Statistical Modeling of Atmospheric Gravity Waves in the Lower Stratosphere

A senior design project submitted in partial fulfillment of the requirements for the degree of Bachelor of Science at Harvard University

Will Fried

S.B. Degree Candidate in Mechanical Engineering

Thesis Advisor: Dr. John Dykema
Faculty Advisor: Prof. Frank Keutsch

Harvard University School of Engineering and Applied Sciences
Cambridge, MA

April 5th, 2019

Abstract

To gain additional time in the fight against climate change, atmospheric scientists are exploring the utility of releasing aerosol particles into the lower stratosphere to reflect incoming solar radiation. Unfortunately, these particles may simultaneously lead to the depletion of ozone levels and pose other environmental threats. To better understand these risks, scientists have developed an atmospheric transport and chemistry model to capture how these particles move around once released. However, in its current form, the model does not account for atmospheric gravity waves, which strongly influence how these particles move and interact with their surroundings. This project addresses this deficiency by first inferring these gravity wave characteristics from 11 years of raw radiosonde data and then applying various statistical and machine learning techniques to construct a joint probability distribution on the seven parameters associated with atmospheric gravity waves.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Project Overview

Atmospheric gravity waves influence both the large and small-scale dynamics of the atmosphere through their vertical transport of horizontal momentum and energy, and through complex wave/turbulence interactions. They are formed when cooler, denser air is forced on top of warmer, thinner air and is restored to its lower altitude by gravity. This situation results in a wave that propagates vertically through the atmosphere while simultaneously oscillating horizontally. The main sources of gravity waves include orography, convection, jet streams and frontal systems.

While the causes and effects of atmospheric gravity waves are well understood, they are difficult to characterize since they occur in an atmosphere where there are many other forces and phenomena at play. Because gravity waves result in a sinusoidal perturbation to the zonal wind, meridional wind, and temperature as they propagate upwards, the main technique of identifying gravity wave activity involves extracting these wind and temperature perturbations. This is traditionally done by using radiosonde data to plot the zonal and meridional winds and temperature as a function of altitude, which yields a sinusoidal scatterplot for each of the three variables, as shown below in Figure 1:

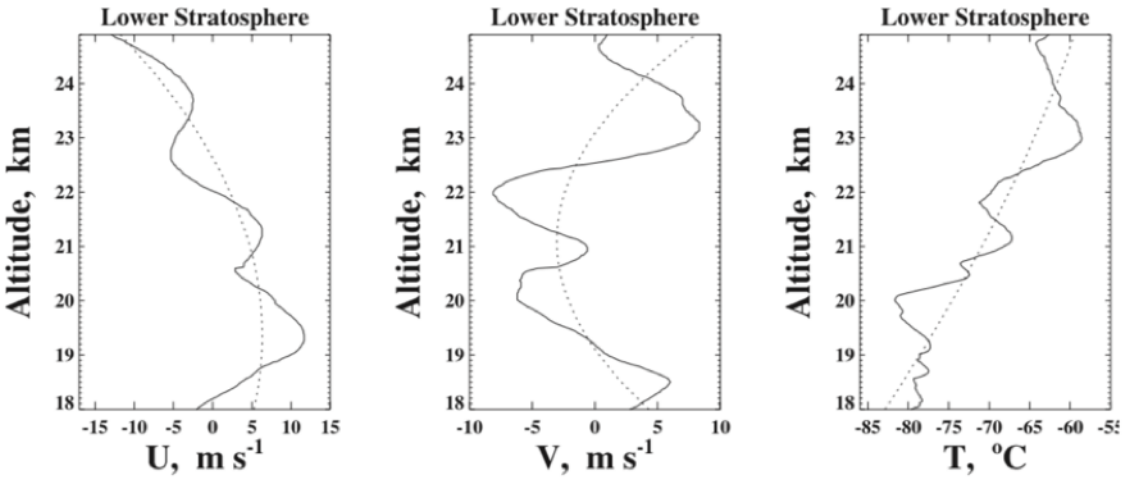


Figure 1: Characteristic sinusoidal profiles of zonal wind, meridional wind and temperature, respectively, in the lower stratosphere. The dashed curve in each plot represents a second-degree polynomial fit to the data. Figure reproduced from [1].

The mean profile of the wind/temperature is then estimated by computing a polynomial fit from the data, and the residuals obtained by subtracting the mean profile from the original data are assumed to be caused by gravity waves. These residuals are then plugged into complex formulas derived from fluid mechanic and thermodynamic theory, which approximate many of the gravity waves parameters including horizontal wavelength, vertical wavelength,

frequency and amplitude.

Many studies have applied the methodology described above to build a general picture of the strength and variation of gravity waves in different geographical regions and during different seasons of the year. For example, several papers describe how gravity wave activity is generally stronger around orographic features, gradually decrease poleward, is stronger in the winter than in the summer and is influenced by El Nino Southern Oscillation [1]. However, since the effect of atmospheric gravity waves is negligible for most applications, little work has been done to accurately characterize gravity wave activity in specific locations in the atmosphere. Therefore, the goal of this project is to devise an approach that uses raw radiosonde data and advanced statistical and computational techniques to comprehensively characterize and predict gravity wave activity in a specific location of the lower stratosphere.

1.2 Client

The direct client for this project is the Stratospheric Controlled Perturbation Experiment (SCoPEx) project team, which is investigating the environmental risk posed by sulfur-containing aerosol compounds that are released into the stratosphere as a form of solar radiation management. While the purpose of these aerosols is to reflect sunlight and mitigate global warming, they unfortunately have the detrimental effect of activating halogen compounds, which directly lead to the depletion of ozone in the stratosphere. Because controlled laboratory experiments do not adequately mimic real atmospheric conditions, the SCoPEx project team is planning to conduct an experiment in the stratosphere over eastern New Mexico to assess the impact of these aerosols on their local environments.

Before running the experiment, however, the SCoPEx project team needs to predict how the plume of aerosol particles will spread and mix with the surrounding air once it is released into the atmosphere. Even though an atmospheric transport and chemistry model has already been developed for this purpose, it neglects to take gravity waves into account. This is a major deficiency in the model, since the constant propagation of gravity waves affects the transport of aerosol particles and the breaking of gravity waves releases energy, which in turn produces areas of turbulent mixing. By learning the parameters of the gravity waves that exist near the experiment site, this project will enable the existing model to resolve gravity waves. This will improve the model's ability to predict the spatiotemporal profile of the aerosol particles, which, in turn, will help determine how these aerosol particles affect the concentration of ozone and halogen species in the stratosphere [2].

1.3 Prior Art

A significant amount of research has been conducted on the phenomenon of atmospheric gravity waves. One fundamental area of research focuses on the dynamics of atmospheric gravity waves and is rooted in the fluid mechanics and thermodynamic principles that govern the behavior of these waves [3]. A particularly active area of research centers on understanding the mechanics that are responsible for the emission of atmospheric gravity waves from common sources such as jet streams and weather fronts [4]. Another major area of research

has concentrated on the development of methodology to infer gravity wave parameters from raw radiosonde data. The two main approaches discussed in the literature are the hodograph method and the energy method, each of which will be described in detail [5, 6, 7]. A related area of research has focused on analyzing the consistency of the parameters inferred from these two methods and examining the degree to which these inferred parameters concur with those simulated from atmospheric models [8]. Finally, researchers have been applying this methodology to raw radiosonde data to build a general picture of how gravity wave activity varies both seasonally and spatially in different layers of the atmosphere [1].

As this academic literature indicates, gravity waves are understood well on a theoretical level as well as on a global scale. However, little work has been done to develop methodology that models gravity wave activity in specific regions of the atmosphere. This is the case since researchers have been focusing on building a general understanding of atmospheric gravity waves rather than detailing the influence of gravity waves in specific situations, such as an in situ experiment in the stratosphere.

1.4 Project Outline

This project can be divided into three main phases, which are outlined below:

1. Infer the gravity wave parameters from thousands of radiosonde soundings using fluid mechanics and thermodynamics principles. These parameters include frequency, vertical wavelength, zonal wavelength, meridional wavelength, zonal wind amplitude, meridional wind amplitude and temperature amplitude.

2. Explore the underlying structure of the inferred gravity wave parameters to gain insight into how to accurately model the joint probability distribution of the seven parameters.

3. Construct a joint probability distribution of the seven gravity wave parameters that best captures the structure of the data.

Each of these phases is discussed in detail in the following sections.

1.5 Modeling Approach

There are two main approaches that are used in statistics and machine learning to model a joint probability distribution. The first is Bayesian inference; the second is frequentist inference. Bayesian inference is appropriate when there exists a prior belief about the parameters in the model before observing the data at hand. This situation can arise if previously observed data has provided information about the parameters, or if there exists domain knowledge that suggests that certain parameters should theoretically be distributed in specific ways. Once the prior distributions on the parameters are established, the final model is then constructed using Bayes' rule, which states that the posterior distribution of each parameter is proportional to the product of the prior distribution and the likelihood function associated with the observed data.

In the particular setting of atmospheric gravity waves, Bayesian inference is not suitable for two reasons. First, there are no studies in the literature that have characterized the distributions of any of the seven atmospheric gravity wave parameters, which means there is no prior belief about any of the distributions from an empirical standpoint. And second, while the mechanisms that generate atmospheric gravity waves are well documented, the specific parameters associated with these gravity waves are not well understood by atmospheric scientists, which means there is no prior belief about any of the distributions from a theoretical standpoint.

Conversely, frequentist inference is used when there are no prior distributions on any of the parameters. This means that the model is created entirely based on the data at hand. Frequentist inference can be split into two subcategories: parametric statistical inference and nonparametric statistical inference. The objective in parametric statistical inference is to use the data to learn the parameters associated with a given model. These include the parameters of probability distributions, the coefficients in regression models and the Pearson correlation coefficient. On the contrary, in nonparametric statistical inference, no assumptions are made about the structure of the data. In an unsupervised learning context such as this project, this means that the data is not assumed to have been generated from particular families of distributions. Examples of nonparametric models include kernel density estimation, order statistics and the Spearman's rank correlation coefficient.

The primary advantage of a nonparametric model is that it provides a flexible framework to model any arbitrary joint distribution because it makes very few assumptions about the data compared to a parametric model. However, this increase in versatility is accompanied by two main drawbacks. The first is that a nonparametric model will have lower predictive power than a parametric model that captures the true data-generating process. For example, if a given random variable is truly normally distributed, a parametric model constructed using a normal distribution with the correct parameters will better capture the ground truth than a nonparametric model constructed using kernel density estimation. This disparity only disappears as the number of data points approaches infinity. The second major downside of nonparametric models is that the amount of data required to construct a model grows exponentially as the number of quantities to model increases. This can be attributed to the curse of dimensionality, which represents the fact that as the dimensionality increases, the volume of the space blows up so rapidly that the data becomes sparse. This, in turn, means that it becomes extremely difficult to confidently make inferences about the underlying data generating process.

