

Towards Forecasts of Seasonal Tropical Cyclone Activity in the North Atlantic using Diffusion Models

Benjamin Ulmer

`ulmerb@alumni.stanford.edu`

University of Cambridge

Charles Powell

University of Cambridge

Robert Rouse

University of Cambridge

Article

Keywords:

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11275019/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Towards Forecasts of Seasonal Tropical Cyclone Activity in the North Atlantic using Diffusion Models

Benjamin Ulmer², Charles Powell², Robert Edwin Rouse²

²Department of Applied Mathematics and Theoretical Physics,
University of Cambridge, Cambridge, United Kingdom

Corresponding author: Benjamin Ulmer (ulmerb@alumni.stanford.edu).

Abstract

Anticipating tropical cyclone activity for an upcoming season is important for managing risks to lives and infrastructure. However, long term monthly seasonal forecasts do not have the resolution needed to assess cyclone activity. We develop a conditional multivariate diffusion model that temporally downscales monthly-mean SEAS5 forecasts for the North Atlantic Ocean into independent weather snapshots. Evaluated on 500 dates from 2022–2024, the model produces representative weather distributions and spatial variability, as well as reproducing physically consistent relationships between fields such as relative vorticity and horizontal winds. The samples recover much of the frequency and seasonal variation of cyclone candidates found in ERA5, but favor extratropical systems and underrepresent candidates with a stricter warm-core criterion. These results demonstrate the potential of diffusion models to translate seasonal forecasts into realistic multivariate weather fields for tropical cyclone risk assessment.

1 Introduction

Tropical cyclones (TCs) are among the most destructive natural phenomena on Earth. Between 1980 and 2024, TCs caused \$1.5 trillion in damage in the United States [1], and resulted in 42% of the United States’ catastrophe-insured losses between 1992 and 2011 [2]. The cost is not only financial: globally, TCs killed 412,644 people, injured 290,654, and affected 466.1 million between 1980 and 2009, with the burden concentrated in South and South-Eastern Asia [3]. Nor is the cost only immediate. In the

United States, the average hurricane causes 24 immediate fatalities, but its effects can last for up to 15 years [4]. Through indirect economic, healthcare-related, and social effects, Young & Hsiang estimate that each hurricane is associated with a further 7,000–11,000 deaths [4]. Unfortunately, while significant uncertainty remains, climate models and recent observations suggest that hurricanes are likely growing more severe as climate change intensifies [5, 6]. Consequently, communities, asset owners, and governmental bodies impacted by TCs will likely find it increasingly valuable to anticipate the severity of the upcoming hurricane season.

In spite of recent advances in medium-range weather forecasting [7, 8], seasonal forecasts of TC activity remain challenging since their genesis and evolution depend on nonlinear interactions across a range of scales, from large-scale atmospheric and oceanic environments to synoptic-scale disturbances and convective-scale dynamics [9, 10]. Seasonal forecast systems, such as the European Centre for Medium-Range Weather Forecasts (ECMWF) SEAS5 system, issue forecasts up to 7 months ahead. They capture the slowly varying, predictable parts of the climate system, such as sea surface temperature (SST) and the El Niño-Southern Oscillation (ENSO) [11]. While SEAS5 is state of the art, these long-term monthly seasonal forecasts do not, nor are intended to, capture the weather variability or storm structures present at shorter timescales.

While the large-scale drivers of TC activity are somewhat predictable at coarse resolution on seasonal timescales [11], the realized activity is strongly modulated by intraseasonal factors such as the Madden-Julian Oscillation, monsoons, and synoptic-scale setups that produce seed disturbances from which TCs grow. Seasonal forecasts can capture these effects but only at the largest spatial scales, and typically only produce monthly-mean predictions. Nonetheless, these large-scale predictions can be used to predict TC activity at longer lead times. Lin et al.’s physics based model connects seasonal predictions directly to cyclone predictions through randomly seeding weak vortices in the Atlantic and evolving their intensities and tracks according to the large-scale environment [12]. This approach is effective, but it hard-codes TC assumptions into its seeding, steering, and intensity models. Another approach is to infer high-resolution realizations of weather that are consistent with the coarse seasonal predictions, a process known as downscaling. The high-resolution weather can then be passed into a detection algorithm such as Okubo–Weiss–Zeta (OWZ), which can identify TC-like structures [13, 14].

Diffusion models have recently been used for weather and climate downscaling. CorrDiff successfully downscales weather to the scale of kilometers, generating fine grained, physically plausible fields from coarse input [15]. Aich et al. built a diffusion model to generate ensembles of high resolution precipitation fields which capture precipitation uncertainty and fine scaled features [16]. These results show that diffusion models can be well suited to producing high-resolution, statistically-representative ensembles of weather fields, which is the capability needed to downscale a seasonal forecast. More recent work uses diffusion to generate coherent weather trajectories from coarse climate data [17], while other generative models make seasonal climate predictions directly [18].

In this research we present a method for downscaling long-range climate forecasts into multivariate weather snapshots from which TC-like structures can be detected. For this study, we focus on the single region of the North Atlantic Ocean. Given monthly mean forecasts of relevant variables, we train a diffusion model to generate samples of physically and statistically plausible weather at six-hourly timesteps by learning a distribution of residuals between the ERA5 historical weather and the SEAS5 forecasted weather. The model is judged by how well its generated samples match the statistics of the ERA5 weather, including the occurrence and distribution of TC-like structures. To be clear, the objective is not to predict individual storms months in advance, but rather to be given a six month forecast and to generate an ensemble of samples of plausible weather conditions that are consistent with the predicted seasonal conditions.

2 Results

The diffusion model converts monthly seasonal forecasts into possible weather snapshots over the North Atlantic. Each snapshot contains twelve fields on a 0.25° grid: horizontal winds, temperature, relative humidity and relative vorticity at 850 and 500 hPa, along with mean sea level pressure (MSLP) and SST. These fields are generated jointly, allowing the model to learn their relationships as well as their individual distributions.

We train the model on training pairs of an interpolated forecast from an SEAS5 monthly mean member six months prior to a date paired with the corresponding ERA5 data for that date. During training, the model learns the distribution of residuals between this background and the corresponding ERA5 weather. At generation time, the model starts from random noise and progressively denoises a residual conditioned on the interpolated forecast and a land/sea indicator. Adding that residual to the background forecast produces a weather snapshot that includes variability not present in the monthly means. Repeating this process with different forecast members and noise draws produces an ensemble of possible outcomes.

We generate twenty samples on each of 500 dates from our test set of 2022–2024. We first examine distributional gains, the recovery of spatial structure, and the ensemble’s probabilistic spread, before using individual fields to illustrate the generated weather and examining the TC specific properties of the samples. The model architecture, training and sampling procedures are described in Methods. We refer to the diffusion-model outputs as samples, the interpolated mean of the 25 SEAS5 members used here as the baseline, and the historical ERA5 fields as ERA5 data. ERA5 is a reanalysis that combines observations with an atmospheric model, rather than a direct observational record.

2.1 Distributions

Marginal distributions for each variable are computed from both the samples and the ERA5 data, and are compared. The samples reproduce the ERA5 marginal distributions more closely than the interpolated seasonal baseline for eleven of the twelve variables (Figure 1). We measure the normalized Wasserstein distance (Table 1) for

Table 1: Metrics for all twelve variables over the 500-date evaluation. Wasserstein distance and CRPS are normalized by their respective ERA5 standard deviations. Lower values are better. The spread-skill ratio uses the finite-ensemble correction described in Methods; a value of 1 indicates equal spread and error.

Variable	Wasserstein distance		Fair CRPS		Spread-skill ratio
	Generated	Baseline	Generated	Baseline	
Zonal wind (850 hPa)	0.0450	0.2382	0.399	0.560	0.834
Zonal wind (500 hPa)	0.0257	0.2534	0.385	0.542	0.867
Meridional wind (850 hPa)	0.0717	0.4876	0.478	0.667	0.848
Meridional wind (500 hPa)	0.0858	0.5263	0.481	0.670	0.842
Temperature (850 hPa)	0.0341	0.0749	0.188	0.274	0.895
Temperature (500 hPa)	0.0405	0.0558	0.192	0.274	0.885
Relative humidity (850 hPa)	0.0799	0.4115	0.461	0.673	1.028
Relative humidity (500 hPa)	0.1081	0.4205	0.497	0.729	1.002
Relative vorticity (850 hPa)	0.0275	0.5003	0.454	0.628	0.912
Relative vorticity (500 hPa)	0.0514	0.5231	0.459	0.636	0.887
MSLP	0.0904	0.2424	0.382	0.544	0.801
SST	0.0459	0.0341	0.066	0.109	0.652
Mean	0.0588	0.3140	0.370	0.525	0.871

each variable, which measures the minimum average change in values needed to transform the samples marginal distribution into the ERA5 distribution, expressed in units of standard deviations. The mean distance falls from 0.3140 for the baseline to 0.0588 for the generated fields, with an improvement for eleven of the twelve variables. The largest gains occur in relative vorticity and meridional wind, consistent with recovering the atmospheric variability missing from the monthly-mean baseline. SST is the sole exception: its marginal distribution becomes slightly less accurate than that of the baseline, which already agrees closely with ERA5. Paired comparisons across dates support improvement for ten variables, with neither SST nor temperature at 850 hPa meeting the adjusted significance threshold (Supplementary Table S6).

2.2 Spatial Variability

Spatial variability in the generated samples is assessed with a radially-averaged power spectral density (PSD) and with mean latitudinal and longitudinal profiles.

