataset scenario, we found that adjusting the epoch distribution (100/300 v/s 200/200) provides better overall results while maintaining the crucial single-frame quality benefits of two-stage training. This suggests that when sufficient data are available, allocating more training time to multi-frame prediction is beneficial, as long as the initial single-frame training stage is preserved.

3 EXPERIMENTS

We evaluated our model using multiple metrics to assess both single-frame quality and temporal coherence. Besides traditional metrics (MAE, PSNR, SSIM, FID) used in previous work, we introduce FVD [13] to better evaluate temporal consistency. All experiments used a learning rate of 3×10^{-4} and are performed on a single NVIDIA L40 GPU, demonstrating remarkable computational efficiency of our approach (in contrast to conventional methods).

3.1 DATA EFFICIENCY STUDY

To investigate model performance in low-data regimes, we trained both models using only cyclones in the North Indian Ocean basin. The results shown in Table 1 demonstrate our model's superior data efficiency:

Table 1: Performance comparison in low-data regime

Method	MAE ↓	PSNR ↑	SSIM ↑	FID ↓	FVD ↓
Baseline (Nath et al. [10])	0.2846	18.07	0.4353	**0.3298**	706.11
Video Diffusion (w/o two-stage)	0.2647	**20.72**	**0.6522**	1.2633	402.98
Video Diffusion (with two-stage)	**0.2300**	20.62	0.6387	0.4955	**402.03**

Impact of Two-Stage Training: The two-stage training strategy proves crucial for maintaining single-frame quality while improving temporal coherence. Without two-stage training, the model achieves similar FVD scores but suffers in single-frame quality (FID increases from 0.49 to 1.26).

3.2 RESULTS AND DISCUSSION

Table 2: Comparison with baseline model on 10-frame prediction task

Method	MAE ↓	PSNR ↑	SSIM ↑	FID ↓	FVD ↓
Baseline (Nath et al. [10])	0.2209	22.49	0.5235	**0.2288**	445.83
Video Diffusion	**0.1781**	**26.13**	**0.7123**	0.2344	**242.41**
Improvement	19.3%	16.2%	36.1%	-2.4%	45.6%

Our model significantly outperforms the baseline (Nath et al. [10]) across most metrics, with particularly strong improvements in temporal coherence as measured by FVD (45.6% reduction). While maintaining similar single-frame quality (FID), our model achieves substantial improvements in MAE (19.3%), PSNR (16.2%), as well as SSIM (36.1%).
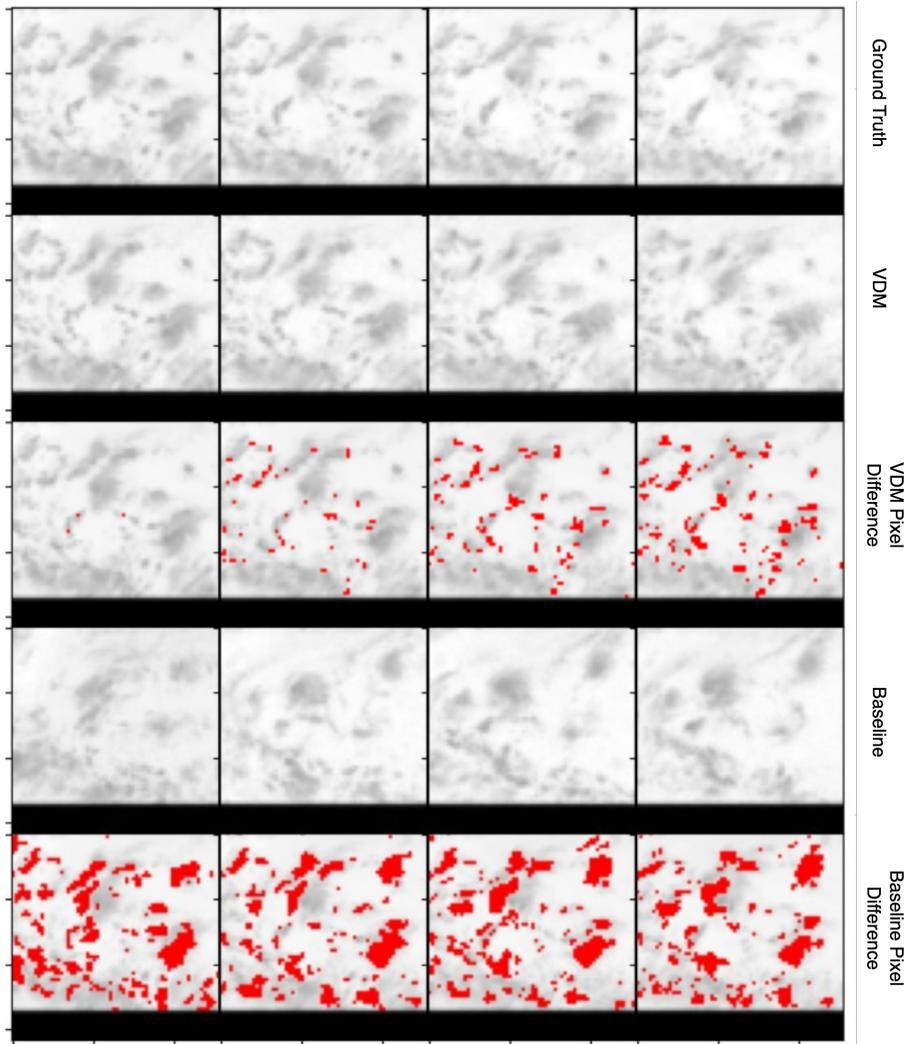
Figure 1: Qualitative comparison of TC forecasting results on the first four frames generated. From top to bottom: (1) Ground truth, (2) our VDM predictions, (3) the difference between VDM prediction and ground truth, (4) Nath et al.'s predictions, and (5) the difference between Nath et al.'s predictions and ground truth. Our VDM method demonstrates improved temporal consistency and more accurate TC evolution patterns.

