Phase Coherence in Wind Data and Simulation

by

Jennifer M. Rinker

Department of Civil and Environmental Engineering
Duke University

Date: _____

Approved:

_____
Henri P. Gavin, Supervisor

_____
John Albertson

_____
Dalia Patiño-Echeverri

_____
Amilcare Porporato

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Civil and Environmental Engineering
in the Graduate School of Duke University

2014

Abstract

Phase Coherence in Wind Data and Simulation

by

Jennifer M. Rinker

Department of Civil and Environmental Engineering
Duke University

Date: _____

Approved:

_____
Henri P. Gavin, Supervisor

_____
John Albertson

_____
Dalia Patiño-Echeverri

_____
Amilcare Porporato

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Civil and Environmental Engineering in the Graduate School of Duke University

2014

Abstract

Novel wind turbine designs are deemed acceptable through a simulation-based certification process that involves generating a synthetic wind record and using it as an input to a computer model of the turbine. Naturally, whether the simulation loads reflect the loads that the turbine would actually experience depends on the accuracy of the wind turbine model and, more importantly, on the accuracy of the method used to generate the synthetic wind record. The simulation methods that are commonly used for this purpose are spectral-based and produce Gaussian, stationary random fields. These methods prescribe a power spectral density (PSD) of the wind velocity, which fixes the magnitudes of the Fourier components, then assumes that the Fourier phase angles are independent and uniformly distributed. An inverse Fast Fourier Transform (IFFT) is then used to transform the wind velocity field to the time domain.

This thesis applies the concept of phase coherence—i.e., Fourier phase angles that are not independent—to the stochastic modeling and simulation of wind velocity fields. Using a large dataset available from the National Wind Technology Center (NWTC), a joint distribution is characterized for the mean wind speed U, turbulence σ_u, Kaimal length scale L, and a metric for the degree of phase coherence in wind data, $\bar{R}$. The correlations between these four parameters, the vertical height, and another phase coherence parameter are presented; only U, σ_u, and L have a significant degree of correlation. The joint distribution is used to generate synthetic wind

records, which are then compared with measured data that have the same parameter values. For data with low to medium coherence values, the synthetic records have a similar qualitative appearance to the data. For high levels of phase coherence, the records simulated with the proposed model were qualitatively different from records with the same parameter values due to the variation of the phase difference spread in the spectral domain. Lastly, the importance of correctly modeling phase coherence is demonstrated by using the data and the synthetic records as inputs to a single-degree-of-freedom (SDOF) oscillator and comparing the peak response statistics and damage equivalent loads (DELs).

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to thank Dr. Andy Clifton at the National Wind Technology Center for providing the wind velocity data, without which this thesis would not exist. I also gratefully acknowledge and thank Duke University and the National Science Foundation (Grant No. 1106401) for the funding that made this thesis possible. Lastly, and most importantly, I thank my research advisor, Professor Henri Gavin, for the inspiration, patience, and knowledge that has brought my research to where it is today.

1

Introduction

Few people will argue that the human race should aim for an energy portfolio that has more renewable resources. Up until recently, we have depended almost exclusively on fossil fuels due to their prevalence, low cost, and facility for conversion to electricity, despite the fact that they produce greenhouse gas emissions that are linked to climate change [1, 2]. It is unlikely that current energy sources are sustainable in terms of the environment, so it is high time that our energy comes from renewable sources.

One such renewable energy source is wind power, which is electricity that is extracted from ambient wind in the atmospheric boundary layer through the use of wind turbines. While wind power alone could not support the nation's energy demands, it is capable of supplying a significant portion of the national energy portfolio: the US Department of Energy (DOE) has published a plan for 20% of the US energy demands to be provided by wind energy by 2030 [3]. For that to occur, wind turbine technology and wind forecasting must be reliable enough such that the volatility of wind power is minimized and its cost is low enough to be competitive with fossil fuels.

Data from Germany in 2013 show that the levelized cost of energy (LCOE) for

onshore wind turbines is the lowest of all of the renewable resources ($33/MWh – $75/MWh), though it is still significantly higher than brown coal ($29/MWh – $40/MWh) [4]. These LCOE values are generally higher than those reported by Lantz *et al.* in the United States [5], which showed the LCOE for onshore wind reaching a minimum of around $60/MWh in 2005, after which it steadily increased to around $80/MWh in 2010. Lantz *et al.* state that this cost increase is tied with wind turbine upscaling, and delineate several research areas where wind turbines might improve. The research area that is most relevant to this thesis is that of resource assessment, which is tied to wind simulation methods and wind turbine design.

Wind turbine designs may be certified by following the recommendations in the IEC 61400 standards, which are a set of design guidelines published by the International Electrotechnical Commission (IEC) [6]. The IEC 61400-1 provides recommendations for modeling the external conditions and for designing the structure, control system, and mechanical systems. The structural design is accomplished through the simulation of different load cases, each corresponding to particular events or conditions that the turbine might experience. In many cases, several different simulations are run for each load case, and a design is deemed acceptable if specified limit states are not exceeded.

The majority of these load cases correspond to a specific type of wind input, often stochastic, and a specific operational configuration of the turbine. If the method used to generate the stochastic wind input creates wind fields that do not accurately reflect real-world conditions, the loads generated in the model will be too high or too low and the turbine will be either over- or under-designed. This is bad either way, as over-designing leads to more expensive turbine production and under-designing leads to unexpected failure and/or maintenance. Thus, the accurate generation of synthetic wind is central to the optimal design of wind turbines.

The two main methods[1] that are currently recommended to generate synthetic turbulent wind fields for wind turbine applications are the Mann turbulence model [16] and the Sandia method [17]. Both methods are stochastic spectral methods in which the discrete Fourier transform (DFT) of the velocity field is generated and then an inverse Fast Fourier Transform (IFFT) is used to construct the field in the time domain. In both cases, the magnitudes of the Fourier vectors are prescribed through an assumed form of the process's power spectral density (PSD) and the phases are uniformly and independently sampled. This produces a Gaussian stationary process that has "equal energy" at all times. (See Appendix A for proof and Section 2 for more discussion.)

The two models differ in their construction of the PSD, their implementation of cross-axis correlations (e.g., the correlation between lateral and vertical wind velocities at a single point), and their implementation of spatial correlations (e.g., the correlation between the downwind velocity at two different heights). In particular, Mann derives a 3D spectral tensor of an isotropic turbulence field that is distorted by vertical shear. This spectral tensor embeds the proper theoretical cross-axis and spatial correlations in its PSD formulation, which makes it the preferred method in the IEC standard. However, its 3D formulation necessitates the use of a 3D Fourier transform. Veldkamp [18] indicates that this is extremely computationally intensive in practice, necessitating the reduction of the number of spatial points, which leads to a loss of frequency resolution and high frequency content. In general, the method is known to be computationally intensive [19], and there appears to be no research clearly indicating its superiority over a modified Sandia method.

The Sandia method in its most basic form includes spatial correlations but not

[1] Note that there are several other techniques for producing turbulent wind fields, including wavelets [7, 8, 9, 10, 11], computational fluid dynamics [12, 13, 14], and Hilbert spectral analyses [15]. This thesis focuses on developing simple modifications to methods currently used in wind engineering, so these other methods are not reviewed here in depth.

cross-axis correlations, though it is possible to add in cross-axis correlations after the field has been simulated [20]. It is substatially less computationally intenstive than the Mann model, as it only requires a series of one-dimensional Fast Fourier Transforms (FFTs). It is the method used in the open-source turbulent field simulator TurbSim, which is developed and maintained by the National Wind Technology Center (NWTC) [21]. Because of its relatively simple formulation and its common use in the wind turbine design process, the simulation method proposed in this thesis is developed as a modification of the Sandia method.

To generate a downwind wind velocity vector at a single point, $u(k)$, the Sandia method only requires the specification of the PSD of the process.[2] The PSD can be directly related to the magnitude of the Fourier component:

$$|U(k)| = \left[\int_{(k-\frac{1}{2})\Delta f}^{(k+\frac{1}{2})\Delta f} \frac{S(f)}{2} df\right]^{1/2}, \tag{1.1}$$

where $S(f)$ is a specified one-sided PSD and Δf is the frequency resolution. For simplicity, the magnitudes are often approximated using a Riemann sum:

$$|U(k)| \approx \sqrt{\frac{S(k\Delta f)\Delta f}{2}}. \tag{1.2}$$

Thus, once the form of the PSD is chosen, the Fourier magnitudes are known. The PSD that is used in the Sandia method is the Kaimal spectrum [22],

$$S(f) = \sigma_u^2 \frac{4L/U}{(1+6fL/U)^{5/3}}, \tag{1.3}$$

which is parameterized by the mean wind speed U (conventionally taken over a 10-minute window), the turbulence σ_u (the standard deviation of the wind velocity,

[2] For simplicity, the proposed simulation modifications will be presented only for the downwind velocity at a single point. However, the proposed method could be applied to a correlated 3D field by using the standard simulation procedure for a 3D field and implementing the proposed model for phase coherence.

usually in a 10-minute window), and the Kaimal length scale, or turbulent length scale, L. To introduce randomness to the simulation method, the Fourier phase angles, $\angle U(k)$, are assumed to be random variables that are independent and uniformly distributed. Once the entire Fourier vector has been prescribed, an IFFT is used to transform it into the time domain.

This thesis examines a new method for characterizing and simulating synthetic wind that does not produce a stationary, Gaussian process. The main contribution of this work is the exploration of the concept of "phase coherence," which has been explored in stochastic earthquake simulations but has not yet been applied to stochastic wind simulations. This thesis uses a large dataset available from the NWTC to fit a joint probability density function (PDF) to several different wind parameters, including a novel parameter that measures the degree of phase coherence. Samples are drawn from the joint distribution to generate synthetic records, which are then compared to data records with the same parameter values to verify that the simulation method mimics the data characteristics. Lastly, the data and synthetic records are used as inputs to a single-degree-of-freedom (SDOF) oscillator to demonstrate the effects of phase coherence on the response of a dynamical system.

An overview of the theory behind phase coherence and its use in earthquake simulations is given in Chapter 2. The details of the dataset used to characterize the joint PDF, including the pre-processing methods required to clean the data, are presented in Chapter 3. The data analysis and characterization of the joint distribution is presented in Chapter 4. The simulation results, comparison with data, and SDOF analysis are given in Chapter 5. Lastly, conclusions are presented in Chapter 6.

2

Phase Coherence

Phase coherence in its most basic form is the concept that the Fourier phase angles are not independent, as is commonly assumed in stochastic simulation methods. In this thesis, the interdependence of the phase angles is modelled by a non-uniform distribution of the phase differences, which are the differences between the angles of adjacent Fourier components. To be more specific, the phase difference values are calculated as

$$\Delta\theta(k) = \angle U(k+1) - \angle U(k), \tag{2.1}$$

where $\angle U(k)$ is the phase of the kth component of the Fourier vector. If the phase difference values are uniformly distributed, then $\Delta\theta(k)$ is equally likely to take any value. Therefore, prescribing the value of $\angle U(k)$ would not affect the probability of $\angle U(k+1)$. However, if the phase difference distribution is not uniform, $\Delta\theta(k)$ will tend towards some mean value, denoted by $\bar{\theta}$. In this case, prescribing the value of $\angle U(k)$ will increase the likelihood that $\angle U(k+1)$ takes a value near $\angle U(k)+\bar{\theta}$. Thus, specifying a non-uniform distribution of the phase differences will produce coherence in the phase angles.

Phase coherence was first investigated by researchers in the earthquake engineer-

ing field because it is closely connected with the simulation of enveloped processes. Section 2.1 provides a literature review of phase coherence as it has been applied to earthquake simulations. Because phase coherence relies on the distributions of angles, which are circular, it is necessary to provide a brief overview of directional statistics, which addresses unique issues that arise with distributions and moments of circular variables (Section 2.2). This chapter also provides a description of the circular distribution that is used to model the phase difference distributions that were found in the wind data, along with various approximation methods that were developed to use the distribution in simulation.

2.1 Background

As mentioned in Chapter 1, the simulation of a stationary, Gaussian process requires assuming that the phase angles in the Fourier domain are independent and uniformly distributed. However, real-world processes are often neither stationary nor Gaussian, so this assumption will produce synthetic records that do not match the experimental data. The connection between phase coherence and non-stationarity in the time domain was first investigated in the earthquake engineering field, where researchers realized that the distribution of differences between adjacent Fourier phases a) was not uniform for real data and b) affected the envelope of the process in the time domain.

Ohsaki [23] was one of the first researchers to examine phase coherence in earthquake simulations. He noted that phase difference distributions in earthquake data looked approximately normal and claimed that their shape was related to the shape of the envelope function for the earthquake record. Nigam [24] took a more theoretical route and proved that a uniformly modulated, Gaussian white noise process has uniformly distributed phase angles. He also derived the PDF for the phase derivative, $d\theta/d\omega$, for a uniformly modulated Gaussian white-noise process, which is not

actually Gaussian, as suggested by Ohsaki.

Thráinsson and Kiremidjian [25] examined earthquake data and fit distributions to the Fourier magnitudes and phases, then derived relationships between the distribution parameters and certain earthquake parameters. One of their most interesting findings was that sorting the phase difference values by Fourier magnitude instead of frequency produced a notable trend in the spread of the phase difference values. Specifically, as the magnitude of the Fourier component decreased, the variance of the phase difference values increased. They binned the phase difference values by their Fourier magnitudes (separated into "small," "medium," and "large"), and fit either a beta distribution or a beta distribution superimposed over a uniform distribution to the bins.

Boore [26] derived a method to calculate the phase derivative, which is a continuous form of the phase difference and is therefore not dependent upon sampling frequencies or durations of the time histories. His examination of the phase derivative, which he refers to as envelope delay, is thorough and well-informed and includes a comparison of the phase derivative values for different earthquake records. Boore proposes a method to modify stochastic earthquake simulation methods and also notes that Ohsaki's conclusions regarding the connection between the envelope and phase difference distributions are innaccurate.

Other researchers such as Wang [27] have connected the concept of phase difference to other methods for simulating nonstationary, enveloped processes, such as wavelets.

Phase difference values are simply the difference in angles between two adjacent Fourier components, but the method used to calculate them can affect the resulting sample of $\Delta\theta$. It should be noted that, because the signals of interest here are real, only half of the Fourier vector is unique: the second half of the vector is the complex conjugate of the first half. Thus, the phase difference values of the second half of

the Fourier vector are the negated phase difference angles of the first half. This symmetry affects the estimated phase difference distributions,[1] so it is important to only use the first half of the Fourier vector.

Additionally, while it is possible to calculate the phase difference values by subtracting the angle from one component from the angle of the adjacent component, this will produce values that are in no particular interval and must then be wrapped to an interval of choice before analysis. A more elegant approach proposed by Krajnik [28] that is used here involves taking a ratio of complex exponentials and then calculating the phase of the result:

$$\angle U(k+1) - \angle U(k) = \angle \left[\frac{\exp[j\angle U(k+1)]}{\exp[j\angle U(k)]} \right]. \tag{2.2}$$

2.2 Directional Statistics

Random variables that wrap, such as the phase difference, feature different statistical rules than those used for classical non-wrapping variables. This section overviews the most relevant aspects of the statistics of wrapping variables, called directional statistics. For a more thorough treatment, interested readers are directed to Mardia and Jupp [29].

For circular random variables, the usual definitions of moments are no longer useful. For example, consider the PDF

$$f_\Theta(\theta) = \frac{1}{2\pi}(1 + \cos\theta) \tag{2.3}$$

for some random variable $\theta \in [0, 2\pi)$, which is plotted in Cartesian coordinates in the left panel of Fig. 2.1. Because this PDF is symmetric about $\theta = 0$ and the variable

[1] For example, consider a signal where all of the sinusoidal components are $\pi/4$ apart. With this signal, the correct phase difference distribution would have only a single non-zero point at $\pi/4$. If the whole Fourier vector is used to calculate phase differences, the resulting distribution will have two non-zero points: one at $\pi/4$ and one at $-\pi/4$.

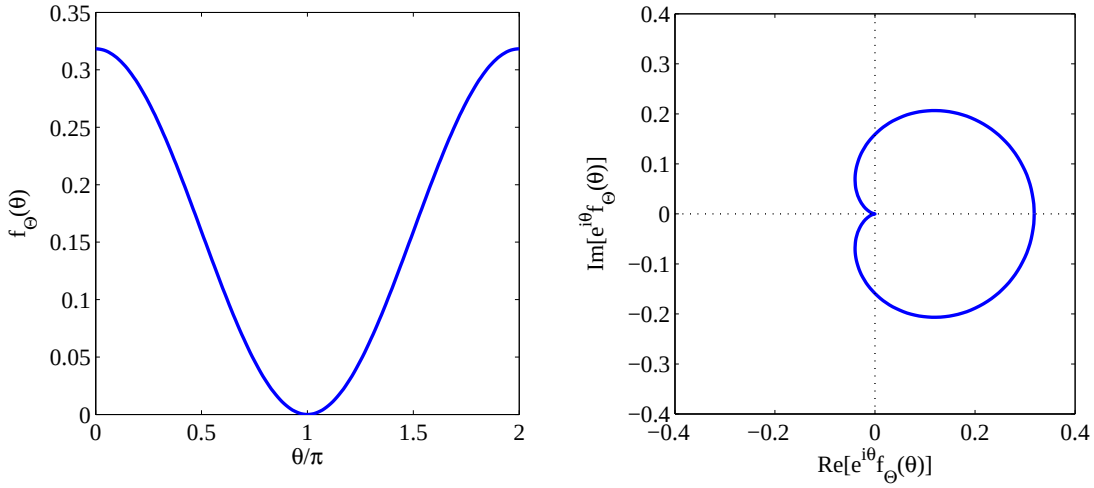

FIGURE 2.1: Example of a circular PDF in Cartesian (left) and polar (right) coordinates.

is circular, the mean value is expected to be $\theta = 0$. However, if we use the standard definition of the first moment,

$$\mathbb{E}\{\theta\} = \int_0^{2\pi} \theta f_\Theta(\theta) d\theta = \pi. \tag{2.4}$$

Not only is this value different from the predicted mean, it corresponds to the angle with the smallest likelihood.

A natural inclination is to examine the distribution in a polar form, which is shown in the right panel of Fig. 2.1. It is apparent from this plot that, due to symmetry, the centroid of the PDF is located along the $\theta = 0$ axis, which is the predicted value for $\mathbb{E}\{\theta\}$. This observation is the basis of the concept of the *mean resultant vector*, which is equivalent to the expected value in non-circular systems. The equation for the mean resultant vector is similar to that for the classical expected value, except that the complex variable $z = e^{i\theta}$ is used in place of the usual kernel:

$$\bar{\rho} = \int_\Omega e^{i\theta} f_\Theta(\theta) d\theta. \tag{2.5}$$

In this equation, Ω is any interval of length 2π and $\bar{\rho}$ is the mean resultant vector,

which is a complex number with magnitude and phase. The magnitude of the mean resultant vector, denoted by $\bar{R}$, is called the mean resultant length (MRL) and is a measure of the non-uniformity of the distribution. If $\bar{R} = 0$, then the distribution is completely uniform.[2] If $\bar{R} = 1$, then the distribution is zero everywhere except for a single value (or that value plus integer multiples of 2π). For example, a distribution that had 50% probability of a random variable being either $\pi/2$ or $5\pi/2$ would have $\bar{R} = 1$. The MRL can also be viewed as a measure of dispersion of the points around some mean angle. The phase of the mean resultant vector, $\bar{\theta}$, is called the mean direction and is the mode of the distribution. The mean direction can be viewed as the expected value of the circular random variable, and it is not affected by the choice of the interval for Ω, unlike the classical definition of the expected value. The mean resultant vector can be approximated from a sample with the following equation:

$$\bar{\rho} \approx \frac{1}{N} \sum_{n=1}^{N} e^{i\theta_n}. \tag{2.6}$$

Applying Eq. (2.6) to the distribution given in Eq. (2.3) yields

$$\bar{\rho} = 0.5, \tag{2.7}$$

which corresponds to $\bar{R} = 0.5$ and $\bar{\theta} = 0$. The mean direction matches the predicted mean value, and the non-zero value for $\bar{R}$ reflects the non-uniformity of the selected distribution.

2.2.1 Effect of mean resultant length and mean direction

It is of interest to examine the effect of the MRL and mean direction on simulations in order to better understand the phenomena they can create in real data. To accom-

[2] Technically, $\bar{R}$ could also equal zero if the set had an even number of samples at evenly spaced angles around the interval (e.g., 100 samples each at 0, $2\pi/3$, and $4\pi/3$). To be completely precise, $\bar{R}$ is a measure of the circular symmetry of $f_\Theta(\theta)$. Because the phase difference values for real data will never fall into angles that are equally spaced, it is acceptable in this application to state that if $\bar{R} = 0$ then the distribution is uniform.

plish this, several different records were generated with different $\Delta\theta$ distributions. The distributions were uniform in all cases, but the values for the mean direction and the MRL were varied individually in order to determine their effects.

The first experiment investigated the effects of the MRL by choosing four uniform distributions with the same mean direction but decreasing widths (i.e., increasing MRLs). The four distributions that were used for the simulations were $\mathcal{U}(0, 2\pi)$, $\mathcal{U}(\pi/4, 7\pi/4)$, $\mathcal{U}(\pi/2, 3\pi/2)$, and $\mathcal{U}(3\pi/4, 5\pi/4)$, which all have a mean direction of π. The MRL values for the four distributions are 0, 0.30, 0.67, and 0.90, respectively, and the simulations are plotted in Fig. 2.2. As can be seen in the figure, narrowing the spread of the $\Delta\theta$ distribution localizes the energy in the system into a single packet that occurs around 300 seconds. Thus, it can be said that increased MRLs lead to a higher degree of non-stationarity in the time domain. Another way to view these results is by noting that, in essence, increasing the MRL produces a "stronger" enveloping function.

The second experiment focused on the effect of the mean direction, $\bar{\theta}$. The $\Delta\theta$ distributions for this experiment were all uniform with width $\pi/2$ ($\bar{R} = 0.90$), but the location of the mean direction was varied from $\pi/4$ to $7\pi/4$ in increments of $\pi/2$. The plots of the simulations are shown in Fig. 2.3. As can be clearly seen in the plots, the mean direction of $\Delta\theta$ changes the location of the packet within the total record. In other words, the mean direction shifts the enveloping function in time.

2.2.2 The von Mises distribution

There are several distributions that have been defined in circular statistics that can be used to describe the probability of a wrapped random variable [29]. One such distribution is the von Mises distribution, which is the maximum entropy distribution for a random variable with a given mean resultant vector. The distribution was first proposed by von Mises in 1918 to describe the variation of measured atomic weights

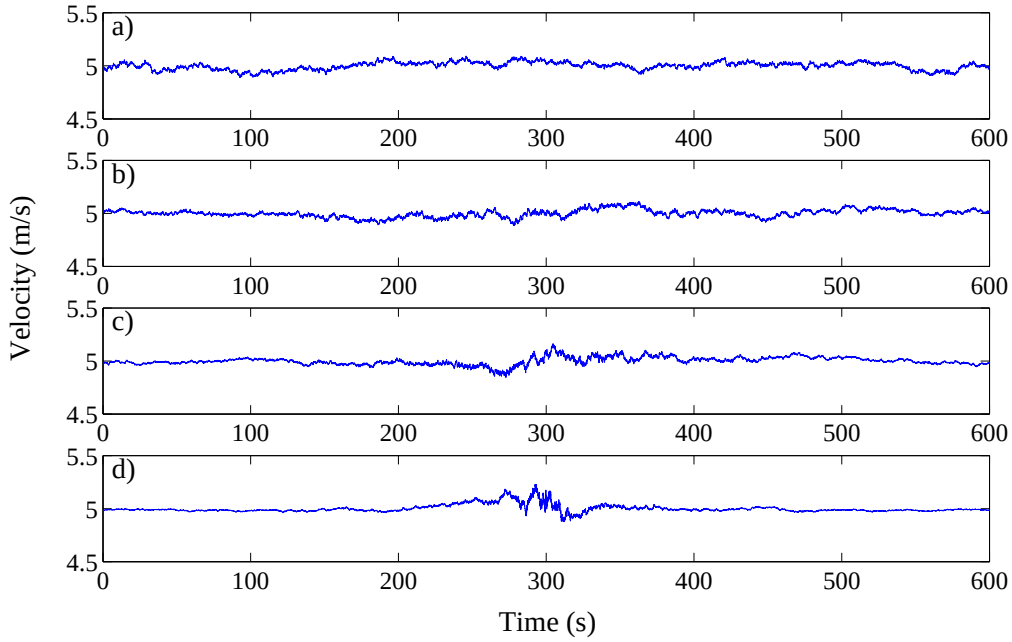

FIGURE 2.2: Demonstration of the effect of the MRL upon simulations ($\bar{\theta} = \pi$). The $\Delta\theta$ distributions for the different subplots are as follows: a) $\mathcal{U}(0, 2\pi)$, b) $\mathcal{U}(\pi/4, 7\pi/4)$, c) $\mathcal{U}(\pi/2, 3\pi/2)$, d) $\mathcal{U}(3\pi/4, 5\pi/4)$.

and is one of the more tractable circular distributions.