Table 1 illustrates the issues that arise as the dimensionality increases. It lists the sample size that is required to guarantee a relative mean squared error less than 0.1 for the zero vector when the density is multivariate normal and the optimal kernel density estimation bandwidth is selected.

Dimension	Sample Size
1	4
2	19
3	67
4	223
5	768
6	2,790
7	10,700
8	43,700
9	187,000
10	842,000

Table 1: Illustration of how the number of data points needed to build an accurate nonparametric model skyrockets as the number of dimensions increases. Table reproduced from [9].

The most promising approach to construct a nonparametric joint probability distribution is to use multivariate kernel density estimation, which extends the idea of approximating the density of a distribution to multiple dimensions. Unfortunately, this proved to be infeasible for several reasons. First, as shown above in Table 1, the curse of dimensionality kicks in when dealing with just seven dimensions, which means that a massive amount of data is needed to create even a modestly accurate model. The other issue is that the packages in R that perform multivariate kernel density estimation cannot do so for more than six-dimensional data due to computational constraints. Because of these limitations, it became apparent that it would be extremely difficult to construct a nonparametric model that could accurately model both the marginal distributions and the correlation structures of the seven gravity wave parameters.

Due of the unsuitability of Bayesian inference and nonparametric statistics, it was clear that the best approach to modeling the joint probability distribution of the seven gravity wave parameters was to build a frequentist, parametric model.

1.6 Model Evaluation and Design Specifications

One of the key components of this project was to quantitatively evaluate the quality of the models that were ultimately developed. Unfortunately, as is typical in an unsupervised learning context, it was impossible to directly measure the accuracy of these models since there is no way of knowing what the ground truth is regarding gravity wave activity in the lower stratosphere over New Mexico – the only information available is the inferred gravity wave parameters. Therefore, alternative methods to assess the reliability of the models needed to be developed.

Two of the most commonly used model selection criteria in unsupervised learning problems are the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). The formulas for each criterion are presented below:

$$AIC = -2 \ln L + 2k \tag{1}$$

$$BIC = -2 \ln L + k \ln n \tag{2}$$

where L represents the likelihood function, k represents the number of scalar parameters in the model and n represents the total number of data points in the observed data. For a given set of candidate models, the model that corresponds to the lowest value of the AIC or BIC is selected.

The formulas for both the AIC and BIC have two terms that take into account different factors. With a negative coefficient, the first term rewards models that have high log-likelihoods, which means that they do a good job at maximizing the probability of observing the empirical data given the model. Meanwhile, the second term, which is positive, encourages simpler models by penalizing the complexity of the models. Without this term, the number of model parameters would tend toward infinity since the value of the log likelihood monotonically decreases as the complexity of the model increases. The lowest AIC/BIC value corresponds to the point where the boost to the log likelihood associated with an increase in the complexity of the model ceases to outweigh the penalty incurred from adding more parameters to the model.

Although it may seem that the exact form of the penalization terms in the formulas for the AIC and BIC are arbitrary, both of these terms in fact stem from statistical theory. As its name suggests, the BIC is derived from a Bayesian inference setting where there exist prior distributions on both a set of candidate models as well as the parameters associated with each of those candidate models. Thus, the BIC is an ideal choice in a situation where it makes sense for the model to be influenced not only by the observed data but also by prior domain knowledge.

Meanwhile, the AIC is derived in a frequentist setting where the model is built entirely based on the observed data. According to statistical theory, the model that minimizes the AIC corresponds to the model that minimizes the expected Kullback-Leibler divergence with respect to the true model that generated the observed data. (The Kullback-Leibler divergence is a measure of the difference between two probability distributions.) Given that Bayesian inference is unsuitable in this situation, as explained above, the AIC is the more appropriate choice to use for this modeling task.

The main shortcoming with this AIC approach is that while it is adept at rating the relative quality of different models, it is unable to evaluate the quality of a particular model in absolute terms. However, this problem is impossible to resolve in a project like this since there are no models or standards that currently exist for atmospheric gravity waves, which

means that there is nothing to compare a new model against.

In addition to formulating a criterion by which to evaluate the relative quality of the models, it was also important to decide when to stop modeling and declare a given model to be the final one. The approach of setting a target value for the AIC could not be used since the AIC value is meaningless by itself – all that matters is a relative increase or decrease in the AIC associated with different models. Rather than coming up with some arbitrary quantitative criteria, a more direct approach was taken. By definition, a joint distribution must model both the marginal distributions of the variables and the correlation between the variables. Therefore, a model would be considered final only once it captured both the marginal and correlation structure of all the gravity wave parameters, and no more obvious steps could be taken to improve the model. At that point, there would be nothing left to add to the model, so it would be considered complete.

At this point, the idea of what constitutes an obvious step that could be taken to improve the model may seem vague and subjective. However, in this project, there is no way of knowing what the data will show beforehand because the gravity wave parameters first need to be inferred from the raw radiosonde data. As a result, it would be premature to present a definition of what would be regarded as an obvious step until the gravity wave parameters are inferred and the modeling process is started. Only then would it become clear how the incomplete models that are built could be modified to better capture the structure of the raw data. One concrete example of what constitutes an obvious step that could be taken to improve the model is described in Section 4.4.2 on polynomial regression.

1.7 Project Flowchart

The general approach to the project is summarized in the flowchart presented below in Figure 2:

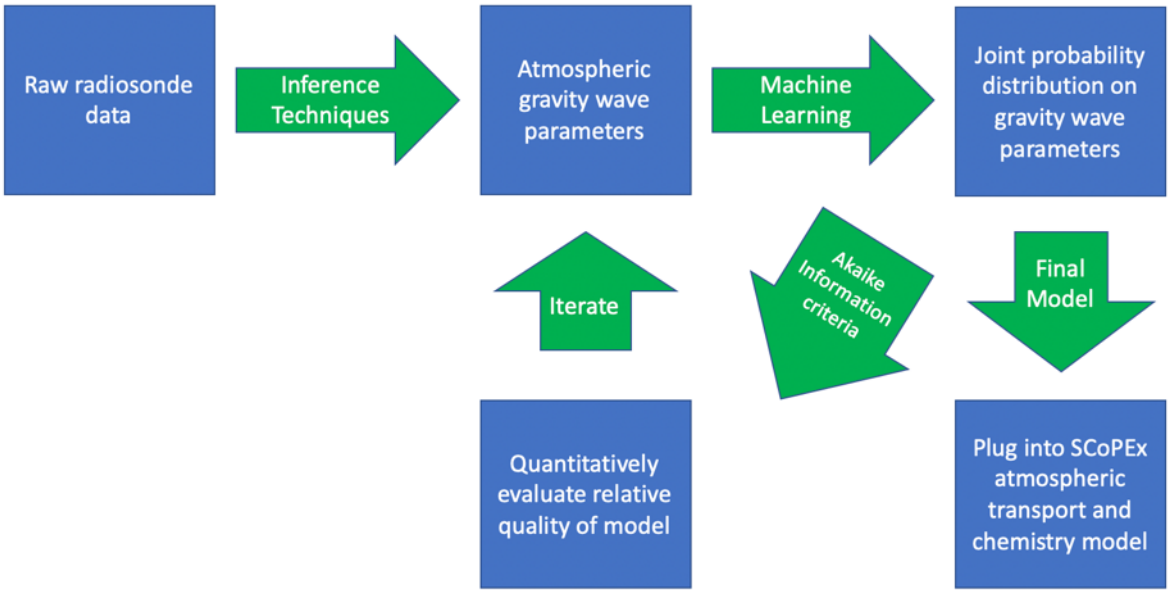


Figure 2: Flowchart laying out the general approach to this project. First, the seven gravity parameters are inferred from the raw radiosonde data. Once the structure of these inferred parameters is explored, the next step is to iteratively construct joint probability distributions on the seven inferred parameters using various machine learning techniques. These models are quantitatively compared to each other using the AIC. Ultimately, the model with the lowest AIC that captures both the marginal and correlation structure of all the gravity wave parameters is considered to be the final model and is incorporated into the SCoPEx atmospheric transport and chemistry model.

2 Inferring Gravity Wave Parameters from Radiosonde Data

2.1 Obtaining Data

Because the experiment on solar radiation management will be taking place over eastern New Mexico, the radiosonde data from the four closest radiosonde stations was used to study the atmospheric gravity wave activity in this vicinity. These stations are located in Santa Theresa, NM, Albuquerque, NM, Midland, TX and Amarillo, TX. The radiosonde data for each of these stations is publicly accessible online and is available from 1998 to today. The data from 1998 to approximately 2008 is called six-second data, which means that the radiosondes recorded the atmospheric conditions every six seconds, while the data from 2008 to today is called one-second data, since the improved radiosondes now record the atmospheric conditions every second.

The six-second data is stored as a compressed .dat.gz, which made it easy to download and

uncompress. However, the one-second data is stored in a BUFR file format, which is used exclusively to store meteorological data and is more difficult to convert to plain text. While an entire Python package called PyBufrKit was developed to decode BUFR files to various text formats, the output text from the BUFR files consisted of a long list of headings but no data. To make matters worse, the Fortran program that was provided to decode the BUFR files was inscrutable and didn't work as described in the instructions. Fortunately, the amount of six-second data available turned out to be sufficient to build the model, so there was no pressing need to obtain the one-second data.

2.2 Cleaning Data

Before inferring the gravity wave parameters, the raw data first had to be cleaned and structured. This involved iterating through all the .DAT files and removing any files that fell into one of two categories. The first category consisted of the soundings that correspond to weather balloons that failed to ascend to an altitude of at least 25 kilometers, which represents the maximum altitude the SCoPEx team is interested in. Although the balloons that ascended to an altitude of between 18 and 25 kilometers certainly recorded some valuable information, there were enough balloons that surpassed the 25-kilometer threshold that it wasn't necessary to analyze the data from those balloons that didn't capture the profile of the entire lower stratosphere. The second category consisted of those soundings that reported at least ten temperature and wind values of 999.9 at altitudes between 18 and 25 kilometers. (This erroneous value of 999.9 signifies that the equipment onboard was malfunctioning at the given moment.) For those soundings that reported fewer than ten values of 999.9, these faulty values were filtered out and the files were kept for analysis.

