

DIFFUSION MODELING FOR HIGH-RESOLUTION WEATHER DOWNSCALING
OVER THE HAWAIIAN ISLANDS

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAII IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE
IN
ATMOSPHERIC SCIENCES

August 2025

By
Aleric R. Krenz

Thesis Committee:

Giuseppe Torri, Chairperson
Peter Sadowski
Pao-Shin Chu

Abstract

High-resolution weather simulations are essential for capturing the complex terrain-driven processes that shape precipitation and wind patterns across the Hawaiian Islands. However, dynamical downscaling using models like WRF at kilometer-scale resolution is computationally expensive. In this study, we explore a generative machine learning approach using a probabilistic score-based diffusion model to statistically downscale ERA5 reanalysis data from 27 km to 1.5 km resolution. The model is trained on hourly ERA5 reanalysis from 2002 to 2009 with WRF simulations as the target, and tested on the 2010–2012 period. We evaluate its performance across key surface variables, including 2-meter temperature, 10-meter wind components, and accumulated precipitation. Evaluation metrics include CRPS, power spectrum analysis, PDF comparisons, and case studies. The model performs well in recreating mesoscale features induced by the topography of the Hawaiian Islands, but struggles to fully capture variability driven by large-scale synoptic forcing. Once trained, the diffusion model produces a high-resolution downscale ensemble in seconds, offering a computationally efficient alternative to dynamical models.

Table of Contents

| | |
|---|----|
| Abstract | ii |
| List of Tables | iv |
| List of Figures | v |
| Chapter 1: Introduction | 1 |
| Chapter 2: Study Area and Datasets | 5 |
| 2.1 Study area and target domains | 6 |
| 2.2 Datasets | 7 |
| 2.2.1 Input Training Datasets for ML Models | 7 |
| 2.2.2 Testing Datasets | 8 |
| Chapter 3: Methods | 10 |
| 3.1 Modeling Methods | 10 |
| 3.1.1 Corrected Diffusion Model | 10 |
| 3.1.2 Inference and Sample Generation | 12 |
| 3.1.3 Control Models | 14 |
| 3.2 Machine Learning Task Setup | 15 |
| 3.3 Testing Parameters | 16 |
| 3.3.1 Testing Parameters: Probability Density Functions | 16 |
| 3.3.2 Testing Parameters: Power Spectrum Analysis | 17 |

| | | |
|------------|---|----|
| 3.3.3 | Testing Parameters: Continuous Ranked Probability score | 19 |
| Chapter 4: | Results | 21 |
| 4.1 | Log Probability | 21 |
| 4.1.1 | Power Spectrum Analysis | 23 |
| 4.2 | Continuous Ranked Probability Score | 25 |
| Chapter 5: | Discussion and Conclusions | 29 |
| 5.1 | Discussion | 29 |
| 5.2 | Conclusions | 31 |

List of Tables

| | | |
|-----|---|----|
| 2.1 | Specifications for each downscaling domain over the Hawaiian Islands. . . . | 5 |
| 2.2 | Summary of variables used from WRF (target) and ERA5 (predictor) datasets. . . . | 5 |
| 3.1 | Training and testing dataset configuration. | 15 |
| 3.2 | Spatial configuration of ERA5 inputs and WRF outputs per domain. | 16 |
| 3.3 | Output pixel counts per timestep by domain. | 16 |
| 4.1 | Comparison of Mean Absolute Error (MAE) and Continuous Ranked Probability Score (CRPS) for Hawai'i County | 26 |
| 4.2 | Comparison of Mean Absolute Error (MAE) and Continuous Ranked Probability Score (CRPS) for Maui County | 26 |
| 4.3 | Comparison of Mean Absolute Error (MAE) and Continuous Ranked Probability Score (CRPS) for Honolulu County | 26 |
| 4.4 | Comparison of Mean Absolute Error (MAE) and Continuous Ranked Probability Score (CRPS) for Kaua'i County | 27 |

List of Figures

| | | |
|-----|--|----|
| 2.1 | Island County Domains represented in Table 2.1 | 6 |
| 3.1 | UNet Regression Step Schematic | 13 |
| 3.2 | UNet Denoising Step Schematic | 14 |
| 3.3 | ERA5 Bilinear Interpolation Step for coarse control model | 14 |
| 4.1 | Log-PDF comparisons for four surface variables (T-2, U-10, V-10, and 1-hour precipitation accumulation) during the 2009–2010 wet season in Hawai‘i County. The diffusion model (green) captures both the central tendency and the tails of the WRF distribution (black) more accurately than ERA5 (light blue dashed) and UNet (dark blue dotted). | 22 |
| 4.2 | Power spectra of surface variables over the Hawai‘i County domain averaged over the full testing dataset. Spectra are shown for WRF target (black), diffusion model (green), UNet regression (dark blue dotted), and interpolated ERA5 (light blue dashed). | 24 |

5.1 Temporal evolution of CRPS (diffusion model) and corresponding WRF precipitation across the Kaua'i domain. The plot illustrates seasonal variation in model performance, with CRPS increasing during wetter periods when forecast uncertainty and rainfall variability are highest. The temporal correlation also reveals that forecast skill degrades notably during high precipitation events, highlighting a potential shortcoming in ensemble spread during these periods. 30

Chapter 1

Introduction

High-resolution atmospheric data is essential for advancing many areas of geosciences in the Hawaiian Islands [3]. This includes regional forecasting and climate impact assessments. However, generating such data is often limited by the high computational cost of running physics-based weather models at fine spatial scales [1][2]. A widely used strategy to address this challenge is *downscaling*, which refers to the process of deriving high-resolution information from coarser-scale data. This is especially useful because global atmospheric models usually operate at resolutions that are too coarse for capturing localized effects like orographic rainfall or coastal winds.

There are two general types of downscaling methods: *dynamical* and *statistical*. Dynamical downscaling uses a high-resolution regional climate model that is driven by coarse global model output. While physically robust, this approach is computationally expensive and limited to small domains. In contrast, statistical downscaling learns relationships between large-scale predictors and local responses, offering a more flexible and less expensive alternative.

Statistical downscaling methods are typically categorized into three main types: weather classification schemes, regression-based models, and weather generators [1].

Among these, regression models are the most widely used due to their straightforward implementation, mapping large-scale predictors directly to local-scale variables. However, such models often underestimate variability and extreme events, and frequently fail to capture important relationships with static features such as terrain. These limitations have been highlighted in previous studies on statistical downscaling of precipitation in Hawai‘i [3].

A persistent challenge across traditional statistical and dynamical downscaling approaches is the accurate representation of extreme events. These events, which reside in the tails of probability distributions, are often inadequately captured by methods that prioritize mean conditions [3]. To address this limitation, deterministic machine learning methods such as neural networks have been introduced for their ability to model nonlinear relationships. For example, 6 km wind speeds over Hawai‘i have been skillfully downscaled using neural networks [13]. While similar efforts have been made for precipitation [13][14], success has been limited. Precipitation exhibits highly nonlinear relationships with other atmospheric variables and remains difficult to model using deterministic neural networks alone. These models often struggle to generate spatially consistent outputs and perform poorly when trained on small or imbalanced datasets.