The von Mises distribution is a two-parameter, symmetric distribution of a circular random variable that has support over any interval of length 2π. It can be used to approximate the wrapped normal and wrapped Cauchy distributions, which are more intuitive because they are the wrapped forms of well-known distributions but are mathematically difficult to implement [29]. The two parameters of the distribution determine the location and spread, and the distribution was found to accurately characterize the phase difference distributions that were found in the data.

The PDF for the von Mises distribution is

$$f(x|\mu, \kappa) = \frac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\kappa)}, \tag{2.8}$$

where $I_0(x)$ is the modified Bessel function of order 0. The parameter μ is the mean direction of the distribution, and κ determines the spread or concentration of

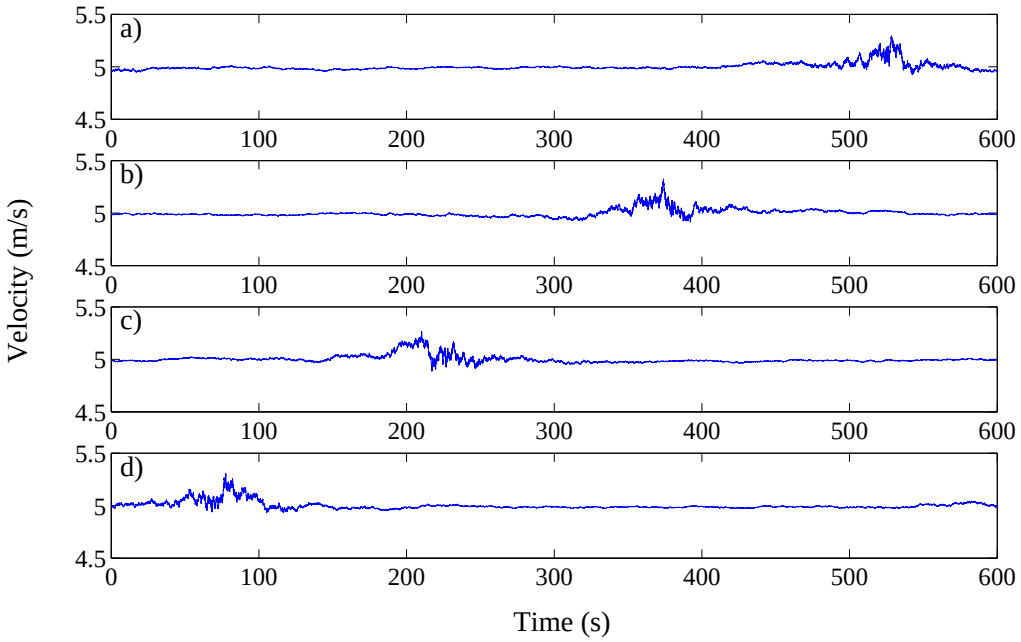

FIGURE 2.3: Demonstration of the effect of $\bar{\theta}$ upon simulations ($\bar{R} = 0.90$). The $\Delta\theta$ distributions for the different subplots are as follows: a) $\mathcal{U}(0, \pi/2)$, b) $\mathcal{U}(\pi/2, \pi)$, c) $\mathcal{U}(\pi, 3\pi/2)$, d) $\mathcal{U}(3\pi/2, 2\pi)$.

the distribution. A plot of the von Mises PDF is shown in the left-hand panel of Fig. 2.4. There is no analytical form of the CDF due to the Bessel functions, but the CDF can be numerically integrated and plotted for demonstrative purposes, as is done in the right-hand panel of Fig. 2.4. As can be seen in the figures, a zero value for κ corresponds to a uniform distribution and increasing κ increases the concentration of the distribution.

This research fits a joint distribution to the MRL and other wind record parameters, which does not require a parametric model for the distribution of $\Delta\theta$ in a record of interest. However, a model for the distribution of $\Delta\theta$ is required in order to simulate a synthetic record with phase coherence; additionally, for comparative purposes it is of interest to be able to determine which von Mises distribution would best fit the phase difference angles from a data record of interest. Thus, it is necessary to be able to a) fit a von Mises distribution to a particular mean resultant vector and

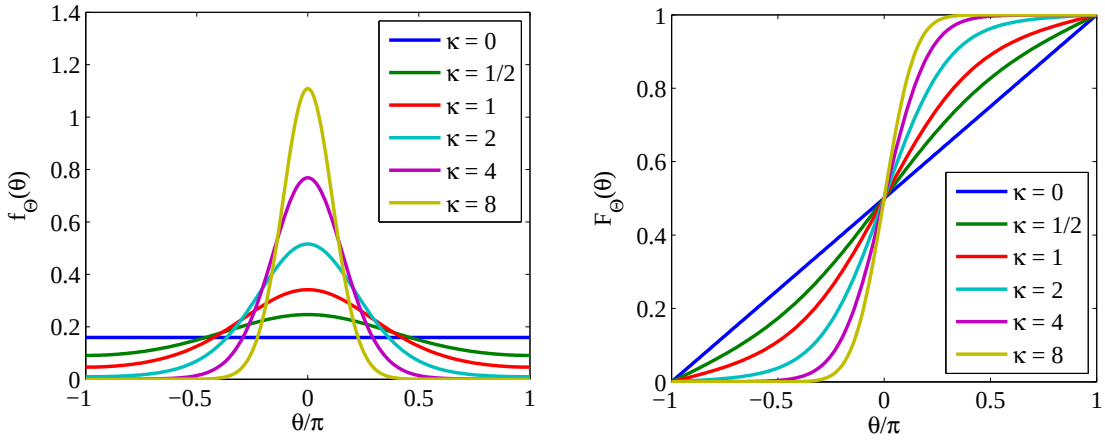

FIGURE 2.4: Plot of the PDF (left) and CDF (right) of a von Mises distribution for different values of κ.

b) generate a sample from the subsequent distribution.

One simple method to determine the distribution parameters is to calculate the mean direction and MRL of the sample and then choose the von Mises distribution that has those values. The mean direction is relatively simple: the value for μ is equal to the mean direction. Determining the value for the concentration is more difficult. There is a closed-form relationship between the MRL and κ,

$$\bar{R} = \frac{I_1(\kappa)}{I_0(\kappa)}, \tag{2.9}$$

but it is not clear how to use this relationship to determine κ from a sample's MRL. The relationship could be inverted numerically, as it is a smooth function, but for computational efficiency a fourth-order polynomial was fit to this relationship and then the Matlab function `roots` was used to invert the approximating polynomial. The fit polynomial parameters are

$$\begin{aligned} a_0 &= \text{-0.01641}, \\ a_1 &= 0.6074, \\ a_2 &= \text{-0.1627}, \\ a_3 &= 0.02032, \\ a_4 &= \text{-0.0009621}, \end{aligned}$$

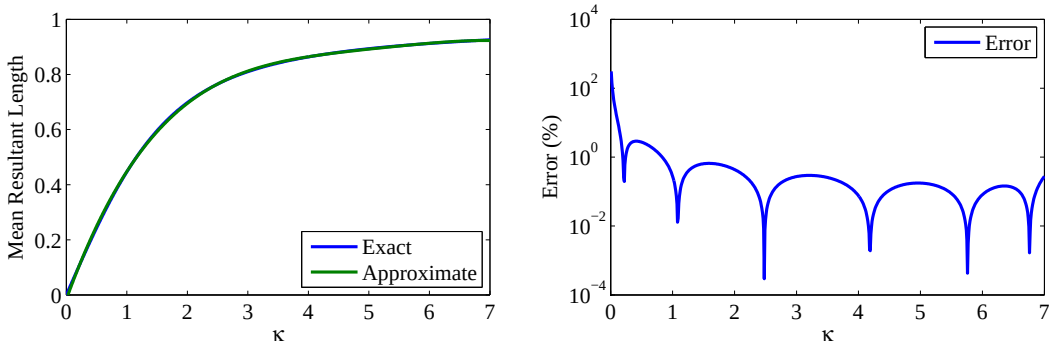

FIGURE 2.5: Comparison of the exact (Eq. (2.9)) and approximate (Eq. (2.10)) relationship between κ and $\bar{R}$ (left), and a plot of the percent error (right).

where

$$\bar{R} = a_0 + a_1\kappa + a_2\kappa^2 + a_3\kappa^3 + a_4\kappa. \tag{2.10}$$

A plot of the exact and approximated relationships between $\bar{R}$ and κ is shown in the left subplot of Fig. 2.5, and the percent error of the approximation is shown in the right subplot of the same figure. The approximation is very good for MRL values up to 0.92, which is sufficiently high for this application. There is a large percent error for $\kappa < 0.17$ due to the small magnitude of the MRL, but the error for all other MRL values is below 3%.

In simulation, it is also necessary to draw samples from a given distribution, which requires a characterization of the inverse CDF. Oftentimes the inverse CDF is known exactly, but because the von Mises distribution does not have an analytic CDF, the analytic inverse CDF does not exist either. An acceptable approximation of the CDF can be generated by approximating the exponential term in the PDF with a Taylor series and integrating the result; doing so with a fourth-order expansion produces a maximum difference between the exact and approximated CDF of 0.0019. However, the resulting analytical expression is still not invertible. Attempts were also made to construct an approximation to the numerical inverse CDF from various sets of basis functions, but no acceptably accurate approximation was found.

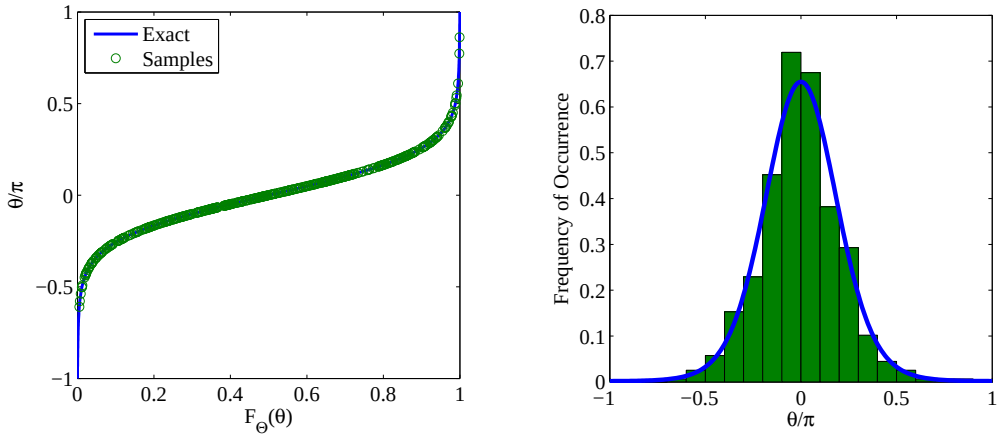

FIGURE 2.6: Plot of the numerical inverse CDF (left) and PDF (right) of an exact von Mises distribution ($\kappa = 3$) and a 500-point sample generated using linear interpolation.

Without an analytical or approximate form for the inverse CDF, the only solution is to sample from the distribution by linearly interpolating points from a numerically generated CDF. A comparison of the CDF and PDF of the exact distribution and a 500-point sample with $\kappa = 3$ is shown in Fig. 2.6. As can be seen in the figure, the distribution of samples closely follows the exact CDF and PDF. This method might not be as accurate as using a numerical solver to invert the approximate CDF, but it has the advantage of being extremely efficient computationally.

3

Wind Data

The data used in this thesis were recorded by sonic anemometers installed on meteorological towers at the NWTC in Louisville, CO, and were available through the generosity of Dr. Andy Clifton at the NWTC. This section describes the experimental system used to acquire the data, the data processing performed by the NWTC, the in-house data processing, and which data were selected to use for the fitting of the joint distribution.

3.1 NWTC Data and Processing

The NWTC has two 135-m tall meteorological towers, named M4 and M5, that are placed upwind of two test turbines on their Colorado campus. These towers are both heavily instrumented with sensors to measure a variety of atmospheric quantities. A detailed description of the instrumentation, data acquisition, and processing can be found in Refs. [30] and [31].

FIGURE 3.1: Picture of the M4 tower with instrumentation boom reproduced from Clifton [30].

3.1.1 Towers and instrumentation

The discussion of the towers and instrumentation here is limited to the M4 tower because only data from that tower were used in this analysis. The M4 tower is located approximately 2 rotor diameters (300 meters) away from a Siemens wind turbine, upwind in the prevailing wind direction. The towers have a lattice structure to further minimize flow effects and feature instrumentation booms to minimize the tower's effect on the sampled data. The main face of the tower is oriented such that the booms extend at an angle into the prevailing wind direction instead of perpendicular to the wind; this is because the booms would not be able to withstand the wind conditions if they were oriented perpendicularly. A photograph of the top of the M4 tower with an instrumentation boom is shown in Fig. 3.1.

The main instruments of interest on the M4 tower are the six 3D sonic anemometers that are located at 15 m, 30 m, 50 m, 76 m, 100 m, and 131 m above the ground. Sonic anemometers have three sets of orthogonal microphone/speaker arms,

and they measure the 3D wind speed by sending a sonic pulse from one arm to another and measuring the time it takes to travel the known distance. The velocities in the x, y, and z directions can then be calculated by dividing the known distance by the measured time. The anemometers are mounted on long booms, which places them approximately 6.4 m away from the outside of the mast leg. The anemometers installed on site are ATI 'K' Type, with a range between -50 °C to $+60$ °C and an accuracy of 0.1 °C. Tests at the NWTC indicate that the sonic anemometers on the long booms measure over 99% of the free-stream wind when the wind is flowing directly towards the booms.

3.1.2 Data acquisition

The anemometers were sampled at a rate of 20 Hz by a rack-mounted data acquisition system (DAQ) consisting of a National Instruments (NI) chassis with PC and PXI cards. The data were recorded in 10-minute segments, which is conventional for wind turbine applications [19]. The overall acquisition process is controlled by a PC running LabVIEW. Once the anemometers have been sampled, the signals are converted to engineering units, then to binary, and then copied to network storage for future processing.

The binary files are processed by code written by NWTC staff, which converts them to a .mat format suitable for Matlab processing. This .mat file is also used to calculate "derived values" that require averaging over the 10-minute period, such as the mean wind speed and turbulence. The 20 Hz data and the 10-minute averages are then written to the web server to be accessed by users.

3.1.3 NWTC processing

The NWTC runs all of the recorded data through several routines for quality control before saving them in the .mat format. The quality control process generally involves

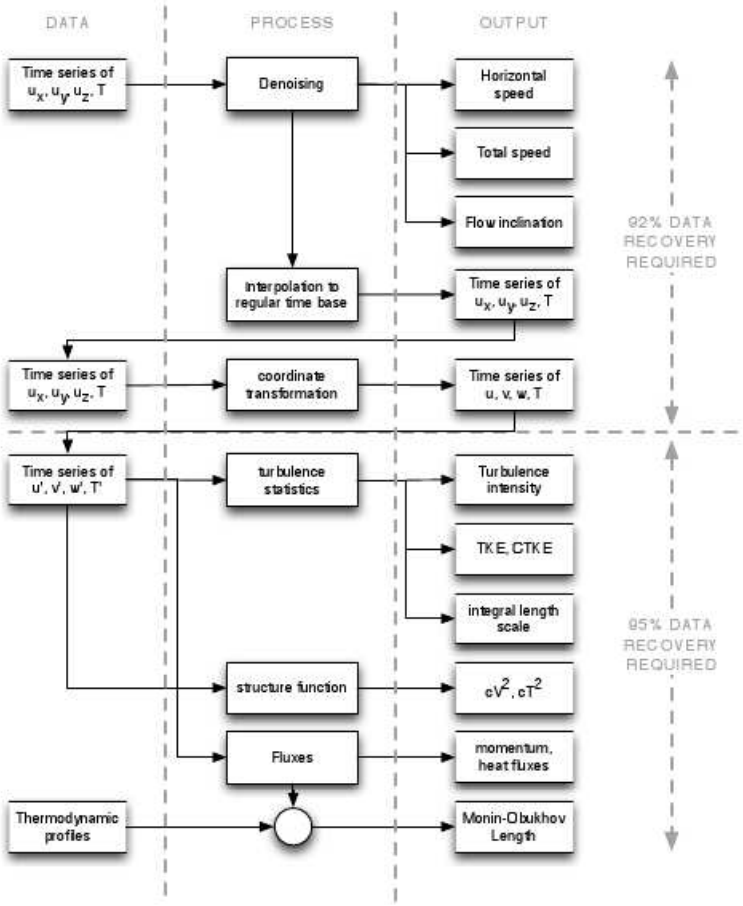

Figure 16. Steps in sonic data processing.

FIGURE 3.2: Diagram of sonic anemometer processing, reproduced from Clifton [30].

checking that the data values do not exceed prescribed manufacturer and user limits to ensure that the signal is within the physical limitations of the sensor and the DAQ and also assigning quality control flags to potentially bad data. Data channels are generally flagged as having failed if there is irregular timing, the data is a constant value (i.e., there has been a malfunction), the channel is empty, the values are all NaNs, or the values are all bad (e.g., -999).

Sonic anemometers undergo even more specific processing to ensure that the recorded data are good. An overview of the sonic anemometer processing utilized by

the NWTC is shown in Fig. 3.2. First, if at least 92% of the original data stream are healthy values, the wind in the three orthogonal directions x, y and z undergo a denoising process. This requires removing "spikes" that occur in the raw wind and/or temperature data. These spikes can be caused by the system instrumentation (if the system loses contact with an instrument, then the channel value is set to -97 in the data stream for that sample) or by insects or particles impacting the sensors. The NWTC processing routine identifies spikes by locating all point-to-point differences that are in the top 1%, then searching for adjacent up-down or down-up jumps. If a spike is detected, it is removed by replacing the value with the average of the points before and after the spike (i.e., linear interpolation). As will be discussed in the next section, this definition of a spike only identifies spikes that are one sample in duration. It was found, however, that many of the data records contain spikes that are more than one sample in length.

In many cases, the DAQ would not sample at exactly 20 Hz. Thus, the next step in the denoising process is to interpolate the recorded data and GPS time stamp from the DAQ into a record with perfectly uniform sampling at 20 samples per second. If there are missing data points in the middle of the record, those points are linearly interpolated; if points are missing at the end of the record and at least 95% of the record was originally error-free, then the mean wind speed is appended to the last missing values. As will be seen below, this appending of the mean creates more spikes in the data records.

Once the records have been cleaned and denoised, they are rotated to generate the downwind, lateral, and vertical components (u, v, and w, respectively). In this rotation process, the cleaned and denoised data are first rotated around the vertical axis such that the average of the lateral wind velocity is zero. The data are then rotated around the new lateral axis until the average of the vertical wind velocity is zero. It should be noted that this rotation only occurs if at least 95% of the original

record is error-free. Note that, with this rotation scheme, the downwind direction is not necessarily horizontal.

From these cleaned, denoised, and rotated records it is then possible to calculate turbulence characteristics, fluxes, and other derived values.

3.2 In-House Processing

It was found that the NWTC data processing methods do not catch and remove all of the data quality issues, so additional processing routines were developed to identify potentially bad data. The main issues present in the data were the presence of quantization and multi-point spikes. The use of linear detrending and its effects on phase coherence are also discussed.

3.2.1 Quantization

It was noted that many records feature some degree of quantization, some of them very severe (see Fig. 3.3). The effects of quantization upon phase difference were initially unknown, so a data flag (3003) was defined to mark those records that featured quantization. A record was defined as quantized if it had 10 quantization "occurrences" in a single record, where a quantization occurrence was defined as having 5 sequential points whose values did not change more than 10^{-12} m/s between each point.

Quantization was a potential concern for this project because it was theorized to have a strong effect on a sample's phase coherence. This hypothesis was based on the facts that a quantized record is essentially a sequence of square waves with different amplitudes and that a square wave has perfect phase coherence because its sinusoidal components must be perfectly aligned to sum up to a square shape. A brief study was conducted to determine if quantized records would skew the phase coherence results.

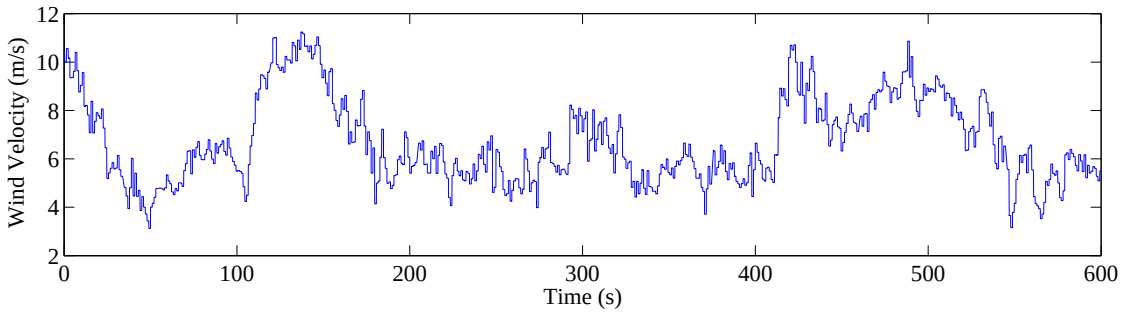

FIGURE 3.3: Plot of wind velocity recorded on October 20, 2013, at 10:20 at a height of 76 m. There is quantization that is evidenced by the rectanguler jumps in the record.

The study started with a Gaussian process generated from a Kaimal spectrum PSD with $L = 340$ m, $U = 5$ m/s, and $\sigma_u = 1$ m/s. Quantization occurrences were introduced at random locations by setting 5 consecutive values equal to the 1.3 times the first value. The number of occurrences and widths were varied to ensure that effects of quantization were thoroughly investigated. It was found that the quantization of records with no phase coherence, even what could be considered severe degrees of quantization, still yields MRLs that are close to zero (on the order of 0.1). However, although the MRL values changed only a litte, there is a visible effect on the distribution. For example, the time history shown in Fig. 3.4 clearly has significant quantization, and a comparison of the histogram of $\Delta\theta$ for the records with and without quantization is shown in Fig. 3.5. The record with no quantization has a $\Delta\theta$ distribution that is essentially uniform, but the distribution of $\Delta\theta$ for the quantized record is clearly non-uniform.

It was additionally found that the method used to detect quantized records could identify several records with sensor malfunctions, such as the record shown in Fig. 3.6. Thus, the method for detecting quantized records could also be used to detect bad data, and therefore any records that were flagged as quantized were removed from subsequent analyses. Doing so left approximately 34,000 records for 15 m, 42,000

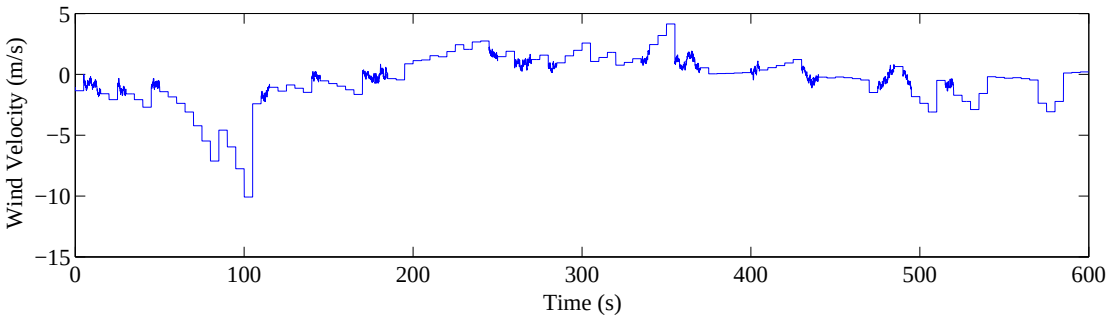

FIGURE 3.4: Synthetic record with severe quantization.

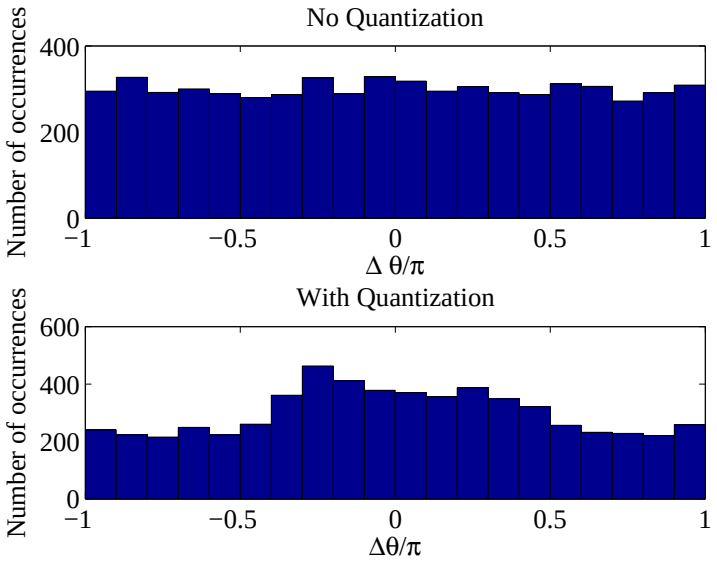

FIGURE 3.5: Histogram of $\Delta\theta$ for a synthetic signal without (top) and with (bottom) quantization.

records for 30 m, 37,000 records for 50 m, 41,000 records for 76 m, 37,000 records for 100 m, and 28,000 records for 131 m.