2.3 Locating Tropopause

The next step was to determine the altitude of the tropopause, which represents the boundary between the upper troposphere and the lower stratosphere. The tropopause can be deduced since the temperature continuously decreases in the troposphere and slightly rebounds in the stratosphere. Therefore, the tropopause can be identified by locating the altitude at which the temperature reaches its minimum value. This altitude lies anywhere between 16 and 18 kilometers for the vast majority of the soundings. Because the reference literature uses the same cutoff value for all soundings rather than individually determining the tropopause altitude for each sounding, and given that the SCoPEx experiment cares only about characterizing gravity wave activity above 18 kilometers, the simplifying assumption was made that the tropopause is located at an altitude of 18 kilometers for all of the soundings. This in turn meant that the altitude interval to analyze for atmospheric gravity waves ranges from 18 to 25 kilometers. This seven-kilometer window provides a complete picture of gravity wave activity in the portion of the atmosphere that the ScoPEx project team is interested in. A typical temperature profile as a function of altitude is illustrated below in Figure 3:

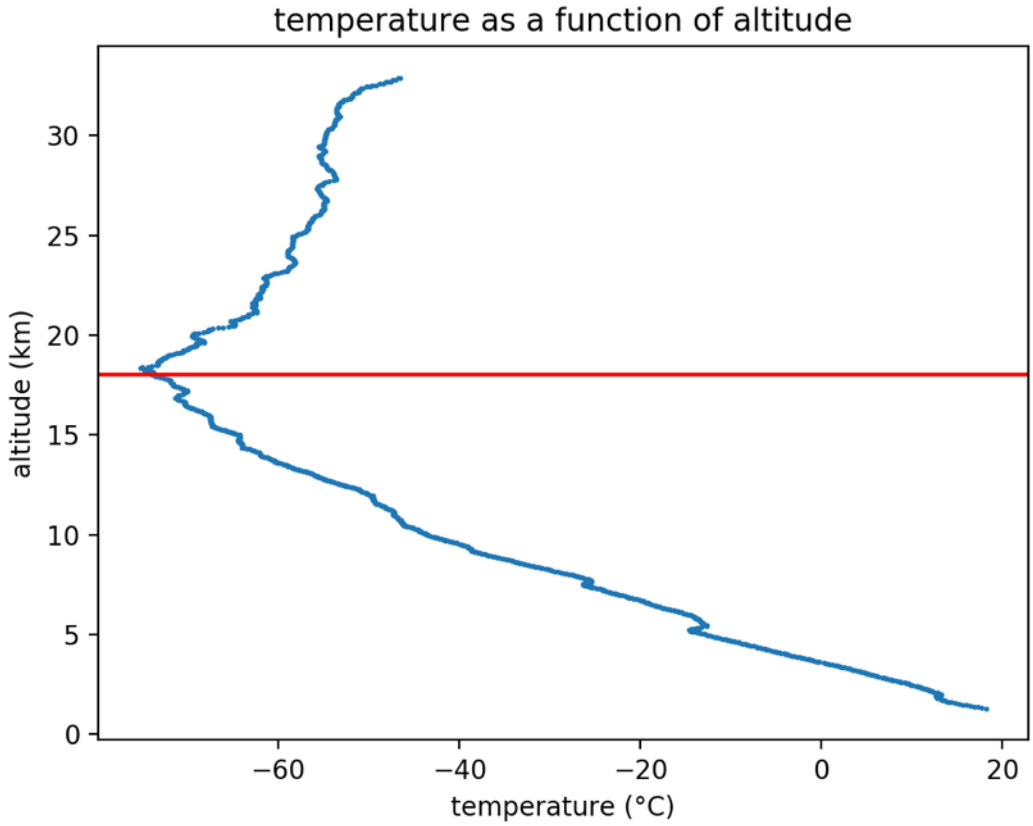


Figure 3: Typical profile of temperature as a function of altitude. The blue curve represents the temperature, while the red line at 18 kilometers approximately represents the minimum temperature and thus denotes the location of the tropopause.

The next step was to infer the gravity wave parameters using two different techniques: the hodograph method and the energy method. The hodograph method is based on the methodology presented in [6], while the energy method is based on the methodology presented in [5]. Each of these techniques is detailed below.

2.4 Hodograph Method

The hodograph method is a multistep procedure that uses fluid mechanics and thermodynamic principles to infer the gravity wave parameters from the raw radiosonde data. The steps involved in the hodograph method are outlined below and are automatically performed by a Python script for each radiosonde sounding:

1. Extract the raw data from a given .DAT file and organize it in a data table where each column corresponds to one of the variables (temperature, wind, pressure, etc.) that are measured every six seconds.

2. Limit analysis to the portion of data corresponding to an altitude between 18 and 25 kilometers.

3. Fit a least-squares second-degree polynomial to the zonal wind, meridional wind and temperature profiles between 18 and 25 kilometers. (The rationale for selecting a second-degree polynomial is discussed in the next section.)

4. Subtract the best fit second-degree polynomial from the raw data to obtain a perturbation for the temperature and zonal and meridional winds. These perturbations are assumed to have been caused by atmospheric gravity waves propagating through the upper atmosphere.

5. Linearly interpolate all of the perturbations at 50 meter increments between 18 and 25 kilometers. For example, if a radiosonde were to record a temperature perturbation of -4°C and -2°C at an altitude of 18.975 km and 19.025km, respectively, then the interpolated temperature perturbation at an altitude of 19km would be -3°C. This step was necessary because the radiosondes record measurements in fixed intervals of time rather than in fixed intervals of altitude, which means that the altitudes at which the recordings are made vary between the soundings.

6. Fit separate sine waves to the perturbations of zonal wind, meridional wind and temperature using a least-squares optimization algorithm. This yields best-guess estimates for the amplitude, wavelength and phase of these perturbation profiles.

7. Calculate the relative standard error of the three best-fit wavelengths that were calculated in step 6, and disregard any radiosonde soundings where the relative standard error is larger than 20%. This 20% threshold is presented in the literature as the optimal cutoff based on previous studies [6]. The reason for this step is that the hodograph method assumes the presence of a coherent, quasi-monochromatic atmospheric gravity wave, as it isn't able to infer multiple gravity waves that are simultaneously propagating through the atmosphere. A strong agreement between the best-fit wavelengths of the three perturbation profiles suggests the presence of a single gravity wave, while high variability between these wavelengths is indicative of either no major atmospheric gravity wave activity or the presence of superposed gravity waves, which cannot be inferred by the hodograph method. Altogether, only around 50% of the radiosonde soundings made it past this step.

8. Refit separate sine waves to the perturbations of zonal wind, meridional wind and temperature using the same least-squares optimization algorithm, but this time set the wavelength to be the average of the wavelengths calculated in step 6. (This average wavelength is assumed to be the vertical wavelength of the atmospheric gravity wave.) This means that the only two parameters to infer are the amplitude and phase of the waves.

9. Declare the best fit amplitudes from step 8 to be the amplitudes of the zonal wind, meridional wind and temperature components of the atmospheric gravity wave propagating through the lower stratosphere.

10. Plot the best fit sine waves corresponding to the zonal wind and meridional wind perturbations on a two-dimensional graph known as a hodograph. This yields an elliptical shape that is centered at the origin as long as the phase difference between these curves is not exactly 0°or 180°.

11. Fit an ellipse to this elliptical shape using a direct least squares method [10] and report the lengths of the ellipse's major and minor axis, as well as the angle that the major axis makes to the horizontal. (The physical significance of the axis lengths is discussed in step 14, while the physical significance of the angle is discussed in step 16.) A typical hodograph and fitted ellipse is illustrated below in Figure 4:

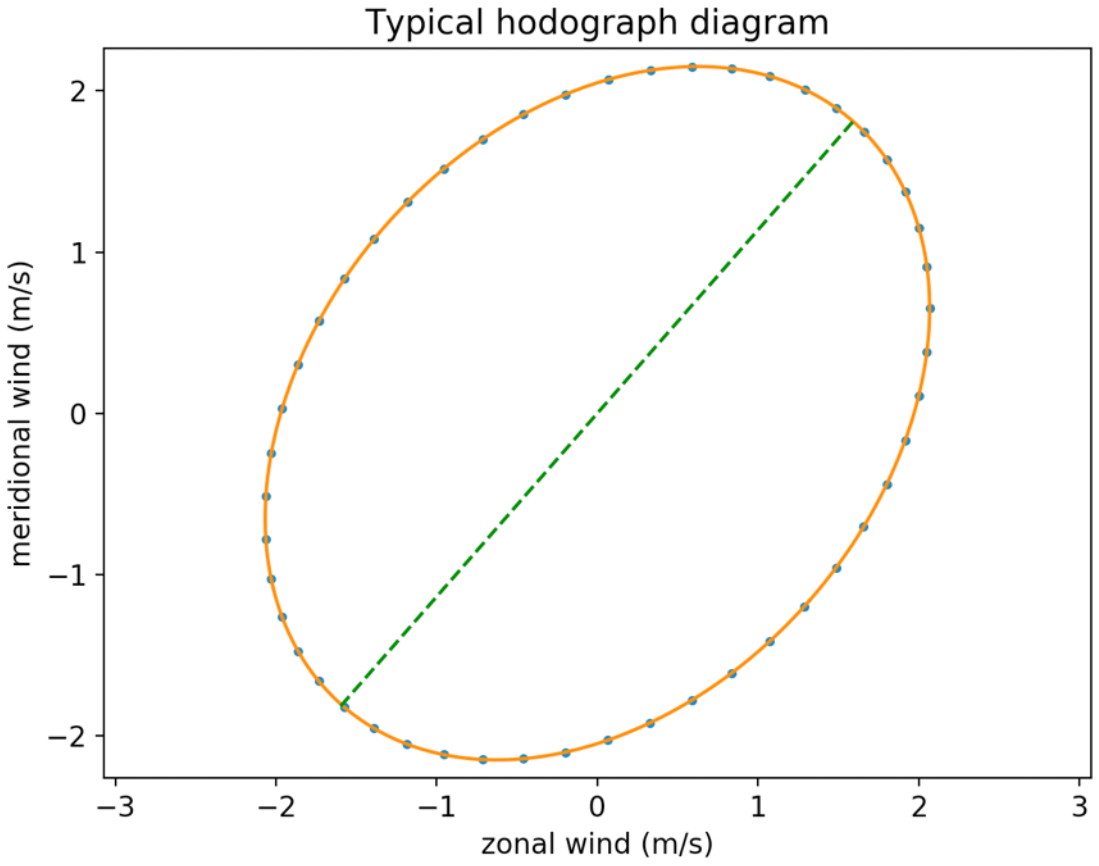


Figure 4: A typical hodograph diagram. The blue dots represent the plot of the zonal wind and meridional wind at equally spaced angles between 0° and 360°. The orange curve depicts the ellipse that was fitted to the blue dots. The green dashed line signifies the direction of the major axis of the fitted ellipse, which, in this case, is 48.7° from the horizontal. This diagram is reproduced from the data recorded by the radiosonde launched from Albuquerque, NM at midnight on January 2nd, 1998.

