

Residual Diffusion Modeling for Km-scale Atmospheric Downscaling

Morteza Mardani^{1,*^a, Noah Brenowitz^{1,*}, Yair Cohen^{1,*}, Jaideep Pathak¹, Chieh-Yu Chen¹, Cheng-Chin Liu², Arash Vahdat¹, Karthik Kashinath¹, Jan Kautz¹, and Mike Pritchard¹}

¹*NVIDIA, Santa Clara, CA. 95050, USA*

²*Central Weather Administration, 64, Gongyuan Road, Taipei 100006, Taiwan*

^{*}*These authors contributed equally*

^a*Corresponding author: mmardani@nvidia.com*

December 12, 2023

Abstract

Predictions of weather hazard require expensive km-scale simulations driven by coarser global inputs. Here, a cost-effective stochastic downscaling model is trained from a high-resolution 2-km weather model over Taiwan conditioned on 25-km ERA5 reanalysis. To address the multi-scale machine learning challenges of weather data, we employ a two-step approach Corrector Diffusion (*CorrDiff*), where a UNet prediction of the mean is corrected by a diffusion step. Akin to Reynolds decomposition in fluid dynamics, this isolates generative learning to the stochastic scales. *CorrDiff* exhibits skillful RMSE and CRPS and faithfully recovers spectra and distributions even for extremes. Case studies of coherent weather phenomena reveal appropriate multivariate relationships reminiscent of learnt physics: the collocation of intense rainfall and sharp gradients in fronts and extreme winds and rainfall bands near the eyewall of typhoons. Downscaling global forecasts successfully retains many of these benefits, foreshadowing the potential of end-to-end, global-to- km-scales machine learning weather predictions.

1 Introduction

Coarse-resolution 25-km global weather prediction is undergoing a machine learning renaissance thanks to autoregressive machine learning models trained on global reanalysis [6, 39, 11, 7, 30, 15, 14]. However, many applications of weather and climate data require kilometer-scale forecasts: e.g., risk assessment and capturing local effects of topography and human land use [22]. Globally, applying ML at km-scale resolution poses significant challenges since training costs are superlinear with respect to the resolution of training data. Moreover, predictions from global km-scale physical simulators are not yet well tuned, so available training data can have worse systematic biases than coarse-resolution or established regional simulations [56, 27], and current data tends to cover short periods of time. Such datasets are also massive, difficult to transfer between data centers and frequently not produced on machines attached to significant AI computing resources like GPUs.

In contrast, for regional simulation, scaling ML to km-scales is attractive. High-quality training data is abundant as many national weather agencies couple km-scale numerical weather models in a limited domain to coarser resolution global models [17] – a process called dynamical downscaling. Since these predictions are augmented by data assimilation from ground-based precipitation radar and other sensors, good estimates of regional km-scale atmospheric states exists [13]. Such dynamical downscaling is computationally expensive, which limits the number of ensemble members used to quantify uncertainties [38]. A common inexpensive alternative is to learn a statistical downscaling from these dynamical downscaling simulations and observations [60].

In this context, ML downscaling enters as an advanced (non linear) form of statistical downscaling with potential to emulate the fidelity of dynamical downscaling. Several ML methods have been previously used

for downscaling [9, 50, 19, 57, 38, 59, 2]. Convolutional Neural Networks, which reduce the input dimensions, have shown promise in globally downscaling climate (100km) data to weather scales (25km) [35, 47, 3, 45]. However, such deterministic ML approaches require interventions to produce useful probabilistic results, such as ensemble inference [47] or predicting the parameters of an assumed distribution [3]).

The stochastic nature of atmospheric physics at km-scale [51] makes the downscaling inherently probabilistic, making it natural to apply generative models at these scales. Generative Adversarial Networks (GANs) have been tested, including for forecasting precipitation at km-scale in various regions [32, 43, 23, 46, 21, 59]; see the latter for a good review. Training GANs, however, poses several practical challenges including mode collapse, training instabilities, and difficulties in capturing long tails of distributions [61, 29, 48].

Alternatively, diffusion models offer sample diversity and training stability [26, 16] alongside demonstrable skill in probabilistically generating km-scales. [1] used a diffusion model for predicting rain density in the UK from vorticity as an input, thus demonstrating potential for channel synthesis. [24] used a diffusion model for downscaling solar irradiance in Hawaii with a 1 day lead time, demonstrating the ability to simultaneously forecast. Moreover, diffusion models have been used directly for probabilistic weather forecasting and nowcasting [31, 33, 36]. See table S1 in 1 for more details.

Building from this work, we turn to our challenge of interest – stochastically downscaling multiple variables simultaneously while also transferring input information to predict a new field (i.e. channel synthesis). If successful, this paves the way towards end-to-end ML downscaling systems that predicts regional high-resolution weather as a postprocessing of global predictions. As a proof of concept we will demonstrate such a ML model trained for the region surrounding Taiwan and show its downscaling abilities from both a global reanalysis and a global forecast model.

Details follow. The key contributions of this paper are:

1. A physics-inspired, two-step approach (CorrDiff) to learn mappings between low- and high-resolution weather data with high fidelity.
2. CorrDiff provides realistic measures of stochasticity and uncertainty, in terms of Continuous Rank Probability Score (CRPS) and by comparing spectra and distributions.
3. CorrDiff reproduces the physics of coherent weather phenomena remarkably well, correcting frontal systems and typhoons.
4. CorrDiff is sample-efficient, learning effectively from just 4 years of data.
5. CorrDiff on a single GPU is at least 22 times faster and 1,300 times more energy efficient than the numerical model used to produce its high resolution training data, which is run on 928 CPU cores.

2 Generative downscaling: Corrector diffusion model

Consider a specific region on Earth, mapped onto a two-dimensional grid. Our input $\mathbf{y} \in \mathbb{R}^{c_{in} \times m \times n}$ is a low-resolution meteorological forecast taken from a 25-km global weather forecasting model (e.g., FourCastNet [39, 30, 7], or the Global Forecast System (GFS) [37]). Here, c_{in} represents the number of input channels and m, n represent the dimensions of a 2D subset of the globe. Our targets $\mathbf{x} \in \mathbb{R}^{c_{out} \times p \times q}$ come from corresponding data aligned in time c_{out} but having higher resolution, i.e. $p > m$ and $q > n$.

In our proof of concept we use the ERA5 reanalysis as input, over a subregion surrounding Taiwan, with $m = n = 36$, $c_{in} = 20$ and $c_{out} = 4$. See Table S2 for details about the inputs and outputs. The target data are 12.5 times higher resolution ($p = q = 448$) and were produced using a radar-assimilating Weather Research and Forecasting (WRF) physical simulator [41] provided by the Central Weather Administration of Taiwan (CWA) [13] (i.e. CWA-WRF). Though imperfect, WRF is a SOTA model for km-scale weather simulations and is used operationally by several national weather agencies.

The goal of probabilistic downscaling is to mimic the conditional probability density $p(\mathbf{x}|\mathbf{y})$. To learn $p(\mathbf{x}|\mathbf{y})$ we employ a diffusion model. Such models learn generative SDEs through the concept of score matching [26, 53, 28, 52, 5], with a forward and a backward processes working in tandem. In the forward process, noise is gradually added to the target data until the signal becomes indistinguishable from noise. The backward process then involves denoising the samples using a dedicated neural network to eliminate the

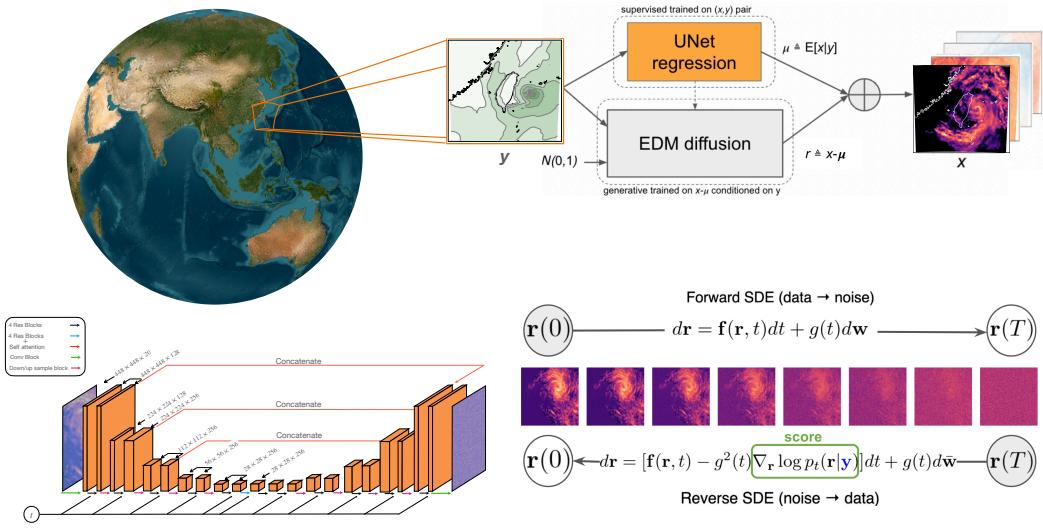


Figure 1: The workflow for training and sampling CorrDiff for generative downscaling. Top: Coarse-resolution global weather data at 25 km scale is used to first predict the mean μ using a regression model, which is then stochastically corrected using EDM diffusion \mathbf{r} , together producing the probabilistic high-resolution 2 km-scale regional forecast. Bottom right: diffusion model is conditioned with the coarse-resolution input to generate the residual \mathbf{r} after a few denoising steps. Bottom left: the score function for diffusion is learned based on the UNet architecture.

noise. Through this sequential denoising process, the model iteratively refines the samples, bringing them closer to the target data distribution. The denoising neural network plays a critical role in this convergence, providing the necessary guidance to steer the samples towards accurate representations of the original data.

The motivation for CorrDiff is that naively applying conditional diffusion models to directly learn $p(\mathbf{x}|\mathbf{y})$ was unsuccessful. The large uncertainty of variables such as radar reflectivity at these scales requires substantial noise levels during the forward process demands a large number of steps in backward process. This impedes learning and leads to poor sample fidelity [54].