3.2.2 Spike removal

It was quickly discovered that the spike detection method implemented at the NWTC did not remove many spikes in the data because they were more than one point in length. The potential effects of spikes on phase coherence were investigated using a similar study as in the previous section, except that a single spike of three points in

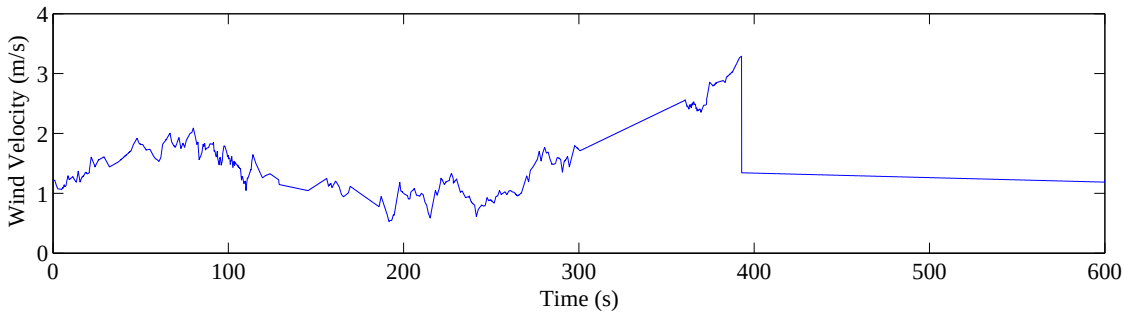

FIGURE 3.6: Example of a sensor malfunction. Recorded on December 8, 2013, at 16:50 at a height of 100 m.

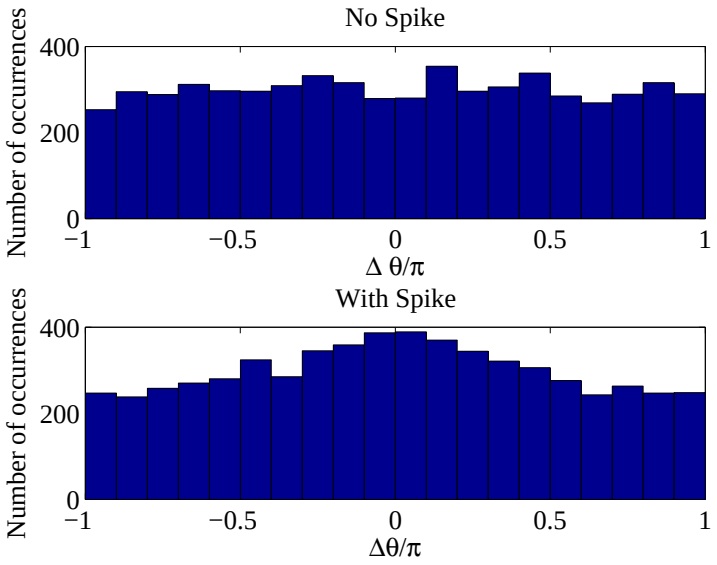

FIGURE 3.7: Histogram of $\Delta\theta$ for a synthetic signal without (top) and with (bottom) a 3-point spike.

duration was added into a Gaussian record generated from a Kaimal spectrum and the change in phase difference values was examined. The presence of spikes was found to have a similar effect on the histogram of $\Delta\theta$ as quantization, as is demonstrated in Fig. 3.7, so it was necessary to determine a method to detect and remove the spikes in the records and flag those records for data quality purposes.

An overview of a selection of spike detection methods is shown in Fig. 3.8. There are two general categories of methods that can be used to detect the presence of spikes

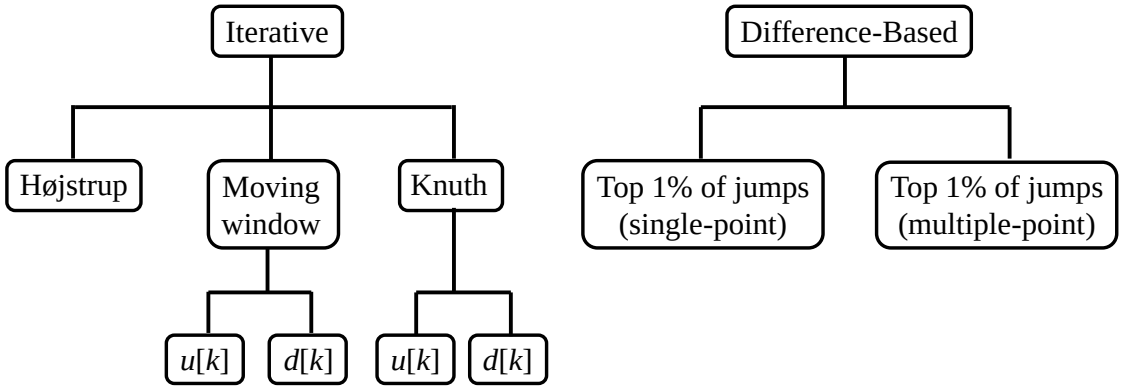

FIGURE 3.8: Overview of a selection of spike detection methods.

in data. The first group scans through the signal and recursively estimates statistics that can be used to identify spikes. Commonly, these statistics are just the mean and standard deviation of the signal, and any points that are more than a few standard deviations from the mean are classified and anomalous and removed either through linear interpolation or with a surrogate model. This group of spike detection/removal methods includes three methods of interest: 1) the procedure prescribed by Højstrup [32], 2) the conventional method of using a moving window, and 3) the author's application of the Knuth algorithm [33] to solve the spike detection problem. Each of the three methods are explained in further detail in the following paragraphs. It should be noted that the moving window and Knuth algorithms can either be applied to the wind data itself, $u(k)$, or the sequence of point-to-point differences, $d(k)$.

The Højstrup method is stated by Vickers and Mahrt [34] to be the basis of their moving average method, so the theory is presented here although it was not considered as a spike detection method. The method has two main components: 1) a recursive, auto-regressive model and 2) recursive estimates of the mean μ_e and standard deviation σ_e of the errors between the predictions and the data. For each data point in the sequence, μ_e and σ_e are used to determine error threshold values corresponding to a specified probability exceedance level (Højstrup used 10^{-5}), assuming

that the errors are normally distributed. Then, the error between the prediction and the data at that iteration is calculated. If the error falls outside the threshold, the point is classified as anomalous and is replaced with the model prediction. The benefit of this method is that there is no need for interpolation, as spike points are simply replaced with the values predicted by the auto-regressive model. This method was ultimately abandoned because it requires two running estimations (one for the auto-regressive model, one for the errors), whereas the other iterative methods only required one.

The moving window method can be found in Vickers and Mahrt [34] and is commonly used for sonic anemometer spike removal. In this recommended method, the mean μ_L and standard deviation σ_L of a moving window of length L are used to calculate a threshold for the wind signal. Thus, if a wind data point is above $\mu_L + \alpha\sigma_L$ or below $\mu_L - \alpha\sigma_L$, the point is identified as a spike. This moving window and threshold can either be applied to the wind signal $u(k)$ or the difference signal $d(k)$. The suggested value for α for $u(k)$ was approximately 3.

The moving window method was tested on both a wind record $u(k)$ and on its difference signal $d(k)$ to evaluate its effectiveness. In both situations, there was a significant variability in the σ_L throughout the record, even when using a window length of 10 seconds (200 points). Because the thresholds were based on the previous data, the method was unable to quickly respond to sudden changes in the signal and it was necessary to make α large to prevent false positives. This, in turn, led to false negatives for smaller spikes that occurred after sudden changes in the wind. The moving window worked marginally better with the difference signal instead of the wind signal, but there were generally many false positives and false negatives. Thus, the method often failed to detect spikes and was therefore abandoned.

Another recursive method that was investigated was the application of the Knuth algorithm [33] to the spike detection problem. The Knuth algorithm recursively

updates the mean and variance of a sample as more points are added; in this case, the sample at time n is simply the first n points of the wind velocity record. The algorithm uses the following equations to update the mean and variance:

$$m(n) = m(n-1) + [u(n) - m(n-1)]/n, \tag{3.1}$$

$$S(n) = S(n-1) + [u(n) - m(n-1)][u(n) - m(n)], \tag{3.2}$$

$$\sigma(n) = \sqrt{S(n)/(n-1)}. \tag{3.3}$$

The threshold at a given point n is then calculated according to $\mu(n) \pm \alpha\sigma(n)$.

The downside of this algorithm is that it was intended for the online estimation of a sample's mean and variance. In other words, it assumes that the signal is stationary and thus is not meant to track changes in the mean. Thus, using this algorithm on $u(k)$, in which the short-term mean varies quite a lot over the record, led to a bias that produced bad estimates of the variance and subsequent threshold values that were too large.

On the other hand, the difference signal $d(k)$ is very close to a zero-mean signal. Thus, the inability of the algorithm to track the mean is not an issue, and it estimates the variance fairly well. However, determining the initial values for the algorithm proved to be difficult. If the initial values for the variance are too large and a spike occurs at the beginning of the record, then the spike is missed. If the values are too small, then non-spike behavior is flagged as a spike. The "best" values for the initial variance estimate vary from record to record with the turbulence, so it was thought to initiate the variance estimate with the value of the variance of the overall record. However, the presence of spikes affected the overall variance, and so this was not a robust method for choosing the initial algorithm values. Thus, the focus shifted away from iterative methods and back to difference-based methods.

The single-point difference method (i.e., $d(k) = u(k+1) - u(k)$) as implemented by the NWTC was unable to detect spikes that were multiple points in length,

nor could it detect spikes that had a point in the middle of a jump. The NWTC spike detection method was therefore modified to use multiple-point differences (i.e., $d(k) = u(k+m) - u(k)$, where m is an integer of choice), which is able to detect spikes with points in the middle of jumps and can be modified to detect spikes that are more than one point long. The method also needed to detect and remove spikes in the first or last few points of the record that were only one jump, not an up-down or down-up pattern. These anomalous spikes could occur by chance—i.e., if the data segment ended mid-spike—but there were also many artificial spikes at the end of records caused by an aspect of the NWTC processing. Specifically, if a record was not 12,000 points long, the mean value was appended to the end of the record.

The ultimate denoising process used in this thesis has the following steps:

1. Calculate the three-point differences $d_3(k) = u(k+2) - u(k)$.

2. Take the absolute value of $d_3(k)$ and determine the 90% threshold value, d_{90}.

3. Construct a boolean vector, $I_{up}(k)$, of the same length as $d_3(k)$ whose elements are 1 if $d_3(k) \geq d_{90}$ and 0 else. Construct a similar vector, $I_{down}(k)$, whose elements are 1 if $d_3(k) \leq -d_{90}$ and 0 else.

4. Find and remove any large jumps in the first or last few points of the record.

5. For each non-zero element in I_{up}:

   - Determine if there is a non-zero element in I_{down} between $k - L_{SW}/2$ and $k + L_{SW}/2$, where L_{SW} is a set maximum width of a spike ($L_{SW} = 6$ points in this thesis).
   - If so, use linear interpolation to replace the values of $u(k)$ between the two non-zero elements in I_{up} and I_{down}.
   - Replace the non-zero elements in I_{up} and I_{down} with zeros.

- If there is no matching non-zero element in I_{down}, delete the non-zero element in I_{up}.

6. Return to 1 and repeat the entire procedure once to ensure that no spikes have been missed.

3.2.3 Detrending

Initially, the records were not linearly detrended after spike removal to ensure that the phase difference values were accurate. However, it was found that records with very small turbulence but a significant linear trend over the 10-minute period had abnormally high phase difference values due to the linear trend. A brief study revealed that detrending records with higher turbulence does not significantly affect the phase difference properties. Records with lower turbulence, however, had phase difference that shifted from high to low after detrending, as expected. Thus, all records were linearly detrended after the spikes were removed from the record.

3.3 Data Used to Fit Joint Distribution

The dataset includes records from 2012, 2013, and 2014, but the concern was raised whether the overlap of certain seasons would overweight certain behaviors. A histogram of the number of records by day of the year is shown in Fig. 3.9, where each subplot corresponds to a different height on the tower. As can be seen in the figure, there are significantly more records for the first 50 days of the year. It was found that these days also correspond to generally higher mean wind speeds and turbulence, so the decision was made to use only the data recorded for the 2013 calendar year. The histogram for the records from 2013 versus the day of year for the 6 different heights is shown in Fig. 3.10. As can be seen in the figure, there is generally good coverage throughout the year, with the exception of 76 m, 100 m, and 131 m for several days in May and June.

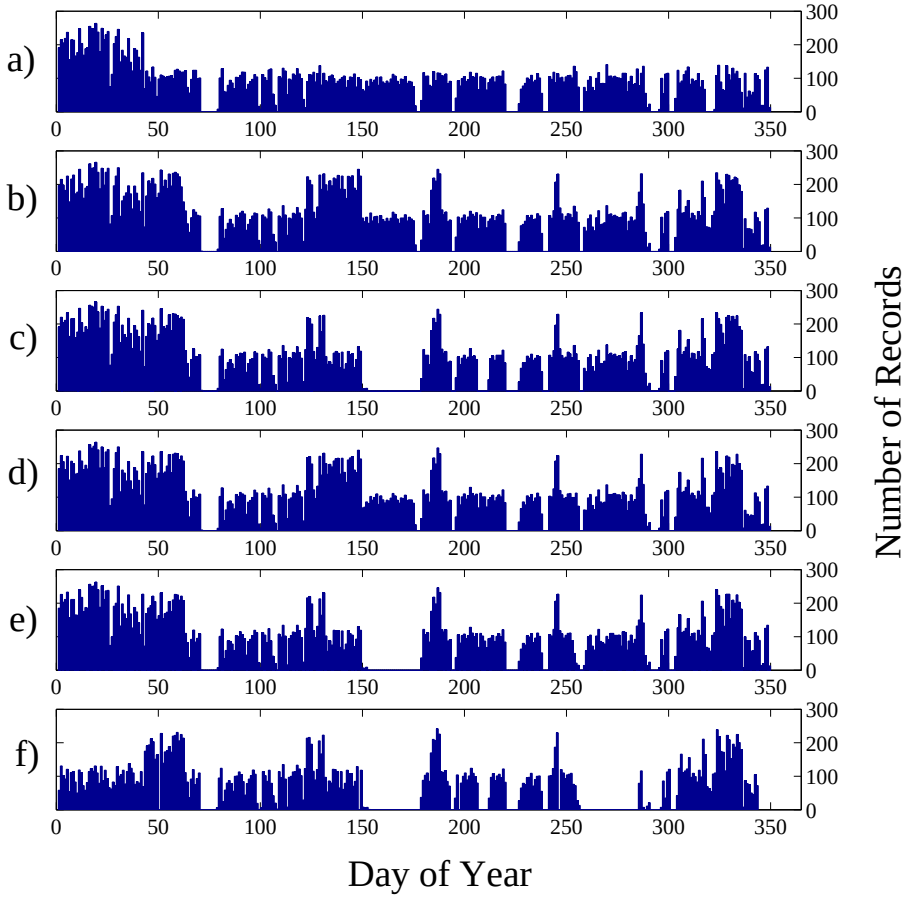

FIGURE 3.9: Number of records per day of year for entire dataset at a) 15 m, b) 30 m, c) 50 m, d) 76 m, e) 100 m, and f) 130 m.

The denoising procedure will not remove multiple spikes that are adjacent to one another, so there were still some spikes that were not detected and as such were present in the processed records. These spikes tended to cause MRL values that were abnormally high; it was also found that some records with obvious sensor malfunctions had high MRL values as well. Thus, all records with MRL values above 0.7 were not used.

In summary, a record was retained in the data set used to fit the joint distribution if it satisfied all of the following characteristics:

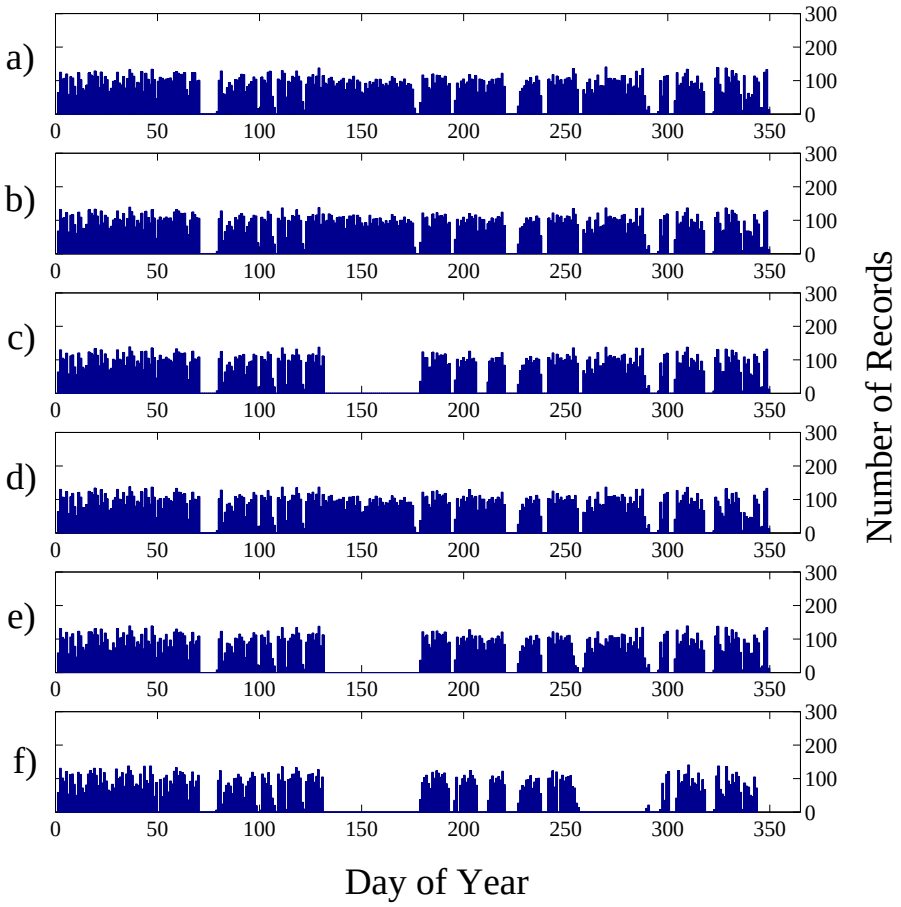

FIGURE 3.10: Number of records per day of year for 2013 at a) 15 m, b) 30 m, c) 50 m, d) 76 m, e) 100 m, and f) 130 m.

- No NaN values;
- 12,000 points in length;
- Recorded in 2013;
- No quantization;
- 20 or fewer spikes that were interpolated;
- Linearly detrended;
- Have $\bar{R} \leq 0.7$.

The total number of records that satisfied these criteria was 159,468, with the height

Table 3.1: Distribution of records in dataset by height.

Height (m)	No. of records
15	29,472
30	30,417
50	25,133
76	29,386
100	24,703
131	20,357

distribution shown in Table 3.1.

4

Data Analysis

The ultimate goal of this research is to develop a parameterized statistical model that can be used to accurately simulate a nonstationary wind velocity rcord, which requires a joint PDF of the variables of interest. The method recommended in the IEC 61400-1 [6] for simulating wind records requires specifying the mean wind speed U, the turbulence σ_u, and the Kaimal length scale L; our proposed modification also requires the specification of the MRL $\bar{R}$ and the mean direction $\bar{\theta}$. The mean direction is assumed to be independent of the other variables, and the joint PDF of the other four parameters can be completely defined by marginal distributions of the variables and by their correlations in a standardized Guassian space. This section discusses an overview of how the parameters were estimated (Section 4.1), the estimation of the Kaimal length scale from the records (Section 4.2), the marginal distributions that were found to best fit the data, (Section 4.3), and the results of the correlation study (Section 4.4).

4.1 Metadata

To fit the marginal PDFs to the data set, it was necessary to process each record and determine the four variables necessary for simulation: the mean wind speed U, turbulence σ_u, Kaimal length scale L, and the MRL $\bar{R}$. The first three variables are necessary to specify the PSD of the process, and the last variable describes the phase coherence of the process. Again, the mean direction only produces a time shift in the envelope, so it is assumed to be independent (this decision is supported by low correlation values, as will be presented later) and is therefore not a parameter in the joint distribution. The processing also extracted and stored the minimum and maximum wind speeds, the mean direction, and the slope of the linear detrending line for possible analysis. All of the parameters, with the exception of the slope of the detrending line, were calculated from a denoised and detrended record. The parameters from each record were stored as a row in a metadata array in Matlab, which contained the time stamp; the quality control flag; the mean, minimum, and maximum wind speeds; the turbulence; the MRL; the mean direction; the Kaimal length scale; and the slope of the detrending line. The Kaimal length scale cannot be directly extracted from the wind velocity record, like the other parameters; the next section specifies how the length scale was determined for each record.

4.2 Kaimal Length Scales

The power spectral density for wind velocity in this study is modeled with a Kaimal spectrum, which is parameterized by the mean wind speed U, the turbulence intensity σ_u, and the Kaimal length scale L. The Kaimal length scale, or turbulent length scale, can be viewed as the spatial length scale of the dominant vortices; smaller Kaimal length scales mean more energy at higher frequencies. The Kaimal spectrum

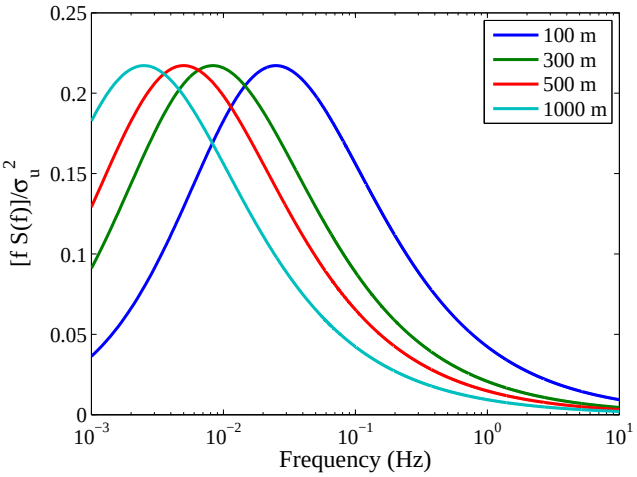

FIGURE 4.1: Plot of nondimensionalized Kaimal spectrum for $L = 100$ m, 300 m, 500 m, and 1000 m.

is often presented in the literature in its nondimensional form:

$$\frac{S(f)f}{{\sigma_u}^2} = \frac{4fL/U}{(1+6fL/U)^{5/3}}. \tag{4.1}$$

A plot of the nondimensionalized version of the Kaimal spectrum for different length scales is shown in Fig. 4.1.

The Kaimal length scale must be determined by fitting the Kaimal spectrum to the data, and there is no discussion in the literature regarding the choice of cost function or best optimization method. Several different cost functions and optimization methods were tested and are discussed in the following subsections, along with the final method that was selected for this thesis.

4.2.1 Cost function

Four cost functions were considered:

$$J = \sum_{i=1}^{N/2} \left[S(f_i)\Delta f - \hat{S}_i \right]^2, \tag{4.2}$$

$$J = \sum_{i=1}^{N/2} \left[\frac{f_i\, S(f_i)\Delta f}{{\sigma_u}^2} - \frac{f_i\, \hat{S}_i}{{\sigma_u}^2} \right]^2, \tag{4.3}$$

$$J = \sum_{i=1}^{N/2} \left[\log[S(f_i)\Delta f] - \log[\hat{S}_i] \right]^2, \tag{4.4}$$

$$J = \sum_{i=1}^{N/4} \left[S(f_i)\Delta f - \hat{S}_i \right]^2. \tag{4.5}$$

In Eqs. (4.2) through (4.5), $S(f_i)\Delta f$ is an estimate of the spectral power at frequency f_i for the Kaimal spectrum and $\hat{S}_i$ is the spectral power at f_i from the record. Equation (4.2) represents the sum-of-squared-error (SSE) of the standard form of the Kaimal spectrum (Eq. (1.3)). Equation (4.3) represents the SSE of the nondimensionalized form of the Kaimal spectrum (Eq. (4.1)). Equation (4.4) is the SSE of the log of the standard Kaimal spectrum. Lastly, Eq. (4.5) is the SSE of the standard Kaimal spectrum, but only for frequencies up to 5 Hz.