To overcome these limitations, researchers have recently turned to deep generative models. One promising class of models is *diffusion models*, which can learn complex and high-dimensional data distributions [4]. These models work by gradually adding noise to data in a forward process and then learning to reverse that process with a neural network. This framework is inspired by thermodynamics and noise-driven processes [5]. Early diffusion models required many iterations and were computationally slow, but recent advancements have made them more efficient.

The strength of diffusion models is their ability to generate probabilistic outputs that match the data distribution. This makes them well-suited for applications in weather and climate, where uncertainty is important. A more advanced type, called *score-based diffusion models*, replaces the Markov noise process with stochastic differential equations (SDEs) [6]. The reverse process is guided by a score function, which is the gradient of the logarithmic probability density, directing samples toward more likely regions in the data space[7].

Despite the growing popularity of generative models, their application to downscaling remains relatively limited. Score-based diffusion models have demonstrated the ability to generate high-resolution, probabilistic forecasts of cloud cover in Hawai'i by super-resolving coarse model outputs, offering a promising example of their use locally [8]. Another recent development is the *Corrected Diffusion* framework by Mardani et al.(2025)[9], which separates the generation task into two parts. First, a UNet model predicts the mean, and then a diffusion model generates the residuals. This approach reduces the complexity of the learning task and better captures fine-scale features like fronts and rainbands.

This thesis builds on the corrected diffusion framework and applies it to the Hawaiian Islands, a region with complex terrain and unique weather patterns. Unlike earlier studies that only used input data below 500 hPa, this work includes predictors up to 200 hPa. This allows the model to capture upper-level features such as cut-off lows and jet interactions, which are important for rainfall events in Hawaii [10][11]. The high-resolution target data comes from a 10-year WRF simulation over the islands[12], and the input predictors are from ERA5 reanalysis[15].

This study also includes high-resolution terrain as an input to help the model capture orographic effects. The model generates four surface variables: 2-meter temperature, 10-

meter U and V wind components, and precipitation. The goal is to create a scalable, efficient, and probabilistic alternative to traditional dynamical downscaling for Hawaii.

Chapter 2

Study Area and Datasets

Table 2.1 Specifications for each downscaling domain over the Hawaiian Islands.

| Domain | Islands Covered | Lat Range | Lon Range |
|-----------------|------------------------|---------------|-----------------|
| Hawai'i County | Hawai'i | 18.74–20.47°N | 154.51–156.35°W |
| Maui County | Maui, Moloka'i, Lana'i | 20.38–21.23°N | 157.57–155.73°W |
| Honolulu County | O'ahu | 21.05–21.91°N | 158.45–157.52°W |
| Kaua'i County | Kaua'i | 21.58–22.44°N | 159.99–159.06°W |

Table 2.2 Summary of variables used from WRF (target) and ERA5 (predictor) datasets.

| Dataset | Variables Used |
|-------------------------|--|
| WRF (Target) | 2-m temperature (T-2), 10-m wind components (U-10, V-10), hourly accumulated precipitation, terrain height |
| ERA5 (Predictor) | <i>Pressure-level:</i> U, V, T, Φ , ω at 1000, 950, 850, 700, 500, 300, 200 hPa <i>Surface-level:</i> 2-m temperature (T-2), 10-m winds (U-10, V-10), total column water vapor (TCWV) |

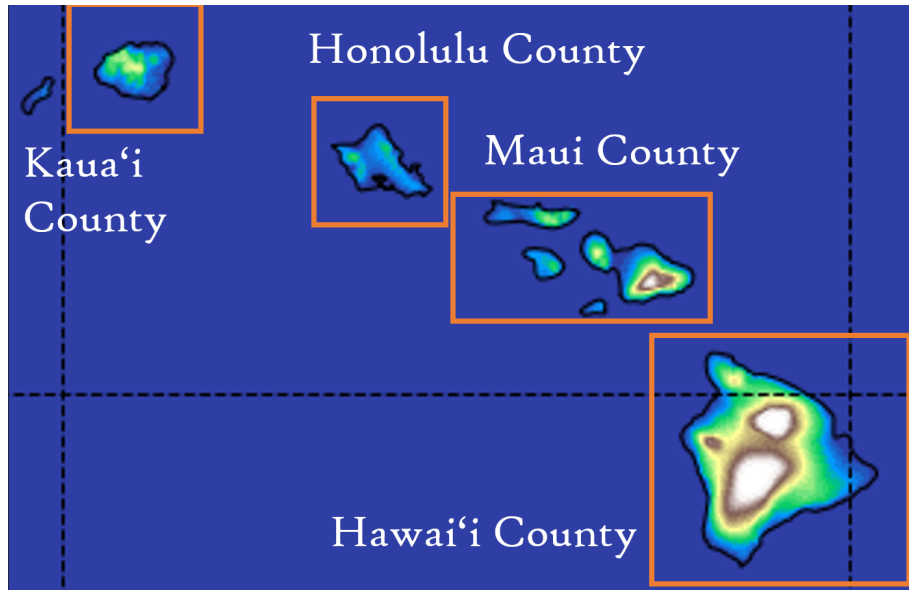


Figure 2.1 Island County Domains represented in Table 2.1

2.1 Study area and target domains

This study focuses on downscaling atmospheric variables over the Hawaiian Islands using the *Corrected Diffusion* model framework. Due to computational constraints associated with high-resolution model training, the state of Hawaii was subdivided into four target domains, corresponding to the four largest counties:

- Hawai'i County domain (Hawai'i)
- Maui County domain, (Maui, Moloka'i, Lana'i, and Kahoolawe)
- Honolulu County domain (O'ahu)
- Kaua'i County domain (Kaua'i)

Each domain was spatially cropped from the full simulation to focus training and evaluation on distinct topographic and meteorological regimes. The approximate latitude and

longitude extents for each domain are shown in Figure 2.1, and the full dimensions are detailed in Table 2.1. These cropped domains correspond to image patch sizes of 128×128 pixels for Hawaii County, 128×64 for Maui County, and 64×64 for both Honolulu and Kauai Counties, based on the native WRF resolution of 1.5 km.

2.2 Datasets

2.2.1 Input Training Datasets for ML Models

Two primary datasets were used to train the diffusion-based statistical downscaling model: a high-resolution simulation from the Weather Research and Forecasting (WRF) model, and coarse-resolution reanalysis from ERA5.

High-resolution WRF dataset (Target Variables) The high-resolution target dataset is derived from a decade-long historical simulation using WRF, spanning from October 2002 to September 2012[12]. This simulation was forced by ERA-Interim boundary conditions and observed sea surface temperatures, and it was configured with a high-resolution vertical coordinate system designed to capture the trade wind inversion. The trade wind inversion is an important feature in Hawaiian meteorology. Properly resolving this inversion helps constrain cloud-top heights and improves precipitation pattern representation.

