DIFFOBS: GENERATIVE DIFFUSION FOR GLOBAL FORECASTING OF SATELLITE OBSERVATIONS

Jason Stock*
NVIDIA Corporation and Colorado State University
stock@colostate.edu

Jaideep Pathak, Yair Cohen, Mike Pritchard, Piyush Garg, Dale Durran, Morteza Mardani & Noah Brenowitz
NVIDIA Corporation
{jpathak,yacohen,mpritchard,piyushg,ddurran, mmardani,nbrenowitz}@nvidia.com

ABSTRACT

This work presents an autoregressive generative diffusion model (DiffObs) to predict the global evolution of daily precipitation, trained on a satellite observational product, and assessed with domain-specific diagnostics. The model is trained to probabilistically forecast day-ahead precipitation. Nonetheless, it is stable for multi-month rollouts, which reveal a qualitatively realistic superposition of convectively coupled wave modes in the tropics. Cross-spectral analysis confirms successful generation of low frequency variations associated with the Madden–Julian oscillation, which regulates most subseasonal to seasonal predictability in the observed atmosphere, and convectively coupled moist Kelvin waves with approximately correct dispersion relationships. Despite secondary issues and biases, the results affirm the potential for a next generation of global diffusion models trained on increasingly sparse, and increasingly direct and differentiated observations of the world, for practical applications in subseasonal and climate prediction.

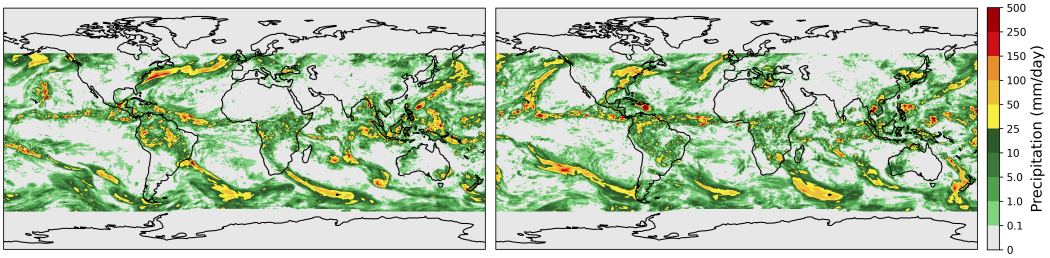
1 INTRODUCTION

As machine learning-driven global forecasting systems exit their infancy and move beyond weather (Pathak et al., 2022; Bi et al., 2022; Chen et al., 2023; Lam et al., 2022) toward climate (Watt-Meyer et al., 2023; Weyn et al., 2019) timescales, whether they can be made to generate realistic convectively coupled tropical disturbances across daily to multi-week simulations becomes an important question. Such atmospheric variability has been a longstanding challenge to capture realistically in physics-based models (Randall, 2013) and is still incompletely understood (Zhang et al., 2020), yet regulates the subseasonal predictability of the Earth System, including important tropical to extratropical teleconnections (Lau et al., 2005).

These dynamics become especially interesting to examine in emerging autoregressive diffusion models (Price et al., 2023; Mardani et al., 2024; Li et al., 2023; Nath et al., 2023) that learn conditional probabilities and are thus well suited to the stochastic character of tropical convective dynamics. Moreover, such methods suggest that diffusion models (Karras et al., 2022; Song et al., 2020) do not require complete information about the atmospheric state, and thus may have the capacity to produce realistic variability even when trained on limited, direct (e.g., univariate) observations. Some recent work (Gao et al., 2023; Leinonen et al., 2023) has explored diffusion models with univariate weather data, but only on short-term scales and in small spatial domains; computational advances in GPU computing allow more ambition today.

In this context, we introduce a computationally ambitious, high-resolution ($0.4°$) global autoregressive diffusion model trained solely on a satellite derived precipitation product. Precipitation is observed globally via microwave sensing satellites (e.g., TRMM, Kummerow et al. (1998) and

*Work done during an internship at NVIDIA.



(a) Observation (b) DiffObs

Figure 1: Example 3-day rollout from Oct 27, 2020 as initial condition.

GPM, Hou et al. (2014)), and prior low-order models have proven skillful at long-range forecasts (Chen et al., 2014). The coupling of both data and our modeling approach is pivotal to the feasibility of our current work. Using our trained model, we perform an in depth analysis of generated tropical variability on 1- to 60-day timescale using domain-informed diagnostics, discovering long-term stability with realistic variability of multiple wave structures.

2 Methodology

We introduce an autoregressive diffusion model, extending the EDM architecture (Karras et al., 2022) with the objective to estimate $p\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right)$ without incorporating additional priors. Achieving this goal assumes a paired spatiotemporal relationship within the underlying distribution to effectively capture the dynamics of the system based solely on the immediate past state. In doing so, our model can rollout predictions by utilizing the estimated next step as the subsequent initial condition.

The design specifics of our model are inspired by the work of prior global diffusion-based weather forecasting models. Specifically, we build upon the work of Mardani et al. (2024), which employs a similar architecture for km-scale downscaling. However, we avoid the use of an intermediate regression model and do not scale down the conditional inputs. Price et al. (2023) present a purely autoregressive diffusion model, but train on a comprehensive state vector given from reanalysis data, where instead we directly estimate a single observational state. Additional details in Appendix A.