When considering the choice between Eq. (4.2) and Eq. (4.3), the nondimensionalized form seems to be a residue of convention and is not based on any scientific reasoning. Thus, there does not seem to be any logical reason to use the nondimensionalized form in the cost function. Additionally, the nondimensional cost function can be viewed as the standard cost function with frequency weighting, and the use of frequency weighting in this application is undesirable because it will increase the weighting on the high-frequency noise. Thus, the SSE of the standard Kaimal spectrum should be chosen over the SSE of the nondimensional form.

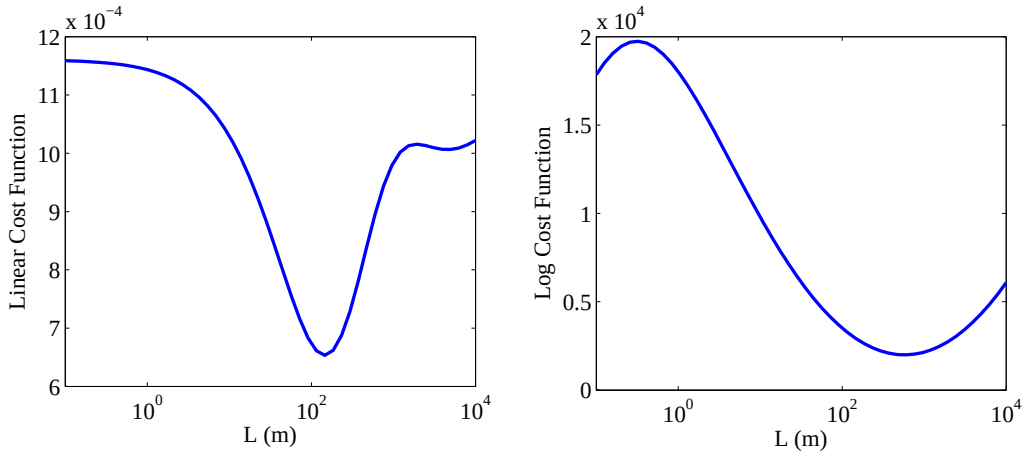

FIGURE 4.2: Comparison of linear (left) and log (right) cost functions versus L for a record recorded at 16:50 on Feb. 21, 2013 at 131 m. The optimal length scale values for the linear and log cost functions are approximately 145 m and 596 m, repsectively.

Second, it was necessary to determine whether to use the "linear" form of the spectrum (see Eq. (1.3)) or to take the log of both sides and define the cost function as the SSE between $\log(S(f))$ and the log of the spectral data. A comparison of both cost functions versus L for a data record is shown in Fig. 4.2. As can be seen in the figure, the choice of cost function affects the optimal length scale—the optimum for the linear cost function occurs around 145 m, and the optimum for the log function is around 596 m. Additionally, note that there are two local minima for the linear cost function of this particular record. This presence of multiple minima in the linear cost function was found in many different records.

In general, the optimal L for the linear cost function was shorter than that for the log cost function, corresponding to more frequency content in the higher frequencies. This is logical, because the log cost function is more affected by the low amounts of power in the higher frequencies, which will shift the optimal solution so there is more power in the lower frequencies. This is demonstrated graphically in Fig. 4.3, which overlays the linear- and log-optimal $S(f)$ over the experimental PSD in a log-log plot.

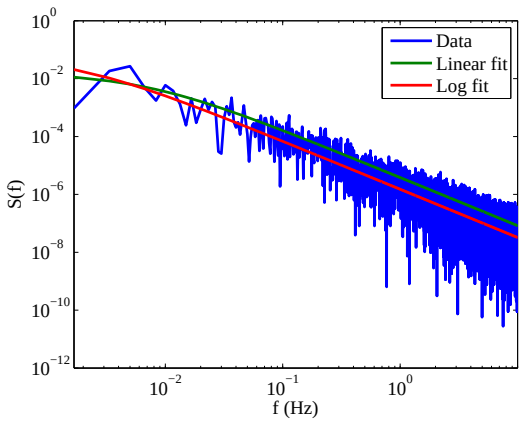

FIGURE 4.3: Comparison of linear and log optimal fits to $S(f)$ for a record recorded at 16:50 on Feb. 21, 2013 at 131 m.

As expected, the fit spectrum from Eq. (4.3) has less power in the higher frequencies. Because the lowest frequency of interest for a large wind turbine is around 1 Hz and because Eq. (4.2) has more energy in that frequency range than Eq. (4.4), the linear SSE is more conservative than the log SSE. It was therefore decided to eliminate Eq. (4.4) as a possible method for calculating L.

The third choice was related to the frequency band used to fit the spectrum. It was possible that low-energy content in the higher frequencies could have a significant effect on the optimal solution when summing over different frequency bands, but a brief investigation found that, with the linear cost function, fitting up to 10 Hz and fitting up to 5 Hz yielded essentially the same optimal solution. Thus, Eq. (4.5) was the final cost function that was eliminated, and it was decided to use Eq. (4.2) to determine the best-fit value for L.

4.2.2 Solution method

The cost functions of interest were sums of the squared errors, so the problem naturally lent itself to the application of least squares optimization techniques. The Kaimal spectrum cannot be transformed such that there is a linear relationship between functions of the spectral value and of the length scale, so linear least-squares

solution methods cannot be used. Thus, an approximately exact optimal solution can only be determined with an iterative solver, such as the Golden Search method or a Gauss-Newton solver, or with a gridding method.

A gridding method was ultimately used to select an approximately optimal L, due to its numerical speed and the smoothness of the cost function. In particular, a grid of lognormally spaced points for L was generated and the value of L that corresponded to the lowest cost was chosen as the optimum. This computation could be vectorized and was sufficiently fast to process the hundreds of thousands of records in the dataset.

4.3 Marginal Distributions

The joint distribution is partially characterized by the marginal distributions, so it was necessary to identify the distributions that best fit the samples of the wind data parameters. The parameters that were investigated were U, σ_u, $\bar{R}$, and L. The details of the distributions for each parameter are discussed in the four subsections below.

For the four parameters, it was found that the majority of the dataset would fit one distribution, but the extremal values would deviate from this main distribution. This deviation in the tail data can be modeled through the use of a generalized Pareto (GP) distribution, which can model exceedance values above a threshold quantile value, P. The use of two distributions does not affect the sampling method: points are drawn from a uniform distribution, and those in (0,P) are mapped using the distribution fit to the main data and those in [P,1) are mapped using the GP distribution. The theory of the GP distribution and how it is fit to the data is presented in the next subsection.

4.3.1 Generalized Pareto distribution

The GP distribution was first introduced by Pickands in 1975 [35] as an alternative to the classic extreme value methods and has been utilized in diverse applications [36, 37, 38, 39, 40]. The classical method for modeling extreme values is to separate the data into groups, select the maximum (or minimum) value, then fit a generalized Extreme Value (GEV) distribution to the set of maximal (or minimal) values. Pickands, on the other hand, proposed a method in which the tail values are transformed into exceedances over a particular threshold and then a GP distribution is fit to the exceedances. The GP distribution has a PDF and CDF of the form

$$f(x) = \begin{cases} \frac{1}{\sigma}\left(1 + k\frac{(x-\theta)}{\sigma}\right)^{-1-\frac{1}{k}} & x > \theta \text{ for } k > 0;\ \theta < x < \theta - \sigma/k \text{ for } k < 0, \\ \frac{1}{\sigma}\exp\left(-\frac{(x-\theta)}{\sigma}\right) & x > \theta \text{ for } k = 0, \end{cases} \tag{4.6}$$

$$F(x) = \begin{cases} 1 - \left(1 + k\frac{(x-\theta)}{\sigma}\right)^{-\frac{1}{k}} & \text{for } k \neq 0, \\ 1 - \exp\left(-\frac{(x-\theta)}{\sigma}\right) & \text{for } k = 0. \end{cases} \tag{4.7}$$

In Eqs. (4.6) and (4.7), the shape parameter k determines the general shape of the distribution, the scale parameter σ affects the spread of the distribution, and the threshold value θ represents the threshold value used to calculate exceedances. The scale parameter must be strictly positive, but the shape and threshold parameters can be any finite, real number. If $k = 0$ and $\sigma = 0$, the GP distribution reduces to the exponential distribution; if $k > 0$ and $\theta = \sigma/k$, it reduces to a Pareto distribution.

There have been multiple proposed methods regarding the fitting of a GP distribution to a sample[41, 42, 43]. Hosking and Wallis [41] looked at the maximum likelihood estimation (MLE) of the parameters with the log likelihood function, which entails choosing a scale and shape parameter such that

$$\log L(\mathbf{x}; \sigma, k) = -n \log \sigma - \frac{(1+k)}{k} \sum_{i=1}^{n} \log\left[1 + k\frac{(x_i - \theta)}{\sigma}\right], \tag{4.8}$$

where the x_i are the data, is maximized[1]. They noted that, if the scale parameter is between $-\frac{1}{2}$ and $\frac{1}{2}$, the parameters derived by MLE are only more accurate than estimators derived from the Method of Moments (MOM) or the method of Probability-Weighted Moments (PWM) if the sample size is larger than 500. Grimshaw [42] investiaged a method to reduce the two-parameter numerical search in MLE to a one-parameter numerical search. Castillo and Hadi [43] noted that the maximum likelihood estimates do not exist for $k < -1$ and may be inaccurate for $k < -0.5$; also, the the use of PWM and MOM is impossible for $k \geq 0.5$ due to the non-existence of moments of order 2 or higher, and they may produce inaccurate answers regardless of the shape parameter. They proposed a method to estimate the GP parameters and quantiles that generally produces more accurate estimations than MLE, MOM, and PWM.

This thesis utilizes the function `gpfit` in Matlab, which uses MLE to determine the best GP parameters. As will be seen in later sections, the fit scale parameter for all of the wind parameters at all heights is between -0.5 and 0.5 and the sample sizes are generally large, so MLE is expected to perform acceptably well in this application. `gpfit` uses the Nelder-Mead algorithm to determine the optimal solution of the negated log-likelihood function:

$$J = n \log \sigma + \frac{(1+k)}{k} \sum_{i=1}^{n} \log \left[1 + k \frac{\hat{x}_i}{\sigma} \right], \tag{4.9}$$

where $\hat{x}_i = x_i - \theta$ and $x_i \geq \theta$.

4.3.2 Mean wind speed

The variability of the mean wind speed is commonly modeled with a Weibull distribution [44, 45, 46, 47, 48, 49]. The Weibull distribution has a PDF and CDF of the

[1] Note that Hosking and Wallis define the shape parameter with the opposite sign in their paper and that there is an error in their equation.

form

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, \tag{4.10}$$

$$F(x) = 1 - e^{-(x/\lambda)^k}, \tag{4.11}$$

where k is the shape parameter and λ is the scale parameter. The IEC standard [6] recommends assuming a Rayleigh distribution for the mean wind speed, which is a Weibull distribution with $k = 2$.

It is possible to transform the Weibull CDF to get a linear relationship between a function of the CDF value F and a function of x. Doing so yields

$$\ln[-\ln(1 - F)] = k \ln x - k \ln \lambda. \tag{4.12}$$

These transformed variables can be plotted in a Weibull plot, which is a special type of quantile plot, in which a sample drawn from a Weibull distribution appears a straight line. Then, an estimation of the distribution parameters k and λ can easily be calculated from a linear fit in this transformed space:

$$k = m, \tag{4.13}$$

$$\lambda = e^{-b/m}, \tag{4.14}$$

where m is the slope of the linear fit and b is the y-intercept. The Weibull plots of the data categorized by height are shown in Fig. 4.4, along with superimposed straight-line fits. As can be clearly seen in the figure, the distribution of mean wind speed at the NWTC for the data available in 2013 are not linear in this transformed space and therefore do not follow a Weibull distribution, as is often assumed. This observation prompted the investigation of other types of distributions.

In the end, the data were found to fit a lognormal distribution very well. The

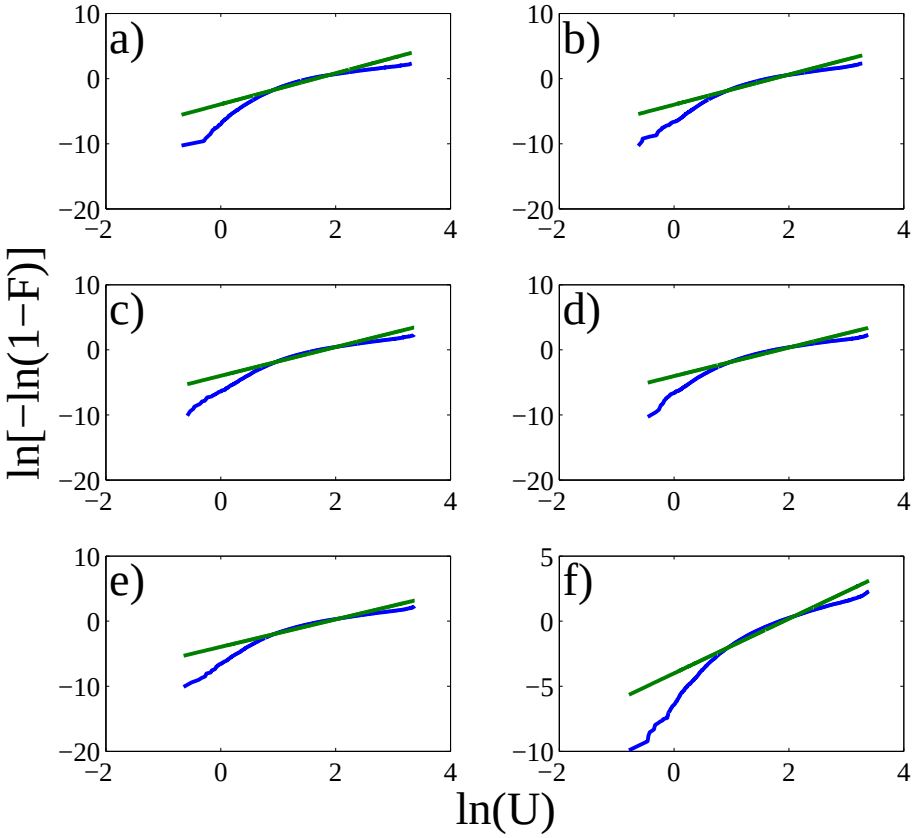

FIGURE 4.4: Weibull plots for U for a) 15 m, b) 30 m, c) 50 m, d) 76 m, e) 100 m, and f) 131 m. The blue line indicates the empirical CDF and the green line is a straight-line fit through the data.

PDF and CDF for a lognormal distribution are given by

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \tag{4.15}$$

$$F(x) = \frac{1}{2} + \frac{1}{2}\text{erf}\left[\frac{\ln x - \mu}{\sqrt{2}\sigma}\right], \tag{4.16}$$

where μ is the mean of $\ln x$, σ is the standard deviation of $\ln x$, and $\text{erf}(x)$ is the standard error function. This CDF can also be transformed in a method similar to that shown for the Weibull plot, which yields

$$\text{erf}^{-1}(2F - 1) = \frac{1}{\sqrt{2}\sigma} \ln x - \frac{\mu}{\sqrt{2}\sigma}. \tag{4.17}$$

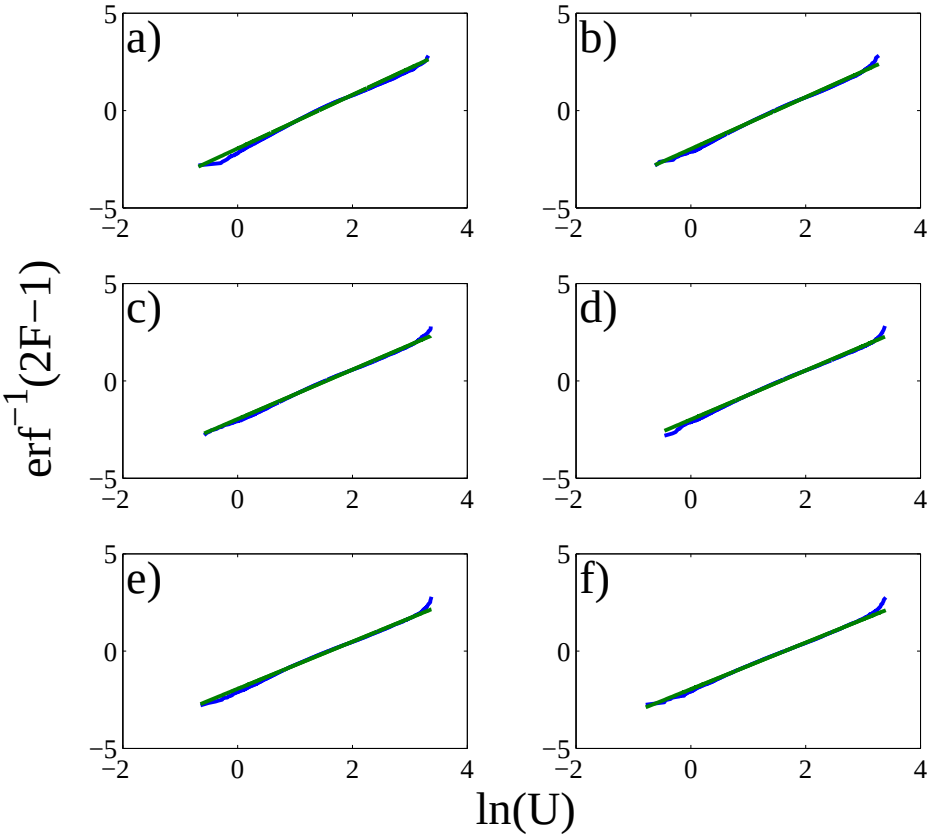

FIGURE 4.5: Lognormal plots for U for a) 15 m, b) 30 m, c) 50 m, d) 76 m, e) 100 m, and f) 131 m. The blue line indicates the empirical CDF and the green line is a straight-line fit through the data.

Transforming the empirical CDF of U according to the above equation and plotting the result yields the plot shown in Fig. 4.5. By comparing Figs. 4.4 and 4.5, it is apparent that the mean wind speed at the NWTC generally follows a lognormal distribution, although the extremal values deviate slightly.

A GP distribution was used to fit the data at the 99% quantile and above. The value for this quantile was chosen by visually determining the quantiles at which the empirical CDF began to diverge from the straight line in the lognormal plot and choosing a value that was well below that point. A comparison of the fit CDF values with and without a GP distribution fit to the tail for the 100 m data is shown in

Fig. 4.6. As can be clearly seen in the figure, using the GP distribution to fit the tail yields a CDF that much better mimics the behavior of the empirical CDF. A plot showing the final bi-distribution fits for the data at the 6 different heights is shown in Fig. 4.7. The parameters for the 6 different bi-distribution fits are listed in Table 4.1. The table also lists the Kolmogorov-Smirnov statistic,

$$D = \sqrt{N} \sup_x |F_{fit}(x) - F_{emp}(x)|, \tag{4.18}$$

where N is the number of points in the sample and x represents the sample points, as a measure of the distribution fit (a lower D indicates a better fit).

In general, the mean wind speed increases with height, which matches the general trend of the wind shear profile commonly assumed [19]. Interestingly, the variation of the mean wind speed also increases with height, which could possibly be due to the presence of nighttime lower-level jets at higher heights.

All of the fit GP shape parameters are between -0.5 and 0, which falls within the region in which the MLE method for GP distributions is generally well-behaved. This is also the region in which the GP distribution has a maximum value of $\theta - \sigma_{GP}/k_{GP}$. The data for 30 m and higher all have roughly the same maximum value (about 11 m/s above the threshold value), but the 15 m tail is closer to exponential ($k_{GP} \approx 0$) so it's maximum value is much larger (about 44 m/s above the threshold value). Lastly, the accuracy of the fit increase with height, but all of the fits are fairly good, considering that N is approximately 25,000 for these data.

4.3.3 Turbulence

The IEC standard [6] recommends the use of a lognormal distribution to characterize the turbulence parameters, and this distribution was found to be the best fit to the data. The turbulence data also had tails that deviated from the general trends, so a GP distribution was once again used to characterize the extreme values. The tails

Table 4.1: Distribution parameters for U bi-distributions and Kolmogorov-Smirnov statistic ($P = 0.99$).

Height (m)	μ_{LN}	σ_{LN}	k_{GP}	σ_{GP}	D
15	1.416	0.5139	-0.06636	2.976	5.203
30	1.485	0.5318	-0.2812	3.492	3.911
50	1.548	0.5575	-0.2680	3.804	3.322
76	1.578	0.5619	-0.3999	4.701	2.782
100	1.601	0.5839	-0.3989	4.3302	2.843
131	1.637	0.5921	-0.3519	3.971	1.716

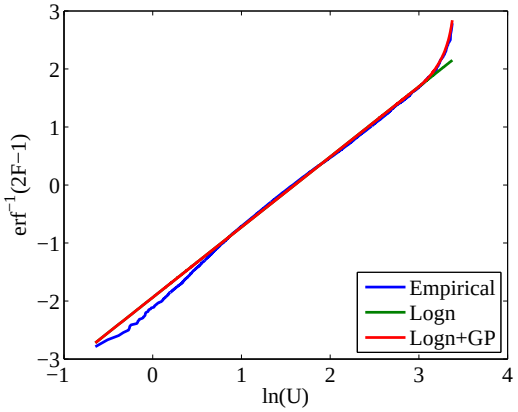

FIGURE 4.6: Comparison of transformed CDFs for the empirical CDF (blue), a best-fit lognormal CDF (green), and a lognormal-GP CDF (red) for the mean wind speed at 100 m.

for the turbulence data start to deviate from lognormal at a lower quantile, so the 95% quantile was used to separate the tail data instead of the 99% quantile.

A plot containing the fit distributions for all of the heights is shown in Fig. 4.8, and the fit parameters are listed in Table 4.2. The fits are generally good, though there is some deviation in the extremely low turbulence values. For this wind parameter, the shape parameters are all close to zero, so it is likely that an exponential distribution could be used to accurately characterize the tails of the turbulence values. A closer inspection (see Fig. 4.9) reveals that, in this case, there is some deviation away from lognormal behavior even in the middle quantiles, which is manifested in the form of a nonlinear relation in these transformed axes. Although it appears significant in

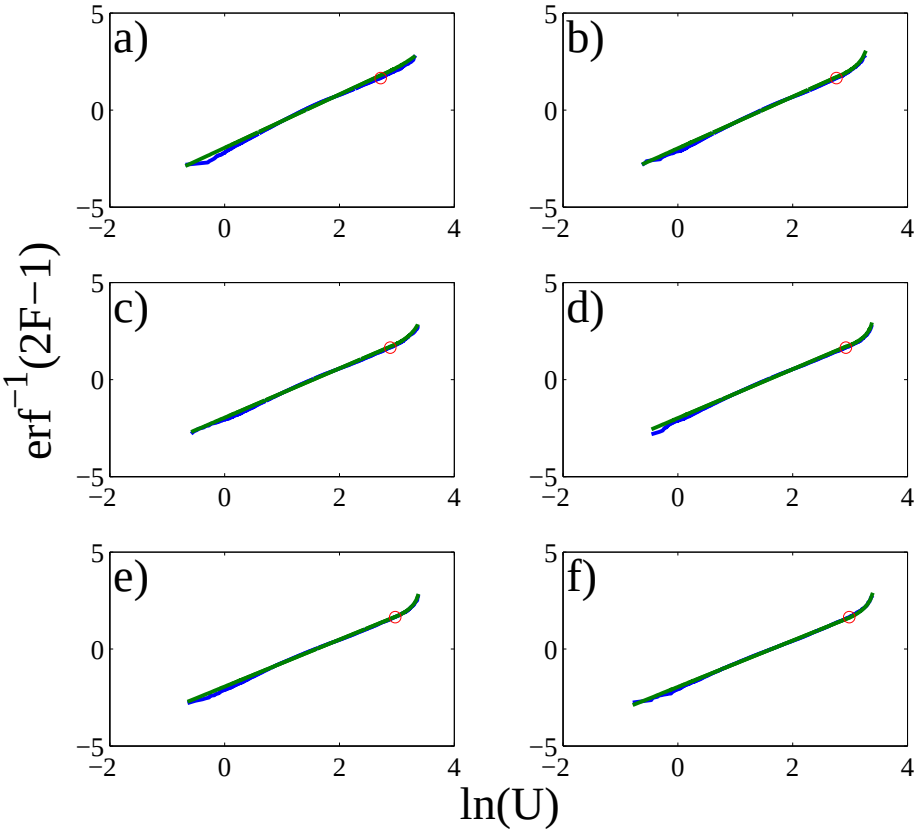

FIGURE 4.7: Lognormal plots for U for a) 15 m, b) 30 m, c) 50 m, d) 76 m, e) 100 m, and f) 131 m. The blue line indicates the empirical CDF, the green line is the transformed CDF calculated with two best fit distributions, and the red point indicates the threshold value used to separate the two distributions (99% quantile).

this transformed space, the Kolmogorov-Smirnov statistic values (see Table 4.2) are still quite good.