The WRF output features a horizontal grid spacing of 1.5 km, with 540 grid points in the zonal direction and 360 grid points in the meridional direction. The model also includes 81 vertical levels and an hourly temporal resolution. From this dataset, the following near-surface variables were selected for downscaling: 2-m temperature (T-2), 10-m zonal wind (U-10), 10-m meridional wind (V-10), and hourly accumulated precipitation.

Additionally, 1.5km resolution terrain height from the WRF static fields was included as a fixed topographic input.

Data from WRF was cropped to match the spatial extent of each county-level domain seen in Figure 2.1 and Table 2.1. These high-resolution target fields served as the ground truth for training the model.

ERA5 reanalysis dataset (Predictor Variables) Coarse-resolution predictor fields were extracted from the ERA5 reanalysis, which provides hourly global atmospheric data from 1940 to near-real time. ERA5 assimilates a broad range of upper-air and surface observations and includes a fully coupled land-surface and wave model[15].

From ERA5, six pressure levels (200, 300, 500, 700, 850, 950, and 1000 hPa) were selected. At each level, the variables used were zonal wind (U), meridional wind (V), temperature (T), geopotential height (Φ) and Vertical velocity (ω). Additionally, surface variables such as 2-m temperature (T-2), 10-m winds (U-10, V-10), and total column water vapor (TCWV) were used.

The selection of predictor variables was informed through consultation with operational meteorologists at the National Weather Service Forecast Office in Honolulu. All ERA5 fields were extracted and interpolated to match the spatial domains described above.

The training dataset consisted of seven years of overlapping WRF and ERA5 data from October 2002 through September 2009.

2.2.2 Testing Datasets

The final three years of data (October 2009 through September 2012) were withheld as an independent testing dataset. This test period included the same set of variables as the training set for both ERA5 and WRF. During testing, ERA5 variables served as inputs to

the trained diffusion model, and the resulting downscaled outputs were compared against WRF data for evaluation.

Chapter 3

Methods

3.1 Modeling Methods

3.1.1 Corrected Diffusion Model

The goal of probabilistic statistical downscaling is to estimate the conditional distribution $p(x \mid y)$, where x in our case denotes high-resolution data and y represents the corresponding low-resolution inputs. To model this relationship, we adopt the *corrected diffusion* (CorrDiff) model framework, which extends score-based diffusion models to improve performance on structured geophysical data[9].

Diffusion models operate by defining a forward and reverse process. In the forward (diffusion) process, Gaussian noise is incrementally added to the target data x over a series of steps until the sample becomes indistinguishable from white noise. The reverse (denoising) process attempts to reconstruct x by learning a series of conditional distributions that gradually remove the noise. This reverse process is learned using a neural network trained to predict the clean signal from noisy inputs at each step.

However, directly learning $p(x \mid y)$ using standard diffusion techniques can be challenging in the context of meteorological data. The presence of static biases, sharp

gradients, multi-scale structures, and significant distributional shifts between training and testing data often results in poor convergence and suboptimal sample quality.

To address these issues, the CorrDiff framework decomposes the learning task into two components:

1. A deterministic prediction of the conditional mean $\mu = \mathbb{E}[x | y]$,
2. A stochastic residual $r = x - \mu$ modeled via diffusion.

The conditional mean is estimated using a UNet-based convolutional neural network (CNN), which provides a coarse but physically consistent approximation of the high-resolution field. UNets are particularly well-suited to spatial prediction tasks due to their ability to capture both local and global context through downsampling and upsampling operations with skip connections[9].

The residual r is then modeled using a diffusion process conditioned on both y and the predicted mean μ . This second branch refines the deterministic prediction by adding spatially coherent stochasticity, allowing the final prediction to better match the distribution of the training data. The full reconstruction is then expressed as:

$$x = \mathbb{E}[x | y] + r = \mu + r$$

This two-branch approach—combining a deterministic mean prediction with a probabilistic residual correction—enables more efficient training, faster convergence, and improved fidelity in downscaled fields compared to traditional conditional diffusion models.

3.1.2 Inference and Sample Generation

Once training is complete, the corrected diffusion model is used to generate high-resolution atmospheric fields from new low-resolution ERA5 inputs. The generation process is designed to mirror training but operates without loss computation, and it consists of two primary stages: a deterministic regression step and a stochastic denoising step.

1. UNet-Regression Step. The first stage involves predicting the conditional mean $\mu = \mathbb{E}[x | y]$, where y is the low-resolution ERA5 input and x is the desired high-resolution output (2-m temperature, U-10, V-10, or precipitation). This conditional mean is estimated using a UNet-based convolutional neural network that takes as input the ERA5 fields and high-resolution static features such as terrain height. ERA5 fields are ingested at the lowest level of the UNet.

This stage provides a "middle-resolution" forecast that serves as a physically informed and spatially coherent baseline for the second stage. The UNet's architecture allows it to ingest multi-scale information and output high-resolution features aligned with terrain and mesoscale dynamics. This stage is entirely deterministic and produces a consistent estimate given a fixed input. The UNet-Regression step is detailed in Figure 3.1.

2. UNet-Denoising Step. The second stage is a diffusion-based generative process that models the residual $r = x - \mu$. It refines the regression output by iteratively de-noising a sequence of noisy latent representations, conditioned on both the ERA5 input and the predicted mean μ .

During inference, this step is initialized with Gaussian noise and proceeds through a fixed number of reverse diffusion steps, progressively transforming noise into a realistic high-resolution residual. Each step of this denoising process is handled by another UNet, which is conditioned on ERA5 data, the terrain field, and the current estimate of the

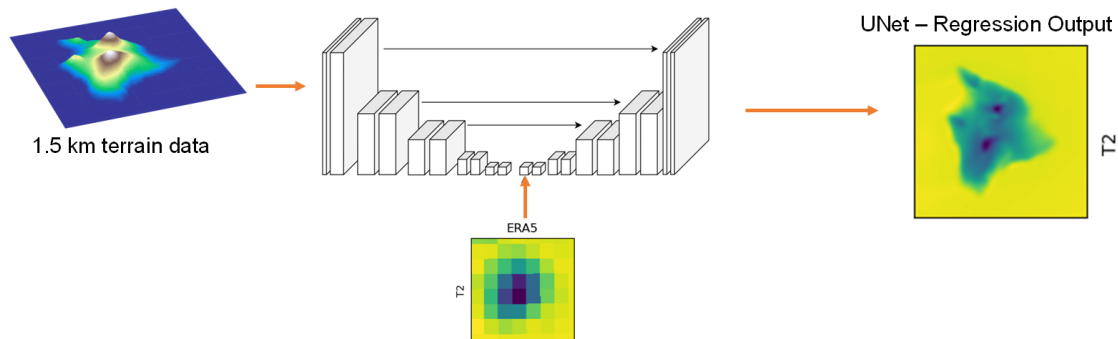


Figure 3.1 UNet Regression Step Schematic

sample. The output from this step is the final residual sample r , which is added to the predicted mean:

$$x = \mu + r$$

The full pipeline thus produces a stochastic, physically realistic high-resolution sample that reflects both large-scale structure (from the regression step) and small-scale variability (from denoising step). The UNet-Denoising step is detailed in Figure 3.2.