We evaluate the PSD over three bands as defined by Orlanski [19]: Macro- β and larger ($\lambda \geq 2000$ km), Meso- α (200–2000 km), and Meso- β (55–200 km). The diffusion model restores much of the spatial variability absent from the interpolated baseline (Figure 2). Mean banded log-spectral distance (LSD) falls from 5.865 to 0.959 dB at the Macro- β and larger scale, and from 21.971 to 1.444 dB at the Meso- α scale (Supplementary Table S1). The gain is largest at the Meso- α scale, where the generated spectra are closer to the ERA5 spectra on all 500 dates for every variable (Supplementary Table S6). This is the band in which the smooth baseline is weakest, and the

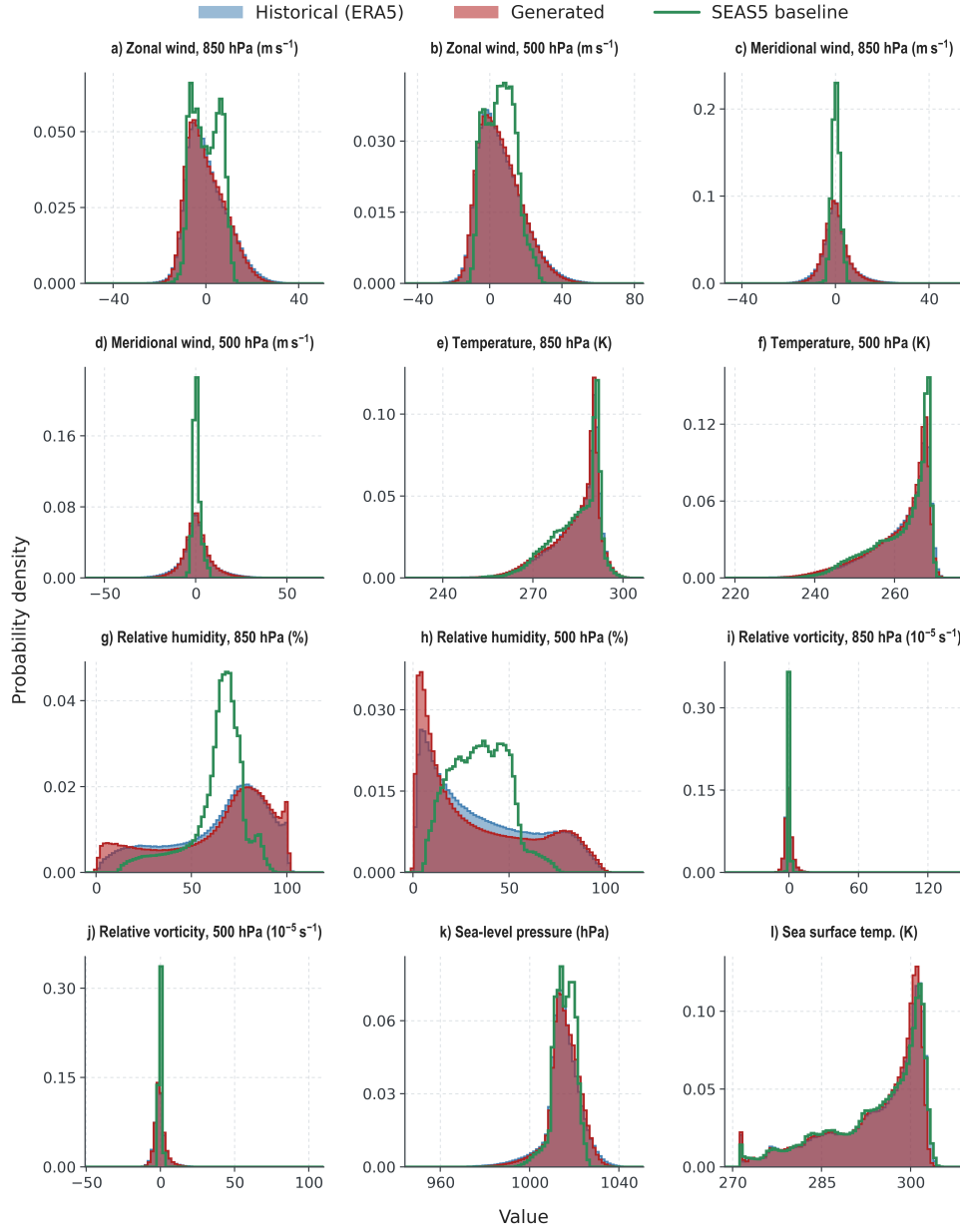


Fig. 1: Variable distributions in samples generated by the diffusion model (red), in the ERA5 samples (blue) and in the baseline interpolated SEAS5 ensemble mean (green). Normalized Wasserstein distances for all twelve variables are given in Table 1.

band that contains the storm-scale circulation the model is intended to recover. At the Meso- β scale, for MSLP and temperature at 500 hPa which tend to be relatively smooth, there is too little resolvable structure to report a meaningful LSD.

To determine if our model had any regional biases, we averaged each variable across samples and latitude and longitude. The mean latitudinal and longitudinal profiles are shown in Supplementary Figures S4 and S5, with their RMSEs reported in Supplementary Table S2. Both the samples and the baseline generally agree well with ERA5, indicating that the model preserves the broad regional patterns captured by the seasonal forecasts while adding finer-scale weather variability.

Our model generates samples with both relative vorticity and horizontal winds, but relative vorticity can also be calculated from the horizontal winds. Regressing the generated vorticity against the derived vorticity gives slopes of 1.0006 and 1.0001, with R^2 values of 0.9989 and 0.9985 at 850 and 500 hPa, respectively, showing the generated winds and vorticity closely satisfy their physical relationship (Supplementary Figure S2).

2.3 Probabilistic Spread

We evaluate the success of the generated samples as a probabilistic forecast using spread-skill scores, rank histograms, and continuous ranked probability scores (CRPS). The main finding is that the ensemble is slightly under-dispersed, although the size and form of the calibration error vary by variable.

The mean corrected spread-skill ratio is 0.871, and ten of the twelve variables fall below one (Supplementary Figure S3, Table 1). SST has the lowest ratio (0.652), followed by MSLP (0.801). Wind-component ratios range from 0.834 to 0.867, and relative vorticity has ratios of 0.912 and 0.887. The overall ratios for relative humidity are close to one (1.028 and 1.002), but Supplementary Figure S3 shows that this comes not from good calibration but from under-dispersion at low spread and over-dispersion at high spread, which cancel in the aggregate.

The rank histograms reveal both dispersion error and systematic bias (Figure 3). Relative vorticity is closest to flat, with reliability indices (RI) of 0.016 and 0.038. SST has the largest departure from uniformity (RI 0.562), but its endpoint frequencies are strongly asymmetric. The spread-skill diagrams and rank histograms show a trend of slight underdispersion across samples, meaning that samples would underestimate uncertainty in the possible weather outcomes.

Mean normalized CRPS falls from 0.525 for the deterministic baseline to 0.370 for the generated ensemble (Table 1). The generated score is lower for every variable. Eleven variables improve on every one of the 500 dates, while SST improves on 497 dates. Full CRPS scores are given in Supplementary Table S4.

2.4 Tropical Cyclone Frequency & Distribution

The generated weather fields are temporally independent, thus we assess TC-like structures using instantaneous detections from the OWZ algorithm. OWZ combines criteria for atmospheric rotation, moisture and vertical wind shear to identify candidate circulations (see Methods for a more complete description). These candidates would usually