2.1 Training Details

Our experiments use the default hyperparameters outlined in Karras et al. (2022), extending the DDPM++ UNet architecture (Song et al., 2020), with the only deviations being the exclusion of self-attention and a reduction in model channels, specifically from $128 \rightarrow 64$. Despite the limitations imposed on the model's receptive field and its ability to capture global synoptic information without self-attention, we find the change is needed to achieve reasonable performance and training stability.

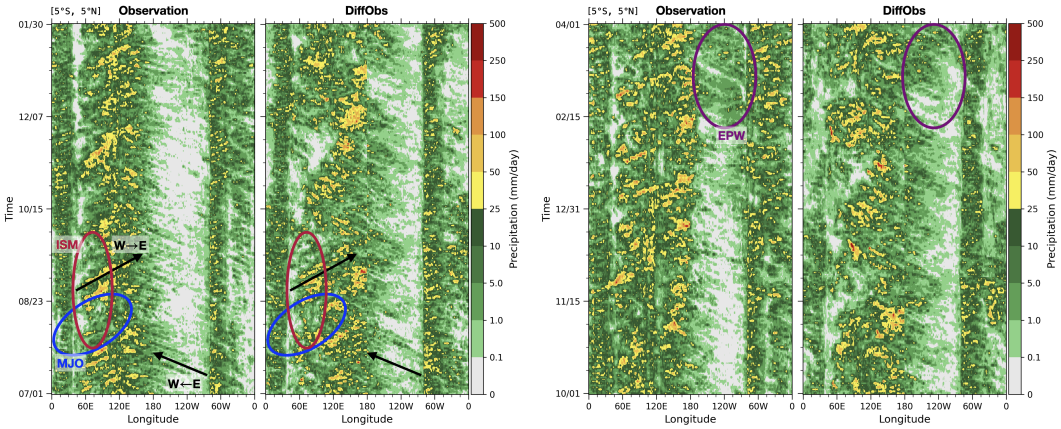
We train our 13.6M parameter model on a cluster with $256\times$ 80 GB H100 NVIDIA GPUs (32 nodes) using a global batch size of $1{,}024$ for 12.5M total steps. End-to-end training takes $4\,\mathrm{h}$ wall-clock time. During generation, we sample with 64 denoising steps using the default noise levels. A single output image is fully generated in $8.5\,\mathrm{s}$ (unoptimized) using 4.6 GB of memory on a single GPU.

3 Experiments

We begin by introducing the dataset for this study in Section 3.1, detailing the preprocessing and partitioning specifics, and in Section 3.2 we showcase and discuss our primary results.

3.1 Dataset Details

This study uses the final precipitation, half hourly Integrated Multi-satellitE Retrievals for Global Perception Measurements (IMERG) L3 Version 06B data (Huffman et al., 2019; 2015). Global estimates are derived through intercalibration, morphing, and interpolating various satellite microwave



(a) Summer to Winter

(b) Boreal Winter

Figure 2: Hovmöller diagrams of observations (left) and DiffObs output (right, generated autoregressively) between 5°N and 5°S for case studies initially conditioned on (a) July 1, 2020 and (b) Oct 1, 2019. Individual colors correspond to the wave propagation directions (W↔E), Indian Summer Monsoon (ISM), Madden–Julian oscillation (MJO), and East Pacific Wavetrain (EPW).

precipitation and infrared retrievals, precipitation gauge analyses, and surface observations (e.g., temperature, pressure, and humidity).

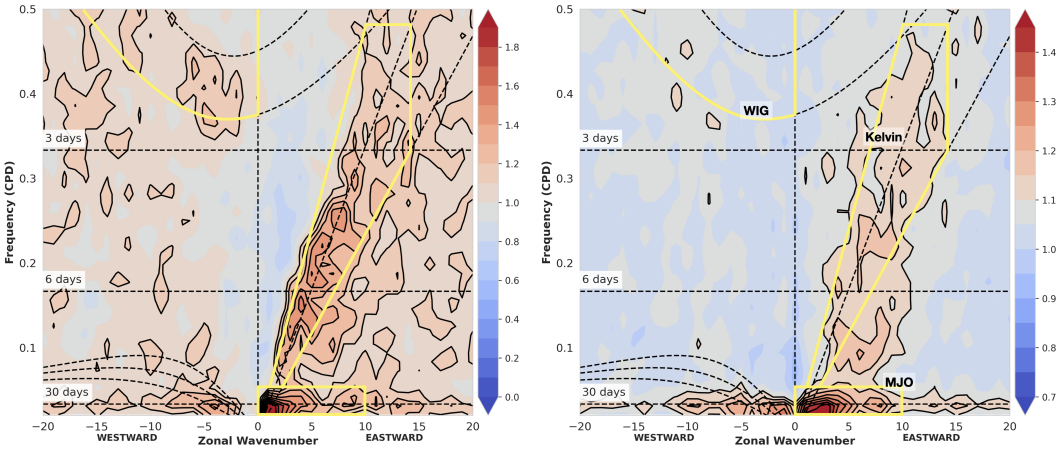
We collect data from June 1, 2000 to Sept 30, 2021 and aggregate all half hour samples for each day into an estimate of total daily precipitation (in $\mathrm{mm/d}$). Thereafter, we spatially coarsen the grid from $0.1° \to 0.4°$ with cropping in the meridional direction between $56.2°$N and $61.8°$S (296 latitudes and 900 longitudes) to avoid masking missing values at the poles. Data are partitioned to the years of 2000–2016 (6,041) for training and 2017–2022 (1,729) for testing, with the total samples in parentheses. Individual sample pairs, $(\mathbf{x}_t, \mathbf{x}_{t-1})$, are on a one-day interval with $\mathbf{x}_{t-1}$ being the condition.