12. Compute the ratio of the major axis to the minor axis. Due to unavoidable errors in the radiosonde's DigiCORA® wind measuring system, a ratio that is too large yields

inferred gravity wave parameters that are unreliable, so any radiosonde soundings that yielded a ratio of over ten were disregarded [11]. This cutoff is presented in the literature as a sensible threshold based on the level of precision of the radiosonde wind speed measuring device [11].

13. Calculate the local Coriolis frequency, denoted by f. The formula for the local Coriolis frequency is:

$$f = 2\Omega \sin \phi \tag{3}$$

where Ω represents the rotational speed of the Earth (7.2921×10^{-5} rad/s) and ϕ represents the latitude (in degrees), which can easily be looked up for each of the four radiosonde stations being considered.

14. Compute the wave intrinsic frequency by multiplying the local Coriolis frequency by the axis ratio. This calculation is derived from the equation that states that the ratio of the gravity wave intrinsic frequency to the local Coriolis frequency is equal to that of the major to minor axes of the hodograph.

15. Calculate the Brunt–Väisälä frequency, denoted by N. The formula for the Brunt–Väisälä frequency is:

$$N = \sqrt{\frac{g}{\theta} \frac{\mathrm{d}\theta}{\mathrm{d}z}} \tag{4}$$

where θ is the potential temperature, g is the gravitational acceleration constant and z is the altitude. Potential temperature represents the temperature that a parcel of air at a given pressure would attain if it were to be adiabatically brought to a standard pressure of 100 kPa. The formula for the potential temperature is:

$$\theta = T \left(\frac{p_0}{p} \right)^{\frac{R}{c_p}} \tag{5}$$

where p_0 is standard reference pressure (100 kPa), p is the pressure, T is the absolute temperature (in K), R is the gas constant of air and c_p is the specific heat capacity of air at constant pressure. The Brunt–Väisälä frequency is calculated at 50 meter increments between 18 and 25 kilometers; the derivative of potential temperature with respect to altitude in (5) is approximated as the change in potential temperature divided by the change in altitude. Because the value of the Brunt–Väisälä frequency is roughly constant throughout the lower stratosphere, the value of the Brunt–Väisälä frequency is taken to be the simple mean of the frequencies calculated at 50 meter increments.

16. Compute the zonal and meridional components of the horizontal wavelength using the simplified dispersion equation for atmospheric gravity waves:

$$k^2_{horizontal} = k^2_{zonal} + k^2_{meridional} = \frac{k^2_{vertical} \left(\Omega^2 - f^2 \right)}{N^2} \tag{6}$$

where $k_{horizontal}$, k_{zonal} , $k_{meridional}$ and $k_{vertical}$ are the horizontal, zonal, meridional and vertical wave numbers, respectively, Ω is the intrinsic frequency of the gravity wave, f is the local Coriolis frequency and N is the Brunt–Väisälä frequency. k_{zonal} and $k_{meridional}$ can then be calculated by multiplying $k_{horizontal}$ by the cosine and sine, respectively, of the angle that the major axis of the fitted ellipse makes to the horizontal.

17. Count the number of missing radiosonde soundings between subsequent soundings. This value will be important in the next phase where the structure of the inferred atmospheric gravity wave parameters is examined.

18. Store all of the radiosonde soundings in a massive dictionary, where the key is the filename (i.e."03020-1998010800.dat" represents the sounding launched from Santa Theresa, NM, which is the station designated by the code 03020, that took place on January 8th, 1998 at midnight). The value is another dictionary containing the inferred frequency, vertical wavelength, zonal wavelength, meridional wavelength, zonal wind amplitude, meridional wind amplitude, temperature amplitude and the number of missing soundings between the given sounding and the previous sounding in which the atmospheric gravity wave parameters were successfully inferred.

2.5 Selecting Degree of Polynomial Fit

The decision to use a second-degree polynomial to remove the background trends is largely arbitrary (even though it's the most common one used in the literature), since there's no scientifically-grounded way to determine which degree polynomial is optimal. One way to gain insight into this seemingly arbitrary decision is to analyze the differences in the gravity wave parameters that are inferred from the hodograph method when different degree polynomials are used to remove the background trends. Any possible discrepancies could then be used to justify the preference for a particular degree polynomial over the other options. For example, if the parameters inferred using a third-degree polynomial turn out to be more consistent across years than those inferred using a fourth-degree polynomial, then this finding would give credence to selecting a third-degree polynomial over a fourth-degree polynomial.

The gravity waves were inferred using second through seventh-degree polynomials, which covers the full range of polynomials considered in the literature. Examining the corresponding kernel density estimates and scatterplots indicated that the gravity wave parameters inferred using different degree polynomials were equally consistent across stations, years, months and time and day. The only discernable trend was that as the degree of the polynomial increased, all the gravity wave parameter distributions, expect for that of frequency, shifted closer to zero. This pattern makes sense because the higher the degree of the polynomial, the more the polynomial fit aligns with the raw data, which, in turn, decreases the perturbation from the background trend. Because these perturbations determine many of the gravity wave parameters, smaller perturbations generally translate into smaller parameter values.

Since the graphs of the inferred parameters were unable to quantitatively distinguish the relative quality of the different degree polynomials, the next best option was to visually

inspect the fits themselves to determine which one(s) seem best. In this case, there are two competing factors at play. On the one hand, the higher the degree of the polynomial, the more flexibility there is to capture the background trend. On the other hand, the lower the degree of the polynomial the lower the risk of overfitting the raw data and interpreting some of the signal as background.

Figure 5 shows how the meridional wind velocity typically varies in the lower stratosphere. In the figure, the meridional wind velocity profile is overlaid with the least-squares polynomial fits corresponding to degrees between two and seven. Overall, the graph indicates that as the degree of the polynomial rises, the least-square curve increasingly aligns with the sinusoidal signal, and thus erroneously considers a portion of the signal to be part of the background trend. Importantly, the second-degree polynomial doesn't suffer from this overfitting, as the curve roughly passes through the equilibrium position of the wave, as desired. This tendency of a second-degree polynomial to align with the equilibrium position of a given sinusoidal wave (the zonal wind velocity, meridional wind velocity and temperature all vary sinusoidally in the lower stratosphere) applies to the vast majority of soundings. Therefore, given the choice between any degree polynomial between two and seven, it seems reasonable to conclude that a second-degree polynomial is the best option overall.

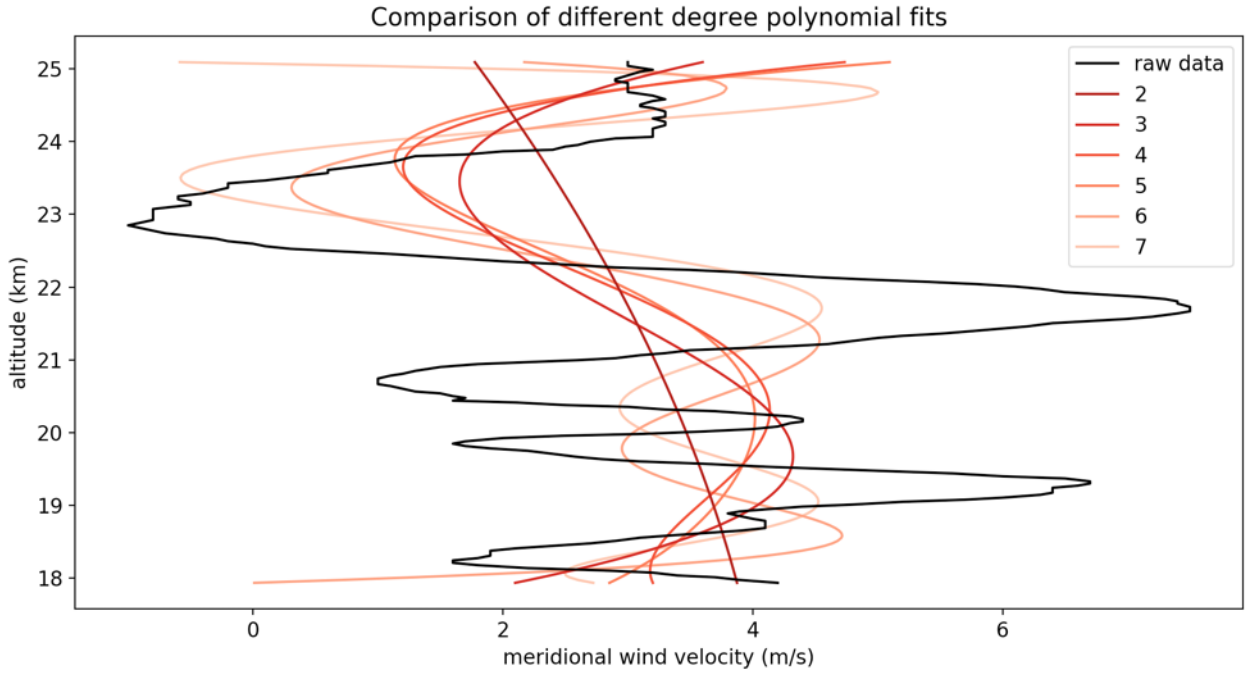


Figure 5: Plot of the meridional wind velocity (black) as a function of altitude in the lower stratosphere. This sinusoidal curve is overlaid with the least-squares polynomial fits of degree two to seven. The higher the degree of the polynomial, the lighter the curve appears and the closer the curve aligns with the meridional wind velocity profile. The notable exception to this trend is the second-degree polynomial (darkest red), which roughly passes through the equilibrium position of the sinusoidal curve.