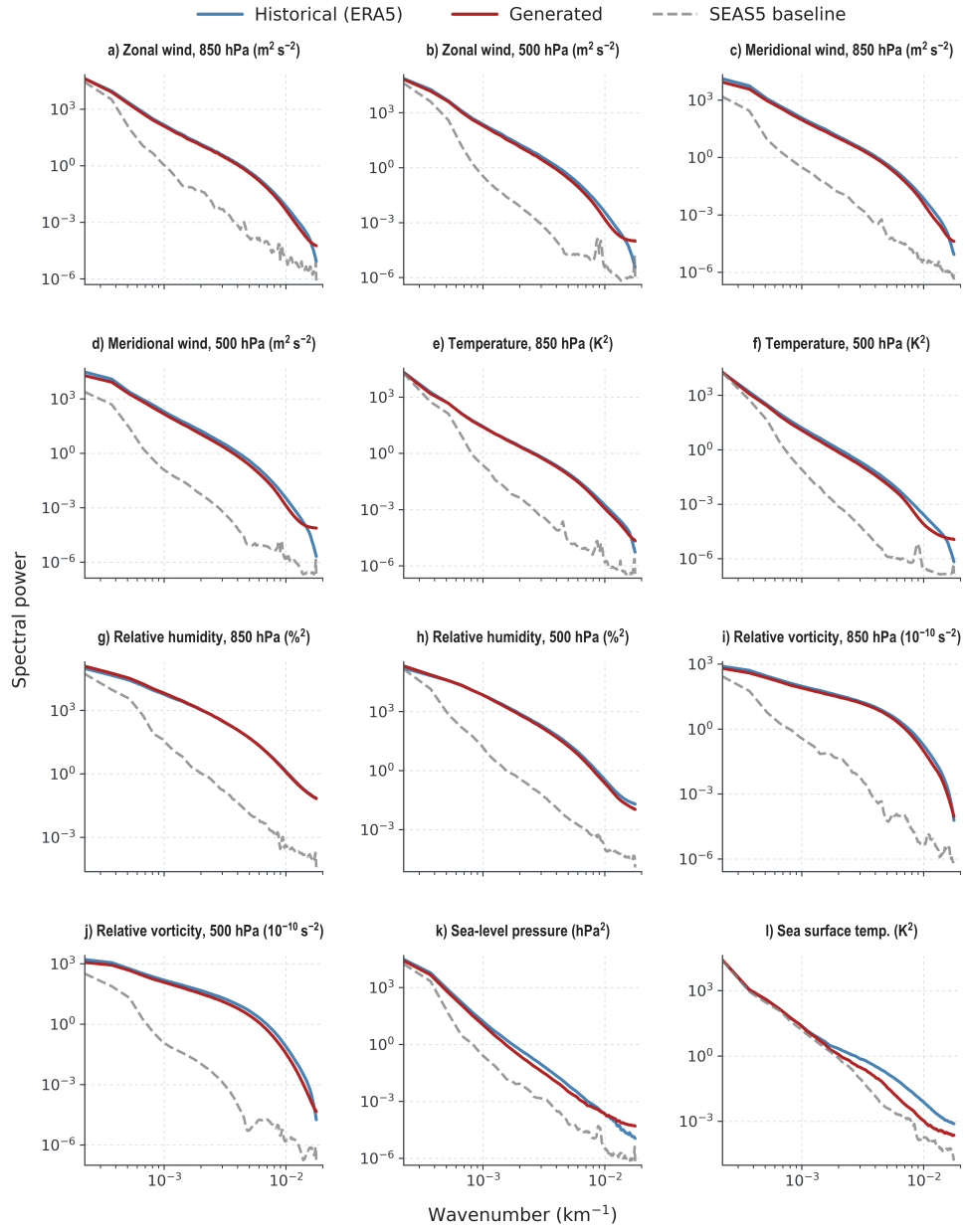


Fig. 2: Radially-averaged PSD of the generated fields, the interpolated SEAS5 baseline, and ERA5 samples for all twelve variables. Here, wavenumber is the inverse of the wavelength. Banded LSD for all twelve variables are given in Supplementary Table S1

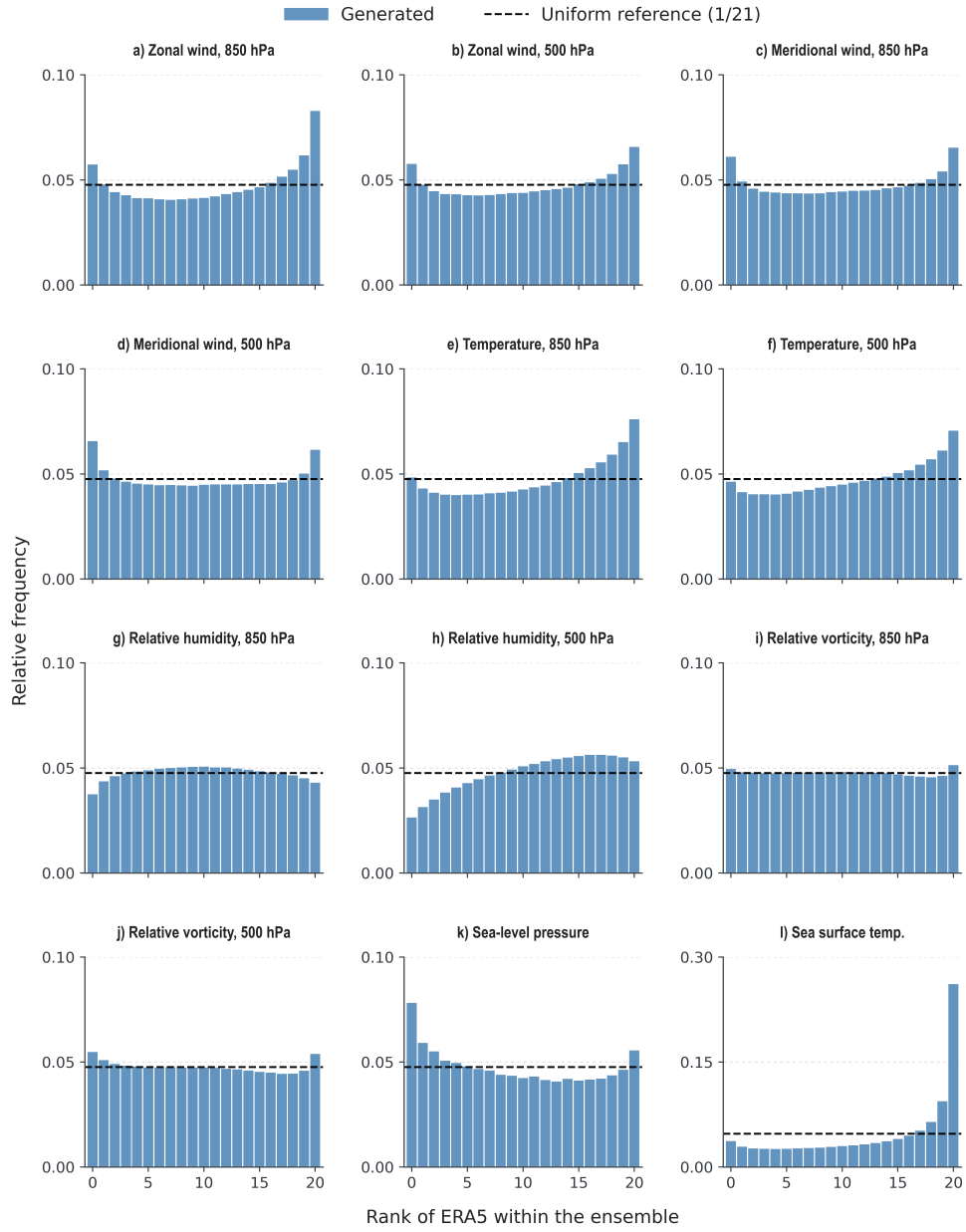


Fig. 3: Rank histograms for all twelve variables in the final twenty-member ensembles against ERA5 over 500 test dates. The dashed line indicates the uniform frequency $1/21$. Scores for all twelve variables are given in Supplementary Table S3.

be linked across adjacent timesteps to assess their persistence and construct storm tracks, which is not possible for the generated samples. Detections therefore represent candidates in individual snapshots, and a single ERA5 storm can contribute detections at several timestamps. Across all 500 evaluated dates, ERA5 contains 1079 candidate detections. Over that same time period, the samples correspond to 1022.4 detections, or 94.8% of ERA5 (Supplementary Table S5).

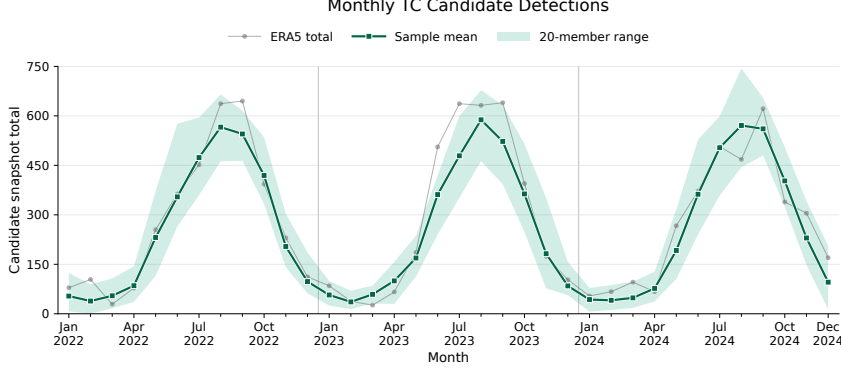
Figure 4a compares the monthly variation in these detections. ERA5 totals include all six-hourly timestamps in each month. For the generated fields, we find the mean number of detections per sampled timestamp and member, then multiply it by the number of six-hourly timestamps in that month. The generated estimates therefore represent the same monthly exposure as ERA5. The generated fields reproduce much of the monthly variation, with a correlation of 0.969 across the 36 monthly comparisons. Both sources peak in late summer in all three years, during the expected period of peak activity in the North Atlantic basin, although the maximum frequency is underestimated in the model compared to ERA5. These results show that the model generates TC-like structures at a frequency close to ERA5.

Figure 5 compares the spatial distribution of TC-like features in the generated samples with ERA5. There is a bias towards extratropical features in the generated samples, likely due to their more frequent presence in the training data. The differences are statistically significant in the northwestern and southeastern Atlantic, indicating a poleward and westward shift in the distribution. Differences also occur in the southwestern part of the Caribbean Sea, near the Isthmus of Panama, which may be an artifact of the detection method mishandling the complex features arising in this region (see figure 6).

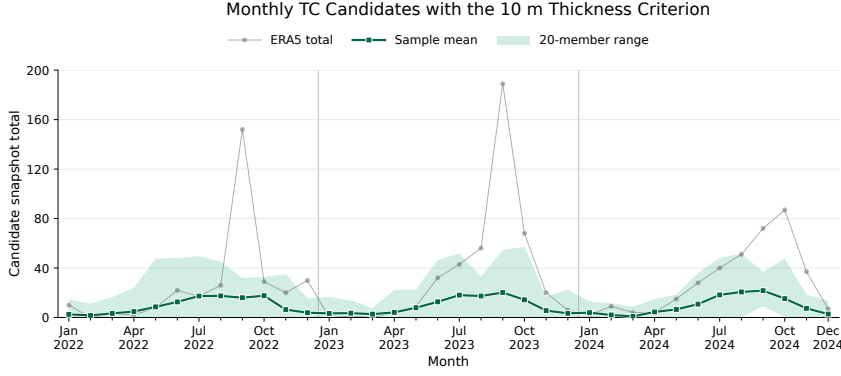
To examine whether the OWZ candidates also contain a warm core expected of TCs, we apply the additional thickness criterion described in Methods. On the same 500 timestamps, the ERA5 snapshots contain 121 detections and the samples contain a mean of 37.8 detections per member. Monthly agreement also weakens, with $r = 0.671$ after filtering (Figure 4b). The model therefore reproduces the broad seasonal occurrence of OWZ candidates more closely than that of candidates with this closed thermal structure. This drop in performance is likely due to a combination of the model’s bias towards the extratropics in its spatial distribution, as well as weaker temperature anomalies, broader vortices, and sensitivity to the choice of threshold in the thickness criterion (which is applied at non-standard pressure levels).