Even with daily-accumulated estimates, the distribution of data is heavily right-skewed and primarily comprised of zero-valued cells with few high, yet critical precipitation values (e.g., in locations with severe weather). We therefore transform the data to be relatively Gaussian, and using statistics from the training data, normalize it between $[-1, 1]$ to align with the assumptions of diffusion models. This is done as $g(x) = 2 \cdot \ln\left(1 + x/\epsilon\right) / x_{max} - 1$, where $\epsilon = 10^{-4}$ and $x_{max} = 17.35$ as found in the transformed data. Computing $g^{-1}(x)$ on model output returns the data to its original units.

3.2 Main Findings

Using our trained model, we can generate forecasts for arbitrary lead times and leverage its inherent probabilistic nature to create an ensemble of forecasts from any initial condition. A representative member of an ensemble is shown in Figure 1, featuring a 3-day rollout initialized with a sample from Oct 27, 2020, to illustrate example output. While exact features should not be expected to match perfectly due to atmospheric chaos, it is notable that the forecast maintains qualitative sharpness, addressing concerns observed in deterministic convolutional networks (Ravuri et al., 2021; Ayzel et al., 2020), and accurately captures the structure of atmospheric conditions, including many high-valued precipitation events. Next, we shift our focus to evaluate long, multi-month rollouts to study multi-scale generated atmospheric variability near the equator.

Figures 2 and 3 illustrate our key findings and capabilities. We first compare how convectively coupled equatorial waves (averaged between 5°S and 5°N) propagate through longitude and time relative to observations with a Hovmöller Diagram (Hovmöller, 1949) — a preferred domain diagnostic — in Figure 2. Qualitatively, a reassuring superposition of eastward- and westward-propagating tropical disturbances are generated at appropriate longitudes, modulated by a large-scale envelope of slow, eastward moving variability characteristic of the Madden–Julian oscillation (MJO, Madden & Julian, 1994), at its expected location spanning the Indian Ocean (50–60°E) to West-Central Pa-



(a) Observation (b) DiffObs

Figure 3: Symmetric / Background Wheeler–Kiladis space-time spectra between 15°N and 15°S. The individually highlighted regions correspond to where the Madden–Julian oscillation (MJO), Kelvin and westward inertio-gravity (WIG) waves are expected to be found.

cific (180°E). Encouraging East Pacific Wavetrain (EPW, Zhou et al., 2012) variability is also found, alongside some artifacts, such as a dateline discontinuity at 180°E/W and a bias towards too much time-mean precipitation generated between (additional comments in Appendix D).

Our second analysis (Figure 3) examines the dispersion relationships revealed in the wavenumber–frequency domain, following the methods in Wheeler & Kiladis (1999); Maier-Gerber (2018). We generate 80 yrs of data on one-day intervals, initially conditioning 1 yr rollouts on Jan 1 for years 2017–2021 and sample with perturbed noise. Temporally concatenating the results within 15°N/S of the equator, we perform spectral analysis to construct a Wheeler–Kiladis diagram, utilizing 96 d windows with a 65 d overlap to isolate the significant spectral peaks.

As a baseline, Figure 3a illustrates the equatorially symmetric signal-to-noise for observations from the 5 years of test data. Key features, familiar to domain scientists, include the dominant power and east-to-west asymmetry on intraseasonal time scales (periodicity longer than 30 days), notably highlighting the MJO, as well as elevated power for an eastward-propagating convectively coupled Kelvin wave (Straub & Kiladis, 2002), spanning wavenumbers 1–14 with periods ranging from 2.5 to 25 days, and exhibiting a quasi-linear dispersion relationship (Kiladis et al., 2005).

Encouragingly, DiffObs reproduces both of these dominant observed features (Figure 3b): the spectral signal of generated power also occurs on intraseasonal time scales, i.e., timescales longer than 30 day periodicity, and across a band of spatial (zonal) wavenumbers 0–9 consistent with a planetary scale, eastward moving mode of variability. Meanwhile, on shorter time scales, the model also generates a moist Kelvin wave spectral power maximum with a qualitatively correct dispersion relationship. Despite other imperfections, such as a tendency for the model to generate too much variability at all wavelengths (Figures 5d and 5f), and an under-representation of power within westward moving tropical wave classes, these are impressive preliminary results. Further discussion of the antisymmetric component and log power spectra can be found in Appendix B. Altogether, these results demonstrate the nontrivial ability to learn and autoregressively generate complex multi-scale tropical weather patterns over extended temporal scales, from precipitation data alone.

4 Conclusion

We have demonstrated an autoregressive, univariate, machine learning diffusion model for predicting daily-accumulated precipitation based only on the previous day's data. The model produces stable long rollouts that exhibit a realistic spectrum of tropical wave modes, including the Madden–Julian oscillation, whose variance dominates on intraseasonal time scales, and which is notoriously difficult to simulate realistically in physics-based models (Zhang et al., 2020). Acknowledging sec-

ondary issues and biases, our overall findings hold promise for the next generation of global diffusion models, emphasizing the impact of training on increasingly sparse and differentiated observations of the world for subseasonal and climate prediction applications.