To sidestep these challenges, we decompose the generation into two steps (Fig. 1). The first step predicts the conditional mean using (UNet) regression (see also 2 and S1 for details), and the second step learns a correction using a diffusion model as follows:

$$\mathbf{x} = \underbrace{\mathbb{E}[\mathbf{x}|\mathbf{y}]}_{:=\mu(\text{regression})} + \underbrace{(\mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{y}])}_{:=\mathbf{r}(\text{generation})}, \quad (1)$$

where \mathbf{y} and \mathbf{x} are the input and target respectively. This signal decomposition is inspired by Reynolds decomposition in fluid-dynamics [40] and climate data analytics. Assuming the regression learns the conditional mean accurately, i.e., $\mu \approx \mathbb{E}[\mathbf{x}|\mathbf{y}]$, the residual is zero mean, namely $\mathbb{E}[\mathbf{r}|\mathbf{y}] \approx 0$, and as a result $\text{var}(\mathbf{r}|\mathbf{y}) = \text{var}(\mathbf{x}|\mathbf{y})$. Accordingly, based on *law of total variance* [10], one can decompose the variance as

$$\text{var}(\mathbf{r}) = \mathbb{E}[\text{var}(\mathbf{r}|\mathbf{y})] + \underbrace{\text{var}(\mathbb{E}[\mathbf{r}|\mathbf{y}])}_{=0} \leq \mathbb{E}[\text{var}(\mathbf{x}|\mathbf{y})] + \underbrace{\text{var}(\mathbb{E}[\mathbf{x}|\mathbf{y}])}_{\geq 0} = \text{var}(\mathbf{x}). \quad (2)$$

That is, the residual formulation ensures targets that have significantly smaller variance. According to (2), the variance reduction is more pronounced when $\text{var}(\mathbb{E}[\mathbf{x}|\mathbf{y}])$ is large, e.g., in the case of Typhoons. For our specific target data we find that the actual variance reduction is significant, especially at large scales; see section 3 and Figure S2. Moreover, the difference \mathbf{r} is very localized and has no significant spatial

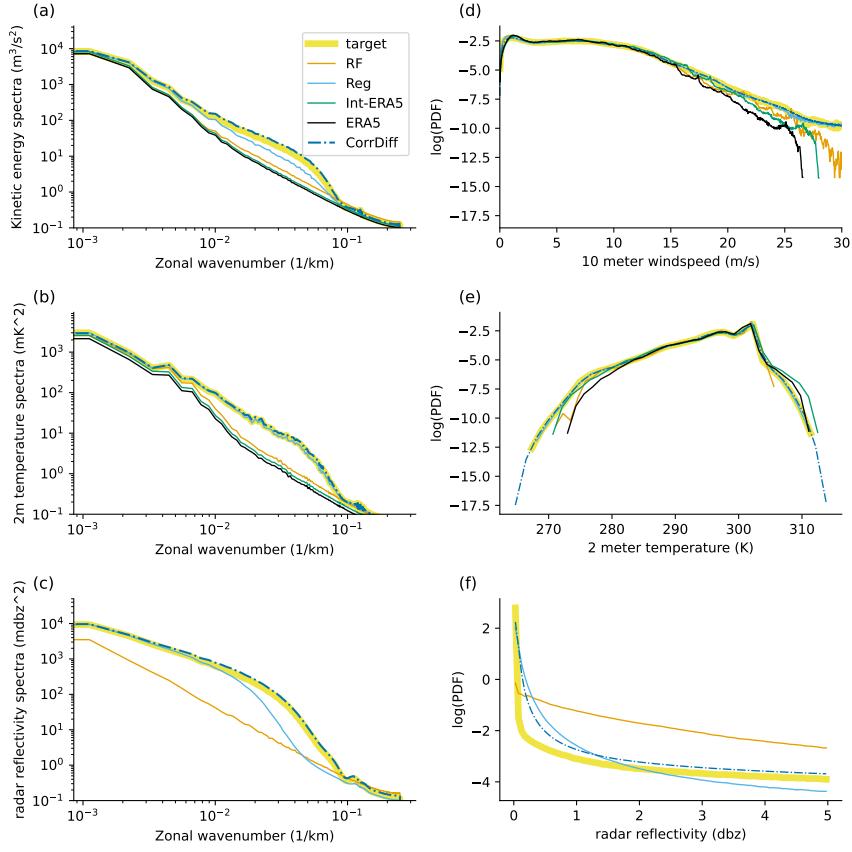


Figure 2: Power spectra and distributions for the interpolated ERA5 input, CorrDiff, RF, Reg, and WRF. These results reflect reductions over space, time and for *CorrDiff* across 192 different samples. Left: Power spectra for kinetic energy (top), 2 meter temperature (middle) and radar reflectivity (bottom). Right: distributions of windspeed, (top), 2 meter temperature (middle) and radar reflectivity (bottom). Radar reflectivity is not included in the ERA5 dataset. We show the log-PDF to highlight the differences at the tails of the distributions. Here wavenumber is the inverse of a wavelength.

auto-correlation at ranges beyond 30 km. In contrast, \mathbf{x} has significant auto-correlation even at a 200 km distance. To sum it up, the main idea of CorrDiff is that learning the distribution $p(\mathbf{r})$ can be much easier than learning the distribution $p(\mathbf{x})$. Since modeling multi-scale interactions is a daunting task in many physics domains, we expect this approach could be widely applied. More details are described in Section 5 and the outline is depicted in Fig. 1.

3 Results

The target data (WRF) span four years (2018-2021) and additional first four months of 2022, at one hour time resolution. We use the first 3 years for training and the rest for testing. The input (course resolution) data is taken from the ERA5 reanalysis for the corresponding dates. See 5 and S2 for details.

CorrDiff downscaling is compared with the input and target data and several baseline models, for deterministic and probabilistic skill scores, distributions and power spectra, and case studies of various coherent atmospheric phenomena.

Method	Metric	u10m	Radar	v10m	t2m
CorrDiff	CRPS	0.29	0.53	0.31	0.14
CorrDiff	MAE	0.40	0.74	0.43	0.19
UNet-regression	MAE	0.45	0.77	0.50	0.24
Random forest regression	MAE	1.15	3.58	1.28	0.81
ERA5 bilinear interpolation	MAE	1.18	-	1.28	0.97

Table 1: Skill scores evaluated from 204 date and time combinations taken randomly from the out-of-sample year (2021). For CorrDiff each evaluation used 192 samples. The table compares CorrDiff, UNet-regression, random forest, and interpolated ERA5 predictions in terms of MAE of the ensemble mean and CRPS for different atmospheric variables. For deterministic predictions, MAE and CRPS are equivalent. A separate RF is fit with scikit-learn for each of the 4 output channels with 100 trees and the default hyperparameters. While crude, this RF provides a simple (and easily tuned) baseline for the performance of Reg.

3.1 Skill

Compared with several baselines, CorrDiff shows the highest deterministic skill (MAE) followed by the UNet regression (Reg), the random forest (RF) and the interpolation of ERA5 (1). The difference between MAE of CorrDiff and that of the UNet reflects the correction of the sample mean by the diffusion step. The consistent improvement in MAE for all the target variables between the UNet and CorrDiff suggests that the generative CorrDiff model (step 2) can correct some of the remaining biases after the UNet-regression (step 1).

3.2 Spectra and distributions

CorrDiff is able to predict the correct power spectra of kinetic energy (KE), 2-meter temperature and radar reflectivity compared with the baselines in Table 1 (Fig. 2 a-c). The benefits of CorrDiff relative to the UNet baseline are stronger for some predicted variables than others. Improvements are most striking for radar reflectivity (Fig. 2 c) for which CorrDiff matches the target power spectra while the UNet-regression is significantly worse. This is expected as radar reflectivity comes from precipitation, which is linked to intrinsically stochastic physics (see [59]), i.e. requires the generation step for skillful prediction. In contrast, for 2-meter temperature (Fig. 2 b) the performance of UNet-regression is comparable to CorrDiff; temperature downscaling is expected to be mostly driven by sub-grid variations in topography that can be learned deterministically from the static grid embeddings.

Good skill is also found when looking at the corresponding probability distributions (Fig. 2 d-f). CorrDiff is able to match the target distribution, including the heavy-tailed structure, which is significantly improved from the prediction based on the UNet alone. These rare extreme values reflect most of the risk. For windspeed (Fig. 2 d), the UNet and the baseline models underestimate the probability of winds faster than 20 m/s. For 2-meter temperatures (Fig. 2 e), the baseline models overestimate the probability of warm extremes and underestimate the probability of cold extremes. Radar reflectivity proves to be the most challenging distribution to reproduce. CorrDiff outperforms the other models in reproducing the reflectivity PDF (Fig. 2f). Although it overestimates the occurrence of reflectivity at lower values (i.e. weakly precipitating systems), the match of slopes at high values attests that the power-law relationship regulating the heavy-tail structure of rain extremes is emulated skillfully.

3.3 Case studies: downscaling coherent structures

Operational meteorologists value case study analysis, since aggregate skill scores and spectra can be more easily gamed or mask symptoms of spatial incoherence. We thus turn our attention to specific weather regimes. Fig. 3 illustrates the variability of the generated radar reflectivity field at four distinct times. When analyzing 200 samples, the standard deviation of radar reflectivity (second column from the left) is roughly 20% of the magnitude of the mean (left column), with the majority of the variance located in and around existing precipitation regions. Such a pattern is anticipated given that the timing and location of precipitation can vary markedly even within a fixed large-scale configuration since convective physics are inherently stochastic.