2.5 Tropical Cyclone Structure

Figure 6 is an example of a generated sample, one of many examined for qualitative accuracy. Supplementary Figure S1 shows an example containing all twelve variables. The generated humidity, wind and vorticity fields contain fine-scale structure that is absent from the interpolated ensemble-mean forecast but comparable to structures present in the corresponding ERA5. The placement of individual features differs from ERA5 as expected, but remains consistent with the large-scale setup captured by the seasonal forecast. We also note consistency between different atmospheric variables that is not enforced in training. For example, both ERA5 and the generated sample concentrate mid-level relative humidity in the extratropics and close to the equator, in



(a) OWZ candidate detections.



(b) OWZ candidate detections with the thickness criterion.

Fig. 4: Monthly OWZ candidate detections from January 2022 to December 2024, (a) without and (b) with the thickness criterion. Grey lines show ERA5 totals over all six-hourly timestamps within each month; green lines show the twenty-member generated mean. Shading shows the minimum-to-maximum range of the twenty member-specific monthly estimates.

agreement with the pressure pattern that features higher pressure around 30° N (indicating subsidence which dries the mid-troposphere) with lower pressure equatorward and poleward. The circulation pattern is also consistent, with strong extratropical zonal flow around 45° N and anticyclonic flow around the high-pressure regions. The generated sample shows two low-pressure systems with vortical features in the relative vorticity field at low levels; both demonstrate good similarity – in terms of both size and structure – with the low-pressure system present in the ERA5 sample. Other vortical features are also consistent in the generated and ERA5 sample, e.g. near mountain ranges at the northern tip of South America.

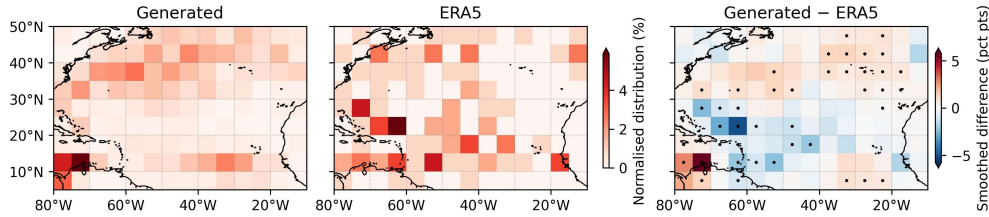


Fig. 5: Comparison of the normalized spatial distribution of TC-like features in the generated samples and ERA5 (500 distinct timestamps, 20 samples each). $5^\circ \times 5^\circ$ bins are used. Stippling marks bins where the pointwise 95% bootstrap confidence interval excludes zero (using 1,000 resamples). Gaussian smoothing with a 0.5-grid-cell standard deviation is applied to the difference to improve clarity.

To further explore the structure of TC-like features in the generated sample, we compare composites of all TC-like features identified by the OWZ detection and thickness criterion in the generated sample and ERA5 in figure 7. The mid-level temperature and MSLP anomaly are calculated relative to the surroundings, showing the depth of the low-pressure systems and the warm core that is indicative of a tropical rather than extratropical cyclone. The central pressure, relative vorticity, and the maximum tangential wind are underestimated, whilst the radius of maximum winds is slightly larger than in ERA5. Owing to the extratropical bias noted in figure 5, the maximum wind orientation is too equatorward compared to ERA5, where the maximum wind occurs in the upper right quadrant as observed in TCs.

3 Discussion

We used a diffusion model to downscale monthly seasonal forecasts of the North Atlantic basin from ECMWF’s SEAS5 model into six-hourly instantaneous snapshots. The model restores much of the distribution and spatial variability expected in daily weather that is not captured by the monthly-mean seasonal forecast baseline, in a physically consistent manner. After identifying TC-like features, we found that the generated samples recover much of the frequency and monthly variation of TC activity present in corresponding ERA5 snapshots. However, the annual peak in monthly candidate counts is underestimated on average, though within the range of the generated ensemble, and the ensembles remain slightly under-dispersed, tending to produce similar samples. Importantly, our model was designed to generate an ensemble of instantaneous weather samples consistent with the forecast seasonal conditions that is representative of the distribution of potential outcomes rather than to reproduce the ERA5 observations on a particular future date; some discrepancy with the ERA5 samples – which represent the particular outcome observed on that date – is therefore expected. The demonstrated distributional and spectral modification of the baseline SEAS5 forecast shows success towards that aim.

In order to detect the small-scale vortical features associated with TCs, it was essential that our model produced physically coherent samples. Close agreement in

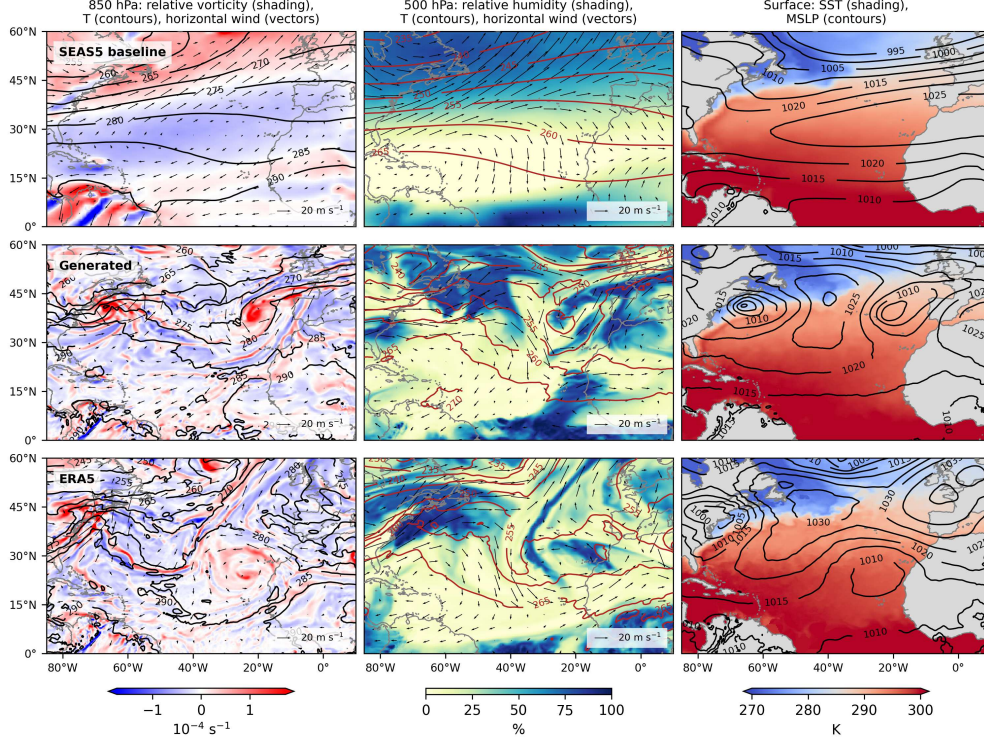


Fig. 6: An example generated sample, with the interpolated SEAS5 forecast reference and ERA5 data for that date. This sample corresponds to the forecast for 6:00 UTC on 26 January, 2023. To be clear, the sample is not intended to replicate ERA5 weather, but rather the aggregate of samples should be representative of the ERA5 weather.

spatial spectra and distributions in individual variables is not sufficient to show that the weather samples are physically coherent, because a set of variables can each be individually realistic while being mutually inconsistent. Our model directly generates the relative vorticity as well as horizontal wind, which was verified to be consistent with the vertical component of the curl of the horizontal wind. Nothing in the loss function or the architecture enforced this consistency; the close agreement (Supplementary Figure S2) shows that the diffusion model learned this relationship. This result is encouraging, as a common critique of data driven machine learning weather prediction models is their lack of physical consistency [20].