One caveat of this work is that the IMERG data we use is a level 3 product that integrates many separate data sources, including ERA5 (Hersbach et al., 2020). Nonetheless, we feel this work takes a bold step away from reanalysis and future work could extend to even lower-level satellite products. Moreover, we would like to explore approaches to explicitly capture the temporal distributions, e.g., use multiple timesteps as the input condition to estimate $p\left(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_{t-2}\right)$, as well as more directly compare to numerical weather prediction and reanalysis data, e.g., the IFS and ERA5.

ACKNOWLEDGMENTS

This work was supported by NVIDIA through an internship with the Climate Simulation Research Group. This work was also supported by NSF Grant No. 2019758, *AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES).*

REFERENCES

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Georgy Ayzel, Tobias Scheffer, and Maik Heistermann. Rainnet v1. 0: a convolutional neural network for radar-based precipitation nowcasting. *Geoscientific Model Development*, 13(6):2631–2644, 2020.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*, 2022.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023.

N Chen, A J Majda, and D Giannakis. Predicting the cloud patterns of the Madden-Julian oscillation through a low-order nonlinear stochastic model. *Geophys. Res. Lett.*, August 2014. ISSN 0094-8276, 1944-8007. doi: 10.1002/2014GL060876.

Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Wang. Prediff: Precipitation nowcasting with latent diffusion models. *arXiv preprint arXiv:2307.10422*, 2023.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Arthur Y Hou, Ramesh K Kakar, Steven Neeck, Ardeshir A Azarbarzin, Christian D Kummerow, Masahiro Kojima, Riko Oki, Kenji Nakamura, and Toshio Iguchi. The global precipitation measurement mission. *Bulletin of the American meteorological Society*, 95(5):701–722, 2014.

Ernest Hovmöller. The trough-and-ridge diagram. *Tellus*, 1(2):62–66, 1949.

George J Huffman, David T Bolvin, Dan Braithwaite, Kuolin Hsu, Robert Joyce, Pingping Xie, and Soo-Hyun Yoo. Nasa global precipitation measurement (gpm) integrated multi-satellite retrievals for gpm (imerg). *Algorithm theoretical basis document (ATBD) version*, 4(26):30, 2015.

G.J. Huffman, E.F. Stocker, D.T. Bolvin, E.J. Nelkin, and Jackson Tan. Gpm imerg final precipitation l3 half hourly 0.1 degree x 0.1 degree v06. Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: March 19, 2022., 2019. 10.5067/GPM/IMERG/3B-HH/06.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

George N Kiladis, Katherine H Straub, and Patrick T Haertel. Zonal and vertical structure of the madden–julian oscillation. *Journal of the atmospheric sciences*, 62(8):2790–2809, 2005.

Christian Kummerow, William Barnes, Toshiaki Kozu, James Shiue, and Joanne Simpson. The tropical rainfall measuring mission (trmm) sensor package. *Journal of atmospheric and oceanic technology*, 15(3):809–817, 1998.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Alexander Pritzel, Suman Ravuri, Timo Ewalds, Ferran Alet, Zach Eaton-Rosen, et al. Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*, 2022.

William KM Lau, Duane E Waliser, and Duane Waliser. *Predictability and forecasting*. Springer, 2005.

Jussi Leinonen, Ulrich Hamann, Daniele Nerini, Urs Germann, and Gabriele Franch. Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. *arXiv preprint arXiv:2304.12891*, 2023.

Lizao Li, Rob Carver, Ignacio Lopez-Gomez, Fei Sha, and John Anderson. Seeds: Emulation of weather forecast ensembles with diffusion models. *arXiv preprint arXiv:2306.14066*, 2023.

Roland A Madden and Paul R Julian. Observations of the 40–50-day tropical oscillation—a review. *Monthly weather review*, 122(5):814–837, 1994.

Michael Maier-Gerber. A python package for the construction of the wheeler–kiladis space-time spectra. `https://github.com/mmaiergerber/wk_spectra`, 2018.

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Residual diffusion modeling for km-scale atmospheric downscaling. *arXiv preprint arXiv:2309.15214*, 2024.

Pritthijit Nath, Pancham Shukla, and César Quilodrán-Casas. Forecasting tropical cyclones with cascaded diffusion models. *arXiv preprint arXiv:2310.01690*, 2023.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Timo Ewalds, Andrew El-Kadi, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*, 2023.

David A Randall. Beyond deadlock. *Geophysical Research Letters*, 40(22):5970–5976, 2013.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

Nigel Roberts. Assessing the spatial and temporal variation in the skill of precipitation forecasts from an nwp model. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, 15(1):163–169, 2008.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Katherine H Straub and George N Kiladis. Observations of a convectively coupled kelvin wave in the eastern pacific itcz. *Journal of the Atmospheric Sciences*, 59(1):30–53, 2002.

Oliver Watt-Meyer, Gideon Dresdner, Jeremy McGibbon, Spencer K Clark, Brian Henn, James Duncan, Noah D Brenowitz, Karthik Kashinath, Michael S Pritchard, Boris Bonev, et al. Ace: A fast, skillful learned global atmospheric model for climate prediction. *arXiv preprint arXiv:2310.02074*, 2023.

Jonathan A Weyn, Dale R Durran, and Rich Caruana. Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, 11(8):2680–2693, 2019.