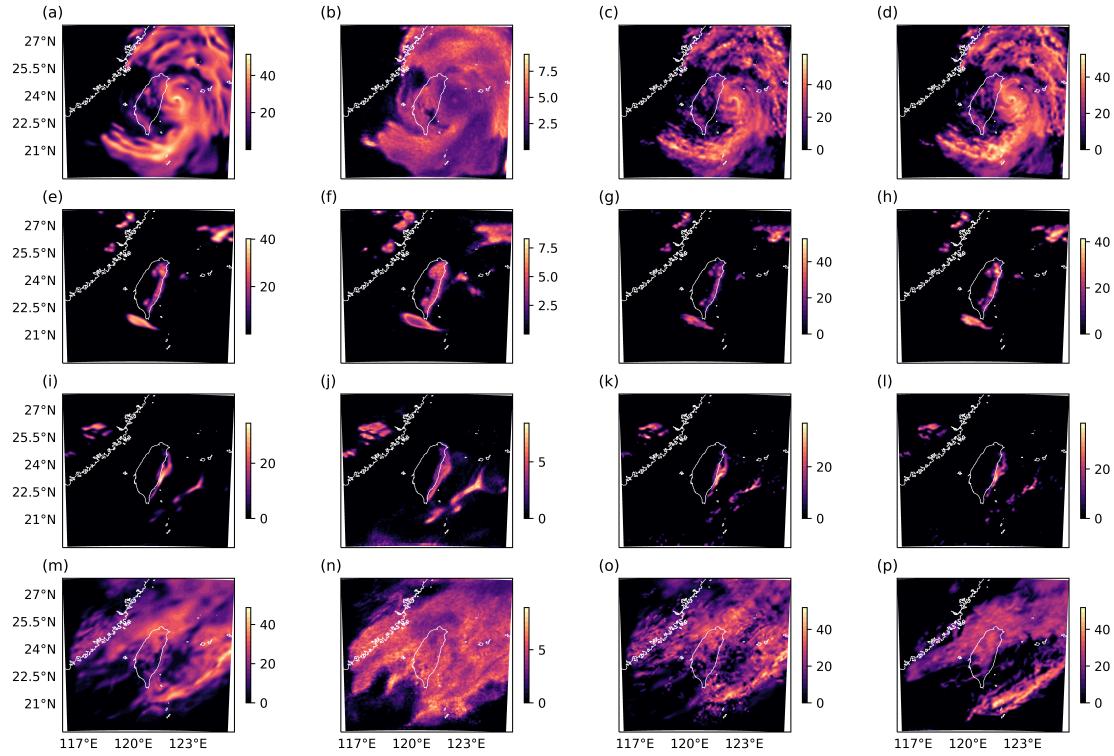


Figure 3: Demonstration of the stochastic prediction of radar reflectivity (in dBZ). Top to bottom: 2021-09-12 00:00, 2021-04-02 06:00, 2021-02-02 12:00 and 2022-02-13 20:00. Left to right: sample mean, sample standard deviation, sample number 200 and the target forecast.

The CorrDiff prediction for an individual sample (number 200, third column from the left) reveals a fine-scale structure akin to the target data (right column). The similarity between panels (a) and (d) highlights the role of the mean prediction in forming large-scale coherent structures, such as typhoon Chanthu (2021), top row, and frontal systems, bottom row. Specifically the typhoon rainbands (spiral bands of clouds that emanate from the typhoon center) are coherent enough to be captured by the mean, but with large variance and smooth structure. The fine-scale structure reflecting the stochastic physics is captured well by the diffusion model, refining the smooth fields of the mean prediction as seen in the third column of Fig. 3. Further comparison across samples is presented in an animation in S3 in 4.

3.3.1 Frontal system case study

Frontal systems are an example of organized atmospheric systems. A cold front is a sharp change in temperature and winds associated with a mature cyclonic storm. As the front moves eastward, the cold air pushes the warm air to its east upward. This upward motion leads to cooling, condensation and ultimately rainfall. That is, these physics should manifest as multi-variate relationships with linked fine scale structures of two wind vector components and temperature that should co-locate with radar reflectivity.

Fig. 4 shows an example of CorrDiff downscaling a cold front. The position of the front is clearly visible in the southeast portion of the domain, where a strong horizontal surface temperature gradient (top) co-locates with strong across-front wind convergence (middle). The along-front components of the wind vector also change abruptly (middle row), which is consistent with the change in temperature. The super resolved gradients in the temperature and winds are encouragingly sharper than the input. The intense rainfall associated with this convergence line can be seen in the radar reflectivity ground truth for the same calendar date (which are shown in bottom row of Fig. 3 above). The generated radar reflectivity is appropriately

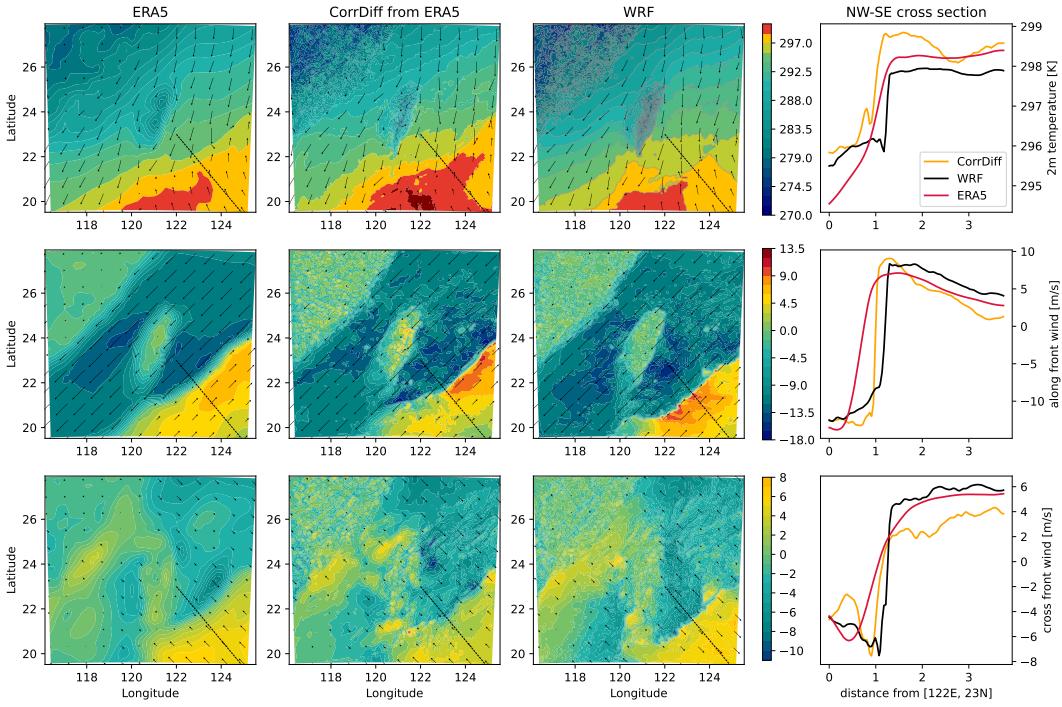


Figure 4: Examining the downscaling of a cold front on Feb 2, 2022 at 20 UTC. Left to right: prediction of ERA5, CorrDiff and Target for different fields, followed by their averaged cross section from 20 lines parallel to the thin dashed line in the contour figures. Top to bottom: 2 meter temperature (arrows are true wind vectors), along front wind (arrows are along front wind component) and across front wind (arrows are across front wind component).

concentrated near the frontal boundary. Both the location of the front, and the magnitude of the horizontal wind and temperature gradients (sharpening of the front) associated with it, are captured well by CorrDiff although some mispositioning of the exact front location is inevitable. These are reassuring signs of learnt physics during the generation task.

3.3.2 Tropical Cyclone case study

Downscaling typhoons (i.e. tropical cyclones) is especially complicated. The average radius of maximum winds of a tropical cyclone is less than 100km, and at 25km resolution of the input data tropical cyclones are only partially resolved, resulting in cyclonic structures that are too wide and too weak compared with high resolution models or observations [8]. A useful downscaling model must simultaneously correct their size and intensity in addition to generating appropriate fine-scale structure.

We find that CorrDiff is able to correct the structure of tropical cyclones accurately. Fig. 5(a)-(f) shows the structure of typhoon Chanthu (2021) on September 12 at 00:00:000 UTC. Compared to the target data (panel c) the poorly resolved typhoon in the ERA5 (panel a) is too wide and does not include a closed contour annulus of winds above 16 m/s surrounding an overly quiescent eye-wall. CorrDiff downscaling (panel b) is able to recover much of the spatial structure of the windspeed compared with the target. The skill of the CorrDiff downscaling compared to interpolating ERA5 can be more clearly quantified by calculating the mean axisymmetric structure of the storms as a function of radius from eye-wall center (panel f). Notably, with downscaling the radius of maximum winds decreases from 75km to about 25km while the windspeed

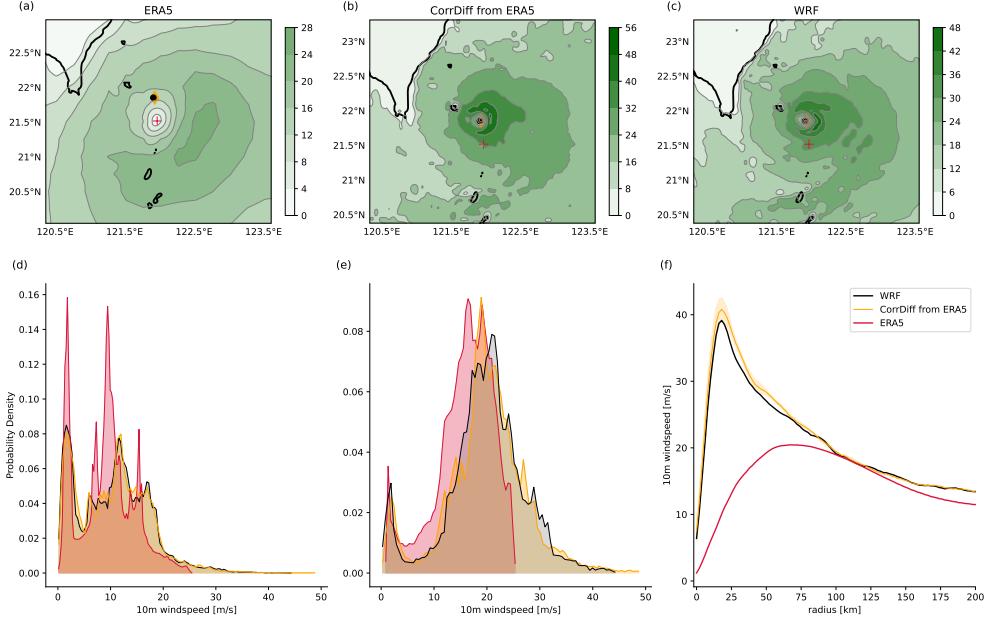


Figure 5: A comparison of the 10m windspeed maps (m/s), distributions and the axisymmetric cross section from typhoon Chanthu (2021) on 2021/09/11:12:00:00UTC. Panels (a),(b),(c) show the 10m windspeed from ERA5, CorrDiff downscaling of ERA5 and the target (WRF), respectively. The CorrDiff panels show the first ensemble member. The solid black contour indicates the Taiwan coastline. Storm center of the ERA5, CorrDiff and WRF are shown in red ‘+’, orange diamond, and the black dot, respectively. Panels (d) and (e) show the distribution shift (normalized PDFs) for the entire CWA domain and for the typhoon selected region in the top panels. Panel (f) shows the axisymmetric structure of the typhoon about its center. For the CorrDiff curves, line is the ensemble mean and the shading shows one standard deviation around the mean.

increases from 20 m/s to 50 m/s – both favorable improvements.