One issue in the generated samples is calibration. The ensemble is slightly under-dispersed, as seen through the spread-skill plots and rank histogram. This under-dispersion matters for TC risk assessment because an over-confident ensemble may understate uncertainty in the weather conditions that influence storm formation and intensity. This problem is also seen in related work. CorrDiff, a similar diffusion-based downscaling model, also reports under-dispersion across most variables [15]. This is unexpected, as diffusion models for image generation are often associated with too much sample diversity and producing hallucinations, rather than too little, and

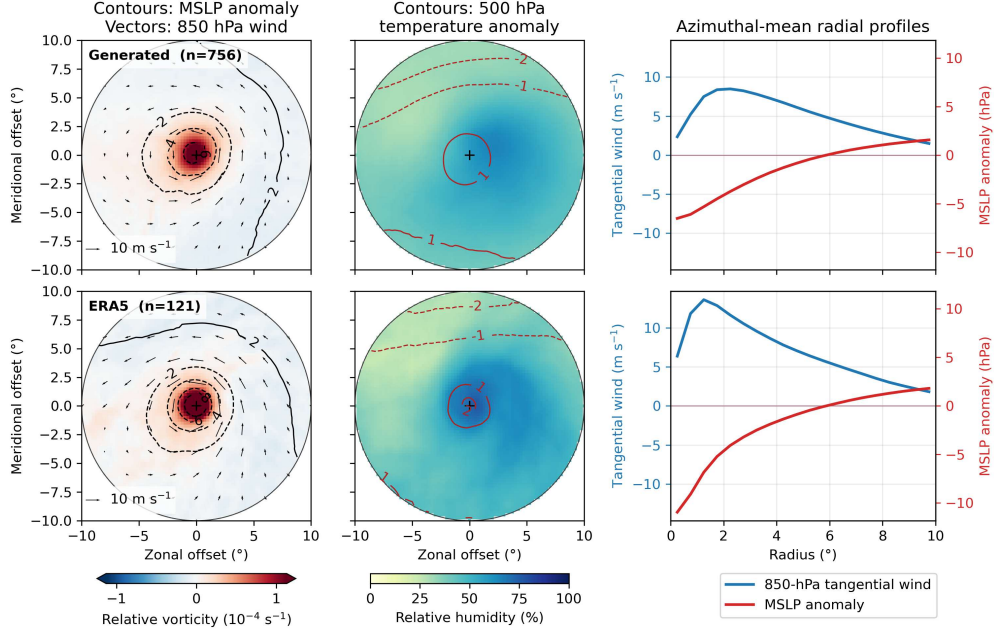


Fig. 7: Comparison of the structure of TC-like features in generated samples and ERA5. Anomalies are calculated relative to an average of the 10° region centered on the feature.

variance reducing techniques are often used during sampling [21, 22]. Several factors may contribute to the underdispersion in our model. Firstly, the cosine noise schedule was chosen as it has been shown to lead to improved sample quality, but may have implications for dispersion. The diversity in the samples produced by a diffusion model is heavily dependent on the denoising of the highest noise level, which the cosine schedule spends less time on. Secondly, a DDIM sampling process was chosen in order to produce samples more efficiently. The DDIM process is governed by a parameter η which, when 0, solves a deterministic ordinary differential equation, and when 1, solves the stochastic chain of a standard DDPM. By setting $\eta = 0.5$, samples were generated more efficiently, but at the cost of losing stochasticity from the DDPM process.

The frequency analysis shows the success of the model in recovering much of the frequency and seasonality of TC-like features. Across the North Atlantic basin, generated counts reach 94.8% of ERA5 counts on the evaluated dates and have correlation of 0.969 with the count of candidates detected on the ERA5 data, as shown in Figure 4. The identified TC-like features show strong similarity to ERA5, including the warm core and wind/humidity structure, though the central pressure deficit and maximum wind strength are underestimated. Nonetheless, the model underproduces samples with the thickness criteria, and the spatial distribution shows some bias towards the extratropics.

There are several limitations to this work. First, each weather snapshot is generated independently. TC-like features are not required to persist, so the samples cannot establish tracks, lifetimes, genesis rates, intensification, or distinct seasonal

storm counts. Second, the model is statistical and has no physical laws or equations enforced in its loss function or on its output. Therefore, it is possible that this model may output physically impossible results, particularly for conditioning that lies outside the training distribution. Third, ERA5 data is used as ground truth, but it is a reanalysis, thus it is a model based on the best estimate of the atmospheric state. Any smoothing or bias in ERA5, particularly in data-sparse regions, is inherited in the model, and so agreement with ERA5 does not establish agreement with actual TC intensities or structures.

To further or continue this work, this model can also be used to analyze the impacts of climate projections. Instead of conditioning on SEAS5, the model could be conditioned on CMIP6 scenarios [23]. By applying this diffusion model to different Shared Socioeconomic Pathway warming scenarios, we could generate samples of synthetic weather to allow analysis of how the intensity and frequency of TCs might shift under different warming scenarios. Another avenue to explore is the generation of temporally coherent sequences of samples. The model could be extended through auto-regressive conditioning on a sample from the prior time period or through a video diffusion model, and thus generate chains of samples rather than a single sample.

This work establishes a first step towards moving from seasonal forecasts to downscaled weather fields and, ultimately, to hurricane prediction. This paper shows diffusion based downscaling could offer a path from long range seasonal forecasts to the plausible and statistically accurate fields required for tropical cyclone risk assessment.

4 Methods

4.1 Problem Formulation

Given monthly forecasts y , the goal is to generate samples of physically plausible weather x at six-hourly timestamps. We take a generative approach, learning the full conditional distribution and sampling from it, as the relationship between a monthly mean and the weather it implies is one-to-many. However, rather than generating the full fields directly, the model generates the residual between the weather and an interpolated forecast. Let $b_m(y)$ be this interpolated baseline for SEAS5 member m . The quantity to sample is then

$$r_m \sim p_\theta(r \mid b_m(y), \ell), \quad x^{(m)} = b_m(y) + r_m,$$

where ℓ is the land/sea mask. During training, member m is drawn uniformly from the 25 SEAS5 members. Its baseline is subtracted from an ERA5 field, whose timestamp may be displaced by a small temporal offset. At generation time, each of the twenty samples uses a different SEAS5 member. The interpolated mean of all 25 members, $\bar{b}(y)$, is kept as the deterministic baseline for evaluation. Figure 8a shows the residual construction.

Modeling the residual rather than the raw fields is an important design decision of the model, inspired by the two stage process used by CorrDiff [15] where it was shown that modeling the residual reduces the variance of the distribution that must be learned by the generating system. When the baseline $b_m(y)$ captures the large scale

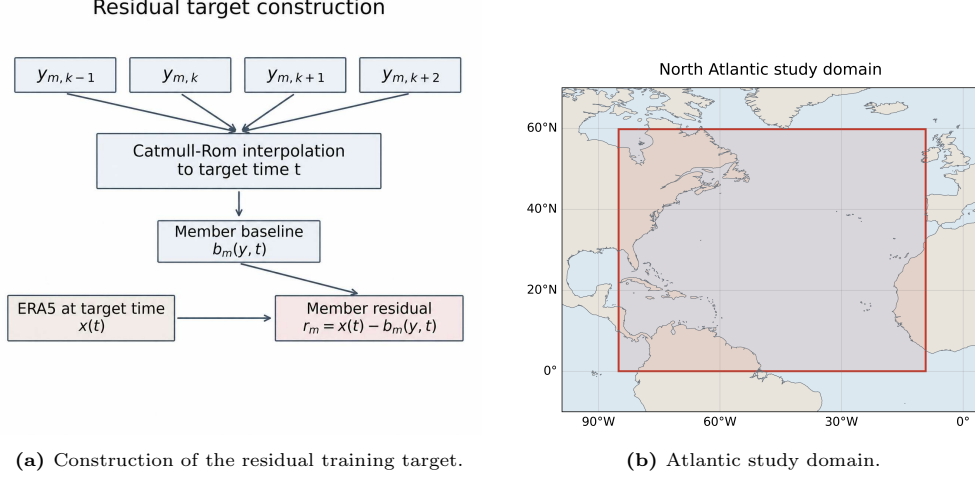


Fig. 8: Residual construction and study domain. (a) Four monthly forecasts for the same SEAS5 member index are interpolated to form the baseline. Subtracting it from ERA5 gives the residual used for training. (b) The red box spans 0°–59.75°N and 85°–9.25°W.

weather patterns reasonably well, the residual can have a smaller variance than x , so the model can spend its effort on resolving fine scale structures rather than re-learning the large-scale patterns like ENSO or the seasonal cycle that the forecast provides. Figure 8 shows how the training target is constructed and the Atlantic domain used.

4.2 Data and Pre-processing

Two datasets are used. Monthly mean forecasts y are taken from SEAS5, the ECMWF seasonal forecasting system [11]. ERA5, the ECMWF reanalysis system, provides the reference weather fields x [24]. ERA5 data up to 2021 is obtained from a corpus hosted in Google Cloud [25], and the remaining data from ECMWF.

The Atlantic domain covers 0°–59.75°N and 85°–9.25°W (Figure 8b). We chose this domain to contain the Main Development Region for hurricanes in the Atlantic [26] as well as their eventual track, which can go as high as 55°N [27]. SEAS5 is bilinearly interpolated from 1.0° to the 0.25° ERA5 grid, giving $H \times W = 240 \times 304$ cells. Training uses 54,852 six-hourly timestamps from 16 June 1981 to 31 December 2018. Validation uses 2019–2021 and testing uses 2022–2024, with 4,384 timestamps in each split.

The model uses twelve variables. SST and MSLP are single-level variables and the remaining ten variables are atmospheric fields at both 500 hPa and 850 hPa: relative vorticity, relative humidity, zonal wind u , meridional wind v , and temperature. We chose these variables to support the work of downstream hurricane detection models.

Relative humidity and relative vorticity use the same definitions for ERA5 and each SEAS5 member. Vorticity is calculated using spherical finite differences,

$$\zeta = \frac{1}{a \cos \phi} \left(\frac{\partial v}{\partial \lambda} - \frac{\partial(u \cos \phi)}{\partial \phi} \right),$$

where a is the Earth’s radius and ϕ and λ are latitude and longitude. Relative humidity uses vapor pressure $e = qp/(0.622 + 0.378q)$ and saturation over water, where q is specific humidity and p is pressure.

The monthly means are placed at the centres of their valid months. The two bracketing months and one additional month on each side are fitted with a Catmull–Rom spline, a cubic interpolation that passes through its control points [28]. The same SEAS5 member index is used at all four nodes, giving $b_m(y)$.