4.3.4 Mean resultant length

The MRL was found to follow a Weibull distribution very closely, including the tails, so it was not necessary to fit a GP distribution to the extreme values. A plot containing the Weibull plots for the 6 different heights is shown in Fig. 4.10, and a list of the distribution fit parameters is shown in Table 4.3.

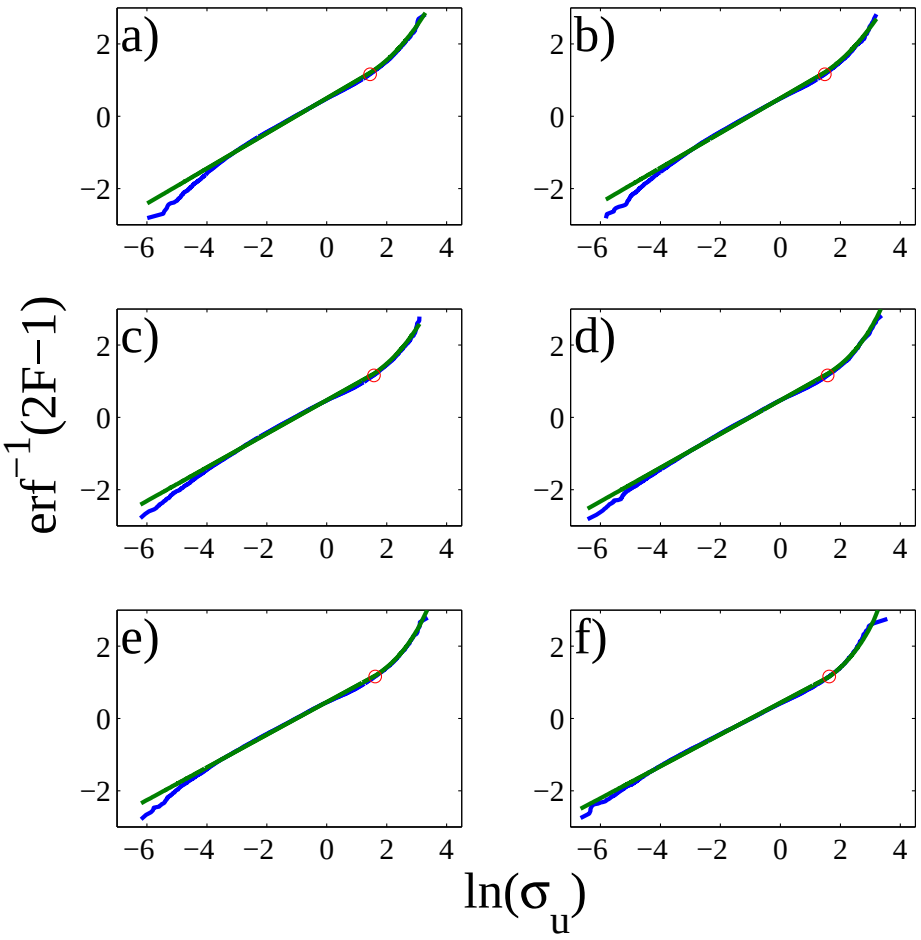

FIGURE 4.8: Lognormal plots for σ_u for a) 15 m, b) 30 m, c) 50 m, d) 76 m, e) 100 m, and f) 131 m. The blue line indicates the empirical CDF, the green line is the transformed CDF calculated with two best fit distributions, and the red point indicates the threshold value used to separate the two distributions (95% quantile).

Table 4.2: Distribution parameters for σ_u bi-distributions and Kolmogorov-Smirnov statistic ($P = 0.95$).

Height (m)	μ_{LN}	σ_{LN}	k_{GP}	σ_{GP}	D
15	-1.029	1.454	0.032	2.689	2.920
30	-1.038	1.469	0.047	2.677	3.063
50	-1.024	1.521	0.011	2.855	3.410
76	-1.011	1.517	-0.026	3.185	3.075
100	-1.004	1.564	-0.037	3.262	3.341
131	-0.990	1.606	-0.035	2.908	2.531

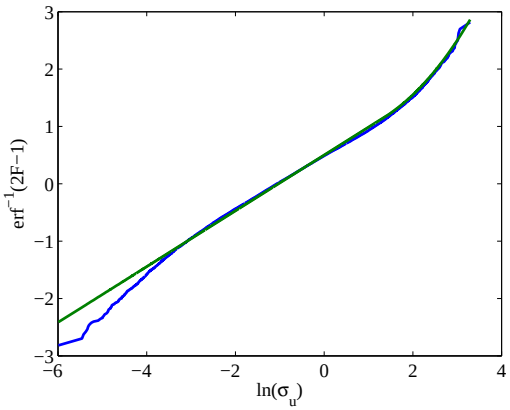

FIGURE 4.9: Magnified plot of the empirical (blue) and fit (green) transformed CDFs of the turbulence for the 15 m data. The data show some variation away from a lognormal distribution in the lower and middle quantiles.

Table 4.3: Distribution parameters for $\bar{R}$ Weibull distribution and Kolmogorov-Smirnov statistic.

Height (m)	λ	k	D
15	0.144	1.752	4.723
30	0.144	1.738	4.466
50	0.145	1.755	4.744
76	0.147	1.728	3.965
100	0.146	1.740	3.943
131	0.148	1.711	3.573

4.3.5 Kaimal length scale

The Kaimal length scales were found to roughly follow a lognormal distribution, although the tail behavior deviated slightly and was not well-captured by a GP distribution. The threshold quantile was chosen to be 95% in order to account for the different tail behavior. A figure containing lognormal plots for the empirical and fit distributions is shown in Fig. 4.11, and the distribution parameters are shown in Table 4.4. The jaggedness of the empirical distribution is due to the gridding method that was used to choose the optimal L. The Kolmogorov-Smirnov statistics shown in Table 4.4 are not as small as the previous fits, but changing the tail fit by small amounts did not affect the statistic, so it is likely that the lower quantile variables

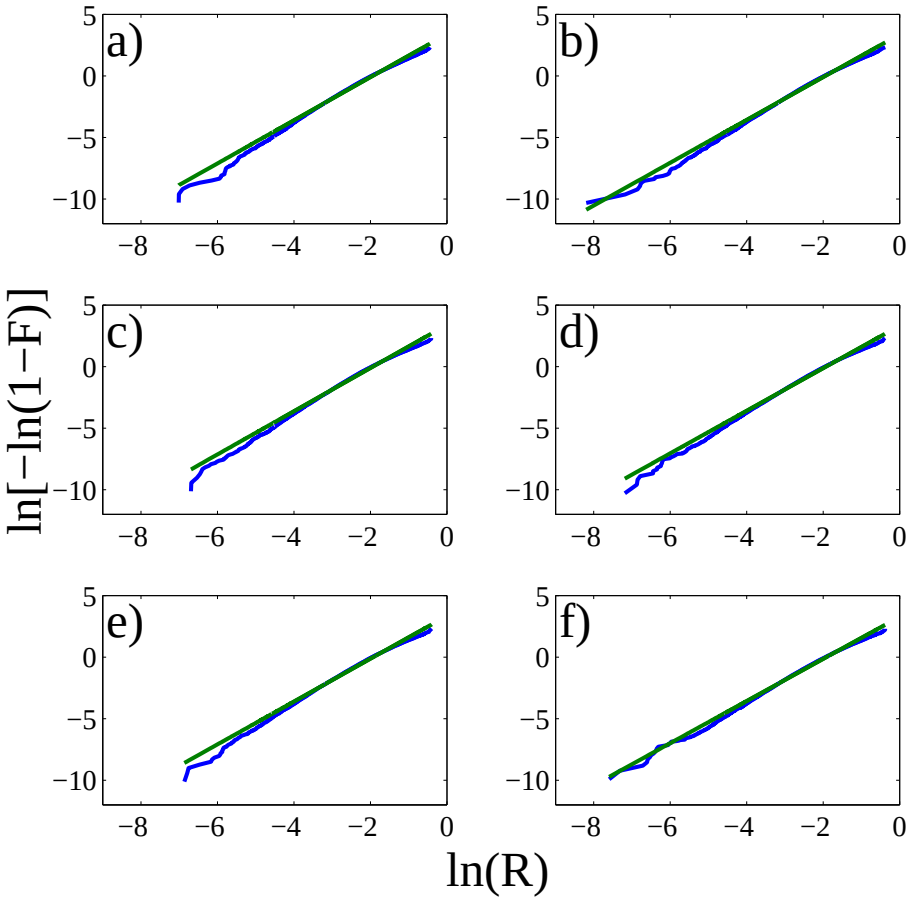

FIGURE 4.10: Weibull plots for $\bar{R}$ for a) 15 m, b) 30 m, c) 50 m, d) 76 m, e) 100 m, and f) 131 m. The blue line indicates the empirical CDF and the green line is the transformed CDF calculated with two best fit distributions.

are the limiting factor in calculating D. Because L is not likely to have a direct effect on wind turbine loads, the authors are content with these fits for the Kaimal length scale.

4.4 Correlations

Two types of correlative studies were completed. First, a simple Spearman correlation of the data for all heights was computed to estimate general trends and to determine any relationship between height and the wind parameters. Second, the

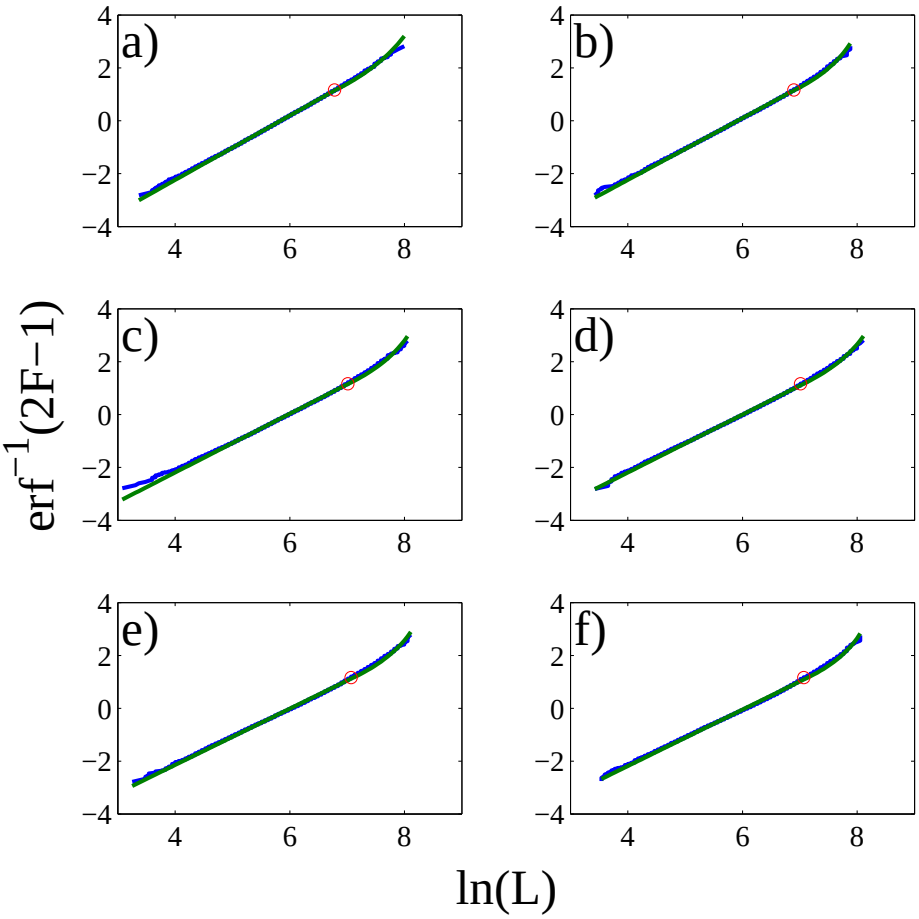

FIGURE 4.11: Lognormal plots for L for a) 15 m, b) 30 m, c) 50 m, d) 76 m, e) 100 m, and f) 131 m. The blue line indicates the empirical CDF, the green line is the transformed CDF calculated with two best fit distributions, and the red point indicates the threshold value used to separate the two distributions (95th quantile).

Table 4.4: Distribution parameters for L bi-distributions and Kolmogorov-Smirnov statistic ($P = 0.95$).

Height (m)	μ_{LN}	σ_{LN}	k_{GP}	σ_{GP}	D
15	5.839	0.581	-0.062	284.781	3.857
30	5.915	0.605	-0.120	321.131	3.914
50	5.973	0.637	-0.102	366.382	3.923
76	6.002	0.646	-0.095	383.525	4.726
100	6.016	0.665	-0.111	410.727	4.809
131	6.049	0.663	-0.137	417.347	4.894

data were categorized by height and a more in-depth analysis was completed to determine the Pearson correlation parameters for the normalized variables that are necessary for the joint distribution.

4.4.1 Spearman correlations

Spearman's rank correlation coefficient, also called Spearman's ρ, is useful in determining the degree of a monotonic relationship between two variables, and it has the advantage over the Pearson correlation in that it does not require the relationship to be linear [50]. A Spearman correlation is in essence the degree of correlation between the CDF values of the different parameters. The value is calculated by first mapping the value of the parameter to its rank in the data set, and identical values are mapped to the average of the ranks (i.e., if u_2 and u_3 would be 5th and 6th, they both take on a rank value of 5.5). Then, the value for the Spearman correlation is calculated by

$$\rho_S = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}, \tag{4.19}$$

where x_i and y_i are the ranks of random variables X_i and Y_i, respectively.

Spearman's ρ is not particularly useful for simulation purposes, unlike the Pearson correlation, but its ability to detect nonlinear monotonic relationships between variables makes it a useful tool for estimating general trends over the entire cleaned dataset and for investigating the relationship between the parameters and height. The Spearman correlation values are shown in Table 4.5, where the significant correlations are in bold. In general, the only parameters that show any significant degree of correlation are U, σ_u, and L. It is interesting to note that U is only slightly correlated with height.

Table 4.5: List of Spearman rank correlation values. The largest correlation value are shown in bold.

	h	U	σ_u	$\bar{R}$	L	$\bar{\theta}$	m
h	1.0000	0.1307	0.0127	0.0115	0.1099	-0.0019	0.0041
U	0.1307	1.0000	**0.6340**	-0.2009	**0.8078**	0.0441	-0.2172
σ_u	0.0127	**0.6340**	1.0000	-0.1026	**0.5409**	0.0123	-0.0700
$\bar{R}$	0.0115	-0.2009	-0.1026	1.0000	-0.0806	0.0200	-0.0533
L	0.1099	**0.8078**	**0.5409**	-0.0806	1.0000	0.0432	-0.1966
$\bar{\theta}$	-0.0019	0.0441	0.0123	0.0200	0.0432	1.0000	-0.1936
m	0.0041	-0.2172	-0.0700	-0.0533	-0.1966	-0.1936	1.0000

4.4.2 Pearson correlations

The ultimate goal of this research was to produce a joint PDF for parameters that can characterize a nonstationary 10-minute wind record. This joint distribution requires the Pearson correlations of the standardized variables. The standardization is done by mapping the variable values to their empirical CDF values, which are uniformly distributed, and then using the inverse standard normal CDF to map the CDF values to a standard Gaussian space. The Pearson correlation, ρ, is given by

$$\rho_S = \frac{\sum_i X_i Y_i}{\sqrt{\sum_i {X_i}^2 \sum_i {Y_i}^2}}, \tag{4.20}$$

where X_i and Y_i are elements of samples of two standardized random variables of interest.

The data were first separated by height, and the correlations were determined for each height; the correlation values for the different heights are shown in Tables 4.6 through 4.11. In general, the only parameters that show any significant correlation to each other for all of the heights are once again U, σ_u, and L; all other correlation values are fairly small. Additionally, comparison plots of the normalized variables for the different heights are shown in Figs. 4.12 through 4.17. These plots only contain 5000 points that were randomly sampled from the available data due to file size constraints on the plots; however, the general trends have been preserved, which are

Table 4.6: Correlation matrix for 15 m data.

	U	σ_u	$\bar{R}$	L	$\bar{\theta}$	m
U	1.0000	**0.7033**	-0.1895	**0.7753**	0.0736	-0.2550
σ_u	**0.7033**	1.0000	-0.0700	**0.5832**	0.0188	-0.0659
$\bar{R}$	-0.1895	-0.0700	1.0000	-0.0416	0.0189	-0.0747
L	**0.7753**	**0.5832**	-0.0416	1.000	0.0626	-0.2278
$\bar{\theta}$	0.0736	0.0188	0.0189	0.0626	1.0000	-0.2307
m	-0.2550	-0.0659	-0.0747	-0.2278	-0.2307	1.0000

Table 4.7: Correlation matrix for 30 m data.

	U	σ_u	$\bar{R}$	L	$\bar{\theta}$	m
U	1.0000	**0.6767**	-0.1939	**0.7956**	0.0392	-0.2465
σ_u	**0.6767**	1.0000	-0.0590	**0.5825**	-0.0004	-0.0594
$\bar{R}$	-0.1939	-0.0590	1.0000	-0.0667	0.0155	-0.0592
L	**0.7956**	**0.5825**	-0.0667	1.0000	0.0346	-0.2204
$\bar{\theta}$	0.0392	-0.0004	0.0155	0.0346	1.0000	-0.1781
m	-0.2465	-0.0594	-0.0592	-0.2204	-0.1781	1.0000

similar for the different heights. Some subplots, such as k, l, and n, correspond to variable pairs with low correlation values but do not feature the expected cloud-like behavior demonstrated in subplots g or h, for example. It is hypothesized that these heart-shaped scatterplots are caused by different mechanisms that occur under different atmospheric conditions. Future work will separate the dataset by atmospheric stability to verify this hypothesis.

The specification of the marginal distributions and correlations completely specify the joint distribution. It is then possible to draw samples from the joint distribution for simulation purposes, as is done in the next chapter.

Table 4.8: Correlation matrix for 50 m data.

	U	σ_u	$\bar{R}$	L	$\bar{\theta}$	m
U	1.0000	**0.6673**	-0.2094	**0.8142**	0.0396	-0.2231
σ_u	**0.6673**	1.0000	-0.1232	**0.5810**	0.0082	-0.0508
$\bar{R}$	-0.2094	-0.1232	1.0000	-0.0865	0.0068	-0.0578
L	**0.8142**	**0.5810**	-0.0865	1.0000	0.0367	-0.1998
$\bar{\theta}$	0.0396	0.0082	0.0068	0.0367	1.0000	-0.1523
m	-0.2231	-0.0508	-0.0578	-0.1998	-0.1523	1.0000

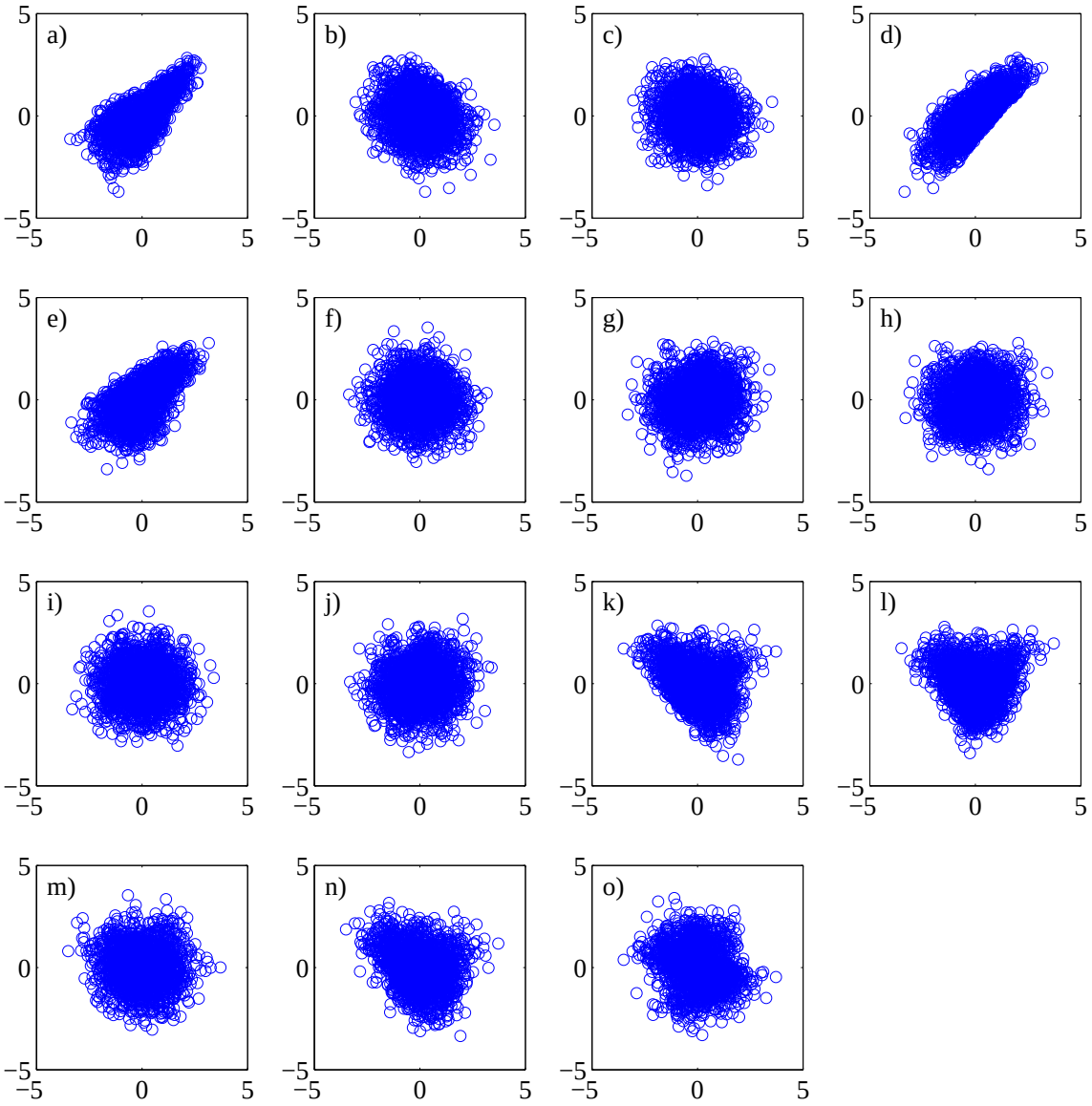

FIGURE 4.12: Plots of 5000 randomly sampled normalized random variables for 15 m. a) U vs. σ_u, b) U vs. $\bar{R}$, c) σ_u vs. $\bar{R}$, d) U vs. L, e) σ_u vs. L, f) $\bar{R}$ vs. L, g) U vs. $\bar{\theta}$, h) σ_u vs. $\bar{\theta}$, i) $\bar{R}$ vs. $\bar{\theta}$, j) L vs. $\bar{\theta}$, k) U vs. m, l) σ_u vs. m, m) $\bar{R}$ vs. m, n) L vs. m, o) $\bar{\theta}$ vs. m.