3. Sample Diversity and Uncertainty Quantification. Unlike deterministic downscaling approaches, this framework enables the generation of multiple plausible realizations for a single input. In this study, 64 unique samples were generated per input condition using different initial noise conditions. These ensemble-like outputs allow for uncertainty quantification and probabilistic evaluation of downscaled forecasts, including metrics like as spatial CRPS.

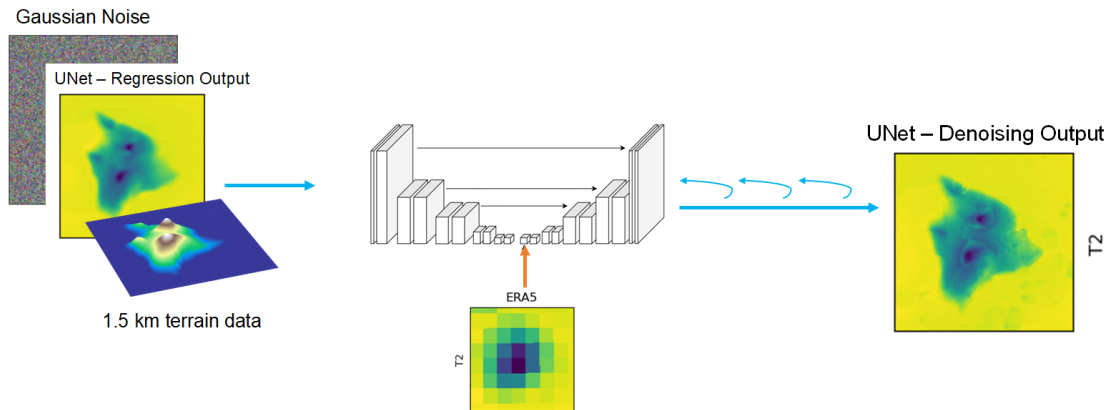


Figure 3.2 UNet Denoising Step Schematic

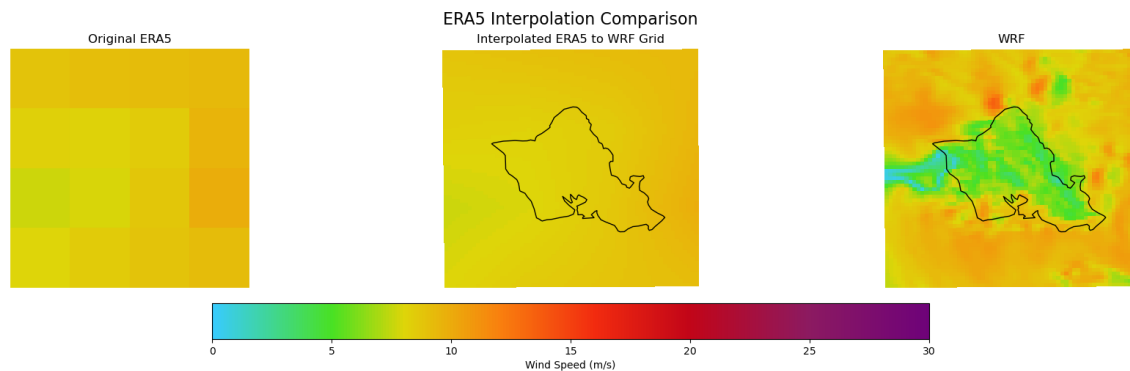


Figure 3.3 ERA5 Bilinear Interpolation Step for coarse control model

3.1.3 Control Models

We use two baselines for comparison: interpolation of the ERA5 data and the regression step of CorrDiff (UNet). Bilinear interpolation was applied to three of the ERA5 surface variables for tests as a control (T-2, U-10, and V10). We did not use precipitation accumulation in the ERA5 dataset, therefore there is no interpolated ERA5 model data in the results for the interpolated ERA5 model. An example of the bilinear interpolation of ERA5 can be seen in Figure 3.3.

3.2 Machine Learning Task Setup

We trained a supervised learning model to downscale coarse ERA5 reanalysis to high-resolution WRF simulations over four Hawaiian Island domains. The dataset was split into a 7-year training period (October 2002–September 2009, 61,368 hours) and a 3-year testing period (October 2009–September 2012, 26,304 hours). Each input sample consisted of 47 channels (seven ERA5 pressure levels, ERA5 surface variables, and static terrain), while the outputs were four WRF surface variables. Spatial coverage varied by domain, from 8×8 ERA5 pixels to 128×128 WRF pixels for Hawai‘i County, down to 4×4 ERA5 pixels to 64×64 WRF pixels for Honolulu and Kaua‘i Counties. For a single diffusion model samples, this configuration yields a total of 131,072 high-resolution output pixels per timestep across all domains. The learning task therefore required the model to reconstruct fine-scale meteorological structure from coarse inputs, preserving the detailed spatial and topographic patterns present in the WRF simulations. Tabular form of the above data can be seen in Tables 3.1 - 3.3.

Table 3.1 Training and testing dataset configuration.

| | Training | Testing |
|-------------------------------|---|---|
| Date Split (10 Years) | Oct '02 – Sep '09 (7 years) | Oct '09 – Sep '12 (3 years) |
| Timestep Count (87,672 total) | 61,368 hours | 26,304 hours |
| Input Channels | 7 pressure levels + surface ERA5 + terrain (47) | 7 pressure levels + surface ERA5 + terrain (47) |
| Output Channels | None | Surface WRF (4) |

Table 3.2 Spatial configuration of ERA5 inputs and WRF outputs per domain.

| Domain | Input ERA5 Pixels | Output WRF Pixels |
|-----------------|-------------------|-------------------|
| Hawai'i County | 8×8 | 128×128 |
| Maui County | 8×4 | 128×64 |
| Honolulu County | 4×4 | 64×64 |
| Kaua'i County | 4×4 | 64×64 |

Table 3.3 Output pixel counts per timestep by domain.

| Domain | Output Per Timestep |
|-----------------|--|
| Hawai'i County | $128 \times 128 \times 4$ channels = 65,536 pixels |
| Maui County | $128 \times 64 \times 4$ channels = 32,768 pixels |
| Honolulu County | $64 \times 64 \times 4$ channels = 16,384 pixels |
| Kaua'i County | $64 \times 64 \times 4$ channels = 16,384 pixels |
| Total | 131,072 pixels |

3.3 Testing Parameters

3.3.1 Testing Parameters: Probability Density Functions

To evaluate how well the full distribution of predicted values aligns with the true data distribution we compute and compare the **probability density functions (PDFs)** of predicted and observed values. This approach allows us to evaluate whether the diffusion model is capturing the *shape, spread, and tails* of the conditional distribution $p(x | y)$, where x represents the high-resolution WRF field and y is the corresponding coarse-resolution ERA5 input.