Data was normalized before training by subtracting the mean and dividing by the standard deviation, computed per channel over the training period (these statistics are kept to transform the output back to physical units). Baseline statistics use the ensemble mean and are shared by members; residual statistics pool all 25 members without temporal smearing. In the diffusion equations, x_0 is the normalized residual:

$$x_0 = \frac{x - b_m(y) - \mu_r}{\sigma_r}, \quad \tilde{b}_m(y) = \frac{b_m(y) - \mu_b}{\sigma_b}.$$

4.3 Diffusion Models

A diffusion model generates samples through two steps, a forward process that incrementally adds noise to data, and a reverse process that incrementally removes noise to create a sample [29, 30]. The normalized residual is noised as

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon,$$

with schedule $\bar{\alpha}_t$ and noise ϵ defined below.

4.4 Adapting the Diffusion Model

4.4.1 Conditioning on the Forecast and Land/Sea Mask

We condition the reverse process on the normalized member baseline $\tilde{b}_m(y)$ and the land/sea mask ℓ , which is 1 over the ocean and 0 over land. The noisy residual, baseline, and mask are concatenated along the channel dimension. With $C = 12$ variables, the network $v_\theta(x_t, t, \tilde{b}_m(y), \ell)$ takes $2C + 1$ channels and outputs C channels of predicted velocity, defined in Section 4.4.3.

4.4.2 Noise Schedule

We use the cosine schedule [31],

$$\bar{\alpha}_t = \frac{f(t)}{f(0)}, \quad f(t) = \cos^2 \left(\frac{t/T + s}{1 + s} \frac{\pi}{2} \right),$$

with $s = 0.008$ and $T = 1000$ steps. The discrete noise variance is capped at 0.999.

4.4.3 v-Prediction

A limitation of predicting the noise ϵ arises at the end of this schedule, where the signal-to-noise ratio approaches zero. Recovering the clean field from a predicted ϵ requires division by $\sqrt{\bar{\alpha}_t}$, so small prediction errors are amplified at the highest noise levels [32, 33]. Instead, the network is trained to predict the velocity [32]:

$$v_t = \sqrt{\bar{\alpha}_t} \epsilon - \sqrt{1 - \bar{\alpha}_t} x_0,$$

from which the clean residual is recovered without that division:

$$\hat{x}_0 = \sqrt{\bar{\alpha}_t} x_t - \sqrt{1 - \bar{\alpha}_t} v_\theta(x_t, t, \tilde{b}_m(y), \ell). \quad (1)$$

Here, velocity denotes a parameterization of the diffusion process, not a physical wind velocity or a time derivative of the weather.

4.4.4 Offset Noise

Standard diffusion models struggle to reproduce the overall mean of a field: because the per-pixel noise is independent, the forward process adds little noise to the field mean, so generated residuals trend towards a zero mean. Offset noise [34, 35] adds to the per-pixel noise a randomized offset that is constant across all pixels:

$$\epsilon = \epsilon_{\text{pix}} + \omega_o \xi, \quad \epsilon_{\text{pix}} \sim \mathcal{N}(0, \mathbf{I}), \quad \xi \sim \mathcal{N}(0, 1).$$

Here, ξ is drawn independently per sample and per channel, and $\omega_o = 0.02$. This modification is useful for downscaling climate forecasts, as the residual can often carry a non-zero mean that the model must be able to reproduce.

4.4.5 Temporal Smearing

To expose the model to short-term weather variability under the same seasonal forecast, we allow a small displacement in the ERA5 timestamp during training. The member baseline at timestamp i is paired with the weather at $i + d$, where

$$d = \text{clip}(\text{round}(\delta), -3, 3), \quad \delta \sim \mathcal{N}(0, 1.5^2),$$

in units of six-hour steps. The maximum displacement is therefore 18 hours in either direction. The baseline remains fixed at i , so the residual target becomes $x_{i+d} - b_{m,i}(y)$.

4.4.6 Loss Function

We modify the standard squared-error objective in three ways. We use an L_1 penalty, which is more forgiving of the large errors that come with predicting extreme events; we add a term that matches the distribution of the generated residual to that of the

true residual, as we want to capture the long tails of the data in order to represent rare events such as hurricanes; and we weight the velocity error according to the noise level.

For each channel c , normalized soft histograms p_c and $\hat{p}_c$ are formed. The distribution penalty is the mean Kullback–Leibler divergence (KLD) across channels:

$$\mathcal{L}_{\text{dist}} = \frac{1}{C} \sum_{c=1}^C \sum_{k=1}^{96} p_c(k) \log \frac{p_c(k)}{\hat{p}_c(k)}.$$

The total loss is the weighted sum of the KLD loss and the L_1 loss:

$$\mathcal{L} = \lambda_\epsilon \mathcal{L}_1 + \lambda_{\text{dist}} \mathcal{L}_{\text{dist}}, \quad \lambda_\epsilon = 0.9, \quad \lambda_{\text{dist}} = 0.1. \quad (2)$$

Min-SNR weighting [36] clamps each timestep’s contribution so that low-noise timesteps, which produce large gradients, do not dominate training. Defining the signal-to-noise ratio at timestep t as

$$\text{SNR}(t) = \frac{\bar{\alpha}_t}{1 - \bar{\alpha}_t}, \quad (3)$$

they propose weighting each timestep’s contribution using the SNR and a truncation constant γ . The loss at timestep t is multiplied by

$$w_t = \frac{\min(\text{SNR}(t), \gamma)}{\text{SNR}(t)} = \min\left(1, \frac{\gamma}{\text{SNR}(t)}\right). \quad (4)$$

We use $\gamma = 5$, as recommended by Hang et al.

4.5 Architecture and Training

The velocity-prediction network is a U-Net with five levels, channel widths of 256, 256, 320, 320, and 384, and four residual blocks per level. Self attention is used at the three deepest levels. The network has approximately 160 million parameters.

The model is trained with AdamW for 60 epochs on four RTX 5090 GPUs, with a batch of 8 per GPU and 32 overall. Training uses bfloat16 precision and gradient clipping at norm 1.0. The learning rate increases linearly to 2×10^{-4} over 500 optimization steps, then follows a cosine decay to 10^{-6} . Weight decay is 10^{-2} . An exponential moving average of the parameters uses decay 0.9999 per step, and the final averaged parameters are used for sampling. Training diffusion timesteps are drawn uniformly.

4.6 Sampling

Once trained, the model can generate samples by reversing the noising process. Running the network at every one of the T timesteps is expensive, so instead we use denoising diffusion implicit model (DDIM) sampling [37], which produces high quality samples in fewer steps. We use 500 timesteps and $\eta = 0.5$, a balance between a

deterministic process ($\eta = 0$) and the standard stochastic reverse process ($\eta = 1$). The sampler reconstructs $\hat{x}_0$ using Eq. 1 and obtains the implied noise from the unmodified velocity output:

$$\hat{\epsilon} = \sqrt{\bar{\alpha}_t} v_\theta + \sqrt{1 - \bar{\alpha}_t} x_t.$$

Both initial noise and stochastic updates use the training offset-noise. The residual is transformed back to physical units and added to its member baseline. Each weather snapshot is generated separately.

4.7 Evaluation

The model is generative, so its objective is to match the distribution of the ERA5 observations. To evaluate this, we look at how well the generated samples reproduce the distribution, spatial structure, and probabilistic structure of the test set.

We sampled 500 dates from 2022–2024 and produced $M = 20$ samples per date, giving 10,000 samples. Each uses one of the first twenty SEAS5 members and is compared with ERA5 on the same date. The interpolated ensemble mean $\bar{b}(y)$ is the baseline.

4.7.1 Qualitative Evaluation

Generated samples were inspected visually for physical plausibility. They were checked for coherent structures, realistic gradients, and consistency across variables.

4.7.2 Distributions

The distribution across all grid points and samples is used to obtain $\hat{P}_c$ for variable c , while P_c is obtained from all representative ERA5 data points. The Wasserstein distance is

$$W(P_c, \hat{P}_c) = \int_{\mathbb{R}} |F_c(z) - \hat{F}_c(z)| dz, \quad (5)$$

where F_c and $\hat{F}_c$ are the respective cumulative distribution functions. To compare variables with different units, the Wasserstein distance is normalized by the standard deviation σ_c of the observational data:

$$\widetilde{W}(P_c, \hat{P}_c) = \frac{W(P_c, \hat{P}_c)}{\sigma_c}. \quad (6)$$

4.7.3 Spatial Structure

Power spectral density

To check variance at different scales, we calculate the power spectrum of each complete rectangular field without area weights. The spatial mean is removed, and a two-dimensional Hanning window is applied. Comparing the spectra shows which scales have too much or too little variance; we quantify the agreement through log-spectral distance:

$$\text{LSD} = \sqrt{\frac{1}{|\mathcal{B}|} \sum_{k \in \mathcal{B}} \left[10 \log_{10} \frac{E_{\text{generated}}(k)}{E_{\text{observations}}(k)} \right]^2}.$$

We evaluate this over three bands as defined by Orlanski [19]: Macro- β and larger ($\lambda \geq 2000$ km), Meso- α (200–2000 km), and Meso- β (20–200 km).

Mean Latitudinal and Longitudinal Profiles

To detect regional bias, we compare mean latitudinal and longitudinal profiles over ocean cells. Averages over latitude use cosine-latitude weights. Reported RMSEs compare time-mean profiles, with latitude weighting for latitudinal RMSE. Paired tests instead use each date’s profile RMSE.

Wind–vorticity consistency

We derive vorticity from the generated winds as the vertical component of the curl of the horizontal wind vector. We regress generated vorticity against wind-derived vorticity and report the slope and R^2 . The RMS difference is divided by the RMS generated vorticity.