Probabilities of damaging typhoon winds shown in panels (d) and (e) in Fig. 5 are significantly improved. In ERA5 (red), occurrence of weak wind speeds is over-estimated and the damaging extreme hurricane winds (above 25m/s) are missing. CorrDiff better predicts the chances of the strong winds most likely to impact society and infrastructure. Further analysis 5 expands beyond this case study to examine generated wind statistics for several hundred typhoons that crossed the domain during 1980 to 2020. Although no target data is available, when comparing the maximum windspeed and radius of maximum windspeed from CorrDiff predictions to a reference from the Japan Meteorological Agency best track dataset [4] CorrDiff is found to correct at least 75% of the error in intensity for windspeed below 50m/s but only 50% of the error for higher windspeeds S4.

3.4 Downscaling a forecast from a global model

An attractive use case for CorrDiff is to directly downscale forecasts from global models. During the writing of this paper, typhoon Koinu (2023) approached Taiwan. Together with CWA we compared the live forecasts of Koinu from GFS, CorrDiff downscaling applied to GFS (here we use the same input channel as before) and WRF (target) data, see Figure 6. Expectation of good results should be tempered by the fact that this exposes CorrDiff to out of sample inputs, by using GFS in place of ERA5. Even without retraining to include the leadtime forecast bias of GFS, we find some encouraging results. An expected enhancement in windspeed intensity is confirmed and it most accurate up to a leadtime of 6 hours (third row) after which it decreases. The radar reflectivity shows some caveats. While the domain averaged reflectivity is corrected well by CorrDiff, the coherent rainband that is exhibited in the target data at lead times larger than 3h is absent in CorrDiff. This rainband is also absent in the input (GFS), and CorrDiff seem to struggle to generate it.

This suggests that CorrDiff could be used for downscaling from a global model at short lead times (12h, for which baseline was provided).

4 Discussion

This study presents a generative diffusion model (CorrDiff) for downscaling coarse-resolution (25-km) global weather state to higher resolution (2km) over a subset of the globe. CorrDiff consists of two steps: regression and generation. The regression step approximates the mean, while the generation step further corrects the mean but also generates the distribution, producing fine-scale details stochastically. This approach is akin to the decomposition of physical variables into their mean and perturbations, common practice in fluid dynamics, e.g. [40].

Through extensive testing in the region of Taiwan, the model is shown to skillfully correct kinetic energy spectra, generate realistic probabilities and downscale coherent structures accurately (with minor caveats such as inaccuracies in the 2-meter temperature on the warm sector of a frontal system (Fig. 4) and over-correction of typhoon sizes in some out-of-sample cases for which we do not have the target (WRF) data, see 5). It is possible that the model’s accuracy could be further improved with a larger training dataset that contains more diverse examples of such rare coherent structures.

The two step approach in CorrDiff also offers the possibility to trade off between the fast inference of the mean using the UNet-regression, and the accurate and probabilistic inference of the CorrDiff. This is particularly useful given that some variables – like the 2 meter temperature – are well predicted by the UNet-regression while others like radar reflectivity depend on the diffusion step for their skill (see Figure 2). Moreover, it could be possible to apply the diffusion step to a mean prediction obtained in a different way (e.g. a numerical model if available) to generate a plausible distribution from a single prediction.

With the current hardware (see 6) and with an unoptimized code (no GPU parallelization or batching utilized), CorrDiff inference is about 22 (40) times faster, and 1,130 (1,800) times more energy efficient than running CWA-WRF on CPU with float32 (mixed) precision.

This paper focused on generation quality, and not on optimal inference speed, for which gains could be easily anticipated. Our prototype of CorrDiff is using a dozen iterations thanks to the initial regression step. However, for future work, we will push to reduce the number of iterations to only a few by using distillation methods [49, 61, 62] and pursuing other performance optimization techniques [34, 58].

Several potential extensions of the proposed method are worth considering:

1. **Downscaling Coarse-Resolution Medium-Range Forecasts:** This requires addressing lead time-dependent forecast errors, enabling a comprehensive evaluation of simultaneous bias correction and downscaling.
2. **Downscaling at Different Geographic Locations:** The primary obstacle here is the scarcity of reliable kilometer-scale weather data. Additionally, addressing the computational scalability of CorrDiff for regions significantly larger than Taiwan is crucial.
3. **Downscaling Future Climate Predictions:** This introduces further complexities related to conditioning probabilistic predictions on various future anthropogenic emissions scenarios and assessing whether the generated weather envelope appropriately reflects climate sensitivity, particularly concerning extreme events.

These extensions have significant potential benefits such as accelerated regional forecasts, increased ensemble sizes, improved climate downscaling, and the provision of high-resolution regional forecasts in data-scarce regions, leveraging training data from adjacent areas.

5 Methods

This section elaborates on the proposed CorrDiff methodology for probabilistic downscaling. It begins with a background on diffusion models to provide the machinery. It then delves into CorrDiff and its associated components. We further detail our experimental setup including the CWA dataset, network architecture, and training protocols. At the end, we briefly discuss evaluation criteria.

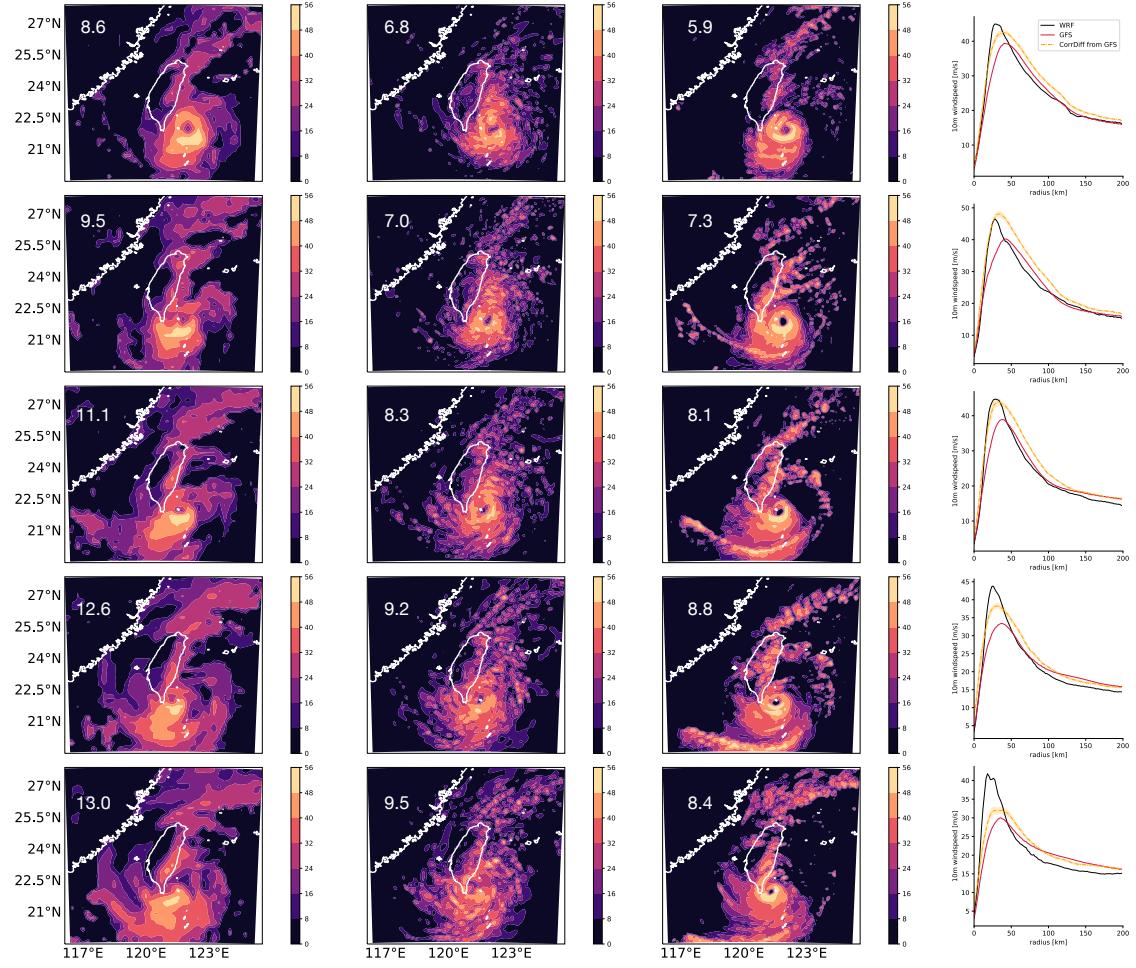


Figure 6: A comparison of a live km-scale forecast made by linking GFS predictions to CorrDiff downscaling, validated against WRF-CWA operational predictions made prior to the typhoon Loinu landfall. Forecast is initialized on 20231004-12:00 UTC, top to bottom show lead time 0, 3, 6, 9, 12 hours. Left to right: forecast from the GFS; CorrDiff downscaling of GFS and the target (WRF-CWA). The fourth column compares the axisymmetric windspeed profile of the typhoon (shading for GFS+CorrDiff shows one standard deviation). The domain mean of the 1-hr maximum radar reflectivity is indicated at the top left of each of the reflectivity maps.