For each variable and region, we aggregate values from WRF and from the diffusion model predictions over a given season (wet or dry). We define the dry season to be from May through October and the wet season from November through April. We compute their empirical PDFs using kernel density estimation (KDE)[17]. For the diffusion model, we compute the PDF of the ensemble-averaged output to match the form used in CRPS

evaluation. We also include baseline PDFs for ERA5 (interpolated to WRF resolution) and the UNet regression model for comparison.

The PDF comparison provides a visual test of how well the diffusion model learns the conditional distribution $p(x | y)$. A good match between the predicted and WRF PDFs, especially in the core and tails, indicates whether or not the model captures the central tendency but also the uncertainty and rare events present in the WRF reference data. Differences in the PDF shape can highlight underdispersion, overdispersion, or structural biases in the generative model.

To improve visibility of tail behavior and highlight discrepancies between models, we plot the logarithm of the estimated PDFs (*log-PDFs*) rather than the raw densities. This transformation accentuates low-probability events and makes it easier to assess whether models properly represent rare but important extremes, such as heavy rainfall or high wind events. In cases where ERA5 assign negligible probability to the extremes, we apply upper and lower value cutoffs to reduce visual distortion and ensure that well-calibrated distributions remain interpretable.

In summary, the PDF analysis provides a distributional perspective that complements the CRPS score. Whereas CRPS evaluates the forecast at each point in space and time, the PDF comparison assesses whether the model learns to generate physically plausible fields whose statistical behavior matches that of the WRF training data.

3.3.2 Testing Parameters: Power Spectrum Analysis

To evaluate the ability of the diffusion model to reproduce realistic spatial structure across scales, we conduct a **power spectrum analysis** of the predicted and reference fields. This technique quantifies the distribution of variance (or “energy”) across spatial frequencies and provides a scale-aware diagnostic of model fidelity[18].

For each variable and timestep, we compute the 2D discrete Fourier transform (DFT) of the WRF field and of the ensemble-mean diffusion output. The squared magnitude of the Fourier coefficients yields the *power spectral density* (PSD), which we then average in spectral space to obtain a 1D power spectrum as a function of spatial wavenumber k . This spectrum describes how variance is distributed across spatial scales, from mesoscale (low k) to submesoscale (high k).

To mitigate sampling noise and highlight systematic behavior, we compute average power spectra across all timesteps in a given evaluation period. For each region and variable, we compare the power spectrum of:

- The high-resolution WRF ground truth (reference),
- The diffusion model (ensemble mean),
- The UNet regression baseline,
- ERA5 interpolated to WRF grid (coarse control).

Spectral comparisons offer insights that pointwise metrics like CRPS or RMSE cannot. For example, underestimation of power at high wavenumbers indicates spatial smoothing and loss of fine-scale structure, while overestimation may reflect artificial noise. A good match between the diffusion model and WRF spectra across the full range of k suggests that the model faithfully reconstructs both large-scale dynamics and small-scale variability.

In our plots, we use log-log scaling to emphasize behavior at both ends of the frequency spectrum. In summary, the power spectrum analysis provides a robust diagnostic for spatial realism. It complements the PDF and CRPS metrics by explicitly evaluating whether the generative model reproduces the *scale-dependent structure* of high-resolution atmospheric fields.

3.3.3 Testing Parameters: Continuous Ranked Probability score

The **Continuous Ranked Probability Score (CRPS)** is a widely used verification metric for evaluating *probabilistic forecasts* of continuous variables. It measures the difference between the forecast *cumulative distribution function (CDF)* and the observed value, generalizing the mean absolute error (MAE) to the probabilistic context.

Formally, for a forecast CDF F and an observation x , the CRPS is defined as:

$$\text{CRPS}(F, x) = \int_{-\infty}^{\infty} [F(y) - 1(y \geq x)]^2 dy \quad (3.1)$$

where $1(y \geq x)$ is the Heaviside step function, which represents the CDF of the observed value. In simpler terms, the CRPS quantifies the integrated squared difference between the forecast distribution and a step function centered at the observation.

Following the formulation by [16], the continuous ranked probability score (CRPS) is defined as

$$\text{CRPS}(F, x) = \mathbb{E}|X - x| - \frac{1}{2}\mathbb{E}|X - X^*|, \quad (3.2)$$

where F is the forecast cumulative distribution, x is the verifying observation, and X , X^* are independent random draws from F . In practice, when using ensemble forecasts $\{x_1, \dots, x_N\}$, we approximate this expression empirically as:

$$\text{CRPS} \approx \frac{1}{N} \sum_{i=1}^N |x_i - x| - \frac{1}{2N^2} \sum_{i=1}^N \sum_{j=1}^N |x_i - x_j|, \quad (3.3)$$

which decomposes into a mean absolute error term and a correction for ensemble spread. This approximation is computed at each grid cell by treating the ensemble as samples from the forecast distribution and evaluating the CRPS as a function of ensemble distance to the truth and internal consistency.

A lower CRPS indicates a *more accurate and sharper* forecast distribution. Unlike deterministic metrics such as RMSE or MAE, CRPS rewards both the *accuracy* and the *reliability* of the forecast. This makes CRPS particularly useful for evaluating ensemble weather prediction systems and data-driven generative models such as diffusion models.

In this study, CRPS is computed beyond just a single score for the diffusion model: as a **time series**.

CRPS Time Series: To capture temporal variations in ensemble skill, we compute the spatially averaged CRPS at each timestep. This results in a time series of CRPS values that highlights how forecast quality varies across different synoptic regimes or seasons. This metric provides a nuanced view of model performance and quantify both where and when ensemble predictions align most closely with high-resolution WRF truth and can identify specific events or periods of systematic model bias.

Chapter 4

Results

To diagnose the skill and performance of the adapted CorrDiff model, we compare an ensemble of UNet-Denoising outputs, hereafter referred to as the *Diffusion Model* with the WRF target, ERA5 interpolated, and the UNet-Regression branch hereafter referred to as the *UNet*. The ERA5 interpolated and the UNet branch serve as examples of baseline deterministic models. The tests that we performed were the continuous ranked probability score on the diffusion outputs versus the mean absolute error of the ERA5 interpolated and UNET, a power spectrum on all of the models, and a logarithmic distribution comparison. Each of these tests were done for each of the target variables.

4.1 Log Probability

To assess the ability of the diffusion model to reproduce the full distribution of target variables, we compare the log-probability density functions (Log-PDFs) of WRF truth, diffusion model samples, UNet regression, and interpolated ERA5. The log-PDF comparison emphasizes differences in the tails of the distributions, which are critical for capturing rare but impactful events.

Figure 4.1 shows the log-PDFs for all four surface variables during the 2009–2010 wet season in the Hawai‘i County domain. This period (November–April) exhibits the greatest variability in precipitation, wind speed, and temperature, making it the most informative for assessing model skill. Each panel aggregates data across all time steps and grid points in the domain, with kernel density estimation used to compute shifted log-PDFs. For brevity, plots from other seasons are omitted, as they exhibit similar qualitative behavior, with the wet season showing the most variance.