4.7.4 Probabilistic Structure

Spread-Skill Ratio

A calibrated ensemble should, on average, be as uncertain as it is inaccurate [38]. In other words, the spread of the samples about their mean should be the same as the typical error of that mean. For member variance s_i^2 with denominator $M - 1$, and weighted averaging $\langle \cdot \rangle_w$ over dates and ocean cells,

$$S = \sqrt{\langle s_i^2 \rangle_w}, \quad \text{RMSE} = \sqrt{\langle (\bar{x}_i - x_i)^2 \rangle_w}, \quad \text{SSRAT} = \sqrt{\frac{M+1}{M}} \frac{S}{\text{RMSE}}.$$

We use the finite-ensemble correction [38], so a ratio below one indicates an over-confident ensemble, and a ratio above one an over-dispersive one. Points are sorted by spread into 15 bins; spread-skill reliability [39] is

$$\text{SSREL} = \sum_k \frac{W_k}{W} \left| \sqrt{\frac{M+1}{M}} S_k - \text{RMSE}_k \right|,$$

where S_k is binwise RMS spread and W_k/W is its fraction of total weight.

Rank Histogram

A well calibrated ensemble gives a flat histogram, since the true value is equally likely to fall in any interval between the samples, whereas a U shape signifies under-dispersion and a dome shape over-dispersion. For the resulting 21 frequencies f_k , reliability is summarized by $\text{RI} = \sum_{k=1}^{21} |f_k - 1/21|$. Lower and upper endpoint ratios are $21f_1$ and $21f_{21}$; their asymmetry also identifies bias.

Continuous Ranked Probability Score

The continuous ranked probability score (CRPS) generalizes the mean absolute error to probabilistic forecasts. We use its finite-ensemble fair form [40, 41]:

$$\widehat{\text{CRPS}}_i = \frac{1}{M} \sum_m |x_i^{(m)} - x_i| - \frac{1}{2M(M-1)} \sum_{m,n} |x_i^{(m)} - x_i^{(n)}|.$$

Scores are normalized by each variable’s test-set ERA5 standard deviation. The fair form reduces finite-ensemble effects, although its assumptions may not hold exactly, since each generated member has different forecast conditioning.

Threshold-Weighted CRPS

Threshold-weighted CRPS [42] focuses on the upper tail using $w(z) = \mathbb{1}[z \geq q_c]$. It is computed by applying the same fair score to transformed members and observations:

$$x_i^{(m)} \mapsto \max(x_i^{(m)}, q_c), \quad x_i \mapsto \max(x_i, q_c). \quad (7)$$

The threshold is the test-set ERA5 90th percentile, held fixed across dates. This retrospective diagnostic uses the same area weighting and normalization as CRPS.

4.8 Tropical Cyclone Occurrence

While the weather metrics measure how well the samples reproduce observed weather patterns, the goal is to support the assessment of hurricane activity from long term forecasts. We apply the Okubo–Weiss–Zeta (OWZ) algorithm separately to generated samples and ERA5. OWZ identifies rotation-dominated flow with moisture and shear criteria [13, 14]. We use a downloaded OWZ package for this task. The OWZ detection algorithm alone identifies many TC-like features that arise from random fluctuations, which are typically filtered out by an additional ‘stitching’ step that identifies temporally coherent features. The samples we generate are time-independent so this is not possible. Instead, we apply an additional criterion based on an approximation of geopotential thickness that reduces the number of identified TC-like features by around 88%.

The OWZ code requires pressure levels that the model does not generate. Using the available 850 and 500 hPa fields, we linearly interpolate winds, temperature and relative humidity in log pressure to 700 hPa and extrapolate them to 950 and 200 hPa. Specific humidity at 950 hPa is calculated from the estimated temperature and relative humidity using the package’s formula. To detect a candidate, all six initial thresholds must be satisfied at the same grid point (Table 2).

After identifying OWZ candidates, we identify TC-like features by approximating the geopotential thickness from 850–500 hPa as

$$Z_{500} - Z_{850} \approx \frac{R_d}{2g} (T_{850} + T_{500}) \ln \left(\frac{850}{500} \right) \quad (8)$$

Table 2: Initial OWZ thresholds. All six criteria must be satisfied at the same point on the 1° detector grid. Missing pressure levels are reconstructed as described in the text. The package’s initial thresholds and sanity bounds are retained.

Field	Threshold
OWZ at 850 hPa	$\geq 5 \times 10^{-5} \text{ s}^{-1}$
OWZ at 500 hPa	$\geq 4 \times 10^{-5} \text{ s}^{-1}$
Relative humidity at 950 hPa	$\geq 70\%$
Relative humidity at 700 hPa	$\geq 50\%$
Specific humidity at 950 hPa	$\geq 10 \text{ g kg}^{-1}$
Smoothed 850–200 hPa wind shear	$\leq 25 \text{ m s}^{-1}$

where the subscripts indicate the pressure level in hPa. We then apply a closed-contour test, requiring that the closed contour 10 m lower than the maximum geopotential thickness within 1° of each OWZ candidate center lies within 6.5° of the center. The parameters were chosen based on the maximum F1 score calculated from the retained candidates compared to identified TCs (including precursor lows and disturbances) in ERA5 in the validation set. This criterion is similar to the UZ algorithm used by TempestExtremes, another common TC tracking algorithm, except with 300–500 hPa thickness.

Data availability

ERA5 hourly reanalysis data on pressure levels and single levels are available from the Copernicus Climate Change Service (C3S) Climate Data Store [43, 44]. The ERA5 data used through 2021 were accessed from the ARCO-ERA5 archive on Google Cloud, documented at <https://github.com/google-research/arco-era5> [25]. ECMWF SEAS5 monthly forecasts are available from the C3S Climate Data Store as seasonal forecast monthly statistics on pressure levels and single levels [45, 46]. Tropical cyclone track data are available from the International Best Track Archive for Climate Stewardship (IBTrACS), [47, 48]; the North Atlantic subset was used in this study.

Code availability

Code for data acquisition, preprocessing, diffusion model training and sampling is publicly available at <https://github.com/benwulmer/atlantic-diffusion> under the MIT licence [49]. The Python implementation of the Okubo–Weiss–Zeta (OWZ) algorithm used to identify tropical cyclone candidates is available at <https://github.com/savinchand/owz-python>.

Acknowledgments

This project was supported by the Institute of Computing for Climate Science at the University of Cambridge. Funding and support were provided by Schmidt Sciences, LLC., and by INIGO Ltd.

Author Information

Author Contributions

Conceptualisation, B.U. and R.E.R.; methodology, B.U., C.P., and R.E.R.; investigation, B.U. and C.P.; writing—original draft, B.U.; writing—review & editing, B.U., C.P., and R.E.R.; resources, B.U.; supervision, R.E.R.

Corresponding Author

Correspondence to Ben Ulmer.

Ethics Declarations

Competing Interests

The authors declare no competing interests.

References

- [1] NOAA Office for Coastal Management: Fast Facts: Hurricane Costs (2026). <https://coast.noaa.gov/states/fast-facts/hurricane-costs.html>
- [2] King, R.O.: Financing Natural Catastrophe Exposure: Issues and Options for Improving Risk Transfer Markets. Technical report, Congressional Research Service, Library of Congress (2013)
- [3] Doocy, S., Dick, A., Daniels, A., Kirsch, T.D.: The Human Impact of Tropical Cyclones: A Historical Review of Events 1980–2009 and Systematic Literature Review. *PLoS Currents* **5** (2013) <https://doi.org/10.1371/currents.dis.2664354a5571512063ed29d25ffbce74>
- [4] Young, R., Hsiang, S.: Mortality Caused by Tropical Cyclones in the United States. *Nature* **635**, 121–128 (2024) <https://doi.org/10.1038/s41586-024-07945-5>
- [5] Climate Central: Study: Ocean warming has intensified recent hurricanes (2024). <https://www.climatecentral.org/climate-matters/hurricane-strength-attribution>
- [6] Gilford, D.M., Giguere, J., Pershing, A.J.: Human-Caused Ocean Warming Has Intensified Recent Hurricanes. *Environmental Research: Climate* **3**(4), 045019 (2024) <https://doi.org/10.1088/2752-5295/ad8d02>