5.1 Background on diffusion models

Consider the data distribution represented by $p_{\text{data}}(\mathbf{x})$. This distribution has an associated standard deviation, denoted by σ_{data} . The forward diffusion process seeks to adjust this distribution, yielding modified distributions denoted by $p_{\text{data}}(\mathbf{x}; \sigma)$. This transformation is achieved by incorporating i.i.d. Gaussian noise with a standard deviation of σ into the data. When σ surpasses σ_{data} by a considerable margin, the resulting distribution approximates pure Gaussian noise.

Conversely, the backward diffusion process operates by initially sampling noise, represented as \mathbf{x}_0 , from the distribution $\mathcal{N}(0, \sigma_{\text{max}}^2 \mathbf{I})$. The process then focuses on the denoising of this image into a series, \mathbf{x}_i , that is characterized by a descending order of noise levels: $\sigma_0 = \sigma_{\text{max}} > \sigma_1 > \dots > \sigma_N = 0$. Each noise level corresponds to a specific distribution of the form $\mathbf{x}_i \sim p_{\text{data}}(\mathbf{x}_i; \sigma_i)$. The terminal image of the backward process, \mathbf{x}_N , is expected to approach the original data distribution.

SDE formulation. To present the forward and backward processes rigorously, they can be captured via stochastic differential equations (SDEs). Such SDEs ensure that the sample, \mathbf{x} , aligns with the designated data distribution, p , over its progression through time [55, 28]. A numerical SDE solver can be used here, where a critical component is the noise schedule, $\sigma(t)$, which prescribes the noise level at a specific time, t . A typical noise scheduler is $\sigma(t) \propto \sqrt{t}$. Based on [28], the forward SDE is given as

$$d\mathbf{x} = \sqrt{2\dot{\sigma}(t)\sigma(t)}d\omega(t), \quad (3)$$

while the backward SDE is

$$d\mathbf{x} = -2\dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t))dt + \sqrt{2\dot{\sigma}(t)\sigma(t)}d\bar{\omega}(t). \quad (4)$$

The term $\dot{\sigma}(t)$ refers to the time derivative of $\sigma(t)$. The forward SDE is a Wiener process, while the backward SDE comprises two terms: a deterministic component representing the probability flow ODE with noise degradation, and noise injection via the Wiener process.

Denoising score matching. An examination of the SDE in Eq. (4) indicates the necessity of the score function, $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma)$, for sampling from diffusion models. Intriguingly, this score function remains unaffected by the normalization constant of the base distribution, regardless of its computational complexity. Given its independence, it can be deduced using a denoising method. If $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma) = (D_{\theta}(\mathbf{x}; \sigma) - \mathbf{x})/\sigma^2$, a denoising neural network, namely $D_{\theta}(\mathbf{x}; \sigma)$, can be trained for the denoising task using

$$\min_{\theta} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} \mathbb{E}_{\sigma \sim p_{\sigma}} \mathbb{E}_{\mathbf{n} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} [\|D_{\theta}(\mathbf{x} + \mathbf{n}; \sigma) - \mathbf{x}\|^2]. \quad (5)$$

Note, the noise variance is also modeled as a random variable that simulates different noise levels in the forward process e.g., based on log-normal distribution; see [28].

5.2 Proposed approach

As discussed in section 2, the high resolution state \mathbf{x} in (1) can be written as the sum of the mean μ and the difference \mathbf{r} , where the latter will be nearly zero mean and exhibits a small distribution shift, which facilitates training diffusion models for correcting the mean prediction. It is worth noting that this two-step method has further implications for learning physics. The UNet-regression step can anticipate many of the physics of downscaling, some of which are deterministic. These include high-resolution topography (which to first order controls the 2-meter temperature variation due to the lapse-rate effect), and the large-scale horizontal wind which combine leading balances in the free atmosphere with the effect of surface friction and topography. Stochastic phenomena such as convective storms that also change temperatures and winds are easier to model as deviations from the mean. Also, cloud resolving models are explicitly formulated using deviations from a larger scale balanced state [42]. In the next section, we discuss the regression and generation step in details.

5.2.1 Regression on the mean

In order to predict the conditional mean $\mu = \mathbb{E}[\mathbf{x}|\mathbf{y}]$, we resort to UNet-regression. A UNet network is supervised with training data $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ to learn the regression. We adopt a UNet architecture that is commonly used in denoising diffusion models. This particular UNet [53] incorporates attention layers and residual layers, allowing it to effectively capture both short and long-range dependencies in the data see section 2 and S1. Mean-Squared-Error (MSE) loss is optimized for training.

5.2.2 Denoising diffusion corrector

Once equipped with the UNet-regression network, we can begin by predicting the conditional mean $\hat{\mu}$, which serves as an approximation of $\mathbb{E}[\mathbf{x}|\mathbf{y}]$. Subsequently, we proceed to train the diffusion model directly on the difference: $\mathbf{r} = \mathbf{x} - \hat{\mu}$. Notably, the difference exhibits a small departure from the target data, allowing for the utilization of smaller noise levels during the training of the diffusion process.

In our approach, we adopt the Elucidated diffusion model (EDM), a continuous-time diffusion model that adheres to the principles of SDEs (in Eq. (3)-(4)) [28] to design the diffusion process and architecture. As a result it has an intuitive and physics driven hyperparameter tuning, which makes it work across different domains. In our case, we want to generate \mathbf{r} by sampling from the conditional distribution $p(\mathbf{r}|\mathbf{y})$ following the SDEs in Eq. (3)-(4). To condition the diffusion model, we concatenate the input coarse-resolution data \mathbf{y} with the noise over different channels. We also learn the score function $\nabla_{\mathbf{r}} \log p(\mathbf{r}|\mathbf{y})$ using the score matching loss in Eq. (5) where the denoiser is now $D_{\theta}(\mathbf{r} + \mathbf{n}; \sigma; \mathbf{y})$ with the conditioning input \mathbf{y} . For the denoiser we again follow the design principles in EDM to use a UNet architecture with skip connections weighted by the noise variance. Architecture details are discussed in Section 5.3.2.

To generate samples from the distribution $p(\mathbf{r}|\mathbf{y})$, we employ the second-order EDM stochastic sampler [28] [Algorithm 2] to solve for the reverse SDE in Eq. (4). Upon sampling \mathbf{r} , we augment it with the predicted conditional mean $\hat{\mu}$ from regression, to generate the sample $\hat{\mu} + \mathbf{r}$. This entire workflow is illustrated in Fig. 1, providing a visual representation of the steps involved in our proposed method.

5.3 Experimental setup

5.3.1 Dataset

The input (conditioning) dataset is taken from ERA5 reanalysis [25], a global dataset at spatial resolution of about 25km and a temporal resolution of 1h, the latter matches the target data listed below. To facilitate training, we interpolate the input data onto the curvirectangular grid of CWA with bilinear interpolation (with a rate of 4x), which results in 36×36 pixels over the region of Taiwan. Each sample in the input dataset consists of 20 channels of information see S2. This includes four pressure levels (500 hPa, 700 hPa, 850 hPa, and 925 hPa) with four corresponding variables: temperature, eastward and northward components horizontal wind vector, and Geopotential Height. Additionally, the dataset includes single level variables such as 2 meter Temperature, 10 meter wind vector, total column water vapor.

The target dataset used in this study is a subset of the proprietary RWRF model data (Radar Data Assimilation with WRFDA ¹). The RWRF model is one of the operational regional Numerical Weather Prediction (NWP) models developed by CWA [13], which focuses on radar Data Assimilation (DA) in the vicinity of Taiwan. Assimilating radar data is a common strategy used in regional weather prediction, which helps constrain especially stochastic convective processes such as mesoscale convective systems and short-lived thunderstorms. The WRF - CWA system uses a nested 2km domain in a 10km that is driven by a global model (GFS) as boundary conditions [13]. By incorporating radar data [12], RWRF improves the short-term prediction of high-impact weather events. The radar observations possess high spatial resolution of approximately 1km and temporal resolutions of 5-10 minutes at a convective scale. These observations provide wind information (radial velocity) as well as hydrometers (radar reflectivity) usefully, with a particular emphasis on the lower atmosphere. The radar data assimilation relies on the availability of reliable and precise observations, which contributes significantly to enhance the accuracy and performance of the applied deep learning algorithms in the context of NWP applications.

The target dataset covers a duration of 52 months, specifically from January 2018 to April 2022. It has a temporal frequency of one hour and a spatial resolution of 2km. We use only the inner (nested) 2km domain, which has 448×448 pixels, projected using the Lambert conformal conical projection method around Taiwan. The geographical extent of the dataset spans from approximately 116.371°E to 125.568°E in longitude and 19.5483°N to 27.8446°N in latitude. We sub-selected 4 channels (variables) as the target variables, that are most relevant for practical forecasting: temperature at 2 meter above the surface, the horizontal winds at 10 meter above the surface and the 1h maximum radar reflectivity - a surrogate of expect precipitation. Notably, the radar reflectivity channel is not present in the input data and needs to be predicted based on the other channels, making its prediction strictly generative. The radar reflectivity data also exhibits a

¹<https://www.mmm.ucar.edu/models/wrfda>

distinct distribution compared to the other output channels, with positive values and a prominent zero-mode consistent with typical non-raining conditions.

Initially, the target data is provided in the NetCDF format, which is the output of the WRFDA assimilation process. Subsequently, vertical interpolated from sigma levels to isobaric levels then save as a custom CWA DMS format. As part of the preprocessing steps, the data is converted to the Hadoop Distributed File System (HDFS) format. Additionally, any missing or corrupted data points represented by "inf" or "nan" values are eliminated from the dataset. This leads to a reduction in the number of samples from 37,944 to 33,813.