Matthew Wheeler and George N Kiladis. Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber–frequency domain. *Journal of the Atmospheric Sciences*, 56 (3):374–399, 1999.

Michio Yanai and Takio Maruyama. Stratospheric wave disturbances propagating over the equatorial pacific. *Journal of the Meteorological Society of Japan. Ser. II*, 44(5):291–294, 1966.

C. Zhang, Á. F. Adames, B. Khouider, B. Wang, and D. Yang. Four theories of the madden-julian oscillation. *Reviews of Geophysics*, 58(3):e2019RG000685, 2020. doi: https://doi.org/10.1029/2019RG000685. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019RG000685. e2019RG000685 2019RG000685.

Putian Zhou, Lingling Suo, Jiacan Yuan, and Benkui Tan. The east pacific wavetrain: Its variability and impact on the atmospheric circulation in the boreal winter. *Advances in Atmospheric Sciences*, 29:471–483, 2012.

APPENDIX

A DIFFUSION DETAILS

Diffusion methods are defined by separate forward and backward processes as represented by stochastic differential equations (SDEs). At a high level, these processes continuously increase or decrease the noise level of an input when moving forward or backward in time, respectively. Concretely, these SDEs evolve a sample, $\mathbf{x}$, to align with some data distribution, p, as it propagates through time (Karras et al., 2022; Song et al., 2020). Leveraging a numerical solver we define a noise scheduler, $\sigma(t)$, to prescribe a given noise level at time t, typically as $\sigma(t) \propto \sqrt{t}$.

The *forward* (drift-removed) SDE, as formulated by Karras et al. (2022); Mardani et al. (2024), is expressed as

$$\mathrm{d}\mathbf{x} = \sqrt{2\dot{\sigma}(t)\sigma(t)}\,\mathrm{d}\omega_t, \tag{1}$$

while the *reverse*-time SDE (Anderson, 1982), sampled iteratively starting from $\mathbf{x}(T) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ for a large $T \ldots 0$ (illustrated in Figure 4), is defined as

$$\mathrm{d}\mathbf{x} = -2\dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{x}} \log p\left(\mathbf{x}; \sigma(t)\right)\,\mathrm{d}t + \sqrt{2\dot{\sigma}(t)\sigma(t)}\,\mathrm{d}\bar{\omega}_t. \tag{2}$$

Here, $\dot{\sigma}(t)$ is the time derivative of $\sigma(t)$ and $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma)$ is the score function (Hyvärinen & Dayan, 2005). The two terms in Equation (2) are the deterministic component representing the probability flow ordinary differential equation (ODE) with noise degradation, and noise injection via the standard Wiener process (denoted by ω_t), respectively.

The significance of the score function in Equation (2) lies in its relevance to sampling from diffusion models. Notably, it has the intriguing characteristic of not relying on the typically intractable normalization constant of the underlying base distribution $p(\mathbf{x}; \sigma)$. Exploiting this independence, we can use a denoising method defined by a neural network that minimizes the expected L_2 loss as a substitute of the score function, i.e., $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma) = \left(D_\theta\left(\mathbf{x}, \sigma\right) - \mathbf{x}\right)/\sigma^2$.

Let D_θ be a denoising model that operates on a noisy input sample given $\mathbf{x}_t \in \mathbb{R}^{c \times h \times w}$ at sample time t and noise level σ, where the previous state or condition, $\mathbf{x}_{t-1}$, is concatenated channel-wise to

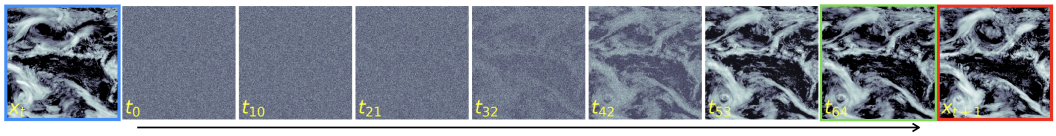


Figure 4: Reverse diffusion of a cropped sample with the input condition, individual sampling steps ($t_0 \rightarrow t_{64}$, inversely labeled), and the next time step estimate and target output.

the input. Concretely, $\mathbf{x}_t$ is the target next-day sample and $\mathbf{x}_{t-1}$ is the previous day input condition. We therefore optimize D_θ in training using

$$\min_\theta \mathbb{E}_{\mathbf{x}_{t,t-1}\sim p_{\text{data}}} \mathbb{E}_{\sigma\sim p_\sigma} \mathbb{E}_{\mathbf{n}\sim\mathcal{N}(\mathbf{0},\,\sigma^2\mathbf{I})} \left[\lambda(\sigma) \| D_\theta\left(\mathbf{x}_t + \mathbf{n}, \mathbf{x}_{t-1}; \sigma\right) - \mathbf{x}_t\|_2^2\right], \tag{3}$$

where the loss weight $\lambda(\sigma) = (\sigma^2 + \sigma_{\text{data}}^2)/(\sigma \cdot \sigma_{\text{data}})^2$, the noise level σ follows a log-normal distribution $\ln(\sigma) \sim \mathcal{N}(-1.2, 1.2^2)$, and $\sigma_{\text{data}} = 0.5$.