- [7] Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T.R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., Willson, M.: Probabilistic weather forecasting with machine learning. *Nature* **637**(8044), 84–90 (2025) <https://doi.org/10.1038/s41586-024-08252-9>
- [8] Schmude, J., Roy, S., Trojak, W., Jakubik, J., Civitarese, D.S., Singh, S., Kuehnert, J., Ankur, K., Gupta, A., Phillips, C.E., Kienzler, R., Szwarcman, D., Gaur, V., Shinde, R., Lal, R., Da Silva, A., Guevara Diaz, J.L., Jones, A., Pfreundschuh, S., Lin, A., Sheshadri, A., Nair, U., Anantharaj, V., Hamann, H., Watson, C., Maskey, M., Lee, T.J., Bernabe Moreno, J., Ramachandran, R.: Prithvi WxC: Foundation Model for Weather and Climate (2024). <https://doi.org/10.48550/arXiv.2409.13598> . <https://arxiv.org/abs/2409.13598>
- [9] Rogers, R.F.: Recent advances in our understanding of tropical cyclone intensity change processes from airborne observations. *Atmosphere* **12**(5) (2021) <https://doi.org/10.3390/atmos12050650>
- [10] Judt, F., Chen, S.S., Berner, J.: Predictability of tropical cyclone intensity: scale-dependent forecast error growth in high-resolution stochastic kinetic-energy backscatter ensembles. *Quarterly Journal of the Royal Meteorological Society* **142**(694), 43–57 (2016) <https://doi.org/10.1002/qj.2626>
- [11] Johnson, S.J., Stockdale, T.N., Ferranti, L., Balmaseda, M.A., Molteni, F., Magnusson, L., Tietsche, S., Decremier, D., Weisheimer, A., Balsamo, G., Keeley, S.P.E., Mogensen, K., Zuo, H., Monge-Sanz, B.M.: SEAS5: The New ECMWF Seasonal Forecast System. *Geoscientific Model Development* **12**, 1087–1117 (2019) <https://doi.org/10.5194/gmd-12-1087-2019>
- [12] Lin, J., Rousseau-Rizzi, R., Lee, C.-Y., Sobel, A.: An Open-Source, Physics-Based, Tropical Cyclone Downscaling Model with Intensity-Dependent Steering (2023). <https://arxiv.org/abs/2302.09455>
- [13] Tory, K.J., Chand, S.S., Dare, R.A., McBride, J.L.: The development and assessment of a model-, grid-, and basin-independent tropical cyclone detection scheme. *Journal of Climate* **26**(15), 5493–5507 (2013) <https://doi.org/10.1175/JCLI-D-12-00510.1>
- [14] Bourdin, S., Fromang, S., Dulac, W., Cattiaux, J., Chauvin, F.: Intercomparison of Four Algorithms for Detecting Tropical Cyclones Using ERA5. *Geoscientific Model Development* **15**, 6759–6786 (2022) <https://doi.org/10.5194/gmd-15-6759-2022>
- [15] Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Nabian, M.A., Ge, T., Subramaniam, A., Kashinath, K., Kautz, J., Pritchard, M.: Residual Corrective Diffusion Modeling for km-Scale Atmospheric Downscaling. *Communications Earth & Environment* **6**, 124 (2025) <https://doi.org/10.1038/s43247-025-02042-5>

- [16] Aich, M., Bathiany, S., Hess, P., Huang, Y., Boers, N.: Diffusion Models for Probabilistic Precipitation Generation from Atmospheric Variables. arXiv preprint arXiv:2504.00307 (2025) <https://doi.org/10.48550/arXiv.2504.00307>
- [17] Schmidt, J., Schmidt, L., Strnad, F.M., Ludwig, N., Hennig, P.: A generative framework for probabilistic, spatiotemporally coherent downscaling of climate simulation. *npj Climate and Atmospheric Science* **8**, 270 (2025) <https://doi.org/10.1038/s41612-025-01157-y>
- [18] Palma, L., Peraza, A., Civantos-Prieto, D., Duarte, A., Materia, S., Muñoz, Á.G., Peña-Izquierdo, J., Romero, L., Soret, A., Donat, M.G.: Data-driven seasonal climate predictions via variational inference and transformers. *npj Climate and Atmospheric Science* **9**, 48 (2026) <https://doi.org/10.1038/s41612-026-01320-z>
- [19] Orlanski, I.: A Rational Subdivision of Scales for Atmospheric Processes. *Bulletin of the American Meteorological Society* **56**(5), 527–530 (1975)
- [20] Bonavita, M.: On Some Limitations of Current Machine Learning Weather Prediction Models. *Geophysical Research Letters* **51**(12), 2023–107377 (2024) <https://doi.org/10.1029/2023GL107377>
- [21] Ho, J., Salimans, T.: Classifier-Free Diffusion Guidance (2022). <https://arxiv.org/abs/2207.12598>
- [22] Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., Laine, S.: Guiding a Diffusion Model with a Bad Version of Itself (2024). <https://arxiv.org/abs/2406.02507>
- [23] Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J., Taylor, K.E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) Experimental Design and Organization. *Geoscientific Model Development* **9**(5), 1937–1958 (2016) <https://doi.org/10.5194/gmd-9-1937-2016>
- [24] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.-N.: The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* **146**(730), 1999–2049 (2020) <https://doi.org/10.1002/qj.3803> <https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803>
- [25] Carver, R.W., Merose, A.: ARCO-ERA5: An Analysis-Ready Cloud-Optimized Reanalysis Dataset. In: 22nd Conference on Artificial Intelligence for Environmental Science, pp. 4–1. American Meteorological Society, Denver, Colorado (2023).

- <https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/415842>
- [26] Goldenberg, S.B., Landsea, C.W., Mestas-Nuñez, A.M., Gray, W.M.: The Recent Increase in Atlantic Hurricane Activity: Causes and Implications. *Science* **293**(5529), 474–479 (2001) <https://doi.org/10.1126/science.1060040>
 - [27] Hart, R.E., Evans, J.L.: A Climatology of the Extratropical Transition of Atlantic Tropical Cyclones. *Journal of Climate* **14**(4), 546–564 (2001) [https://doi.org/10.1175/1520-0442\(2001\)014<0546:ACOTET>2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)014<0546:ACOTET>2.0.CO;2)
 - [28] Catmull, E., Rom, R.: A Class of Local Interpolating Splines. In: Barnhill, R.E., Riesenfeld, R.F. (eds.) *Computer Aided Geometric Design*, pp. 317–326. Academic Press, New York (1974)
 - [29] Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. PMLR, vol. 37, pp. 2256–2265 (2015). <https://proceedings.mlr.press/v37/sohl-dickstein15.html>
 - [30] Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851 (2020). <https://arxiv.org/abs/2006.11239>
 - [31] Nichol, A., Dhariwal, P.: Improved Denoising Diffusion Probabilistic Models (2021). <https://arxiv.org/abs/2102.09672>
 - [32] Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: *International Conference on Learning Representations* (2022). <https://arxiv.org/abs/2202.00512>
 - [33] Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5404–5411 (2024). https://openaccess.thecvf.com/content/WACV2024/html/Lin_Common_Diffusion_Noise_Schedules_and_Sample_Steps_Are_Flawed_WACV_2024_paper.html
 - [34] Guttenberg, N.: Diffusion with Offset Noise (2023). <https://www.crosslabs.org/blog/diffusion-with-offset-noise> Accessed 2026-06-30
 - [35] Kutsuna, T.: A Probabilistic Formulation of Offset Noise in Diffusion Models (2026). <https://arxiv.org/abs/2412.03134>
 - [36] Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., Guo, B.: Efficient Diffusion Training via Min-SNR Weighting Strategy (2024). <https://arxiv.org/abs/2303.09556>
 - [37] Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models (2022). <https://arxiv.org/abs/2010.02502>

[//arxiv.org/abs/2010.02502](https://arxiv.org/abs/2010.02502)

- [38] Fortin, V., Abaza, M., Anctil, F., Turcotte, R.: Why Should Ensemble Spread Match the RMSE of the Ensemble Mean? *Journal of Hydrometeorology* **15**(4), 1708–1713 (2014) <https://doi.org/10.1175/JHM-D-14-0008.1>
- [39] Haynes, K., Lagerquist, R., McGraw, M., Musgrave, K., Ebert-Uphoff, I.: Creating and Evaluating Uncertainty Estimates with Neural Networks for Environmental-Science Applications. *Artificial Intelligence for the Earth Systems* **2**(2), 220061 (2023) <https://doi.org/10.1175/AIES-D-22-0061.1>
- [40] Gneiting, T., Raftery, A.E.: Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102**(477), 359–378 (2007) <https://doi.org/10.1198/016214506000001437>
- [41] Zamo, M., Naveau, P.: Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts. *Mathematical Geosciences* **50**, 209–234 (2018) <https://doi.org/10.1007/s11004-017-9709-7>
- [42] Gneiting, T., Ranjan, R.: Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics* **29**(3), 411–422 (2011) <https://doi.org/10.1198/jbes.2010.08110>
- [43] Copernicus Climate Change Service: ERA5 hourly data on pressure levels from 1940 to present. Copernicus Climate Change Service Climate Data Store (2018). <https://doi.org/10.24381/cds.bd0915c6> . <https://doi.org/10.24381/cds.bd0915c6>
- [44] Copernicus Climate Change Service: ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service Climate Data Store (2018). <https://doi.org/10.24381/cds.adbb2d47> . <https://doi.org/10.24381/cds.adbb2d47>
- [45] Copernicus Climate Change Service: Seasonal forecast monthly statistics on pressure levels. ECMWF, Copernicus Climate Change Service Climate Data Store (2018). <https://doi.org/10.24381/cds.0b79e7c5> . <https://doi.org/10.24381/cds.0b79e7c5>
- [46] Copernicus Climate Change Service: Seasonal forecast monthly statistics on single levels. ECMWF, Copernicus Climate Change Service Climate Data Store (2018). <https://doi.org/10.24381/cds.68dd14c3> . <https://doi.org/10.24381/cds.68dd14c3>
- [47] Knapp, K.R., Kruk, M.C., Levinson, D.H., Diamond, H.J., Neumann, C.J.: The international best track archive for climate stewardship (IBTrACS): Unifying tropical cyclone best track data. *Bulletin of the American Meteorological Society* **91**, 363–376 (2010) <https://doi.org/10.1175/2009BAMS2755.1>
- [48] Gahtan, J., Knapp, K.R., Schreck, C.J. III, Diamond, H.J., Kossin, J.P., Kruk,

M.C.: International Best Track Archive for Climate Stewardship (IBTrACS) Project, Version 4r01. NOAA National Centers for Environmental Information. North Atlantic subset (2024). <https://doi.org/10.25921/82ty-9e16> . <https://doi.org/10.25921/82ty-9e16>

- [49] Ulmer, B.: Atlantic residual diffusion. Zenodo. Version 1.0.0 [Software] (2026). <https://doi.org/10.5281/zenodo.23149631> . <https://doi.org/10.5281/zenodo.23149631>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [articlehurricanes4.pdf](#)