To avoid over-fitting we divide the data into training and testing sets. Three years of data 2018-2020 are used for training (2,4154 samples total). For testing we use the full fourth year, 2021, as well as the first four months (January to April) of 2022.

5.3.2 Network architecture and training

The CorrDiff method has two step learning approach. To ensure compatibility and consistency, we employ the same UNet architecture used in EDM (Elucidated Diffusion Models) for both the regression and diffusion networks(see S1). This architecture is based on the UNet model proposed in [53]. We enhance the UNet by increasing its size to include 6 encoder layers and 6 decoder layers. The base embedding size is set to 128, and it is multiplied over channels according to the pattern [1,2,2,2,2]. The attention resolution is defined as 28. For time embedding in the diffusion process, we utilize Fourier-based position embedding. However, in the regression network, time embedding is disabled. No data augmentation techniques are employed during training. Overall, the UNet architecture comprises 80 million parameters. Additionally, we introduce 4 channels for sinusoidal positional embedding.

During the training phase, we use the Adam optimizer with a learning rate of 2×10^{-4} , $\beta_1 = 0.9$, and $\beta_2 = 0.99$. Exponential moving averages (EMA) with a rate of $\eta = 0.5$ are applied, and dropout with a rate of 0.13 is utilized. The regression network solely receives the input conditioning channels. On the other hand, the diffusion training incorporates the 20 input conditioning channels from the coarse-resolution ERA5 data, which are concatenated with 4 noise channels to generate the output for each denoiser. For diffusion training, we adopt the Elucidated Diffusion Model (EDM), a continuous-time diffusion model. During training, EDM randomly selects the noise variance such that $\ln(\sigma(t)) \sim \mathcal{N}(0, 1.2^2)$ and aims to denoise the samples per mini-batch. EDM is trained for 100 million steps, while the regression network is trained for 30 million steps. The training process is distributed across 16 DGX nodes, each equipped with 8 A100 GPUs, utilizing data parallelism and a total batch size of 512. The total training time for regression and diffusion models was 7 days that amounts to approximately 21,504 GPU-hours.

For sampling purposes, we employ the second-order stochastic sampler provided by EDM. This sampler performs 18 steps, starting from a maximum noise variance of $\sigma_{\max} = 800$ and gradually decreasing it to a minimum noise variance of $\sigma_{\min} = 0.002$. We adopt the rest of hyperparameters from EDM as listed in [28].

5.4 Evaluation criterion

Probabilistic predictions aim to maximize sharpness subject to calibration [44]. Qualitatively, calibration means that the likelihood of observing the true value is the same as observing a member drawn from the ensemble. A necessary condition for calibration is that the spread-error relationship be 1-to-1 when averaged over sufficient samples [18]. Calibration also manifests as a flat rank-histogram. A simple metric used below is the root-mean-squared error of the sample mean. In the large sample limit, the sample mean becomes deterministic. So we expect this error to be comparable for generative and deterministic models.

Instead of considering both calibration and spread separately, it can be easier to use proper scoring rules like the continuous-ranked-probability score (CRPS) [20]. Let x be a scalar observation and F be the cumulative distribution of the probabilistic forecast (e.g., the empirical CDF of generated samples). Then, CRPS is defined as

$$CRPS(F, x) = \int_{-\infty}^{\infty} (F(y) - \mathbb{1}_{\{y \geq x\}})^2 dy.$$

The F which minimizes CRPS is the true cumulative distribution of x . For a deterministic forecast, $F(y) = \mathbb{1}_{\{y \geq x_0\}}$ where x_0 is the forecast value, CRPS is equivalent to the mean absolute deviation.

6 Acknowledgements

We extend our profound appreciation to the Central Weather Administration (CWA) of Taiwan², a premier government meteorological research and forecasting institution, for granting us access to the invaluable operational Numerical Weather Prediction (NWP) model dataset and for their expert guidance on data consultation. Our gratitude also extends to the AI-Algo team at NVIDIA, especially Kamyar Azizzadenesheli, Anima Anandkumar, Nikola Kovachki, Jean Kossaifi, and Boris Bonev for their insightful discussions. Additionally, we are indebted to David Matthew Hall, Dale Durran, Chris Bretherton for their constructive feedback on the manuscript.

References

- [1] Henry Addison, Elizabeth Kendon, Suman Ravuri, Laurence Aitchison, and Peter AG Watson. Machine learning emulation of a local-scale uk climate model. *arXiv preprint arXiv:2211.16116*, 2022.
- [2] Rilwan A Adewoyin, Peter Dueben, Peter Watson, Yulan He, and Ritabrata Dutta. Tru-net: a deep learning approach to high resolution prediction of rainfall. *Machine Learning*, 110:2035–2062, 2021.
- [3] Jorge Baño-Medina, Rodrigo Manzanas, and José Manuel Gutiérrez. Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, 13(4):2109–2124, 2020.
- [4] Monika Barcikowska, Frauke Feser, and Hans Von Storch. Usability of best track data in climate statistics in the western north pacific. *Monthly Weather Review*, 140(9):2818–2830, 2012.
- [5] G. Batzolos, J. Stanczuk, C.-B. Schönlieb, and C. Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.
- [6] Zied Ben-Bouallegue, Mariana C A Clare, Linus Magnusson, Estibaliz Gascon, Michael Maier-Gerber, Martin Janousek, Mark Rodwell, Florian Pinault, Jesper S Dramsch, Simon T K Lang, Baudouin Raoult, Florence Rabier, Matthieu Chevallier, Irina Sandu, Peter Dueben, Matthew Chantry, and Florian Pappenberger. The rise of data-driven weather forecasting, 2023.
- [7] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, pages 1–6, 2023.
- [8] Gu-Feng Bian, Gao-Zhen Nie, and Xin Qiu. How well is outer tropical cyclone size represented in the era5 reanalysis dataset? *Atmospheric Research*, 249:105339, 2021.
- [9] Tobias Bischoff and Katherine Deck. Unpaired downscaling of fluid flows with diffusion bridges. *arXiv preprint arXiv:2305.01822*, 2023.
- [10] Joseph K Blitzstein and Jessica Hwang. *Introduction to probability*. Crc Press, 2019.
- [11] Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical fourier neural operators: Learning stable dynamics on the sphere. *arXiv preprint arXiv:2306.03838*, 2023.
- [12] Pao-Liang Chang, Jian Zhang, Yu-Shuang Tang, Lin Tang, Pin-Fang Lin, Carrie Langston, Brian Kaney, Chia-Rong Chen, and Kenneth Howard. An operational multi-radar multi-sensor qpe system in taiwan. *Bulletin of the American Meteorological Society*, 102(3):E555–E577, 2021.
- [13] I-Han Chen, Jing-Shan Hong, Ya-Ting Tsai, and Chin-Tzu Fong. Improving afternoon thunderstorm prediction over taiwan through 3dvar-based radar and surface data assimilation. *Weather and Forecasting*, 35(6):2603–2620, 2020.

²<https://www.cwa.gov.tw/eng/>

- [14] Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6:190, 2023.
- [15] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [16] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *Proceedings of NeurIPS*, volume 34, pages 8780–8794, 2021.
- [17] Devajyoti Dutta, Ashish Routray, D Preveen Kumar, and John P George. Regional data assimilation with the ncmrwf unified model (ncum): impact of doppler weather radar radial wind. *Pure and Applied Geophysics*, 176:4575–4597, 2019.
- [18] V Fortin, M Abaza, F Anctil, and R Turcotte. Why should ensemble spread match the RMSE of the ensemble mean? *J. Hydrometeorol.*, 15(4):1708–1713, August 2014.
- [19] Andrew Geiss and Joseph C Hardin. Radar super resolution using a deep convolutional neural network. *Journal of Atmospheric and Oceanic Technology*, 37(12):2197–2207, 2020.
- [20] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.*, 102(477):359–378, March 2007.
- [21] Aofan Gong, Ruidong Li, Baoxiang Pan, Haonan Chen, Guangheng Ni, and Mingxuan Chen. Enhancing spatial variability representation of radar nowcasting with generative adversarial networks. *Remote Sensing*, 15(13):3306, 2023.
- [22] William J Gutowski, Paul Aaron Ullrich, Alex Hall, L Ruby Leung, Travis Allen O’Brien, Christina M Patricola, RW Arritt, MS Bukovsky, Katherine V Calvin, Zhe Feng, et al. The ongoing need for high-resolution regional climate models: Process understanding and stakeholder information. *Bulletin of the American Meteorological Society*, 101(5):E664–E683, 2020.
- [23] Lucy Harris, Andrew TT McRae, Matthew Chantry, Peter D Dueben, and Tim N Palmer. A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, 14(10):e2022MS003120, 2022.
- [24] Yusuke Hatanaka, Yannik Glaser, Geoff Galgon, Giuseppe Torri, and Peter Sadowski. Diffusion models for high-resolution solar forecasts. *arXiv preprint arXiv:2302.00170*, 2023.
- [25] 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.
- [26] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proceedings of NeurIPS*, volume 33, pages 6840–6851, 2020.
- [27] Cathy Hohenegger, Peter Korn, Leonidas Linardakis, René Redler, Reiner Schnur, Panagiotis Adamidis, Jiawei Bao, Swantje Bastin, Milad Behravesh, Martin Bergemann, et al. Icon-sapphire: simulating the components of the earth system and their interactions at kilometer and subkilometer scales. *Geoscientific Model Development*, 16(2):779–811, 2023.
- [28] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.
- [29] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017.
- [30] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 0(0):ead12336, 2023.