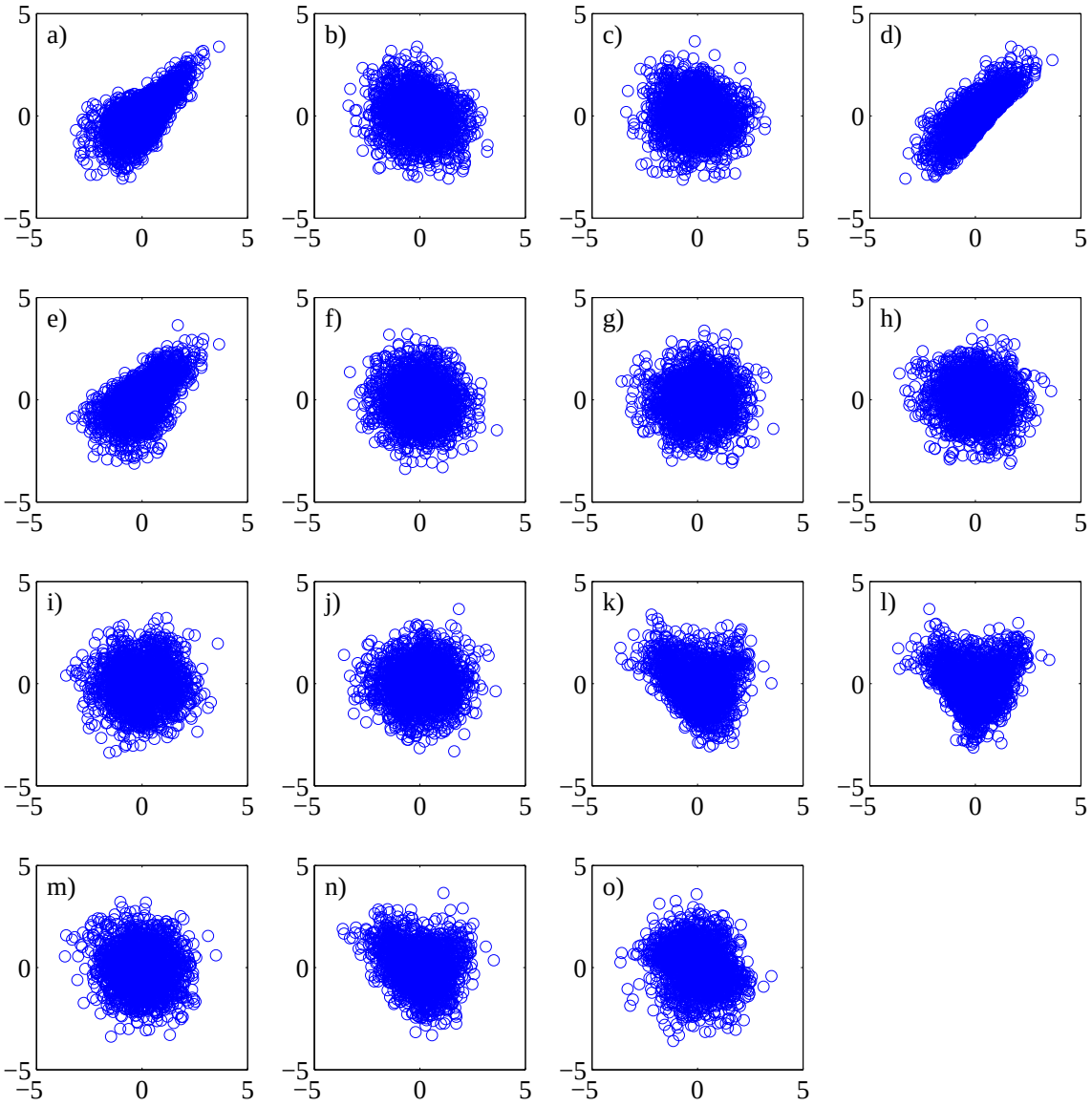

FIGURE 4.13: Plots of 5000 randomly sampled normalized random variables for 30 m. a) U vs. σ_u, b) U vs. $\bar{R}$, c) σ_u vs. $\bar{R}$, d) U vs. L, e) σ_u vs. L, f) $\bar{R}$ vs. L, g) U vs. $\bar{\theta}$, h) σ_u vs. $\bar{\theta}$, i) $\bar{R}$ vs. $\bar{\theta}$, j) L vs. $\bar{\theta}$, k) U vs. m, l) σ_u vs. m, m) $\bar{R}$ vs. m, n) L vs. m, o) $\bar{\theta}$ vs. m.

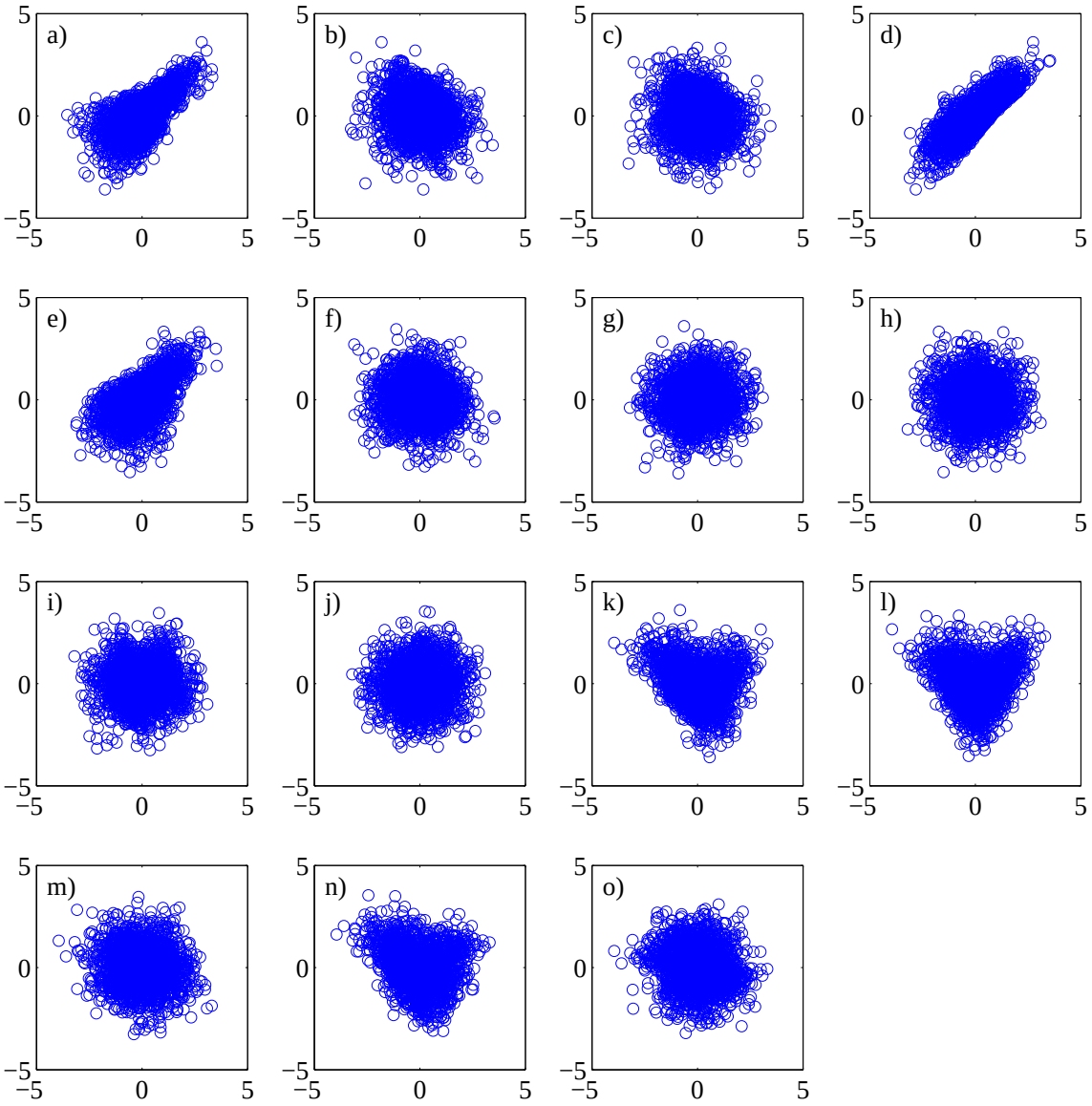

FIGURE 4.14: Plots of 5000 randomly sampled normalized random variables for 50 m. a) U vs. σ_u, b) U vs. $\bar{R}$, c) σ_u vs. $\bar{R}$, d) U vs. L, e) σ_u vs. L, f) $\bar{R}$ vs. L, g) U vs. $\bar{\theta}$, h) σ_u vs. $\bar{\theta}$, i) $\bar{R}$ vs. $\bar{\theta}$, j) L vs. $\bar{\theta}$, k) U vs. m, l) σ_u vs. m, m) $\bar{R}$ vs. m, n) L vs. m, o) $\bar{\theta}$ vs. m.

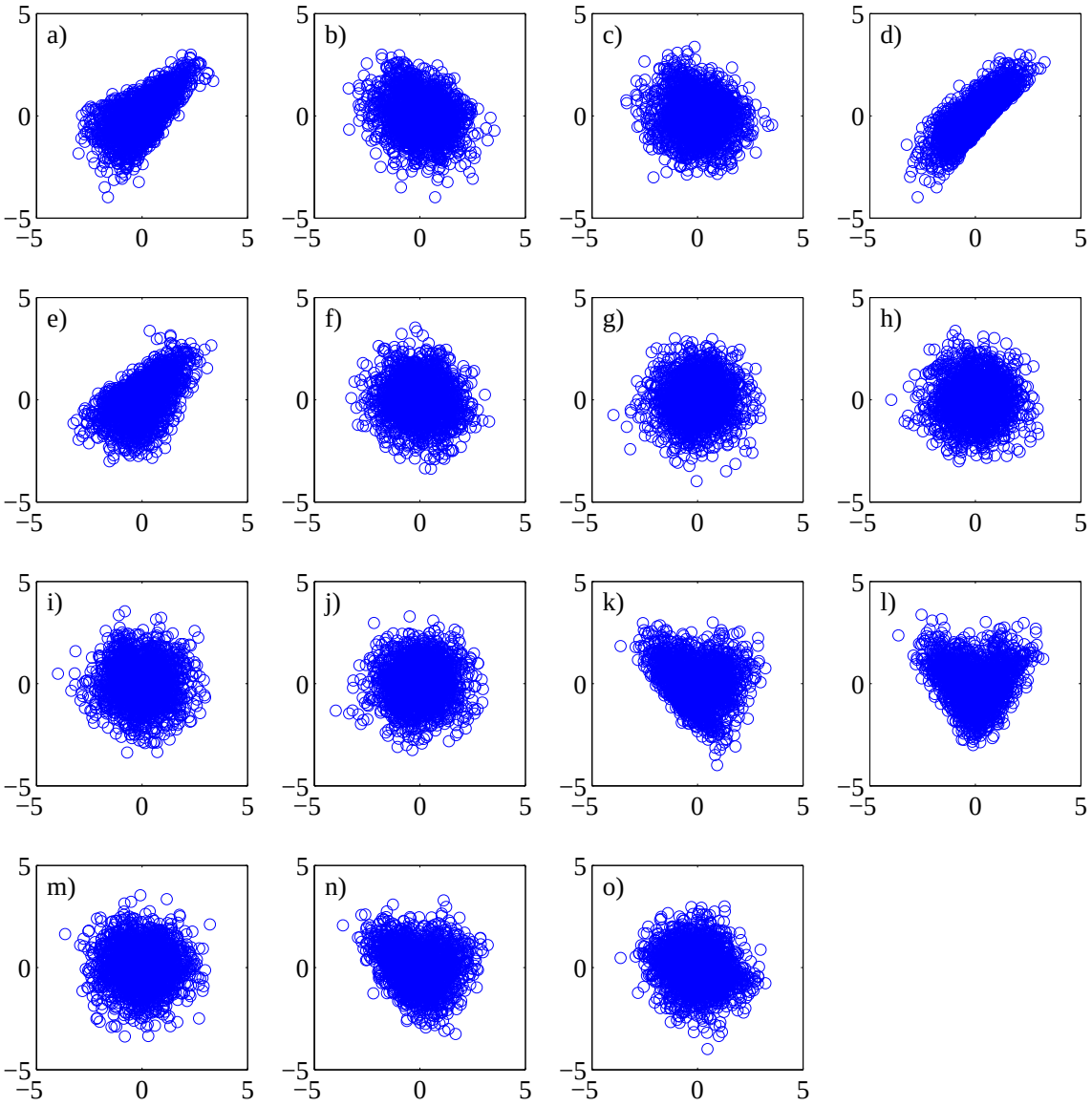

FIGURE 4.15: Plots of 5000 randomly sampled normalized random variables for 76 m. a) U vs. σ_u, b) U vs. $\bar{R}$, c) σ_u vs. $\bar{R}$, d) U vs. L, e) σ_u vs. L, f) $\bar{R}$ vs. L, g) U vs. $\bar{\theta}$, h) σ_u vs. $\bar{\theta}$, i) $\bar{R}$ vs. $\bar{\theta}$, j) L vs. $\bar{\theta}$, k) U vs. m, l) σ_u vs. m, m) $\bar{R}$ vs. m, n) L vs. m, o) $\bar{\theta}$ vs. m.

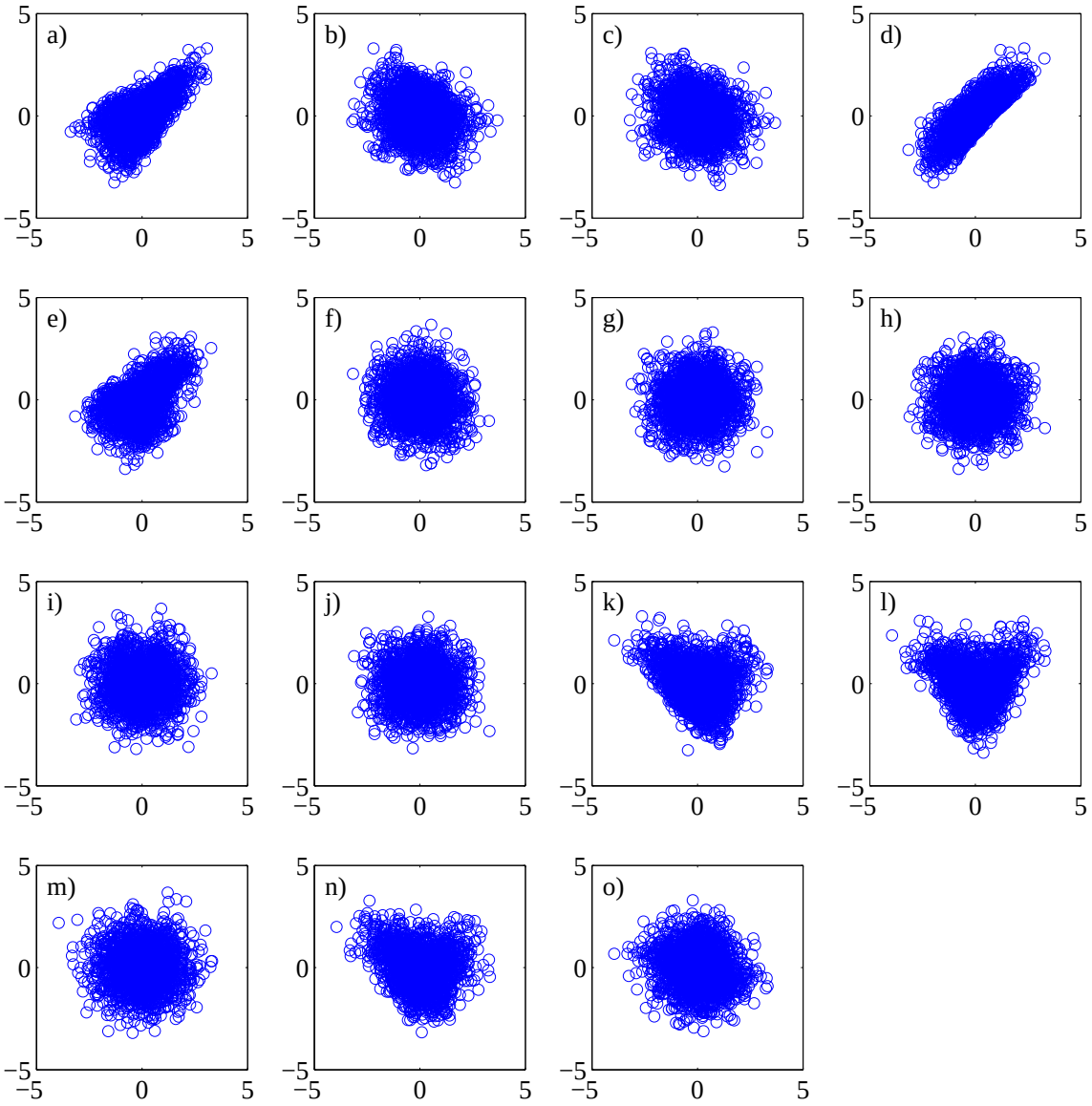

FIGURE 4.16: Plots of 5000 randomly sampled normalized random variables for 100 m. a) U vs. σ_u, b) U vs. $\bar{R}$, c) σ_u vs. $\bar{R}$, d) U vs. L, e) σ_u vs. L, f) $\bar{R}$ vs. L, g) U vs. $\bar{\theta}$, h) σ_u vs. $\bar{\theta}$, i) $\bar{R}$ vs. $\bar{\theta}$, j) L vs. $\bar{\theta}$, k) U vs. m, l) σ_u vs. m, m) $\bar{R}$ vs. m, n) L vs. m, o) $\bar{\theta}$ vs. m.

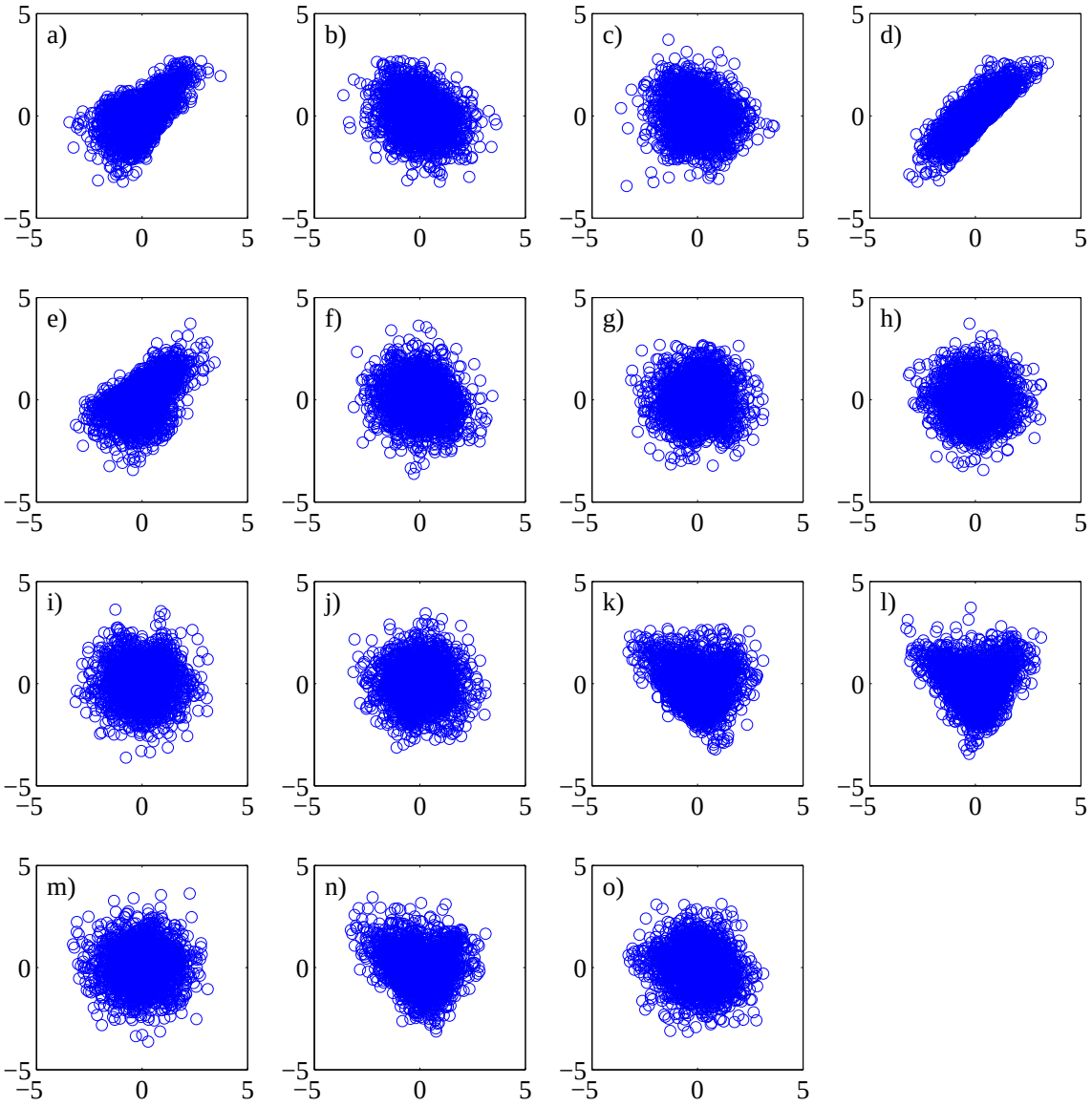

FIGURE 4.17: Plots of 5000 randomly sampled normalized random variables for 131 m. a) U vs. σ_u, b) U vs. $\bar{R}$, c) σ_u vs. $\bar{R}$, d) U vs. L, e) σ_u vs. L, f) $\bar{R}$ vs. L, g) U vs. $\bar{\theta}$, h) σ_u vs. $\bar{\theta}$, i) $\bar{R}$ vs. $\bar{\theta}$, j) L vs. $\bar{\theta}$, k) U vs. m, l) σ_u vs. m, m) $\bar{R}$ vs. m, n) L vs. m, o) $\bar{\theta}$ vs. m.

Table 4.9: Correlation matrix for 76 m data.

	U	σ_u	$\bar{R}$	L	$\bar{\theta}$	m
U	1.0000	**0.6193**	-0.1992	**0.8194**	0.0374	-0.2215
σ_u	**0.6193**	1.0000	-0.1128	**0.5367**	0.0041	-0.0594
$\bar{R}$	-0.1992	-0.1128	1.0000	-0.0786	0.0159	-0.0525
L	**0.8194**	**0.5367**	-0.0786	1.0000	0.0380	-0.1988
$\bar{\theta}$	0.0374	0.0041	0.0159	0.0380	1.0000	-0.1591
m	-0.2215	-0.0594	-0.0525	-0.1988	-0.1591	1.0000

Table 4.10: Correlation matrix for 100 m data.

	U	σ_u	$\bar{R}$	L	$\bar{\theta}$	m
U	1.0000	**0.6262**	-0.2290	**0.8274**	0.0350	-0.2067
σ_u	**0.6262**	1.0000	-0.1585	**0.5314**	0.0148	-0.0514
$\bar{R}$	-0.2290	-0.1585	1.0000	-0.1043	0.0174	-0.0552
L	**0.8274**	**0.5314**	-0.1043	1.0000	0.0362	-0.1907
$\bar{\theta}$	0.0350	0.0148	0.0174	0.0362	1.0000	-0.1501
m	-0.2067	-0.0514	-0.0552	-0.1907	-0.1501	1.0000

Table 4.11: Correlation matrix for 131 m data.

	U	σ_u	$\bar{R}$	L	$\bar{\theta}$	m
U	1.0000	**0.6215**	-0.2844	**0.8287**	0.0176	-0.1912
σ_u	**0.6215**	1.0000	-0.1739	**0.5190**	0.0031	-0.0460
$\bar{R}$	-0.2844	-0.1739	1.0000	-0.1614	0.0195	-0.0421
L	**0.8287**	**0.5190**	-0.1614	1.0000	0.0250	-0.1739
$\bar{\theta}$	0.0176	0.0031	0.0195	0.0250	1.0000	-0.1274
m	-0.1912	-0.0460	-0.0421	-0.1739	-0.1274	1.0000

5

Simulation

This section provides an overview of the proposed modified Sandia method and compares the sampled data to the actual dataset, both in terms of marginal distributions and comparisons in the time domain. Records generated with the standard Sandia method are used as input to an SDOF oscillator, and its response is compared to the response generated when a data record with the same parameters is used as input.

5.1 Method

The parameters for the detrend slope value m and the mean direction $\bar{\theta}$ are not included in the generation of a correlated sample of wind parameters because a) they both have low correlations values with the other parameters and b) their effects on a system's fatigue or ultimate response are likely to be minimal. To be more specific, a gradual increase in the mean wind speed over a 10-minute period will not cause large load excursions in a system; similarly, $\bar{\theta}$ only affects the temporal location of packets of energy, not the shape or magnitude of the packets, and should therefore have a minimal effect on the system response. Thus, in the proposed simulation method, the slope is set to zero and $\bar{\theta}$ is independently drawn from a uniform distribution. The

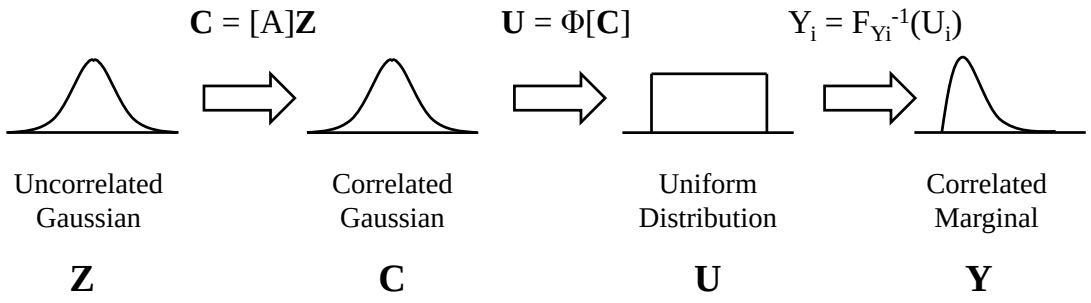

FIGURE 5.1: Schematic of the process for sampling the wind parameters from the joint distribution. Four independent normal variables are correlated, transformed to a uniform space, then to the correct marginal form to produce a sample of U, σ_u, L, and $\bar{R}$.

remaining correlated variables, U, σ_u, L, and $\bar{R}$, are drawn from the joint distribution characterized by the marginals and correlations matrices presented in the previous chapter.