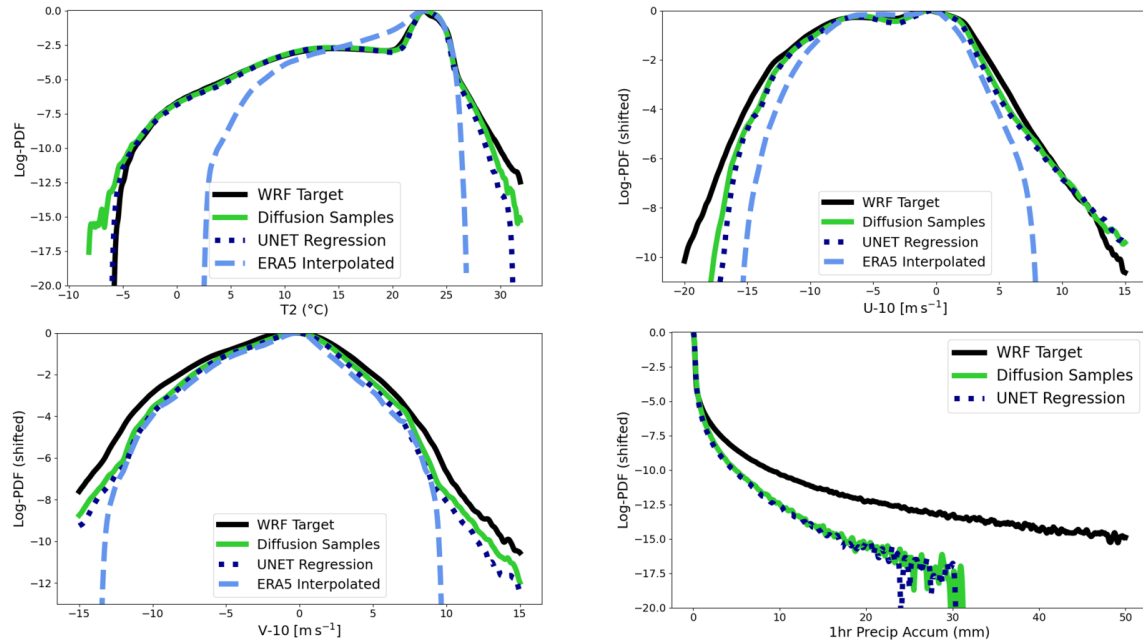


Figure 4.1 Log-PDF comparisons for four surface variables (T-2, U-10, V-10, and 1-hour precipitation accumulation) during the 2009–2010 wet season in Hawai‘i County. The diffusion model (green) captures both the central tendency and the tails of the WRF distribution (black) more accurately than ERA5 (light blue dashed) and UNet (dark blue dotted).

The diffusion model closely tracks the shape of the WRF distribution across all variables with some distinctions:

- **T-2 (2-meter temperature):** The diffusion model accurately captures both the mode and the spread of the distribution but seems to overestimate and underestimate the cool and warm ends of the spectrum respectively. ERA5 interpolated as expected underestimates cold extremes. The UNet keeps up well but also falls off at the far warm end, but tracks WRF in the cool end of the tail.
- **U-10 and V-10 (10-meter winds):** Both ERA5 and UNet show narrower distributions, under-representing high wind values. The diffusion model maintains better agreement with WRF in both directions, but still underestimates at the farthest tails of the distributions.
- **Precipitation:** The diffusion model and UNet both capture the exponential shape of the precipitation distribution, but both fail to retain higher density in the lower tail.

These results suggest that the corrected diffusion model produces ensembles with better calibrated uncertainty and more realistic extremes for temperature and wind but seems to struggle with the highest rainfall. The deterministic UNet also performs well in the same way. In contrast, interpolated ERA5 exhibits the most over-smoothed distributions, particularly in tail regions.

4.1.1 Power Spectrum Analysis

To assess how well the models capture spatial structure across scales, we compute the 2D power spectra for each variable across the full testing period and average the spectra over wavenumber bins. This yields a 1D power spectrum that reflects the distribution of spatial variance from mesoscale to sub-mesoscale.

Figure 4.2 shows the average power spectra for all four surface variables T-2, U-10, V-10, and hourly precipitation over the Hawai'i County domain. These spectra are computed using the full testing dataset from October 2009 to September 2012. For each variable, the diffusion model, UNet, interpolated ERA5, and WRF target spectra are compared.

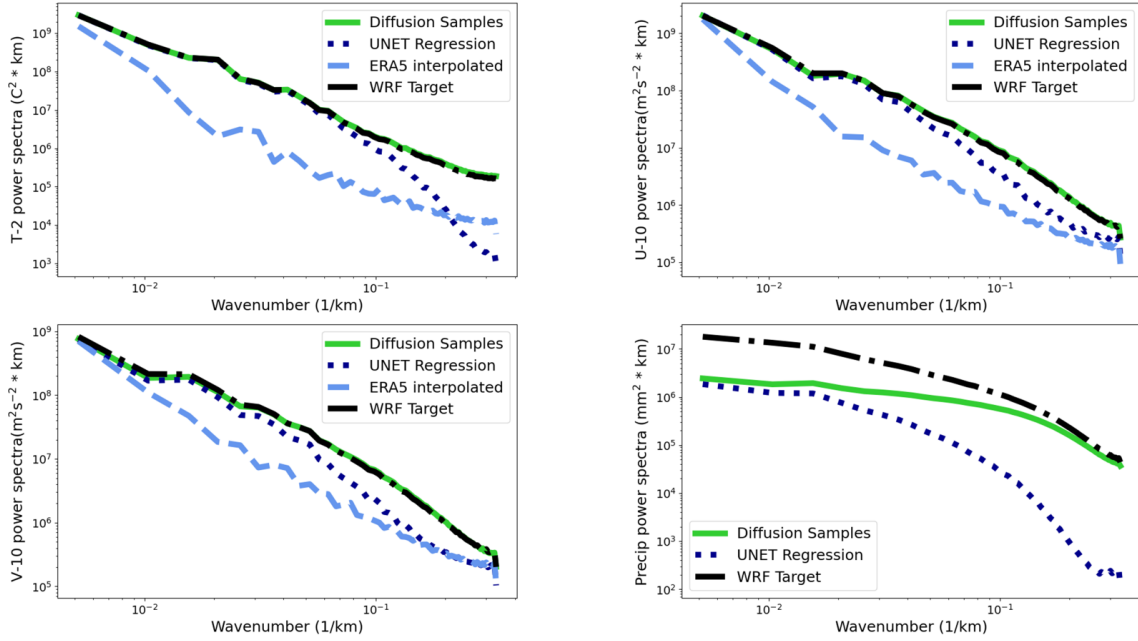


Figure 4.2 Power spectra of surface variables over the Hawai'i County domain averaged over the full testing dataset. Spectra are shown for WRF target (black), diffusion model (green), UNet regression (dark blue dotted), and interpolated ERA5 (light blue dashed).