2.6 Energy Method

The energy method represents a second way to infer the gravity wave parameters from the raw radiosonde data. The energy method uses fluid mechanics and thermodynamic principles to first infer the energy associated with atmospheric gravity waves and then deduce the gravity wave parameters from these energy values. The steps involved in the energy method are outlined below:

1. Perform steps 1-5 of the hodograph method.

2. Calculate the Brunt–Väisälä frequency using the same procedure outlined in step 15 of the hodograph method.

3. Approximate the ascent rate of the radiosonde as follows:

$$W(t) = \frac{dZ(t)}{dt} \approx \frac{Z(t + \Delta t) - Z(t - \Delta t)}{2\Delta t} \tag{7}$$

   where Z is the altitude of the radiosonde calculated hydrostatically and Δt is the sampling interval of the radiosonde (six seconds in this case). The difference in hydrostatic altitude at two points in time (denoted by $k-1$ and k) is calculated as follows:

$$\Delta Z = \frac{R}{g} \frac{T_{v,k-1} + T_{v,k}}{2} \ln \frac{p_{k-1}}{p_k} \tag{8}$$

   where R is the universal gas constant for air, g is the gravitational acceleration constant, T_v is the virtual air temperature and p is the pressure. The virtual air temperature is, in turn, calculated as follows:

   (a) Calculate the saturation pressure of water vapor according to the following empirically-derived equation:

$$\ln p_{sat} = 54.8 - \frac{6763}{T} - 4.2 \ln T + 0.000367T + \tanh 0.0415(T - 218.8) * (54 - \frac{1331}{T}) - 9.4 * \ln T + 0.014T \tag{9}$$

   where T is the temperature measured in Kelvins.

   (b) Multiply p_{sat} by the relative humidity, which is recorded by the radiosonde, to obtain the partial pressure of water vapor (PP_{water}).

   (c) Divide PP_{water} by the atmospheric pressure to obtain the mixing ratio (r).

   (d) Calculate the virtual temperature as follows:

$$T_v = T(1 + 0.61r) \tag{10}$$

4. Linearly interpolate the ascent rate of the radiosonde at 50 meter increments between 18 and 25 kilometers.

5. Remove a centered running average of height five kilometers from the vertical ascent rate profile to obtain a vertical ascent wind perturbation. This perturbation is considered to be the vertical velocity fluctuation. This five kilometer window is presented in the literature as the optimal width [12].

6. Compute the energy density associated with the atmospheric gravity wave at each of the 50-meter increments from 18 to 25 kilometers. The total energy is comprised of a kinetic energy and a potential energy component, which are calculated as follows:

$$E_{total} = E_{kinetic} + E_{potential} = \frac{1}{2}\left(u'^2 + v'^2 + w'^2\right) + \frac{1}{2}\frac{g^2 T'^2}{N^2 \bar{T}^2} \tag{11}$$

where u' , v' , w' and T' are the zonal wind, meridional wind, vertical wind and temperature perturbations, respectively, g is the gravitational acceleration constant, N is the Brunt–Väisälä frequency and $\bar{T}$ is the background temperature.

7. The quantities just calculated can then be used to determine many of the gravity wave parameters. For example, the frequency (Ω) can be inferred by evaluating the following expression:

$$\Omega^2 = f^2 \frac{E_{kinetic} + E_{potential}}{E_{kinetic} - E_{potential}} \tag{12}$$

where f is the local Coriolis frequency. Meanwhile, the vertical wavenumber ($k_{vertical}$), which is needed to infer the zonal and meridional wavelengths, can be derived from:

$$k_{vertical} = -\frac{\left(N^2 - \Omega^2\right) w' \bar{\rho}}{\Omega p'} \tag{13}$$

where N is the Brunt–Väisälä frequency, Ω is the intrinsic frequency, w' is the vertical velocity fluctuation, $\bar{\rho}$ is the background density and p' is the pressure perturbation.

At this point, there's no reason to continue with the energy method because there are three main issues that limit its usefulness and credibility. First, the energy method cannot infer the zonal wind amplitude, meridional wind amplitude or temperature amplitude. Second, the inferred energy density varies wildly at different altitudes, as illustrated below in Figure 6. Without a reliable value for the total energy density, any gravity wave parameters inferred from the energy method would be statistically meaningless. For example, given the enormous variation, it would be foolish to compute the mean of the specific energy and assume that value to be the true energy density associated with a given atmospheric gravity wave. Finally, the vertical wind perturbation is extremely inconsistent across different altitudes, as illustrated below in Figure 7. Just as with regard to energy density, Figure 7 indicates that the inferred vertical wind perturbations have little internal consistency and are thus unreliable. This, in turn, means it would impossible to use the energy method to accurately compute $k_{vertical}$ in Equation 9, since it is directly proportional to the vertical wind perturbation, w'.

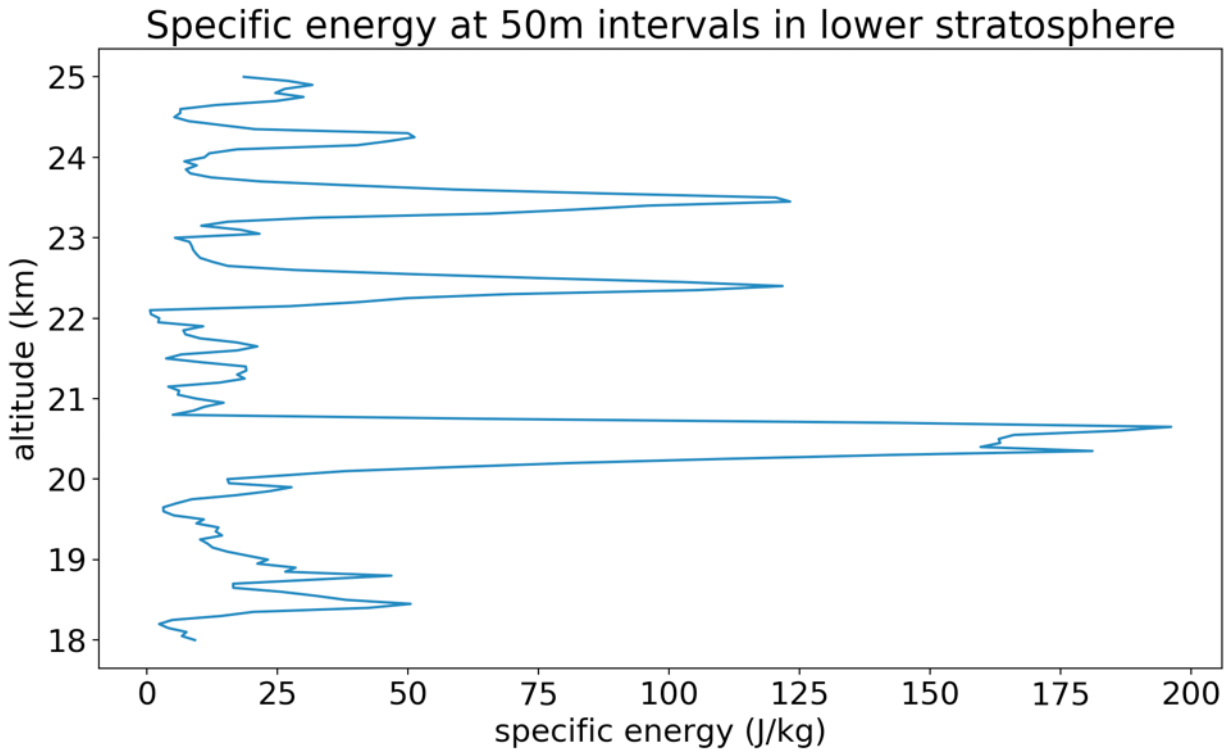


Figure 6: Specific energy as a function of altitude in lower stratosphere for a typical radiosonde sounding. The spiky plot illustrates that the inferred specific energy is extremely erratic. In this sounding the mean specific energy is 35.1 J/kg and the standard deviation is 43.5 J/kg.

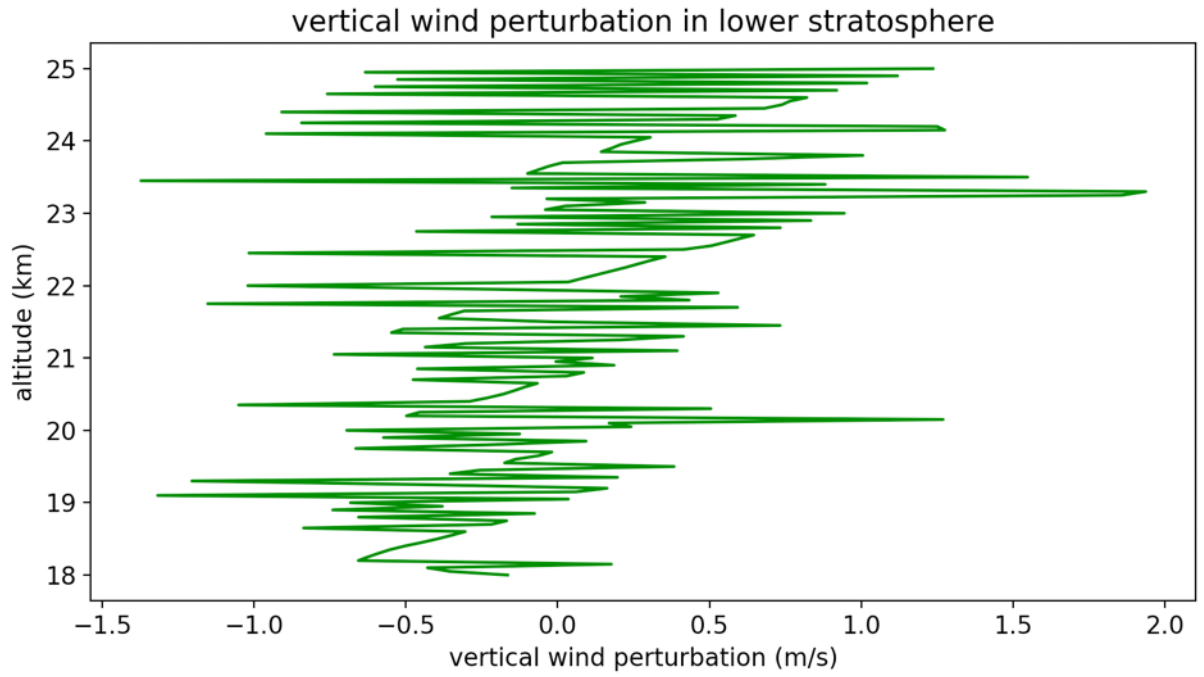


Figure 7: Vertical wind perturbation as a function of altitude in lower stratosphere for a typical radiosonde sounding. The spiky plot illustrates that the inferred vertical wind perturbations are extremely erratic.

Overall, these flaws show that while the fluid mechanics and thermodynamic underpinnings of the energy method are correct, the procedure yields unreliable results when applied to actual data. Therefore, the energy method was discarded and the hodograph method was used exclusively to infer the gravity wave parameters from the radiosonde data.