- [31] 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.
- [32] Jussi Leinonen, Daniele Nerini, and Alexis Berne. Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7211–7223, 2020.
- [33] 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.
- [34] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.
- [35] Bin Mu, Bo Qin, Shijin Yuan, and Xiaoyun Qin. A climate downscaling deep learning model considering the multiscale spatial correlations and chaos of meteorological events. *Mathematical Problems in Engineering*, 2020:1–17, 2020.
- [36] Pritthijit Nath, Pancham Shukla, and César Quilodrán-Casas. Forecasting tropical cyclones with cascaded diffusion models. *arXiv preprint arXiv:2310.01690*, 2023.
- [37] National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce. Ncep gfs 0.25 degree global forecast grids historical archive, 2015.
- [38] Nidhi Nishant, Sanaa Hobeichi, Steven C Sherwood, Gab Abramowitz, Yawen Shao, Craig Bishop, and Andy J Pitman. Comparison of a novel machine learning approach with dynamical downscaling for australian precipitation. *Environmental Research Letters*, 2023.
- [39] 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.
- [40] Stephen B Pope. *Turbulent flows*. Cambridge university press, 2000.
- [41] Jordan G Powers, Joseph B Klemp, William C Skamarock, Christopher A Davis, Jimy Dudhia, David O Gill, Janice L Coen, David J Gochis, Ravan Ahmadov, Steven E Peckham, et al. The weather research and forecasting model: Overview, system efforts, and future directions. *Bulletin of the American Meteorological Society*, 98(8):1717–1737, 2017.
- [42] Kyle G Pressel, Colleen M Kaul, Tapio Schneider, Zhihong Tan, and Siddhartha Mishra. Large-eddy simulation in an anelastic framework with closed water and entropy balances. *Journal of Advances in Modeling Earth Systems*, 7(3):1425–1456, 2015.
- [43] Ilan Price and Stephan Rasp. Increasing the accuracy and resolution of precipitation forecasts using deep generative models. In *International conference on artificial intelligence and statistics*, pages 10555–10571. PMLR, 2022.
- [44] Adrian E Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.*, 133(5):1155–1174, May 2005.
- [45] Neelesh Rampal, Peter B Gibson, Abha Sood, Stephen Stuart, Nicolas C Fauchereau, Chris Brandolino, Ben Noll, and Tristan Meyers. High-resolution downscaling with interpretable deep learning: Rainfall extremes over new zealand. *Weather and Climate Extremes*, 38:100525, 2022.
- [46] 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.

- [47] Eduardo R. Rodrigues, Igor Oliveira, Renato L. F. Cunha, and Marco A. S. Netto. Deepdownscale: a deep learning strategy for high-resolution weather forecast. *arXiv preprint arXiv:1808.05264*, 2018.
- [48] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- [49] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2021.
- [50] Badr-eddine Sebbar, Saïd Khabba, Olivier Merlin, Vincent Simonneaux, Chouaib El Hachimi, Mohamed Hakim Kharrou, and Abdelghani Chehbouni. Machine-learning-based downscaling of hourly era5-land air temperature over mountainous regions. *Atmosphere*, 14(4):610, 2023.
- [51] Tobias Selz and George C Craig. Upscale error growth in a high-resolution simulation of a summertime weather event over europe. *Monthly Weather Review*, 143(3):813–827, 2015.
- [52] Jiefeng Song, Chaoyue Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [53] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [54] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.
- [55] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Ashish Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [56] Bjorn Stevens, Masaki Satoh, Ludovic Auger, Joachim Biercamp, Christopher S Bretherton, Xi Chen, Peter Düben, Falko Judt, Marat Khairstdinov, Daniel Klocke, et al. Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, 6(1):1–17, 2019.
- [57] B Teufel, F Carmo, L Sushama, L Sun, MN Khalil, S Bélair, A Shamseldin, D Nagesh Kumar, and J Vaze. Physics-informed deep learning framework to model intense precipitation events at super resolution. *Geoscience Letters*, 10(1):19, 2023.
- [58] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.
- [59] Emily Vosper, Peter Watson, Lucy Harris, Andrew McRae, Raul Santos-Rodriguez, Laurence Aitchison, and Dann Mitchell. Deep learning for downscaling tropical cyclone rainfall to hazard-relevant spatial scales. *Journal of Geophysical Research: Atmospheres*, page e2022JD038163, 2023.
- [60] Robert L Wilby, TML Wigley, D Conway, PD Jones, BC Hewitson, J Main, and DS Wilks. Statistical downscaling of general circulation model output: A comparison of methods. *Water resources research*, 34(11):2995–3008, 1998.
- [61] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *Internatiodal Conference on Learning Representations (ICLR)*, 2022.
- [62] Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In *International Conference on Machine Learning*, pages 42390–42402. PMLR, 2023.

Citation	Architecture	Resolutions	Pixels	Variables
Addison et al., (2022) [1]	Diffusion	input: 60km target: 8.8km	64^2	precipitation
Harris et al., (2022) [23]	GANs + ensemble forecast	input: 10km target: 1km	940^2	precipitation
Hatanaka et al., (2023) [24]	Cascaded diffusion	input: 30km target: 1km	128^2	day-ahead solar-irradiance
Leinonen et al., (2020) [32]	GANs	input 8km target: 1km	128^2	precipitation
Leinonen et al., (2020) [32]	GANs	input 16km target: 2km	128^2	cloud optical-thickness
Price and Rasp, (2022) [43]	Corrector GAN	input 32km target 4km	128^2	precipitation
Vosper et al., (2023) [59]	WGAN	input 100km target 10km	100^2	precipitation from tropical-cyclones
Current work	CorrDiff	input 25km target 2km	448^2	10 meter winds 2 meter temperature radar-reflectivity

Table S1: Downscaling models presented in the most relevant works we could find with respect to the current study. We highlight the resolution ratios, the pixel size of the high resolution prediction, predicted variables and architecture.

Supplementary Information

1 Our position with respect to existing works

To highlight the novel component of our work we expand our review. Table S1 above present a shortlist of the most relevant works that perform weather downscaling. Some of the successes of previous downscaling works are by performing state vector inflation as [1], on a relatively larger domain [23], with a large resolution ratio, e.g. [24, 1] and downscaling precipitation from tropical cyclones [59]. However, most of these related works focus on a single variables per model (note that [32] provided two models each per variable and is thus listed twice in the table). The variables of interest in all these works are related to properties of cloud and precipitation. While [59] showed a successful super resolution (recovering 10km from data coarsened to 100km) of tropical cyclone precipitation. To the best of our knowledge ML downscaling of tropical cyclones, which requires accounting for different physics, across many channels and channel synthesis was not shown before.

The combined prediction of selected dynamical, thermodynamical and microphysical (cloud related) variables in concert marks a new capability of such models. Its utility is demonstrated here by examining coherent structures and how all variables jointly downscaled in a physically consistent manner.

2 Descriptions of the architecture and the channels

Table S2 lists the input and output channels for the CorrDiff model. Note that the two differ both by pixel size and by the channels themselves. For single level variables, the input includes total column water vapor but lacks the maximum radar reflectivity which is present in the output, and vice versa. The input also includes pressure level variables at the 925, 850, 700 and 500 (hpa) levels.

	Input	Output
Pixel side	36 x 36	448 x 448
Single level channels	Total column water vapor Temperature at 2 meter Eastward wind at 10 meter Northward wind at 10 meter	Maximum radar reflectivity Temperature at 2 meter Eastward wind at 10 meter Northward wind at 10 meter
Pressure level channels	Temperature Geopotential Eastward wind Northward wind	

Table S2: A list of the input and output resolutions and channel for the CorrDiff downscaling model. Input channels include the both single level variables and pressure level variables, the latter are used at 925, 850, 700 and 500 (hpa) levels.

Figure S1 shows the architecture of the Unet in CorrDiff.

3 Localization by two-step formulation

Figure S2 (left) demonstrates the role of the two steps associated with CorrDiff as a function of spatial scales. From the target data (blue), it is seen that the regression step learns mostly large spatial scales, leaving the small scales almost completely for the diffusion step. In addition, from Fig S2 (right), it is observed that the residual is quite localized. It seems that to fully resolve the residual at each spatial location, only 40km radius around that is needed, while for the target data there is long-range correlations up to 200km. This has important implications for training and sampling efficiency of diffusion models since one can deploy diffusion models with smaller UNet desnoising architectures to aggregate the local information. We leave further study of this for future research.

4 Examining sample diversity of CorrDiff

In order to examine the realism of individual samples from CorrDiff and their quality compared with the target data figure shows an animation of the target data (left) Reg (middle) and 20 generated samples of the CorrDiff prediction of maximum radar reflectivity.

5 Comparison of simulated Typhoons with historical records

Although our out-of-sample data is limited to 2021, we can compare CorrDiff simulated typhoons in the CWA region with historical records. The Japan Meteorological Agency best track data (JMA tracks) [4] includes the maximum windspeed (intensity) and radius of maximum windspeed (size) of typhoons in west pacific for several decades. During the years 1980 to 2020 for (which we have ERA5 data available) and within the CWA domain, we identified 648 instances of typhoons with intensities of 30 m/s or greater in this time period. Panels (a) and (b) of figure S4 display the storm size and intensity, respectively, revealing the expected correction for ERA5 typhoons achieved through the application of CorrDiff. One imperfection of CorrDiff is that it shrinks all storms, including those with the correct size or those that were too small in the ERA5 input data (panel a). The main benefit is an improvement in the windspeeds, removing most of the error between the ERA5 and the observed records for windspeeds up to 50 m/s (panel b); stronger storms have room for improvement. CorrDiff generates a five-fold increase in the probability of hurricane-force winds exceeding 33 m/s (panel c). To the extent that JMA tracks can serve as ground truth, such distribution shift has significant societal implications as these low-probability, high-impact events represent a substantial portion of the overall risk, underscoring the importance of accurate modeling and forecasting.