Visual inspection (see Figure 1) of the samples reveals that our model handles complex cloud patterns more effectively than the baseline. While the baseline often struggles with dense cloud regions, our approach maintains better cloud cover consistency with the ground truth. This improvement can be greatly attributed to the model's improved ability to consider temporal context during generation.

3.3 LONG-HORIZON FORECASTING

Following Nath et al. [10], we performed similar experiments to forecast the entire duration of all cyclones in the testing split, organised region-wise. Nath et al. [10] identified a reliable forecasting horizon of 36 hours, beyond which sharp declines in forecast accuracy, as indicated by SSIM charts, are observed. Integrating our model into their cascaded pipeline extends this reliable horizon from 36 to 50 hours, as shown by SSIM charts (see charts in Appendix C). Furthermore, the minimum SSIM values predicted by our model consistently remain higher than those of the baseline, suggesting a better preservation of cyclone structure over extended periods.

4 CONCLUSION

We present a novel application of video diffusion models for forecasting TCs that significantly improves both temporal coherence as well as forecast horizon. Our two-stage training strategy successfully addresses the challenge of maintaining single-frame quality while optimising for temporal dynamics. Experimental results demonstrate substantial improvements across multiple metrics, with particularly strong gains in temporal consistency (45.6% reduction in FVD) while preserving single-frame fidelity with a competitive FID of 0.23. Most notably, our method extends the reliable forecasting horizon from 36 to 50 hours, representing a significant step forward in enhancing long-term TC prediction capabilities.

For future work, several promising directions emerge, including integration of multiple satellite channels (beyond IR bands), increasing the number of frames generated in a single forward pass beyond the current 10-frame limit, investigation of physics-informed loss functions to better preserve cyclone dynamics (ensuring physical consistency), and using VDMs as extension to other crucial weather forecasting applications in need of improvements in temporal coherence capabilities.

REFERENCES

[1] Knutson T, Camargo SJ, Chan JCL, Emanuel K, Ho CH, Kossin J, et al. Tropical Cyclones and Climate Change Assessment: Part II: Projected Response to Anthropogenic Warming. Bulletin of the American Meteorological Society. 2020 Mar;101(3):E303-22. Available from: https://journals.ametsoc.org/view/journals/bams/101/3/bams-d-18-0194.1.xml.

[2] Krichene H, Vogt T, Piontek F, Geiger T, Schötz C, Otto C. The social costs of tropical cyclones. Nature Communications. 2023 Nov;14(1):7294. Available from: https://www.nature.com/articles/s41467-023-43114-4.