This demonstration of the simulation method will be for a height of 30 m. The full correlation matrix for this example—which is equal to the covariance matrix $[\Sigma]$ due to the unitary variance of the normalized random variables—is presented in Table 4.7. The steps for the simulation process, which involves sampling from the joint distribution and then generating a time series, are given below. A diagram of the sampling process is given in Fig. 5.1.

1. Sample $\bar{\theta}$ from $\mathcal{U}(0, 2\pi)$.

2. Independently sample four standard normal variables, one each for U, σ_u, L, and $\bar{R}$. Place these into a column vector $\mathbf{Z}$.

3. Correlate the Gaussian variables through a linear transformation,

$$\mathbf{C} = [A]\mathbf{Z}, \tag{5.1}$$

where $[A][A]^T = [\Sigma]$ and $[A]$ is the Cholesky decomposition of $[\Sigma]$.[1] This pro-

[1] It is also possible to use the eigenvalues and vectors of $[\Sigma]$ to perform this correlation instead of the Cholesky decomposition. Then $\mathrm{C} = [V][D]^{1/2}\mathrm{Z}$.

duces four correlated Gaussian variables with zero mean and unitary variance.

4. Use the standard normal CDF to transform the correlated variables into a uniform probability space:

$$\mathbf{U} = \Phi\left(\mathbf{C}\right). \tag{5.2}$$

5. Transform the variables to their correct marginal form by using the inverse CDF of each marginal distribution:

$$U = {F_U}^{-1}(U_1), \tag{5.3}$$

$$\sigma_u = {F_{\sigma_u}}^{-1}(U_2), \tag{5.4}$$

$$L = {F_L}^{-1}(U_3), \tag{5.5}$$

$$\bar{R} = {F_{\bar{R}}}^{-1}(U_4). \tag{5.6}$$

6. Calculate the Kaimal PSD from U, σ_u, and L and use it to specify the magnitudes of the Fourier vector.

7. Use Eq. (2.10) and `roots` to determine the value of κ from $\bar{R}$. Set $\mu = \bar{\theta}$, then use the method outlined in Section 2.2.2 to draw a sample of $N/2$ phase difference values from a von Mises distribution with the calculated μ and κ. Use `cumsum` to calculate the corresponding phase angles and construct the first half of the Fourier vector.

8. Assemble the full Fourier vector from the unique first half, then call `ifft` to produce a time history.

9. Scale and shift the record appropriately to ensure it has the correct mean wind speed and turbulence.

The last step in this simulation process is necessary due to the continuous form of the Kaimal PSD; choosing $|X(f_i)| = \sqrt{S(f_i)\Delta f/2}$ will not produce a time series

whose standard deviation is equal to σ_u. Thus, normalizing the IFFT result by its standard deviation, multiplying by σ_u, then adding U to the record will ensure the record has the correct mean and variance (assuming the original IFFT result is zero-mean).

5.2 Comparison of Marginals with Data

To verify that the sampled values have similar distributions to the dataset, 10,000 samples were drawn from the joint distribution for the 30 m height and compared with the 30 m dataset. The plots of the sample marginals for the four different parameters are compared with the marginals for the entire dataset in Fig. 5.2. Each subplot corresponds to a different parameter (clockwise from top left is U, σ_u, L, and $\bar{R}$), and the plots are either Weibull or lognormal plots to allow ease of comparison for the main distribution and the tail. The marginals for the sample and for the entire dataset fit well, so the sampling method is verified.

5.3 Comparison of Time Records

In addition to verifying that the correct marginal distributions are recreated with the sampling method, it is also important to verify that the simulations in the time domain have a similar appearance to the data. Many records were found to be qualitatively similar to the data, especially for smaller MRL values. The simulations for certain records with higher MRL values, however, were not found to be visually similar to the data record with the same parameters.

An example record was chosen, recorded at 03:30 on July 11, 2013 at a height of 50 m, to demonstrate this difference between data and simulation and to investigate the possible origins of the differences between the records. This particular record has a very high degree of phase coherence ($\bar{R} = 0.66$) and therefore a high degree of nonstationarity, as demonstrated by the time history shown in the top plot of

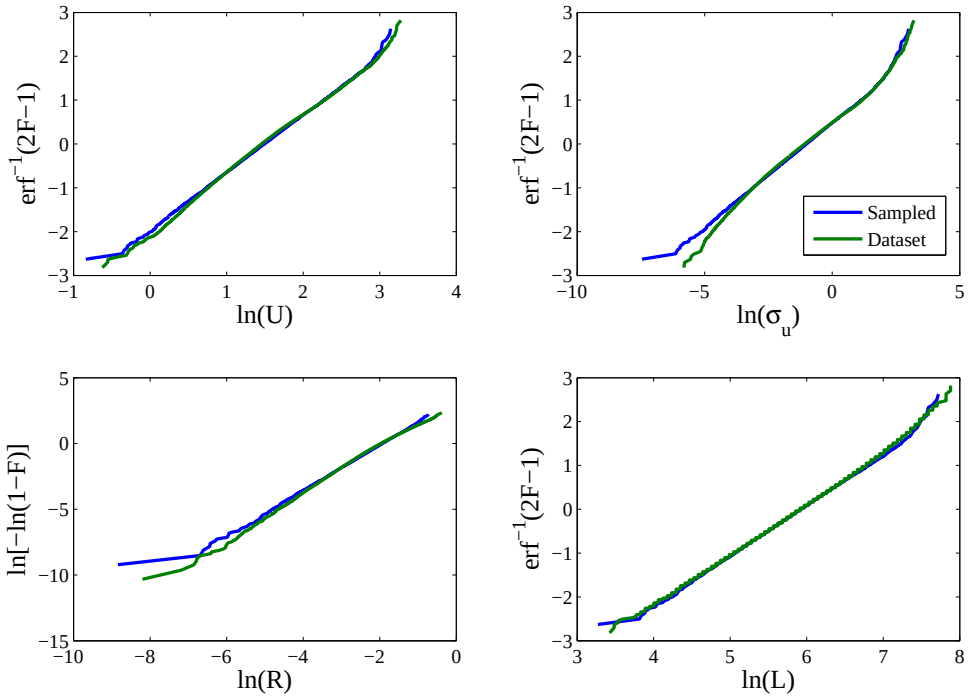

FIGURE 5.2: Monte Carlo analysis (10,000 samples) to demonstrate that the sampling method creates the correct marginal distributions for the four parameters. The blue lines indicate the marginals from the entire dataset; the green lines correspond to the sampled values.

Fig. 5.3. The figure features the data in the top subplot and a synthetic record, generated with the method described in Section 5.1, in the bottom plot. Both the data and the synthetic records have the same U (6.796 m/s), σ_u (1.056 m/s), $\bar{R}$ (0.657), and L (988 m) values. As can be clearly seen in the figure, the simulation method does not produce the same degree of nonstationarity as is present in the recorded data; there is more energy at all points in time instead of concentrated in the time period from 50 s to 250 s.

It was hypothesized that these differences in the time domain were caused by a mismatch between the fit von Mises distribution and the actual distribution of phase difference values that were found in the data. Histograms of the phase difference dif-

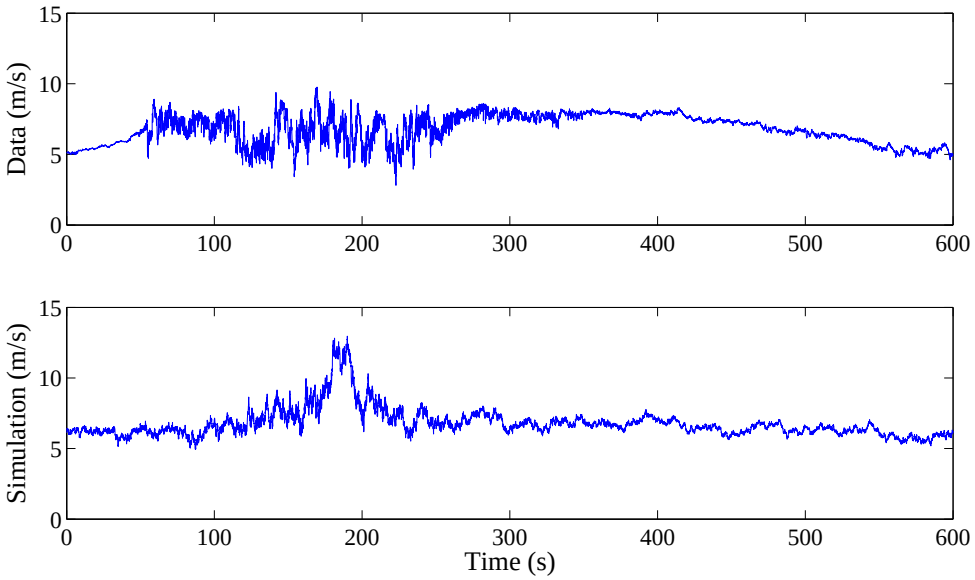

FIGURE 5.3: Time histories of data record recorded at 03:30 on July 11, 2013 at a height of 50 m (top) and a synthetic record (bottom) with the same values for U (6.796 m/s), σ_u (1.056 m/s), $\bar{R}$ (0.657), and L (988 m).

ference values for the data and synthetic record are shown in Fig. 5.4. The von Mises distribution that has the same MRL value produces a histogram that has a wider and shorter peak and fewer values that are sampled at the antimode. However, this difference in distributions does not fully explain the temporal variation, as drawing $\Delta\theta$ from the empirical histogram produced a time history that was very similar to the lower subplot in Fig. 5.3.

Further differences between the spectral characteristics of the data and simulation were observed by following the observation made by Thráinsson and Kiremidjian [25] that sorting by Fourier magnitude generated trends in the spread of $\Delta\theta$ in the spectral domain. Plotting the phase difference values as a function of frequency did not yield any significant trends (see Fig. 5.5); however, sorting the phase difference values by the inverse of the Fourier magnitude[2] produced a distinct trend in the

[2] Sorting by the inverse Fourier magnitude, instead of the Fourier magnitude, was selected because of the inverse relationship between wind velocity PSDs and frequency. Thus, sorting by inverse

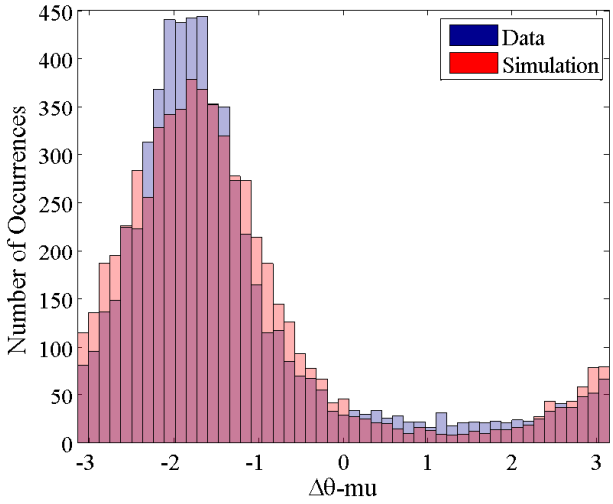

FIGURE 5.4: Histograms for the phase difference values for the data (light blue) and simulation (pink). The fit distribution has a wider and shorter peak and fewer values at the antimode.

data (see Fig. 5.6). In particular, the data's spectral coefficients with the largest Fourier magntiudes (towards the left-hand side of the plot) show a much tighter distribution of $\Delta\theta$, and the variance of $\Delta\theta$ generally increases as the magnitudes of the Fourier coefficients decrease. For the Fourier magnitudes with the very smallest magnitudes (towards the right-hand side of the plot), the phase difference values are almost uniformly distributed. This variation in the $\Delta\theta$ spread for different Fourier magnitudes is not captured in the proposed simulation methods; the spread of $\Delta\theta$ is the same for all simulated Fourier magnitudes (see bottom subplot).

These qualitative trends can be quantified by binning the phase difference values, sorted by Fourier magnitude, then calculating the MRL for each bin. Doing this for a bin size of 150 points and overlaying the MRL values over the sorted $\Delta\theta$ values produces the plot shown in Fig. 5.7. The plot clearly displays the high coherence values for large magnitudes and the steady decrease in coherence with decreasing Fourier magnitude.

Fourier magnitude is the same as sorting by frequency for synthetic records.

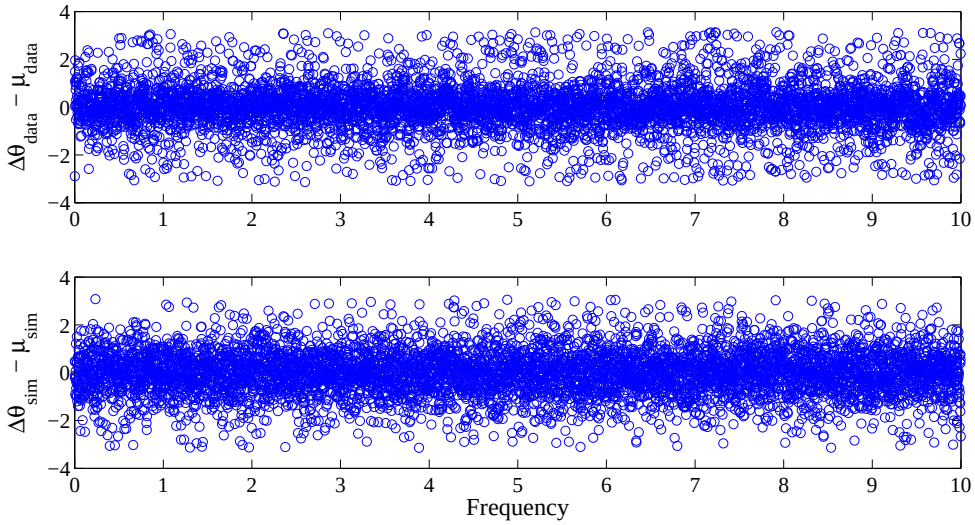

FIGURE 5.5: Scatterplot of $\Delta\theta$ for data (top) and simulation (bottom), sorted by frequency. The phase difference values have been shifted in this plot so that the mean direction is at 0.

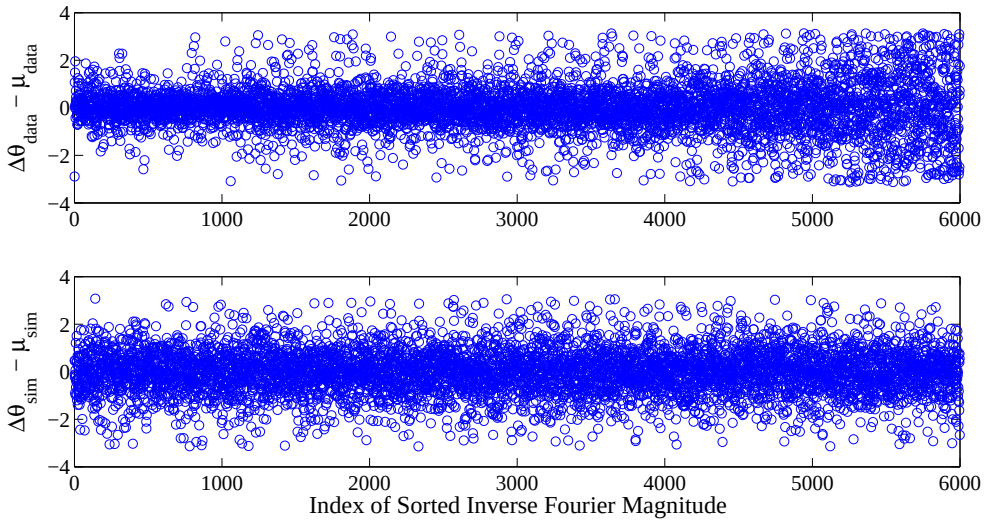

FIGURE 5.6: Scatterplot of $\Delta\theta$ for data (top) and simulation (bottom), sorted by inverse Fourier magnitude. The phase difference values have been shifted in this plot so that the mean direction is at 0.

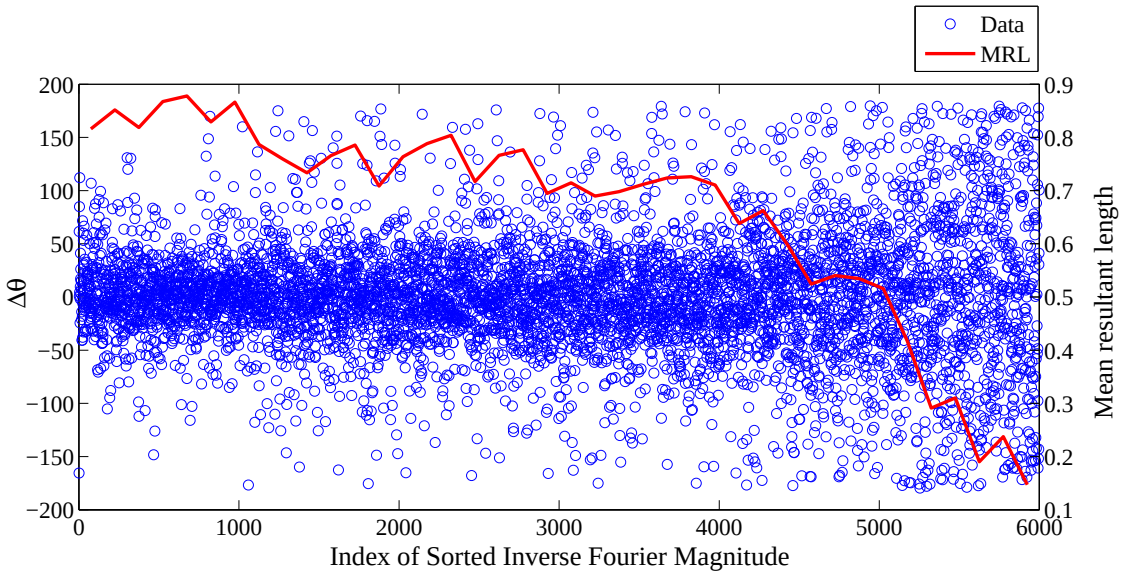

FIGURE 5.7: Scatterplot of $\Delta\theta$ for data with binned vales for $\bar{R}$ overlayed (150 points per bin), sorted by inverse Fourier magnitude.

This trend in the variation of the phase difference spread was observed in many records that had high degrees of phase coherence and displayed a high degree of non-stationarity in the time domain. Thus, developing a simulation method that would exactly mimic this nonstationary behavior would require the modeling of this variation of the phase difference values in the spectral domain. This problem, however, is extremely complicated due to the variation in the shape of the MRL trend from record to record and due to the variability of in Fourier magnitudes from record to record. Thus, the modeling of the spectral variation of the phase difference distributions is left for further study.

5.4 Response from an SDOF Oscillator

This study is motivated by the implicit assumption that the improper modeling of phase coherence in synthetic records will lead to an improper wind turbine design. The full verification of this assumption should and will be carried out using the

National Renewable Energy Laboratory (NREL) 5 MW reference FAST model in the future [51]. However, the effects of phase coherence on a dynamical system can be examined on a simple SDOF oscillator ($\zeta = 0.01$ and $\omega_n = 5$ rad/s) to illustrate two points: 1) how the system responds to input generated with the proposed method for different degrees of phase coherence, and 2) how the system responds to the data, to an input generated with the standard Sandia method and to an input generated with the proposed method.

5.4.1 Simulations with varying phase coherence

For this experiment, the input signal is assumed to have a Kaimal spectrum for the PSD and the phase difference values are drawn from von Mises distributions with $\bar{\theta} = \pi$ and varying values of κ. Figure 5.8 shows the time histories of the input signal $x(t)$ in the left column and the oscillator response $y(t)$ in the right column for $\kappa =$ 0, 1, and 4 (first, second, and third rows). As expected, increasing the value of κ produces a concentration of signal energy in the time domain. This, in turn, excites the oscillator to a higher degree and produces larger excursions.

This qualitative observation can be quantified by performing a simple Monte Carlo analysis in which the quantities of interest are the record's maximum value, minimum value, and damage equivalent load (DEL). The first two parameters can be viewed as the required yield strength of the material in tension and compression, respectively, assuming $y(t)$ is the stress in the system. The DEL for a stress series is the stress range that would generate the same level of fatigue damage as the original record, assuming a constant stress range and an average number of cycles per record. DELs can be calculated using the Palmgren-Miner linear damage accumulation rule, a log-log SN relationship, and an average number of cycles per record N_{avg} [18]:

$$\mathrm{DEL} = \left[\sum_{i=1}^{N} \frac{n_i {S_i}^m}{N_{avg}} \right]^{1/m}, \tag{5.7}$$

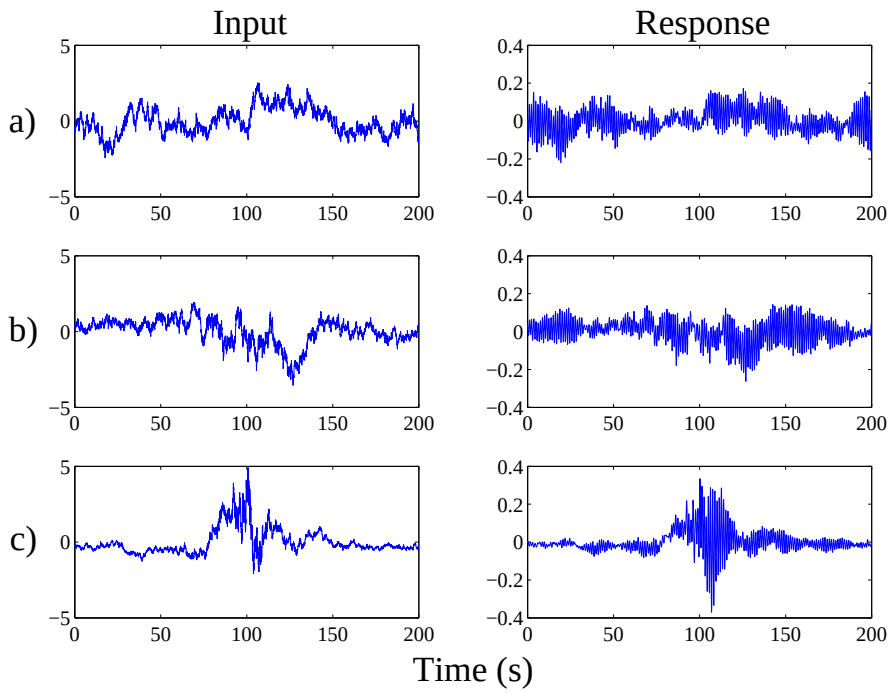

FIGURE 5.8: Plots of the input signal and response of SDOF oscillator for a) $\kappa = 0$, b) $\kappa = 1$, and c) $\kappa = 4$. The plots in the left column are the input to the oscillator and the plots in the right column are the time histories of the oscillator response.

where n_i is the number of cycles at stress range S_i and m is the SN curve slope. For this experiment, the value for m was taken to be 3, which is commonly assumed for components made of metal. The value for N_{avg} was 480 cycles, which was calculated by rainflow counting 5 records, calculating the number of cycles for each record, and rounding the average result to the nearest integer.

The results of the Monte Carlo analysis are shown in Figure 5.9, which were generated by simulating 5000 individual realizations for each of the three different values for κ. As can be clearly seen in the figure, an increase in phase coherence leads to "more dangerous" responses: the maximum response increases, the minimum response decreases, and the DEL increases. In addition, the presence of phase coherence leads to distributions of these parameters that have a larger variance, implying that more simulations would be necessary to ensure the design is robust.

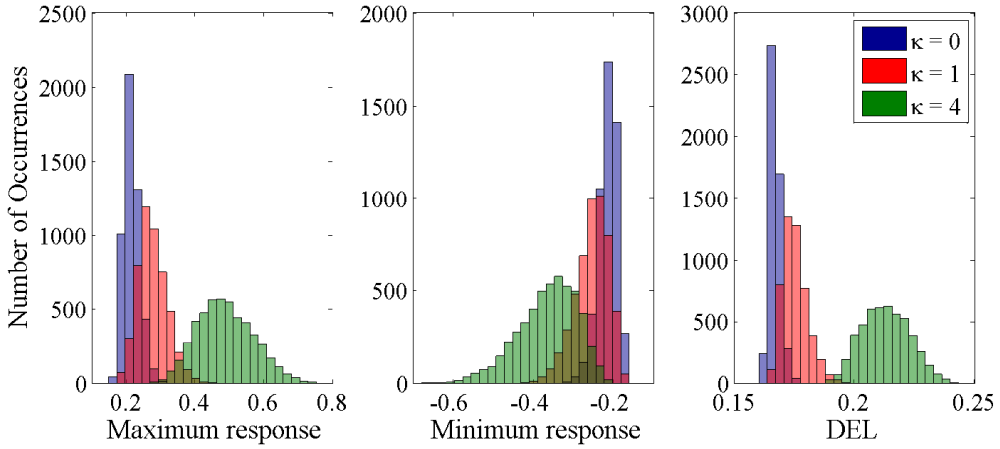

FIGURE 5.9: Histograms of the maximum response, minimum response, and DEL for varying levels of κ (5000 samples).