3 Exploring Inferred Gravity Wave Parameters

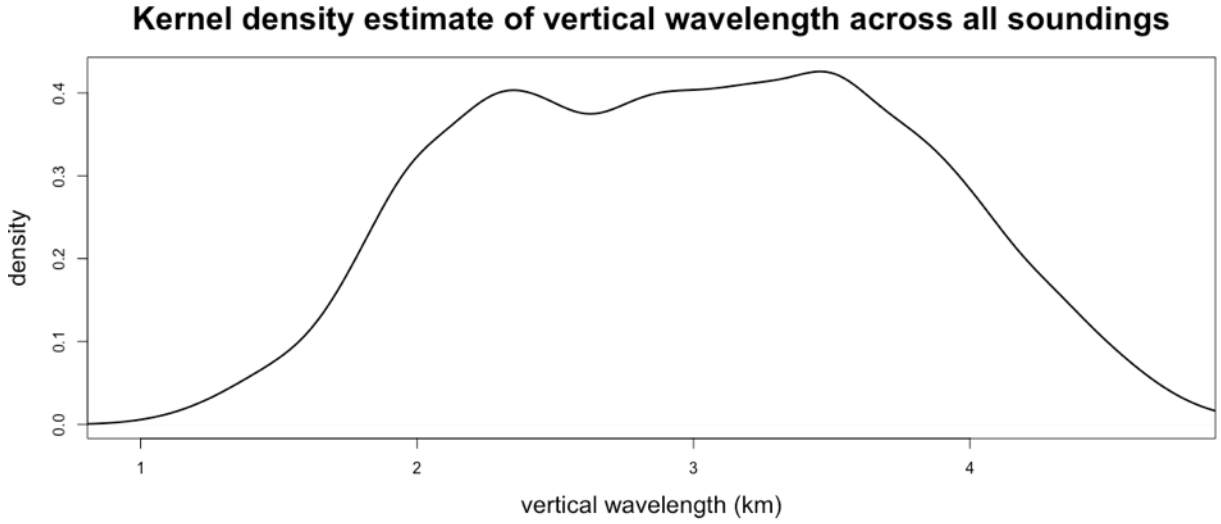
Now that the gravity wave parameters had been inferred, the next step was to explore the underlying structure of the inferred data to gain insight into how to model the joint probability distribution of the parameters. This process involved looking into several critical features of the data that are each outlined below:

3.1 Overall Structure

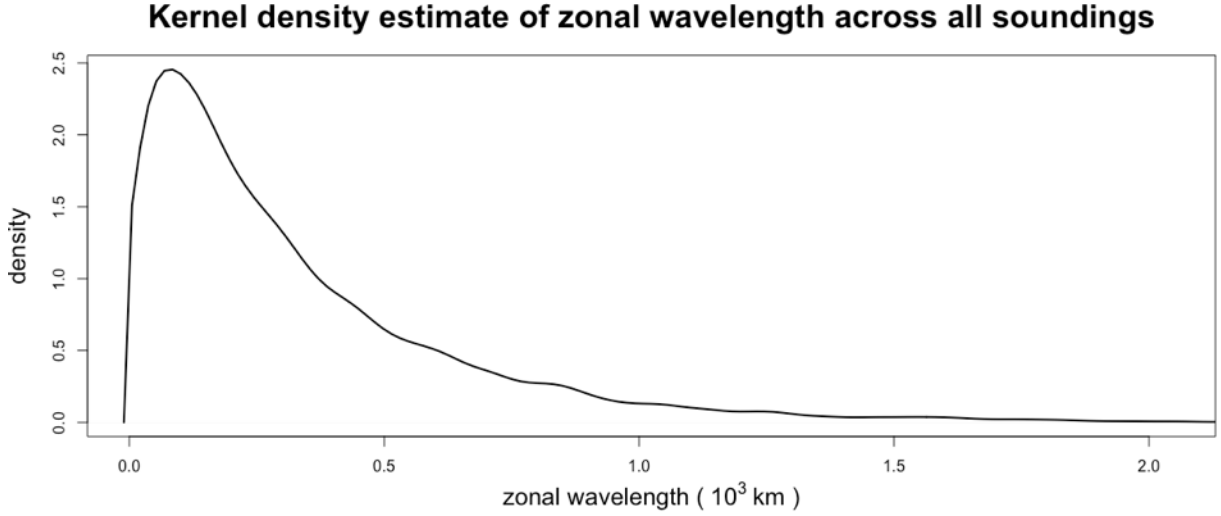
Before digging deeper into the inferred data, the first step was to get a sense of what the data looks like as a whole. Because a joint distribution needs to accurately model both the marginal distributions of the parameters as well as the correlation structure between each pair of parameters, this can best be done by exploring the kernel density estimates of each of the parameters as well as the scatterplots between each pair of parameters. These graphs are presented in Figures 8, 9 and A.1.

The goal of kernel density estimation is to approximate the probability distribution function of a random variable. It essentially works by counting the number of data points in an interval centered at a given location and then dividing this quantity by the length of the interval times the sample size. This procedure is repeated for many locations in the support of the distribution to produce a smooth density curve. This means that some of the density associated with values that are close to zero spreads into negative territory, which is an issue because all the gravity wave parameters can only take on positive values. Therefore, a modification was made to the kernel density estimation procedure that sets the density of any negative values to zero and nonuniformly adds that density to positive values such that the total area under the curve remains one. This adjustment is applied to all the kernel density estimates presented in the graphs [13].

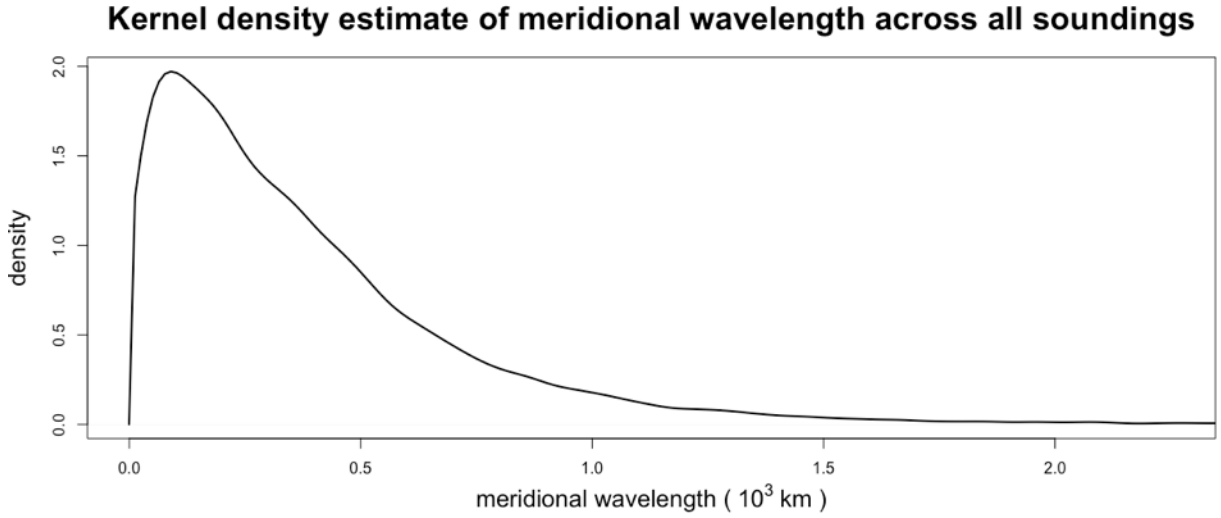
Meanwhile, the Spearman's rank correlation coefficient, which measures how well the relationship between two variables can be described using a monotonic function, is used to determine the correlation between pairs of parameters. The Spearman's rank correlation coefficient is equal to 1 and -1 for a monotonically increasing and decreasing relationship, respectively, and takes on an intermediate value when the relationship between the two variables is not strictly increasing or decreasing. This measure of correlation is used instead of the better-known Pearson correlation coefficient because it is a more robust method that isn't nearly as influenced by outliers and skewed random variables as is the Pearson correlation coefficient.


Kernel density estimate of vertical wavelength across all soundings
density
0.0
0.1
0.2
0.3
0.4
1
2
3
4
vertical wavelength (km)

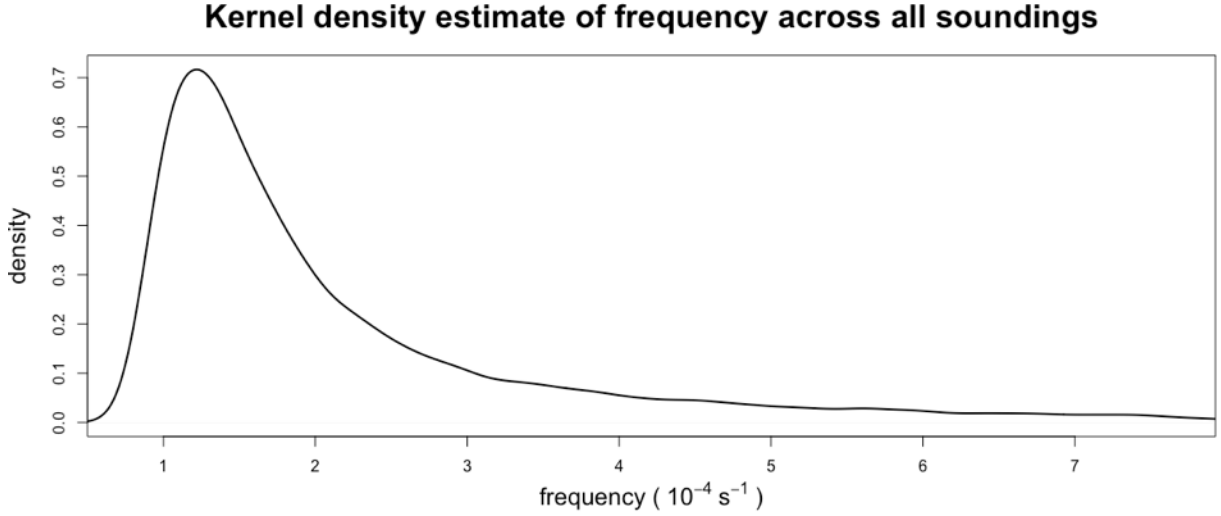
(a)


Kernel density estimate of zonal wavelength across all soundings
density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.5
1.0
1.5
2.0
zonal wavelength (10^3 km)