[3] Emanuel K. Response of Global Tropical Cyclone Activity to Increasing CO2: Results from Downscaling CMIP6 Models. Journal of Climate. 2021 Jan;34(1):57-70. Available from: https://journals.ametsoc.org/view/journals/clim/34/1/jcliD200367.xml.

[4] Schalkwijk J, Jonker HJJ, Siebesma AP, Van Meijgaard E. Weather Forecasting Using GPU-Based Large-Eddy Simulations. Bulletin of the American Meteorological Society. 2015 May;96(5):715-23. Available from: https://journals.ametsoc.org/doi/10.1175/BAMS-D-14-00114.1.

[5] Huang C, Mu P, Bai C, Watson PA. TCP-Diffusion: A Multi-modal Diffusion Model for Global Tropical Cyclone Precipitation Forecasting with Change Awareness. arXiv; 2024. ArXiv:2410.13175. Available from: http://arxiv.org/abs/2410.13175.

[6] Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, et al. A Survey on Deep Learning: Algorithms, Techniques, and Applications. ACM Computing Surveys. 2019 Sep;51(5):1-36. Available from: https://dl.acm.org/doi/10.1145/3234150.

[7] Accarino G, Donno D, Immorlano F, Elia D, Aloisio G. An Ensemble Machine Learning Approach for Tropical Cyclone Detection Using ERA5 Reanalysis Data. arXiv; 2023. ArXiv:2306.07291. Available from: http://arxiv.org/abs/2306.07291.

[8] Kovordányi R, Roy C. Cyclone track forecasting based on satellite images using artificial neural networks. ISPRS Journal of Photogrammetry and Remote Sensing. 2009 Nov;64(6):513-21. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0924271609000434.

[9] Leinonen J, Hamann U, Nerini D, Germann U, Franch G. Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. arXiv; 2023. ArXiv:2304.12891. Available from: http://arxiv.org/abs/2304.12891.

[10] Nath P, Shukla P, Wang S, Quilodrán-Casas C. Forecasting Tropical Cyclones with Cascaded Diffusion Models. arXiv; 2024. ArXiv:2310.01690. Available from: http://arxiv.org/abs/2310.01690.

[11] Ho J, Salimans T, Gritsenko A, Chan W, Norouzi M, Fleet DJ. Video Diffusion Models. arXiv; 2022. ArXiv:2204.03458. Available from: http://arxiv.org/abs/2204.03458.

[12] Blattmann A, Dockhorn T, Kulal S, Mendelevitch D, Kilian M, Lorenz D, et al.. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv; 2023. ArXiv:2311.15127. Available from: http://arxiv.org/abs/2311.15127.

[13] Unterthiner T, Steenkiste Sv, Kurach K, Marinier R, Michalski M, Gelly S. Towards Accurate Generative Models of Video: A New Metric & Challenges. arXiv; 2019. ArXiv:1812.01717. Available from: http://arxiv.org/abs/1812.01717.

[14] Ho J, Salimans T. Classifier-Free Diffusion Guidance. arXiv; 2022. ArXiv:2207.12598. Available from: http://arxiv.org/abs/2207.12598.