Table 5.1: Wind parameters for four data records used to evaluate simulation method.

Record	U (m/s)	σ_u (m/s)	$\bar{R}$	L (m)	$\bar{\theta}$ (rad)
1	3.494	0.246	0.672	523	2.704
2	3.257	0.897	0.673	415	2.270
3	6.796	1.056	0.657	988	-1.838
4	3.473	0.739	0.668	110	-0.760

Thus, it is likely that the characterization of phase coherence has a significant impact upon the design of wind turbines.

5.4.2 Response comparison for data and simulation

It is also important to compare the oscillator response when the input is data versus a simulated input. For this experiment, the SDOF responses were compared for three different inputs: the data; a synthetic record generated with a modified Sandia method with phase coherence; and a synthetic record generated with the standard Sandia method.

Four records were chosen that featured high coherence values, and the parameters for the records are shown in Table 5.1. The mean wind speed was used to calculate the Kaimal PSD, but all of the records were shifted to a zero mean before being input

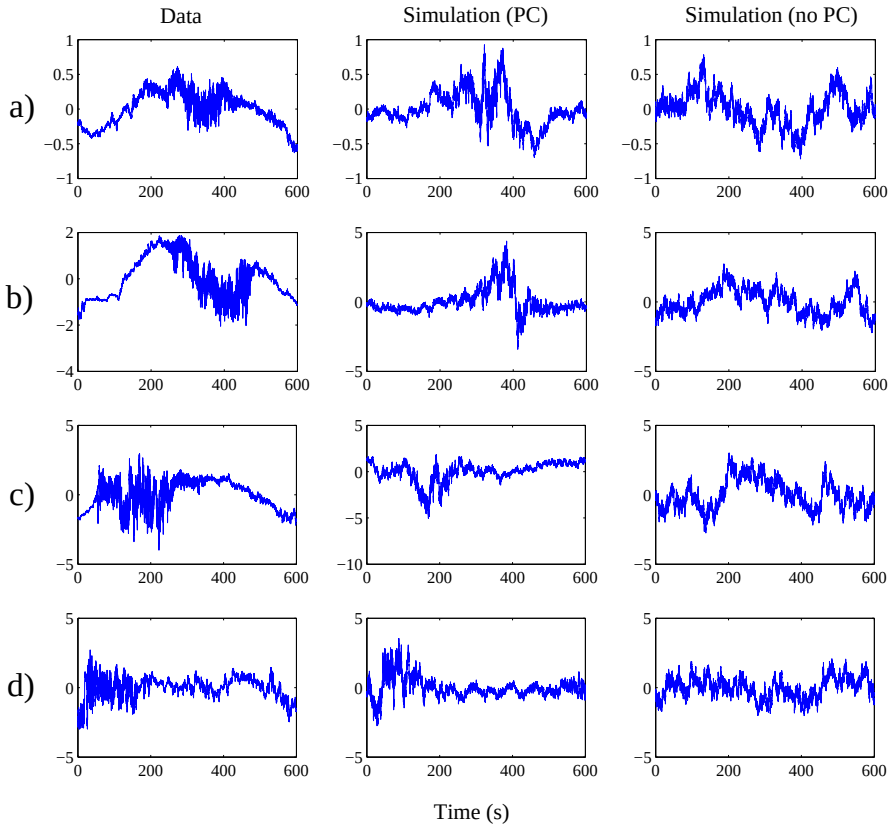

FIGURE 5.10: Comparison of the data, simulation with phase coherence, and simulation without phase coherence for four different records (a-d) used as input to the SDOF oscillator.

to the oscillator to facilitate comparison between the results. Example realizations for the synthetic inputs to the system are compared with the detrended real data in Fig. 5.10. As noted in the previous section, the simulation generally does not produce the same degree of nonstationarity or the same type of envelope as the data. Additionally, as expected, the simulation with no phase coherence does not have any concentration of energy in the time domain, so it is expected that the SDOF response will be generally smaller for the simulation with the standard Sandia method.

The output response of the SDOF oscillator is shown in Fig. 5.11. As predicted, the response to the simulation with no phase coherence is generally smaller in ampli-

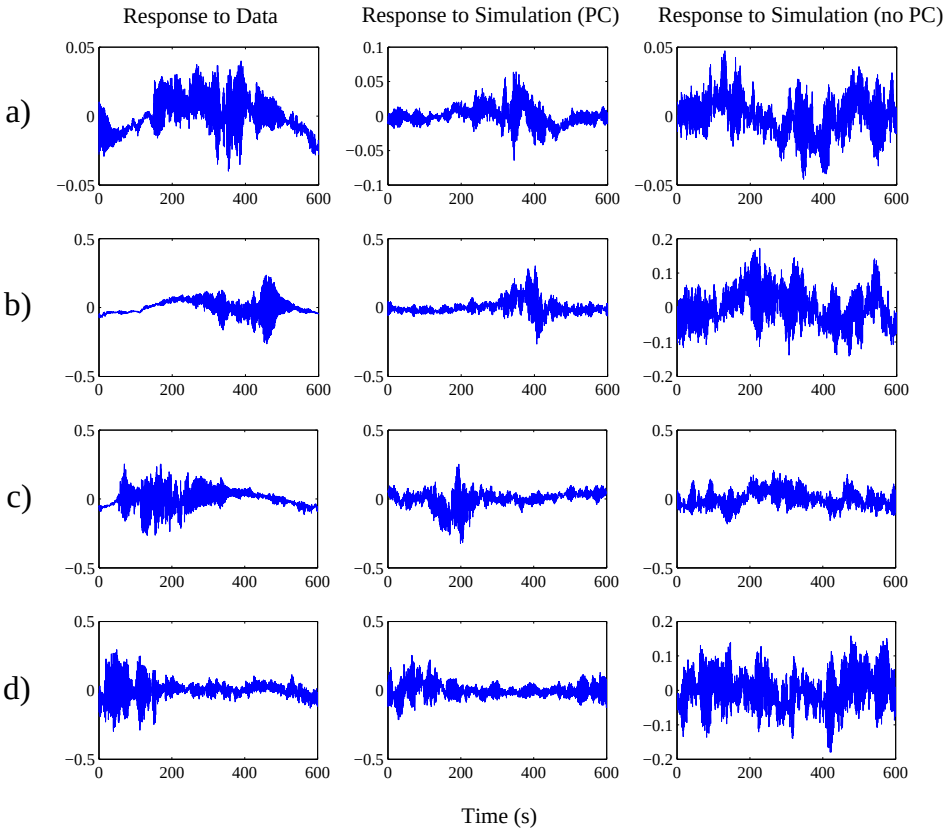

FIGURE 5.11: Comparison of the output from the SDOF oscillator for the data, simulation with phase coherence, and simulation without phase coherence for four different records (a-d).

tude than the simulation with phase coherence (note the difference in y-axis limits between the two columns). The response to the synthetic input with phase coherence is more similar to the response to data than the response without phase coherence, but it does not mimic the correct behavior perfectly. In particular, the extremal values are generally larger, but the computed responses tend to have more cycles with slightly smaller amplitudes. This will create differences in the load spectra generated by the two methods and will likely affect DEL calculations.

These qualitative observations can be quantified with a Monte Carlo simulation. Five thousand different realizations were generated to determine the mean and stan-

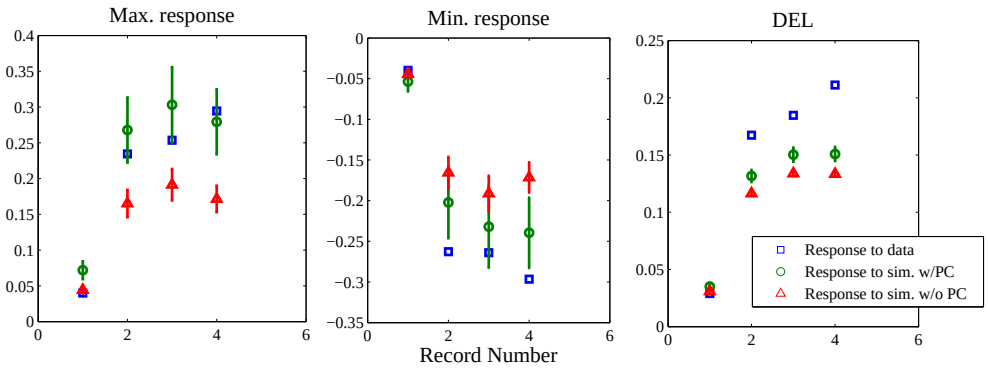

FIGURE 5.12: Comparison of the oscillator's maximum response, minimum response, and DEL for the the data (blue square), modified Sandia method (green circle), and standard Sandia method (red triangle) for the four different data records. The error bars indicate the standard deviation of the parameter.

dard deviation of the maximum response, minimum response, and DEL for responses to the data and to simulations with and without phase coherence. The results are shown in Fig. 5.12, where the record number is plotted along the x-axis and the value for the parameter is plotted along the y-axis. The blue circles indicate the SDOF response from the data, the green are for the response to the modified Sandia method, and the red are from the standard Sandia method; the error bars indicate the standard deviation of the results. As can be seen in the figure, the modified Sandia method provides more accurate results than the standard Sandia method for all three quantities, but there is still a significant difference between the response to data and the response to the modified Sandia method. These results indicate that the modeling of the variation of the $\Delta\theta$ spread in the spectral domain is essential to accurately determining the loads that real wind will generate in a wind turbine.

6

Conclusion

Spectral methods for wind simulation such as the Sandia method assume that the phase angles of a single-point wind field simulation are independent and uniformly distributed. This thesis introduces the concept of phase coherence for wind modeling as a simple method to expand the Sandia method to generate nonstationary simulations.

A joint PDF is fit to 6 different wind parameters (mean wind speed, turbulence, mean resultant length, mean direction, turbulent length scale, and slope of detrend line) that were calculated from a large dataset provided by the NWTC. A presented correlation study indicates that only U, σ_u, and L have any significant degree of correlation. The distributions that best fit the marginals of U, σ_u, $\bar{R}$, and L are lognormal/GP, lognormal/GP, Weibull, and lognormal/GP, respectively, and their best fit parameters are given. The sampling method is explained and demonstrated, and the generation of the correct marginal distributions is verified with a Monte Carlo simulation. Synthetic wind velocity records generated with the proposed method are compared with data records with the same parameters to verify that the simulation record is acceptably adequate. Simulations with lower MRL values are found to be

qualitatively similar to the data, but some simulations with higher degrees of phase coherence do not have the same degree of nonstationarity as the original data. These differences are explained through the description of the spectral variation of phase coherence.

The proposed simulation method is used as input to an SDOF oscillator to demonstrate the effects of phase coherence on a dynamical system. In one study, the oscillator responses to synthetic records with varying degrees of phase coherence are compared, and a presented Monte Carlo analysis of the maximum response, minimum response, and DEL reveals that all three quantities increase in magnitude and variability. In a second study, the oscillator response to data, to simulations with the modified Sandia method, and to simulations with the standard Sandia method are examined for four different data records. The proposed method generates better results than the standard Sandia method, though it does not match the results from the data perfectly.

Appendix A

Miscellaneous Proof

Theorem 1. *The inverse discrete Fourier transform of a vector of the form*

$$U = [P_1 e^{j\theta_1} \ \cdots \ P_N e^{j\theta_N}]^T, \tag{A.1}$$

where P_k are deterministic coefficients and θ_i are independent and uniformly distributed, is wide-sense stationary (WSS) and approximately Gaussian. Furthermore, if $P_k = 1$, the transformed vector is white noise.

Proof. The inverse discrete Fourier transform is given by

$$u(n) = \sum_{k=1}^{N} U(k) \exp\left[j\frac{2\pi n}{N}k\right] = \sum_{k=1}^{N} P_k \exp\left[j\left(\frac{2\pi n}{N}k + \theta_k\right)\right]. \tag{A.2}$$

Thus, $u(n)$ is a scaled sum of independent, identically distributed (i.i.d.) random variables, and by the Central Limit Theorem it approaches a Gaussian distribution as $N \to \infty$.

Note that

$$\mathbb{E}\{u(n)\} = \mathbb{E}\left\{\sum_{k=1}^{N} P_k \exp\left[j\left(\frac{2\pi n}{N}k + \theta_k\right)\right]\right\} \tag{A.3}$$

$$= \sum_{k=1}^{N} P_k \exp\left[j\frac{2\pi n}{N}k\right] \mathbb{E}\left\{\exp[j\theta_k]\right\}. \tag{A.4}$$

Because θ_k is uniformly distributed,

$$\mathbb{E}\left\{\exp[j\theta_k]\right\} = \int_0^{2\pi} \frac{1}{2\pi} \exp[j\theta_k] d\theta_k = 0. \tag{A.5}$$

Thus,

$$\mathbb{E}\{u(n)\} = 0. \tag{A.6}$$

Consider the autocovariance function:

$$\mathbb{E}[u(n)u(m)^*] = \mathbb{E}\left\{\sum_{k=1}^{N} P_k \exp\left[j\left(\frac{2\pi n}{N}k + \theta_k\right)\right] \sum_{l=1}^{N} P_l \exp\left[-j\left(\frac{2\pi m}{N}l + \theta_l\right)\right]\right\} \tag{A.7}$$

$$= \mathbb{E}\left\{\sum_{k=1}^{N}\sum_{l=1}^{N} P_k P_l \exp\left[j\frac{2\pi}{N}(nk - ml) + (\theta_k - \theta_l)\right]\right\} \tag{A.8}$$

$$= \sum_{k=1}^{N}\sum_{l=1}^{N} P_k P_l \exp\left[j\frac{2\pi}{N}(nk - ml)\right] \mathbb{E}\left\{\exp[j\theta_k]\exp[-j\theta_l]\right\} \tag{A.9}$$

Because the θ_k are independent and uniformly distributed, the expected value term is zero unless $k = l$. Therefore,

$$\mathbb{E}[u(n)u(m)^*] = \sum_{k=1}^{N} P_k^2 \exp\left[j\frac{2\pi k}{N}(n - m)\right]. \tag{A.10}$$

Note that this expression is only a function of the difference between n and m. Thus, because the mean is zero and the autocovariance only depends on the time lag between the signals, this signal is WSS.

If $P_k = 1$, then Eq. (A.10) reduces to

$$\mathbb{E}[u(n)u(m)^*] = \sum_{k=1}^{N} \exp\left[j\frac{2\pi k}{N}(n-m)\right] = \sum_{k=1}^{N}\left(\exp\left[j\frac{2\pi}{N}(n-m)\right]\right)^k. \quad \text{(A.11)}$$

If $n = m$, then this reduces to N. If $n \neq m$, applying the equation for a geometric series reveals

$$\mathbb{E}[u(n)u(m)^*] = \frac{\exp\left[j\frac{2\pi}{N}(n-m)N\right]-1}{\exp\left[j\frac{2\pi}{N}(n-m)\right]-1} = \frac{1-1}{\exp\left[j\frac{2\pi}{N}(n-m)\right]-1} = 0. \quad \text{(A.12)}$$

Therefore, if $P_k = 1$,

$$\mathbb{E}[u(n)u(m)^*] = N\delta_{nm}, \quad \text{(A.13)}$$

where δ_{nm} is the Kronecker delta, which proves that $u(n)$ is a white vector.

□

Bibliography

[1] G. Boer, N. McFarlane, and M. Lazare, "Greenhouse gas-induced climate change simulated with the CCC second-generation general circulation model," *Journal of Climate*, vol. 5, no. 10, pp. 1045–1077, 1992.

[2] N. Nakicenovic and R. Swart, "Special report on emissions scenarios," tech. rep., Intergovernmental Panel on Climate Change, Caimbridge, UK, July 2000.

[3] S. Lindenberg, "20% wind energy by 2030: Increasing wind energy's contribution to US electricity supply," tech. rep., US Department of Energy, 2008.

[4] C. Kost, J. N. Mayer, J. Thomsen, N. Hartmann, C. Senkpiel, S. Phillips, S. Nold, S. Lude, N. Saad, and T. Schlegl, "Levelized cost of electricity: Renewable energy technologies," tech. rep., Fraunhofer Institut for Solar Energy Systems, November 2013.

[5] E. Lantz, M. Hand, and R. Wiser, "The past and future cost of wind energy," Tech. Rep. NREL/P-6A20-54526, National Renewable Energy Laboratory, August 2012.

[6] I. E. Committee *et al.*, "IEC 61400-1: Wind turbines part 1: Design requirements." Online, 2005.

[7] K. Gurley and A. Kareem, "Applications of wavelet transforms in earthquake, wind and ocean engineering," *Engineering structures*, vol. 21, no. 2, pp. 149–167, 1999.

[8] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological society*, vol. 79, no. 1, pp. 61–78, 1998.

[9] T. Kitagawa and T. Nomura, "A wavelet-based method to generate artificial wind fluctuation data," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 91, no. 7, pp. 943–964, 2003.

[10] M. Farge, "Wavelet transforms and their applications to turbulence," *Annual Review of Fluid Mechanics*, vol. 24, no. 1, pp. 395–458, 1992.

[11] N. D. Kelley, R. M. Osgood, J. T. Bialasiewicz, and A. Jakubowski, "Using wavelet analysis to assess turbulence/rotor interactions," *Wind Energy*, vol. 3, no. 3, pp. 121–134, 2000.

[12] J. Franke, C. Hirsch, A. Jensen, H. Krüs, M. Schatzmann, P. Westbury, S. Miles, J. Wisse, and N. Wright, "Recommendations on the use of cfd in wind engineering," in *Cost Action C*, vol. 14, p. C1, 2004.

[13] S. Murakami, "Overview of turbulence models applied in cwe–1997," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 74, pp. 1–24, 1998.

[14] T.-H. Shih, W. W. Liou, A. Shabbir, Z. Yang, and J. Zhu, "A new k-ϵ eddy viscosity model for high reynolds number turbulent flows," *Computers & Fluids*, vol. 24, no. 3, pp. 227–238, 1995.

[15] L. Wang and A. Kareem, "Modeling and simulation of transient winds in downbursts/hurricanes," in *Proceedings of the 10th American Conference on Wind Engineering*, (Baton Rouge, LA), 2005.

[16] J. Mann, "Wind field simulation," *Probabilistic engineering mechanics*, vol. 13, no. 4, pp. 269–282, 1998.

[17] P. S. Veers, "Three-dimensional wind simulation," Tech. Rep. SAND88-0152, Sandia National Labs., Albuquerque, NM (USA), March 1988.

[18] H. F. Veldkamp, *Chances in wind energy: a probalistic approach to wind turbine fatigue design*. PhD thesis, Technische Universiteit Delft, 2006.

[19] T. Burton, N. Jenkins, D. Sharpe, and E. Bossanyi, *Wind Energy Handbook*. John Wiley & Sons, 2011.

[20] N. D. Kelley, "Full vector (3-D) inflow simulation in natural and wind farm environments using an expanded version of the SNLWIND (VEERS) turbulence code," Tech. Rep. NREL/TP-442-5225, National Renewable Energy Laboratory, Golden, CO, January 1993.

[21] B. J. Jonkman, "Turbsim user's guide: Version 1.50," 2009.

[22] J. Kaimal, J. Wyngaard, Y. Izumi, and O. Coté, "Spectral characteristics of surface-layer turbulence," *Quarterly Journal of the Royal Meteorological Society*, vol. 98, no. 417, pp. 563–589, 1972.

[23] Y. Ohsaki, "On the significance of phase content in earthquake ground motions," *Earthquake Engineering & Structural Dynamics*, vol. 7, no. 5, pp. 427–439, 1979.

[24] N. Nigam, "Phase properties of a class of random processes," *Earthquake Engineering & Structural Dynamics*, vol. 10, no. 5, pp. 711–717, 1982.

[25] H. Thráinsson and A. S. Kiremidjian, "Simulation of digital earthquake accelerograms using the inverse discrete fourier transform," *Earthquake engineering & structural dynamics*, vol. 31, no. 12, pp. 2023–2048, 2002.

[26] D. M. Boore, "Phase derivatives and simulation of strong ground motions," *Bulletin of the Seismological Society of America*, vol. 93, pp. 1132–1143, June 2003.

[27] L. Wang, *Stochastic Modeling and Simulation of Transient Events*. PhD thesis, University of Notre Dame, 2008.

[28] E. Krajnik, "A simple and reliable phase unwrapping algorithm," in *Signal Processing VI: Theories and Applications* (R. Vandewalle, R. Boite, M. Moonen, and A. Oosterlinck, eds.), (Amsterdam), pp. 917–919, Elsevier, 1992.

[29] K. V. Mardia and P. E. Jupp, *Directional statistics*. Chichester, New York: Wiley, 2 ed., 2000.

[30] A. Clifton, "135-m meteorological masts at the National Wind Technology Center: Instrumentation, data acquisition, and processing." Online, December 2013.

[31] A. Clifton, "135-m meteorological towers at the National Wind Technology Center: an unofficial guide to data products." Online, November 2013.

[32] J. Højstrup, "A statistical data screening procedure," *Measurement Science and Technology*, vol. 4, no. 2, p. 153, 1993.

[33] D. E. Knuth, *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms*, vol. 2. Addison-Wesley, 1997.

[34] D. Vickers and L. Mahrt, "Quality control and flux sampling problems for tower and aircraft data," *Journal of Atmospheric and Oceanic Technology*, vol. 14, no. 3, pp. 512–526, 1997.

[35] J. Pickands III, "Statistical inference using extreme order statistics," *the Annals of Statistics*, pp. 119–131, 1975.

[36] Q. Wang, "The POT model described by the generalized Pareto distribution with Poisson arrival rate," *Journal of Hydrology*, vol. 129, no. 1, pp. 263–280, 1991.

[37] R. L. Smith *et al.*, "Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone," *Statistical Science*, vol. 4, no. 4, pp. 367–377, 1989.

[38] J. Holmes and W. Moriarty, "Application of the generalized Pareto distribution to extreme value analysis in wind engineering," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 83, no. 1, pp. 1–10, 1999.

[39] C. De Michele and G. Salvadori, "A generalized Pareto intensity-duration model of storm rainfall exploiting 2-Copulas," *Journal of Geophysical Research: Atmospheres (1984–2012)*, vol. 108, no. D2, 2003.

[40] M. Gilli *et al.*, "An application of extreme value theory for measuring financial risk," *Computational Economics*, vol. 27, no. 2-3, pp. 207–228, 2006.

[41] J. R. Hosking and J. R. Wallis, "Parameter and quantile estimation for the generalized Pareto distribution," *Technometrics*, vol. 29, no. 3, pp. 339–349, 1987.

[42] S. D. Grimshaw, "Computing maximum likelihood estimates for the generalized Pareto distribution," *Technometrics*, vol. 35, no. 2, pp. 185–191, 1993.

[43] E. Castillo and A. S. Hadi, "Fitting the generalized Pareto distribution to data," *Journal of the American Statistical Association*, vol. 92, no. 440, pp. 1609–1620, 1997.

[44] C. Justus, W. Hargraves, A. Mikhail, and D. Graber, "Methods for estimating wind speed frequency distributions," *Journal of applied meteorology*, vol. 17, no. 3, pp. 350–353, 1978.

[45] J. Seguro and T. Lambert, "Modern estimation of the parameters of the Weibull wind speed distribution for wind energy analysis," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 85, no. 1, pp. 75–84, 2000.

[46] A. Garcia, J. Torres, E. Prieto, and A. De Francisco, "Fitting wind speed distributions: A case study," *Solar Energy*, vol. 62, no. 2, pp. 139–144, 1998.

[47] J. P. Hennessey Jr, "Some aspects of wind power statistics," *Journal of Applied Meteorology*, vol. 16, no. 2, pp. 119–128, 1977.

[48] S. Rehman, T. Halawani, and T. Husain, "Weibull parameters for wind speed distribution in Saudi Arabia," *Solar Energy*, vol. 53, no. 6, pp. 473–479, 1994.

[49] M. Stevens and P. Smulders, "The estimation of the parameters of the Weibull wind speed distribution for wind energy utilization purposes," *Wind engineering*, vol. 3, pp. 132–145, 1979.

[50] A. Lehman, *JMP for basic univariate and multivariate statistics: a step-by-step guide*. SAS Institute, 2005.

[51] J. M. Jonkman, S. Butterfield, W. Musial, and G. Scott, "Definition of a 5-MW reference wind turbine for offshore system development," Tech. Rep. NREL/TP-500-38060, National Renewable Energy Laboratory, Golden, CO, February 2009.