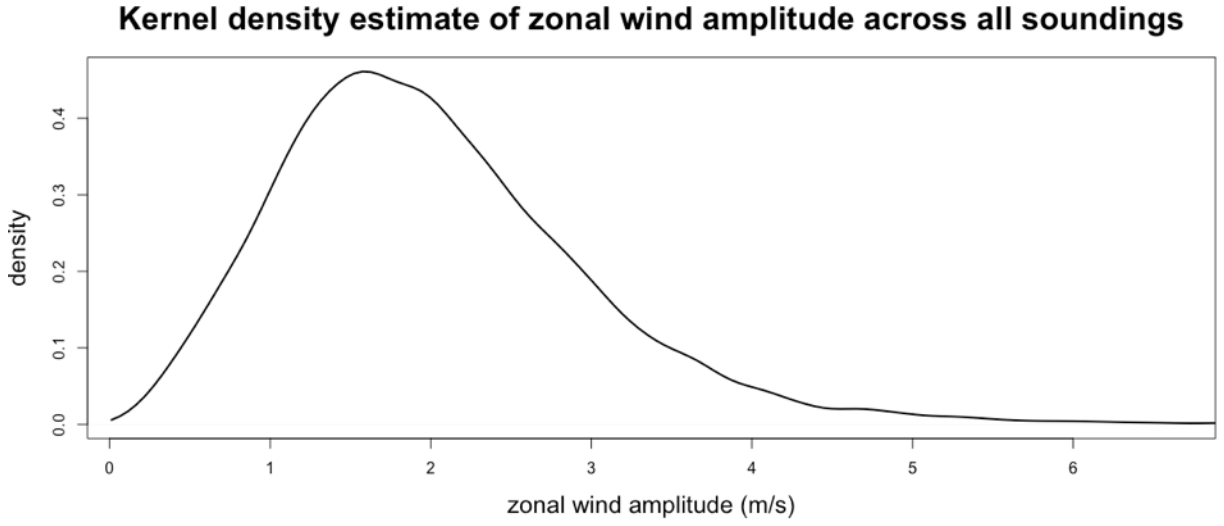
(b)


Kernel density estimate of meridional wavelength across all soundings
density
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0
meridional wavelength (10^3 km)

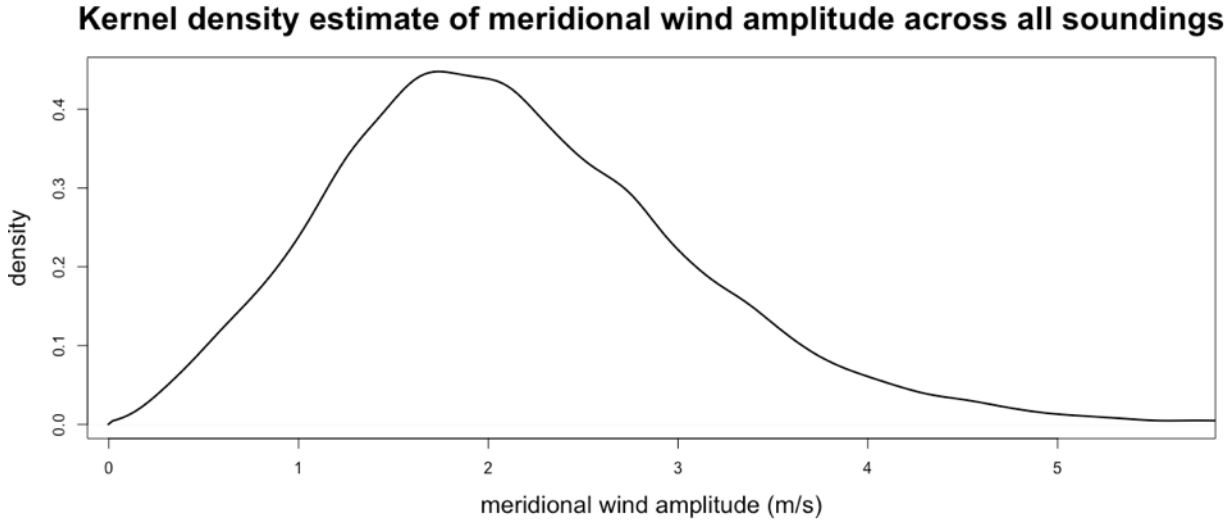
(c)


Kernel density estimate of frequency across all soundings
density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
3
4
5
6
7
frequency (10^{-4} s^{-1})

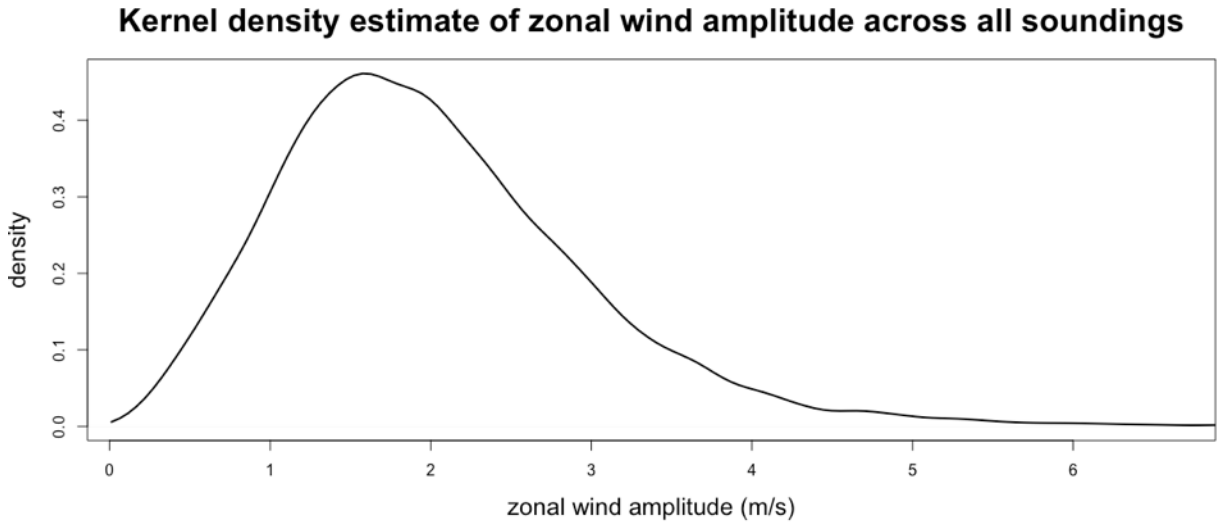
(d)


Kernel density estimate of zonal wind amplitude across all soundings
density
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
6
zonal wind amplitude (m/s)

(e)

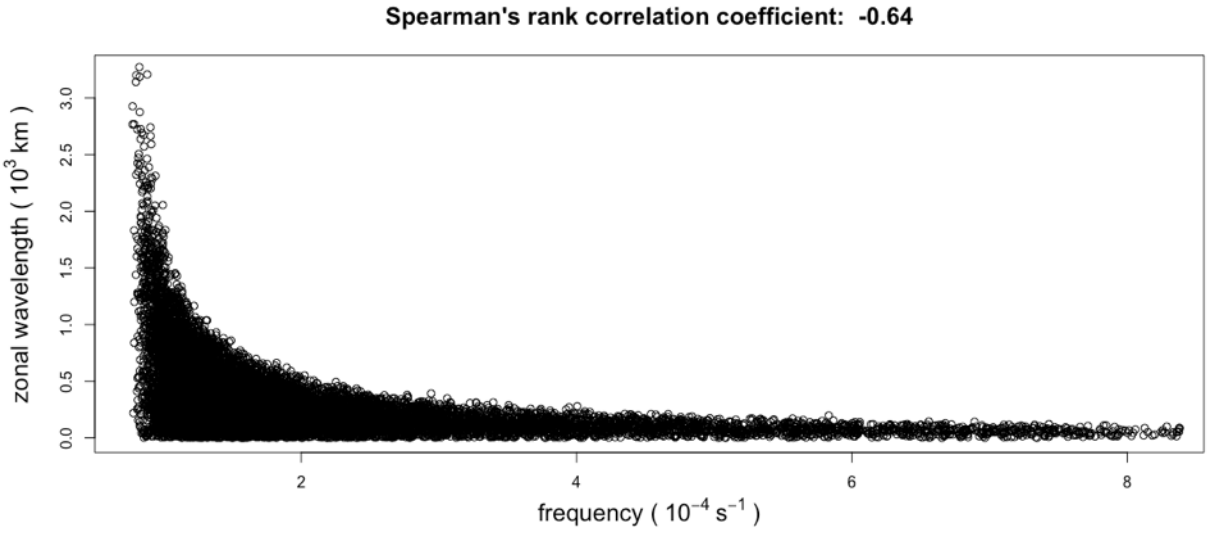

Kernel density estimate of meridional wind amplitude across all soundings
density
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
meridional wind amplitude (m/s)

(f)

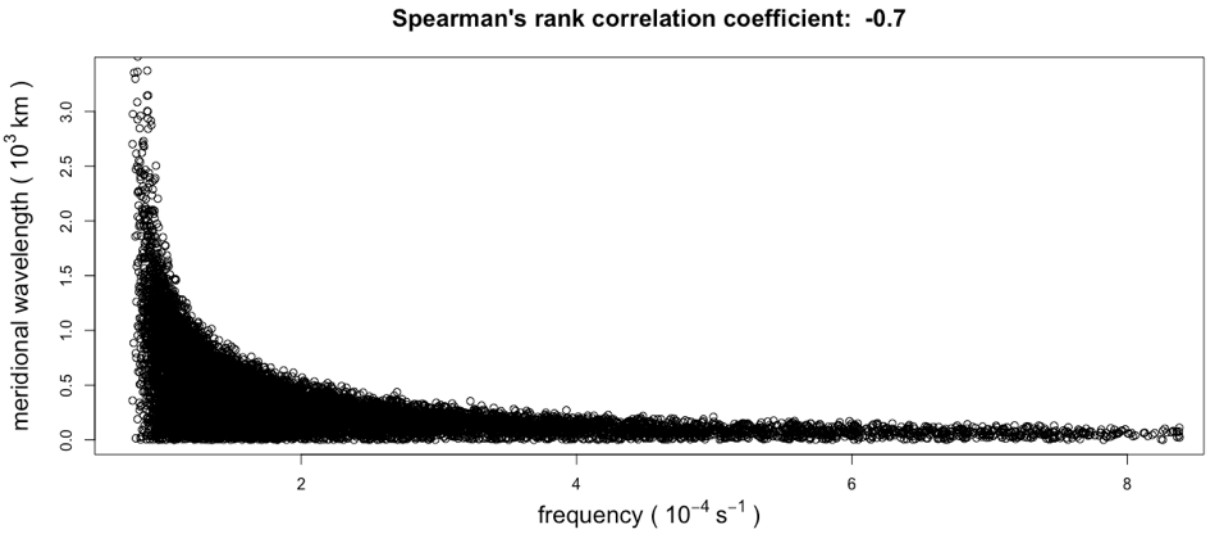


(g)