Key findings from each panel include:

- **T-2 (2-m temperature):** The diffusion model matches the WRF reference closely across all wavenumbers, retaining both large- and fine-scale variability. UNet matches moderately well at larger scales but loses power at higher wavenumbers. ERA5 is significantly underpowered across the spectrum, especially at mesoscale ranges.

- **U-10 and V-10 (10-m winds):** Both wind components show similar behavior. The diffusion model preserves much of the WRF spectral shape, particularly at intermediate and small scales. UNet again underestimates small-scale variability, and ERA5 fails to capture high-frequency power beyond synoptic scales.
- **Precipitation (Hourly Accumulated rainfall):** The diffusion model recovers substantially more high-wavenumber power than UNet and offers a closer match to WRF. ERA5 is not shown here due to lack of 1-hour accumulation fields for the training and testing set. Despite not fully matching the intensity of the WRF spectrum, the diffusion model demonstrates a marked improvement in capturing sub-mesoscale rainfall structure.

Overall, the diffusion model outperforms both deterministic baselines in representing spatial variance across scales, particularly for fields influenced by topography and localized processes. The spectral analysis confirms that the model is not only producing accurate marginal distributions (as seen in PDF plots), but also generating realistic spatial textures consistent with high-resolution WRF simulations.

4.2 Continuous Ranked Probability Score

To evaluate the probabilistic skill of the corrected diffusion model, we compute the Continuous Ranked Probability Score (CRPS) for each surface variable across the entire testing dataset. CRPS is a proper scoring rule that measures the accuracy and sharpness of probabilistic forecasts by comparing the predicted cumulative distribution function (CDF) to the observed outcome. It generalizes the Mean Absolute Error (MAE) to the probabilistic setting and provides a unified metric for comparing ensemble-based models

like the diffusion model with deterministic baselines such as ERA5 and the UNet regression model.

Because the CRPS simplifies to the MAE for deterministic forecasts, it allows a direct comparison between the probabilistic diffusion model and the deterministic ERA5 interpolation and UNet predictions. For each forecast variable, we compute CRPS (for the probabilistic model) and MAE (for the deterministic models) at each grid point, then average over space and time to produce a single representative score. Tables 4.1 - 4.4 show the comparison between the MAE and CRPS across county domains and target variables.

Table 4.1 Comparison of Mean Absolute Error (MAE) and Continuous Ranked Probability Score (CRPS) for Hawai'i County

| Variable | MAE (UNET) | MAE (ERA5) | CRPS (Diffusion) |
|--------------------|-------------------|-------------------|-------------------------|
| T2 (°C) | 0.522 | 1.409 | 0.405 |
| U10 (m/s) | 1.207 | 1.871 | 0.941 |
| V10 (m/s) | 1.125 | 1.660 | 0.884 |
| Precipitation (mm) | 0.098 | — | 0.070 |

Table 4.2 Comparison of Mean Absolute Error (MAE) and Continuous Ranked Probability Score (CRPS) for Maui County

| Variable | MAE (UNET) | MAE (ERA5) | CRPS (Diffusion) |
|--------------------|-------------------|-------------------|-------------------------|
| T2 (°C) | 0.387 | 1.409 | 0.313 |
| U10 (m/s) | 1.371 | 2.637 | 1.096 |
| V10 (m/s) | 1.197 | 1.839 | 0.946 |
| Precipitation (mm) | 0.066 | — | 0.047 |

Table 4.3 Comparison of Mean Absolute Error (MAE) and Continuous Ranked Probability Score (CRPS) for Honolulu County

| Variable | MAE (UNET) | MAE (ERA5) | CRPS (Diffusion) |
|--------------------|-------------------|-------------------|-------------------------|
| T2 (°C) | 0.364 | 0.734 | 0.304 |
| U10 (m/s) | 1.084 | 1.681 | 0.877 |
| V10 (m/s) | 0.966 | 1.179 | 0.771 |
| Precipitation (mm) | 0.086 | — | 0.065 |

Table 4.4 Comparison of Mean Absolute Error (MAE) and Continuous Ranked Probability Score (CRPS) for Kaua‘i County

| Variable | MAE (UNET) | MAE (ERA5) | CRPS (Diffusion) |
|--------------------|------------|------------|------------------|
| T2 (°C) | 0.420 | 0.933 | 0.347 |
| U10 (m/s) | 1.089 | 2.032 | 0.876 |
| V10 (m/s) | 1.037 | 1.307 | 0.827 |
| Precipitation (mm) | 0.126 | — | 0.098 |

The following summarizes the CRPS results for each surface variable across the four counties:

- **T-2 (2-meter temperature):**
 - The diffusion model consistently shows lower CRPS than the UNET-Regression MAE and ERA5 across all counties.
 - Performance is strongest in Maui and Honolulu counties.
- **U-10 (10-meter zonal wind):**
 - The diffusion model consistently shows lower CRPS than the UNET-Regression MAE and ERA5 across all counties.
 - Performance of the diffusion model is strongest in Honolulu and Kaua‘i counties.
- **V-10 (10-meter meridional wind):**
 - Like U10, the diffusion model consistently shows lower CRPS than the UNET-Regression MAE and ERA5 across all counties.
 - Performance of the diffusion model is strongest in Honolulu and Kaua‘i counties.
- **Precipitation (Hourly Accumulated Rainfall):**
 - The diffusion model consistently shows lower CRPS than the UNET-Regression MAE and ERA5 across all counties.

- Performance is strongest in Maui and Honolulu counties.

These findings confirm the diffusion model's advantage in probabilistic skill, and demonstrate its competitive performance on deterministic metrics for temperature, wind and precipitation.

Chapter 5

Discussion and Conclusions

5.1 Discussion

Evaluation of the performance of the diffusion model using CRPS, Log-PDF analysis and power spectrum analysis provides a comprehensive view of its strengths and limitations in different weather variables and regions.

The model performs well for temperature and wind (T-2, U-10, V-10), consistently outperforming both interpolated ERA5 and the UNet baseline in CRPS and mean absolute error (MAE). These improvements are mirrored in the Log-PDF distributions, even during the most variable season with temperature and wind matching the WRF target most closely. The power spectra further supports this by showing that the diffusion model closely matches the spectra across wavenumbers in the WRF target fields throughout the test period, especially for wind components. ERA5 and UNet predictions exhibit an inability to reproduce the WRF target patterns in temperature and wind.