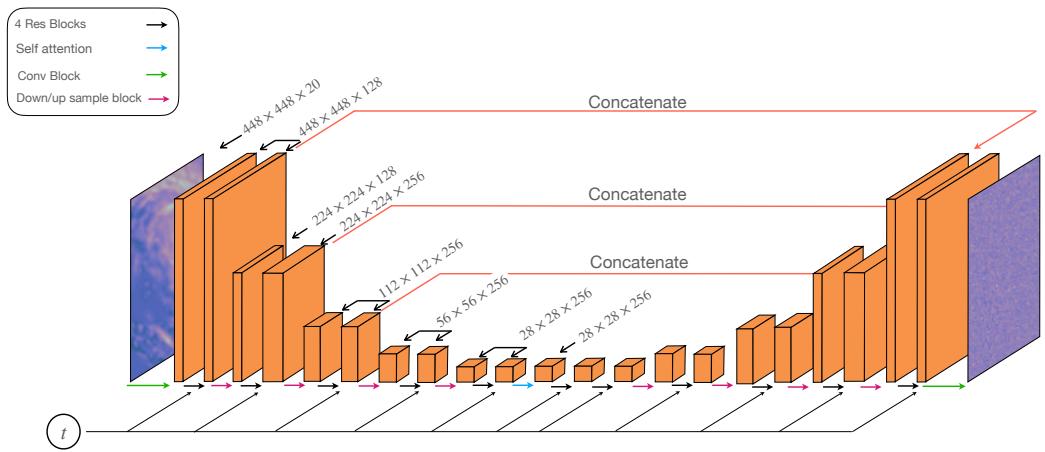


Figure S1: A sketch of the hierarchical UNet architecture adopted in both the regression model and the denoising diffusion model. Note, in the regression stage, time embedding is not used.

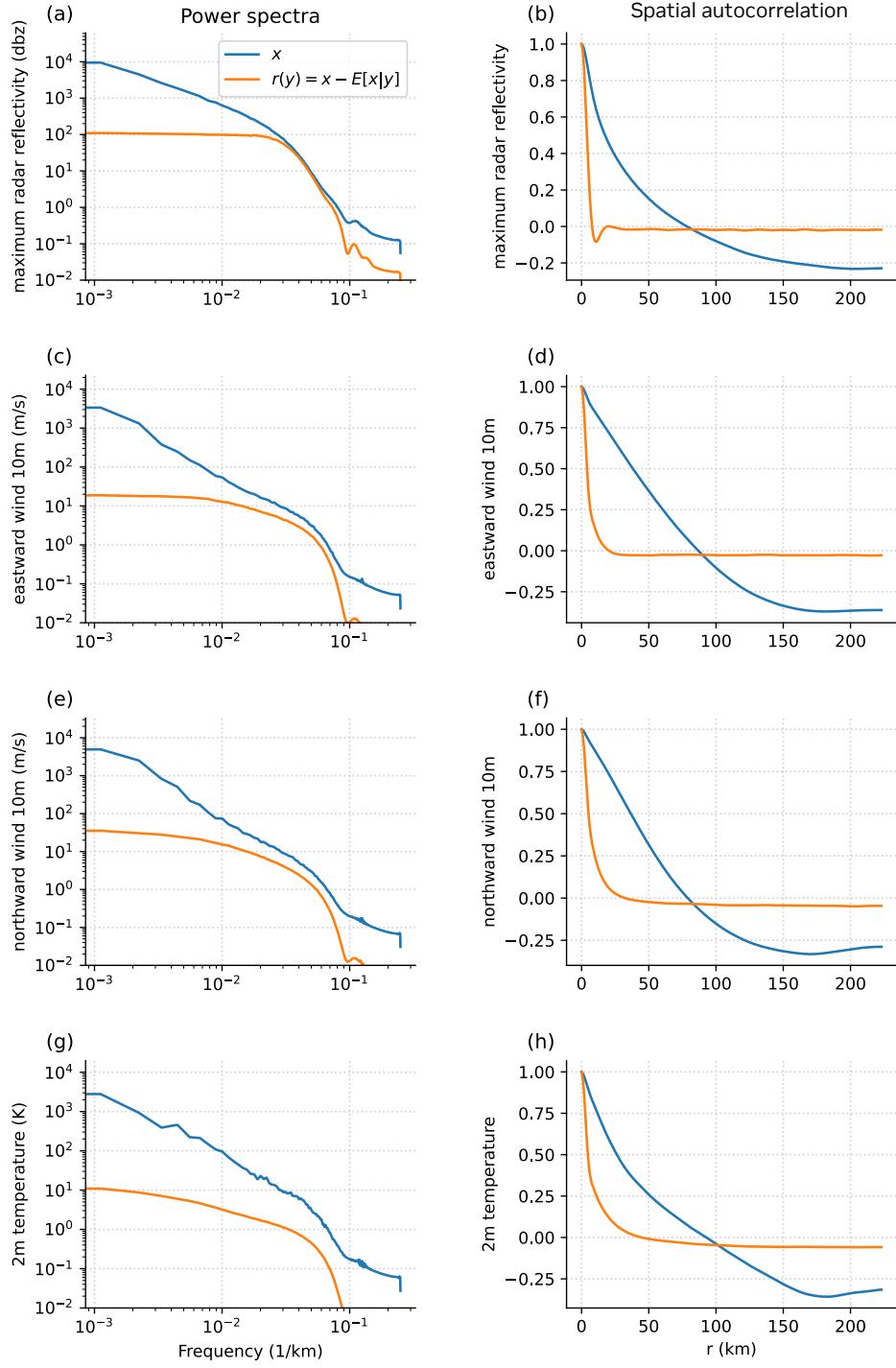


Figure S2: Left column: power spectra, right column: spatial auto-correlation. Top to bottom: maximum radar reflectivity, 10m eastward wind, 10m northward wind and 2m temperature. This figure compares the original target x and the difference $r = x - \mathbb{E}[x|y]$. The difference has greatly reduced variance at large-scales and equivalently removes the long-range auto-correlations.

Figure S3: Comparison of many samples of maximum radar reflectivity (right) with the prediction from the regression (Reg) and the truth (WRF) for several cloud regimes.

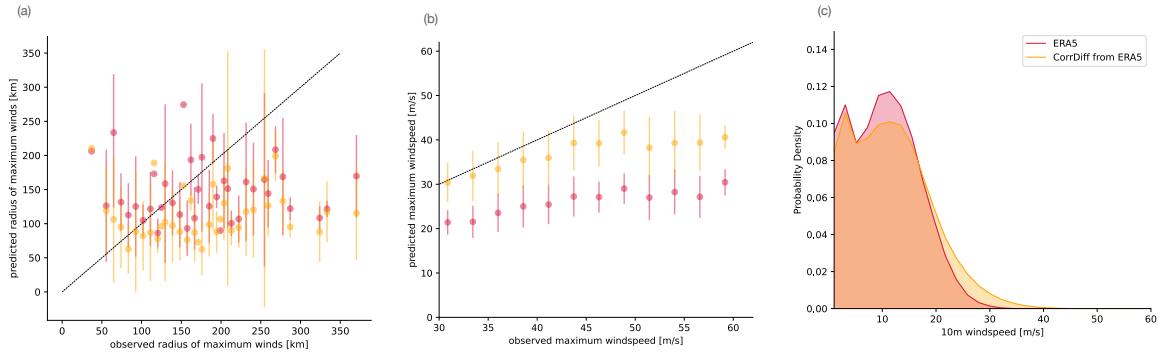


Figure S4: Comparison between ERA5, CorrDiff (from ERA5) and observed records for all typhoons found in the domain with windspeed of 30 m/s or more in the years 1980-2020. Panel (a): abscissa: mean (dot) and one standard deviation (bars) of predicted radius of maximum winds for the ERA5 (red) and CorrDiff (orange) storms, grouped by their observed values; ordinate: corresponding observed values. Panel (b) same as (a) but the maximum windspeed. Panel (c): The PDFs of all instances with typhoon of windspeed 30 m/s or greater in the years 1980-2020 in the CWA domain.

6 Energy efficiency and latency of CorrDiff downscaling inference compared to WRF simulations

Comparing the performance of a statistical downscaling like CorrDiff with a dynamical downscaling like the WRF-CWA is not straightforward. The former produces maps of a subset of channels at time $t + 1$ from a low resolution global forecast at $t + 1$, while the latter produces maps of the full state vector at time $t + 1$ from a low resolution global forecast at t using a numerical time-stepper where each time step is of the order of seconds.

Nonetheless, given that a global model like GFS is available, we can compare the compute time and energy required to obtain high resolution 1 hourly maps in Taiwan from the two approaches. Doing so, we compare the speed of CorrDiff against the operational WRF run by CWA on their respective hardware. The CWA-WRF is run on Fujitsu FX-100 system, where each node is equipped with 32 SPARC64 Xifx CPU cores. A 13 hour deterministic CWA-WRF forecast (excluding data assimilation) is run on 928 CPU cores (across 29 nodes with a maximum system memory of 6.9GB per node) and takes about 20 minutes. CorrDiff inference is run on a single NVIDIA RTX 6000 Ada Generation GPU, which takes 4.1 sec per downscaling sample. Given a global model 1 hour leadtime forecast, the CorrDiff statistical downscaling on a single GPU is about 22 times faster than the dynamic downscaling that runs CWA-WRF on 928 CPUs. Moreover, CorrDiff it is about 1130 times more energy efficient. We note that our current implementation of CorrDiff is far from optimized and does not utilize GPU parallelization and batching for generating many samples independently. A simple change to mixed-precision (float 16) gains almost another factor 2 speedup, without any degradation of the results, leading to about 38 \times speed up and about 1,800 \times energy efficiency compared with WRF-CWA. Furthermore, given global model data (which both systems need) CorrDiff can be run for the 13h on 13 GPUs obtaining about a 13x speedup for the 13h forecast used, over the above results (but with the same energy efficiency). We believe that GPU parallelization and batched inference will likely lead to even more significant speed-up and energy improvements.

	Hardware	Latency (sec/FH)	Power (J/sec)	Energy (kJ/FH)
WRF-CWA	928 CPUs	91.38	15.15	1285.46
CorrDiff, float32	1 GPU	4.1	300	1.23
CorrDiff, float16	1 GPU	2.4	300	0.72

Table S3: A comparison of running the WRF model on the CWA system with CorrDiff inference on a single NVIDIA RTX 6000 Ada Generation GPU. Latency is given per Forecast Hour (FH) and Power is given in Joule/sec (W) per a single hardware unit (a CPU or a GPU), while Energy is for the entire forecast system (928 CPU for WRF-CWA) per FH.