The denoising model, D_θ, with an underlying trainable network, F_θ, is preconditioned following

$$D_\theta\left(\hat{\mathbf{x}}_t, \mathbf{x}_{t-1}; \sigma\right) = c_{\text{skip}}(\sigma)\hat{\mathbf{x}}_t + c_{\text{out}}(\sigma) F_\theta\left(\left[c_{\text{in}(\sigma)}\hat{\mathbf{x}}_t, \mathbf{x}_{t-1}\right]; c_{\text{noise}}(\sigma)\right), \tag{4}$$

where the noisy input $\hat{\mathbf{x}}_t = \mathbf{x}_t + \mathbf{n}$, and $c_*(\sigma)$ (Karras et al., 2022, Table 1) are preconditioning variables to scale and modulate the individual components. The previous timestep (condition) is concatenated channel-wise by $[\cdot]$ and $c_{\text{noise}}(\sigma)$ is an additional latent condition for F_θ.

To generate samples from our diffusion model, we leverage Equation (2) with our trained denoising network D_θ. Specifically, we iteratively solve this using the stochastic EDM sampler that combines a second-order deterministic ODE integrator with stochastic Langevin-like churn. The sampled output is used autoregressively as the condition for the next sample time step.

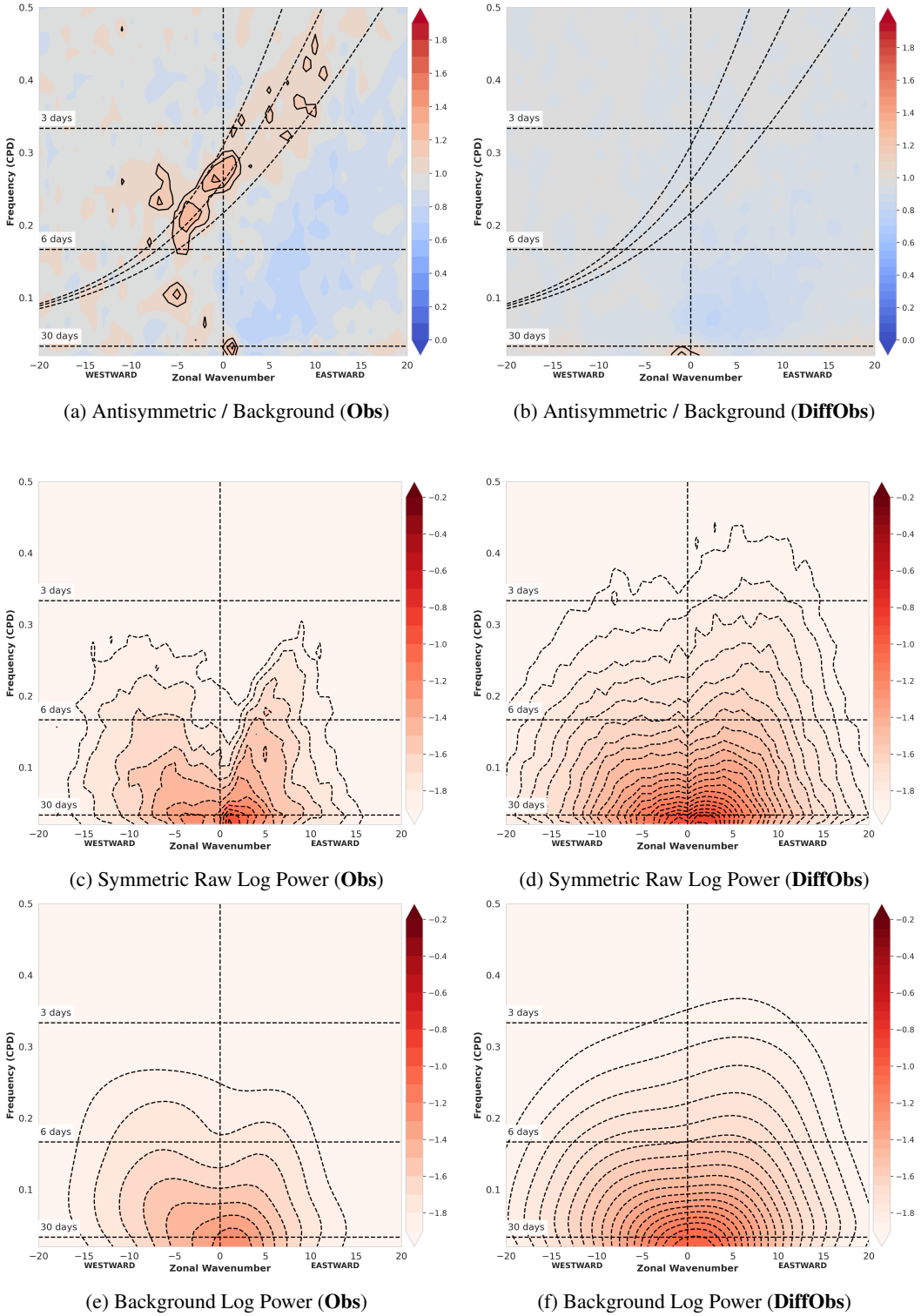
B Wheeler–Kiladis Space-Time Spectra

In their seminal work, Wheeler & Kiladis (1999) showed how fast and slowly oscillating atmospheric waves, some of which arise in idealized theories (Yanai & Maruyama, 1966), can be observed through spectral analysis of satellite-observed outgoing longwave radiation, including the Madden–Julian oscillation. This form of two dimensional spectral analysis has became one of the fundamental techniques for evaluating numerical models' representation of tropical waves that regulate predictability on synoptic to subseasonal timescales. Computing the Wheeler–Kiladis diagram reveals that DiffObs captures many, if not all, of the predominant observed modes and tropical wave signals. This could be viewed as a surprising property of a machine learning model trained only on a single variable, given that dynamical theories of such waves encompass several atmospheric state variables operating in concert — e.g., the vorticity and divergence of the horizontal winds are required for modeling in the equatorial Rossby and Kelvin waves, respectively, while the MJO requires complex non linear advection and moisture effects.