For precipitation, the performance is more mixed. While the diffusion model generates realistic light-to-moderate rainfall patterns, it struggles to reproduce extreme events. This is evident in the log-probability density functions (log-PDFs), which show under-

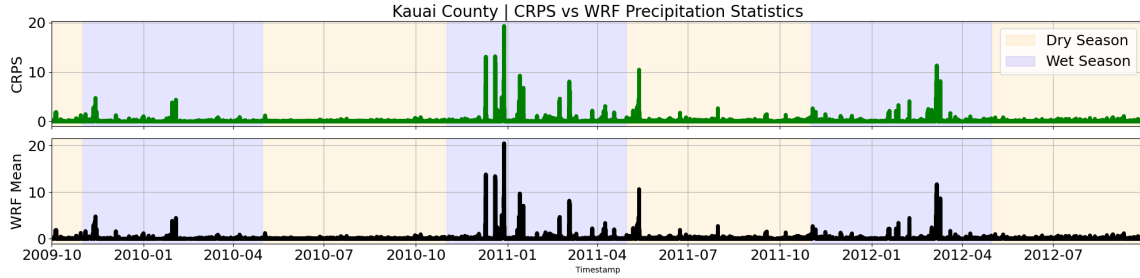


Figure 5.1 Temporal evolution of CRPS (diffusion model) and corresponding WRF precipitation across the Kaua‘i domain. The plot illustrates seasonal variation in model performance, with CRPS increasing during wetter periods when forecast uncertainty and rainfall variability are highest. The temporal correlation also reveals that forecast skill degrades notably during high precipitation events, highlighting a potential shortcoming in ensemble spread during these periods.

representation in the upper tail. The power spectra for precipitation confirm a lack of high-wavenumber energy, indicating insufficient spatial variability during heavy rainfall.

To further investigate this limitation, we introduced a time-resolved version of the CRPS. By computing domain-averaged CRPS at each hourly timestep and comparing it to the corresponding WRF precipitation, we were able to trace specific periods of degraded performance. Figure 5.1 illustrates this for Kaua‘i: CRPS increases sharply during episodes of intense precipitation. This suggests that ensemble spread remains too narrow during dynamically active periods, and confirms that static skill metrics alone may mask event-driven failures.

This diagnostic tool not only complements the spatially and statistically averaged results but also helps isolate specific failure modes of the model, namely, a lack of uncertainty representation during extremes. These insights will directly inform future improvements.

5.2 Conclusions

This study demonstrates the capability of score-based diffusion models to downscale coarse ERA5 inputs into realistic, high-resolution weather fields across the Hawaiian Islands. The diffusion model performed particularly well in replicating near-surface temperature and wind fields, as confirmed by low CRPS values and good agreement with WRF benchmarks in both Log-PDF and power spectrum evaluations. For precipitation, the average CRPS values of the model were also low in the island domains and significantly outperformed the MAE of the UNet, indicating that the probabilistic approach provides greater fidelity to capture rainfall variability. However, performance declined during the wet season, where the CRPS scores increased and the PDF/power spectrum analysis revealed a tendency to underpredict heavy rainfall events. To better understand these discrepancies, we introduced temporally resolved CRPS plots and examined their relationship to WRF rainfall. The time series shows a clear seasonal signal, with elevated CRPS values aligning with periods of intense rainfall, reinforcing that model skill diminishes during synoptically active regimes. This additional diagnostic provides crucial insight into where and when the diffusion model struggles.

To address current limitations and improve the performance of the model in synoptically active regimes, several paths forward are proposed. First, expanding the input space to include additional predictors of large-scale forcing may improve the model’s ability to capture rare and extreme rainfall events. Conditioning the diffusion model on multiple previous ERA5 timesteps ($t = -1, -2, \dots$) may also better capture the evolving atmospheric context, especially for precipitation. Once the model is tuned, shifting from reanalysis data to forecast datasets, such as GFS or ECMWF operational products, would support the model’s application in real-time forecasting environments. Furthermore, incorporating training data that integrates radar and station-based rainfall observations

from the Hawaiian Islands would improve the accuracy of the ground truth for rainfall fields. These developments would enable the diffusion model to generate ensemble rainfall and wind fields more robustly for hazard assessment, and support future integration into operational systems. Ultimately, these improvements would broaden the utility of this approach for forecasting and risk management due to the speed and probabilistic nature of diffusion models.

Bibliography

1. Wilby, R.L., Charles, S.P., Zorita, E., et al. "Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods.", 27, 1-27, 2004.
2. Wilby, R. L., Wigley, T. M. L. "Downscaling general circulation model output: a review of methods and limitations." *Progress in Physical Geography*, 21(4), 530-548, 1997.
3. Norton, Chase W., Pao-Shin Chu, and Thomas A. Schroeder. "Projecting Changes in Future Heavy Rainfall Events for Oahu, Hawaii: A Statistical Downscaling Approach." *Journal of Geophysical Research* 116, no. D17, September 14, 2011.
4. Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models." arXiv, December 16, 2020.
5. Sohl-Dickstein, Jascha, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics." arXiv, November 18, 2015.
6. Song, Yang, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. "Score-Based Generative Modeling through Stochastic Differential Equations." arXiv, February 10, 2021.

7. Karras, Tero, Miika Aittala, Timo Aila, and Samuli Laine. “Elucidating the Design Space of Diffusion-Based Generative Models.” arXiv, October 11, 2022.
8. Hatanaka, Yusuke, Yannik Glaser, Geoff Galgon, Giuseppe Torri, and Peter Sadowski. “Diffusion Models for High-Resolution Solar Forecasts.” arXiv, January 31, 2023.
9. Mardani, Morteza, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, et al. “Residual Corrective Diffusion Modeling for Km-Scale Atmospheric Downscaling.” *Communications Earth and Environment* 6, no. 1, February 24, 2025.
10. Kodama, Kevin R., and Steven Businger. “Weather and Forecasting Challenges in the Pacific Region of the National Weather Service.” *Weather and Forecasting* 13, no. 3: 523–46, September 1998.
11. Morrison, Ian, and Steven Businger. “Synoptic Structure and Evolution of a Kona Low.” *Weather and Forecasting* 16, no. 1: 81–98, February 2001.
12. Xue, Lulin, Yaping Wang, Andrew J. Newman, Kyoko Ikeda, Roy M. Rasmussen, Thomas W. Giambelluca, Ryan J. Longman, Andrew J. Monaghan, Martyn P. Clark, and Jeffrey R. Arnold. “How Will Rainfall Change over Hawai‘i in the Future? High-Resolution Regional Climate Simulation of the Hawaiian Islands.” *Bulletin of Atmospheric Science and Technology* 1, no. 3–4: 459–90, December 2020.
13. Liu, Guangpeng, Brian Powell, and Tobias Friedrich. “Climate Downscaling for Regional Models with a Neural Network: A Hawaiian Example.” *Progress in Oceanography* 215, July 2023.

14. Hatanaka, Yusuke, "Machine Learning Based Statistical Downscaling for Rainfall on Hawaiian Islands." Master's Thesis, University of Hawai'i at Mānoa, December 2022.
15. Copernicus Climate Change Service (2023): ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS)
16. Gneiting T, Raftery AE. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association.*;102(477):359-378, 2007.
17. Wilks, D. S. "Statistical Methods in the Atmospheric Sciences" (4th ed., Chapter 3.3.6: Kernel Density Smoothing). Elsevier. 2019.
18. Wilks, D. S. "Statistical Methods in the Atmospheric Sciences" (4th ed., Chapter 10.5: Frequency Domain–II. Spectral Analysis). Elsevier. 2019.