APPENDIX A   MODEL ARCHITECTURE
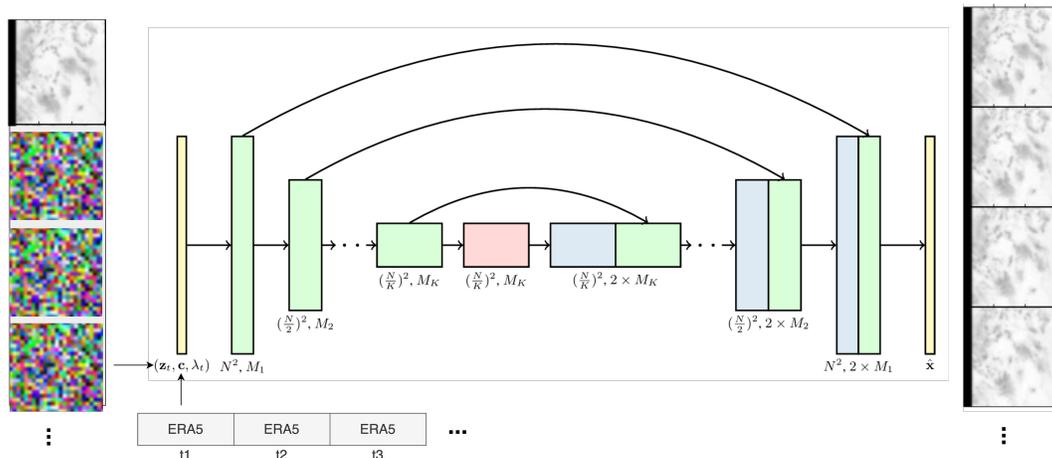

Figure A.1: Illustration of the model pipeline. Our VDM model takes as input a noisy image $\mathbf{z}_t$, conditioning variables $\mathbf{c}$, and a timestep embedding λ_t, and progressively denoises the sample using a U-Net-based architecture with skip connections. ERA5 data from multiple timesteps (t_1, t_2, t_3) is used as conditioning information. Output $\hat{\mathbf{x}}$ represents the denoised prediction.

APPENDIX B   IMPLEMENTATION DETAILS

Table B.1: Hyperparameters used in model training

Hyperparameter	Global Dataset	North Indian Ocean Region
Batch Size	1	1
Sequence Length	10	10
Learning Rate	3×10^{-4}	3×10^{-4}
Guidance Scale	3.0	3.0
Epochs Stage 1 (Single-frame)	100	200
Epochs Stage 2 (Multi-frame)	300	200

APPENDIX C   ADDITIONAL RESULTS

C.1   FORECAST SSIM

(a) Mocha
(North Indian Ocean)

(b) Ida
(North Atlantic Ocean)

(c) Roslyn
(Eastern Pacific Ocean)

(d) Molave
(Western Pacific Ocean)

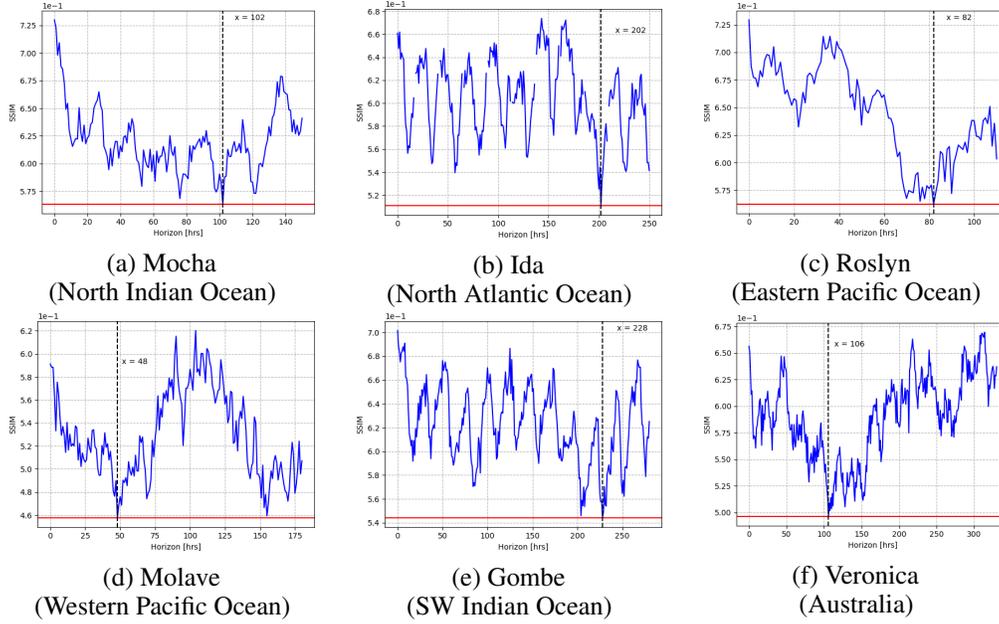(e) Gombe
(SW Indian Ocean)

(f) Veronica
(Australia)

Figure C.2: SSIM values over the entire cyclonic duration. The dashed lines indicate the hourly marks at which the minimum SSIM values are obtained for each cyclone.