Our main finding is the discovery of Kelvin wave and MJO spectral signals within the signal-to-noise ratio of the equatorially-symmetric component of the space-time spectra (see Figure 3b). However, this is only one view of the analysis, and we glean more details from the additional components of the analysis. Specifically, in the antisymmetric component of observations (Figure 5a), there is evidence of a mixed Rossby-gravity (MRG) and eastward inertio-gravity (EIG) waves that are not apparent in Figure 5b. In Figures 5d and 5f, we show the intermediate computations, where the background spectra of model output is more similar to the raw power spectra than that of the observations. This suggests that DiffObs has a strong background signal that is similar to red-noise and that overall DiffObs generates too much variance.

C Short-Term Predictability

While short-term predictability is not the focus of this work, we find it is important to assess for comprehension and to be relatable to prior work. Therefore, we evaluate quantitative model predictability using the Root-Mean-Squared Error (RMSE, Price et al., 2023) relative to forecasts derived by



(a) Antisymmetric / Background (**Obs**)

(b) Antisymmetric / Background (**DiffObs**)

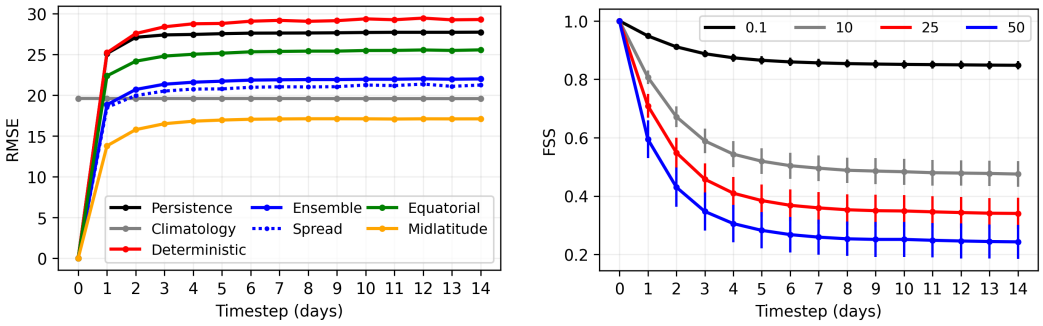
(c) Symmetric Raw Log Power (**Obs**)

(d) Symmetric Raw Log Power (**DiffObs**)

(e) Background Log Power (**Obs**)

(f) Background Log Power (**DiffObs**)

Figure 5: Additional Wheeler–Kiladis components and power spectra of observations (left column) and model output (right column) that support Figure 3.



(a) RMSE

(b) FSS

Figure 6: Short-term predictability for 14-day forecasts, initiated from each day between Jan 1, 2020 and Sept 15, 2021 with 5 ensemble members. Panel (a) shows the RMSE and (b) shows the ensemble mean FSS with different thresholds ($\mathrm{mm/d}$) at a $8.4°$ neighborhood (21 pixels).

the observations and the Fractions Skill Score (FSS, Roberts & Lean, 2008; Roberts, 2008). While the Continuous Ranked Probability Score (CRPS) and Brier Skill Score (BSS) are common metrics to evaluate forecast performance, we currently do not have other datasets to make for a meaningful comparison. As such, FSS is defined for a given neighborhood size as

$$\mathrm{FSS} = 1 - \frac{\frac{1}{n}\sum_n \left(P_y - P_t\right)^2}{\frac{1}{n}\left[\sum_n P_y^2 + \sum_n P_t^2\right]}, \tag{5}$$

where P_y is the forecast fraction and P_t is the target observed fraction (exceeding a certain threshold), and n is the number of spatial windows over the domain. We define the ensemble mean RMSE similar to Price et al. (2023) as

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_m \frac{1}{|G|}\sum_i a_i \left(t_{i,m} - \bar{y}_{i,m}\right)^2}, \tag{6}$$

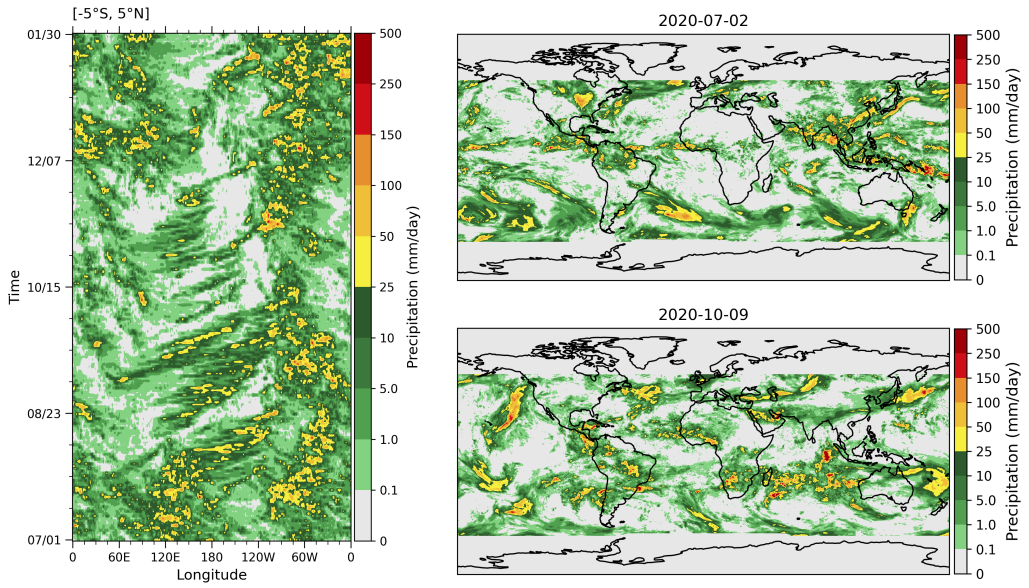
where $y^n_{i,m}$ denotes the $n \in N$ ensemble for the $m \in M$ forecast for a lead time at a latitude and longitude indexed by $i \in G$, $t_{i,m}$ is the target observation, and $\bar{y}_{i,m} = 1/N \sum_n y^n_{i,m}$ is the ensemble mean of model predictions. We use a latitude weighting derived from the area mean on a sphere, normalized to have unit mean as defined by

$$a_i = \frac{\cos\left(\mathrm{lat}\left(i\right)\right)}{\frac{1}{N_{\mathrm{lat}}}\sum_j \cos\left(\mathrm{lat}\left(j\right)\right)}. \tag{7}$$

We compare individual scores to the *persistence forecast* taken as the initial condition repeated over the forecast duration (14-days) as well as *climatology* found by a two week average window of mean daily conditions for years 2000–2022. Additionally, we compute the ensemble mean errors individually at the midlatitudes (between $30°$N/S and $60°$N/S) and tropics (between $23.5°$N and $23.5°$S) without any latitude weighting. We also show the deterministic error and the ensemble spread as the square root of $(n+1)/n$ times the average over all forecasts of the ensemble variance.

Figure 6a shows the forecast errors plateau quickly after a 3-day lead time, yet there is consistency with the persistence and deterministic errors approaching the ratio of $\sqrt{2}$ with lead time for climatology and the ensemble, respectively. Even though the ensemble error is better than persistence, given this consistency and that the model is under-dispersive (spread $<$ error), we can deduce that the ensemble mean is relatively poor. By computing errors at varying latitudes, we see the greatest error with the estimations in the tropics, where there is also a bias toward high precipitation values.

In evaluating our model using FSS, various neighborhood sizes were considered (not shown within), revealing that the $8.4°$ neighborhood size effectively captures large-scale atmospheric structure. The outcomes of experimenting with different thresholds are presented in Figure 6b. Notably, lower thresholds have higher skill, which continuously decreases with high-valued estimates, and the highest skill is at early lead times (out to 5 days), where skill plateau thereafter as seen with RMSE.



(a) Hovmöller

(b) Individual Output Samples

Figure 7: Experimental DiffObs output (Appendix D), generated autoregressively when initially conditioned on July 1, 2020. Panel (a) is the Hovmöller between $5°$S to $5°$N and (b) are individual output samples with their corresponding steps shown by the date.

D Additional Experiments

In our experiments, we observe a dateline discontinuity at $180°$E/W, evident down the center of the Hovmöller in Figure 2 (right). Ideally, we would like our model to be consistent around the globe and allow for periodic wave propagation. We aim to address this by modifying our network architecture and by including additional conditions, as outlined in Appendix D.1. However, we find the changes to be suboptimal, showing in Appendix D.2 further inconsistencies and worse performance relative to the baseline model.

D.1 Modifications

For our network to have rotational equivariance, we modify the convolutions to use circular-padding in the zonal direction and zero-padding in meridional direction (due to the cylindrical structure of our data). This effectively removes any spatial bindings given by the dateline. As such, we include a two-channel static condition (concatenated channel-wise to the existing condition) of $\cos(\text{lon})$ and $\sin(\text{lon})$, repeated over the meridional directions. These additional conditions, spatially aligned with the input, *should* maintain spatial coherence.

In addition to the padding and coordinate conditions, we also include the zonal average of the cosine of solar zenith angle as a function of the condition date and latitudes to account for temporal variability. We compute this for each latitude, ϕ, at time t as,

$$\cos\theta_s = \sin\phi\sin\delta + \cos\phi\cos\delta\cos h, \tag{8}$$

where δ is the declination of the sun and h is the hour angle. Given that our data is daily-accumulated, we integrate time by zonally averaging at UTC+0 and repeat the value for each latitude. The result is again concatenated channel-wise and *should* be effective to provide seasonal cycle conditioning. It is important to note that during training, we use the date associated with the condition (i.e., the previous timestep), iterating $\cos\theta_s$ in time during a rollout.

D.2 PRELIMINARY RESULTS

We train our updated model with the same hyperparameters and training specifics detailed in Section 2 and repeat similar evaluations from Section 3. While the periodicity is preserved across the dateline, we find the results to inadequately represent the atmospheric dynamics. The most salient representation of this is illustrated in Figure 7a when comparing to observation (Figure 2, left). Notably, no landmasses are identified, eastward-propagating waves traverse the dateline, and oscillating wave signals are not captured. While it is not abundantly clear as to why these modifications yield worse results, we note that there should be careful considerations when iterating on future work.