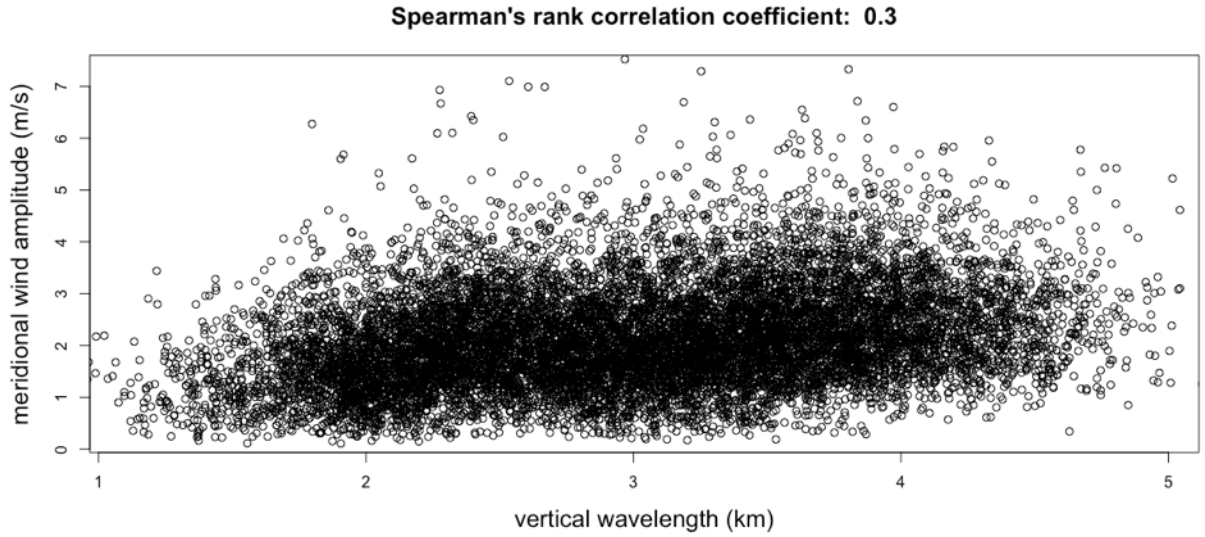
Figure 8: Kernel density estimates for each of the seven gravity wave parameters. All seven graphs are presented since they form the basis of the model that will be developed.


Spearman's rank correlation coefficient: -0.64
zonal wavelength (10^3 km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
4
6
8
frequency (10^{-4} s^{-1})

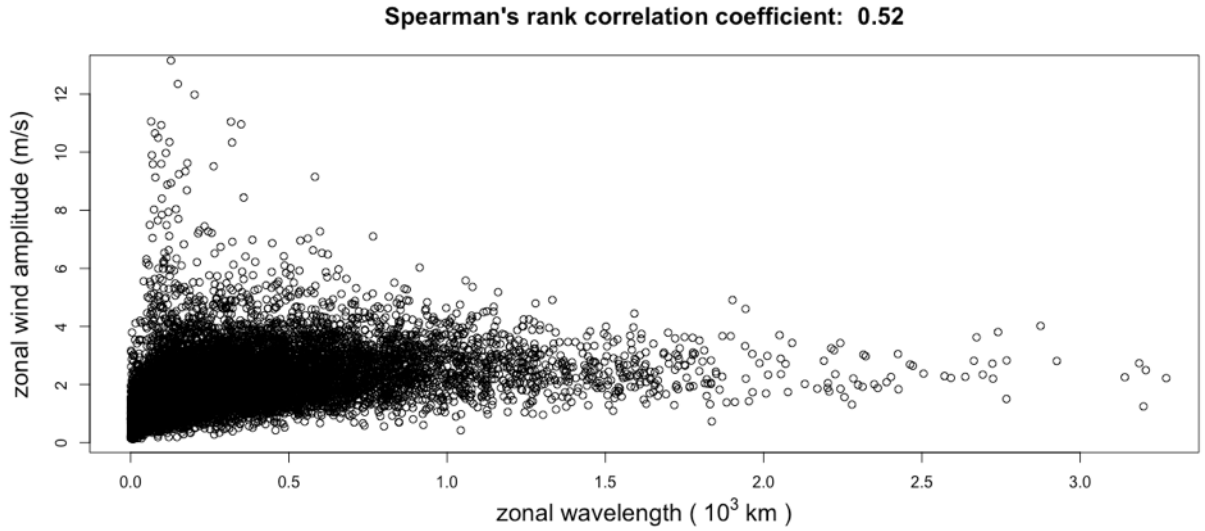
(a)


Spearman's rank correlation coefficient: -0.7
meridional wavelength (10^3 km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
4
6
8
frequency (10^{-4} s^{-1})

(b)



(c)



(d)

Figure 9: Scatterplots for select pairs of gravity wave parameters. The Spearman's rank correlation coefficient is reported in the title of each scatterplot.

Figures 9(a) and 9(b) illustrate a highly nonlinear, inverse relationship between frequency and zonal wavelength and frequency and meridional wavelength. Meanwhile, Figures 9(c) and 9(d) indicate a linear relationship between each pair of parameters. (The seemingly nonlinear skewness displayed in Figure 9(d) can be attributed to the fact that the marginal distributions of the two parameters are different and does not detract from the linear relationship between the two.) In fact, all 19 scatterplots outside of Figures 9(a) and 9(b) show a weak to moderate linear correlation between each pair of parameters. The rest of these scatterplots can be found in Figure A.1.

3.2 Sounding-to-sounding Variation

While the kernel density estimates and scatterplots described above provide a lot of insight into the overall structure of the data, they fail to capture how the inferred gravity wave parameters vary sounding to sounding. The first possibility is that the gravity wave parameters inferred in a given sounding depend on the gravity wave parameters inferred in the previous sounding. The second possibility is that the two sets of inferred parameters are independent of each other. Ideally, there would be a sounding every 12 hours (once at midnight and once at noon), but due to various issues with many of the soundings (i.e. radiosonde wasn't launched, radiosonde recorded too many false measurements, radiosonde didn't reach a high enough altitude, etc.), only roughly 40% of the soundings had their gravity wave parameters successfully inferred using the hodograph method. Therefore, the time between consecutive soundings where the gravity wave parameters were able to be inferred ranges anywhere from 12 hours (i.e. no missing soundings in between) to upwards of 150 hours (i.e. more than 12 missing soundings in between).

To determine whether the parameters inferred from consecutive soundings can be modeled independently of each other or must be modeled together, the absolute difference between the parameters inferred from consecutive soundings was calculated along with the number of missing soundings in between those consecutive soundings. This process was repeated for all of the seven parameters across all four stations. For each parameter, all consecutive soundings corresponding to a gap between the soundings of a given multiple of 12 hours were grouped together and the mean value was computed. The bar chart below illustrates these calculations for one of the gravity wave parameters. (The six other bar charts can be found in Figure A.2 and share the same overall structure.)

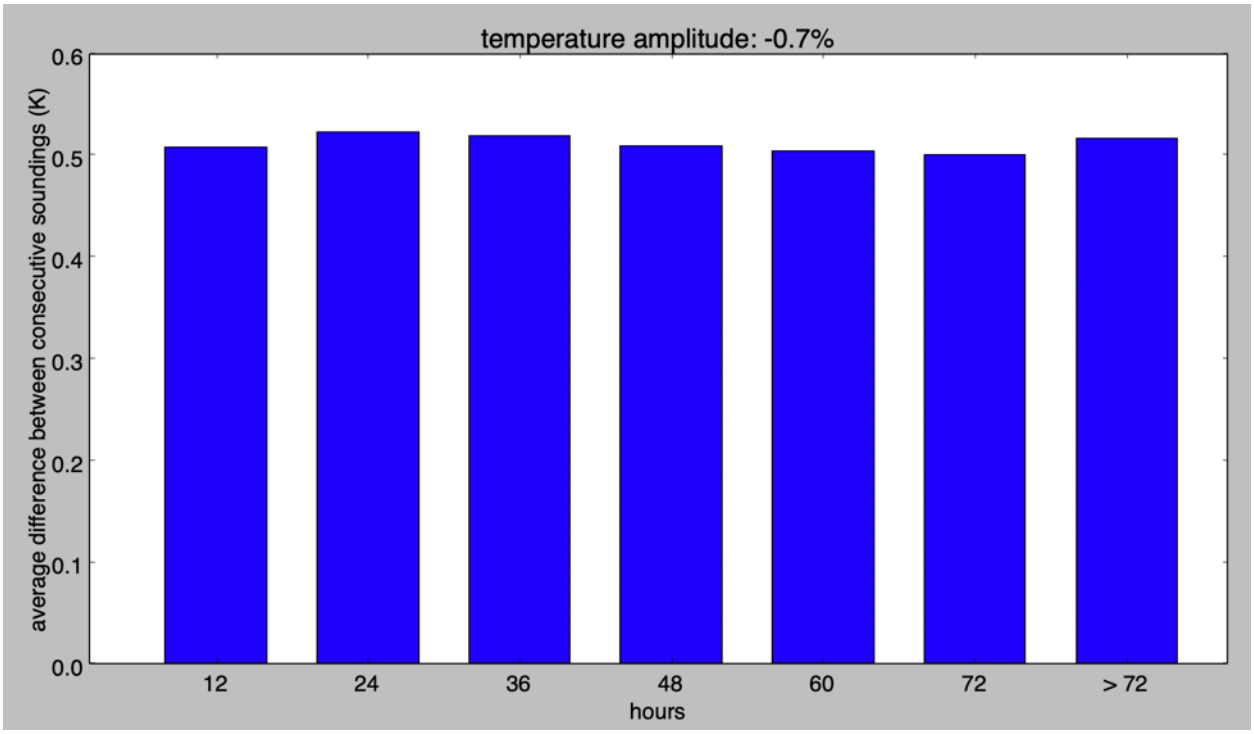


Figure 10: Bar chart for temperature amplitude that illustrates how the average difference between consecutive soundings varies as a function of the number of hours between consecutive soundings. The percentage presented in the title of the graph indicates the percent difference between the average difference between consecutive soundings that take place 12 hours apart and the average difference between consecutive soundings across all soundings regardless of the number of hours between consecutive soundings.

If the parameters inferred in radiosonde soundings that were launched just hours apart were highly correlated with each other, the heights of the bars in the graphs above would increase from right to left as the number of hours between consecutive soundings rises. However, Figure 10 clearly illustrate that this is not the case, as the heights of the bars effectively remain constant as the number of hours between consecutive readings increases. Quantitatively, the data shows that the average absolute difference in the parameters inferred from consecutive soundings where there are zero missing soundings in between is just 1.7% lower (averaged across all seven parameters) than the average absolute difference in the parameters inferred from consecutive soundings where there are any number of missing soundings in between. This small discrepancy indicates that the soundings are predominantly independent of each other, which means that any given sounding can be modeled independently of all the other soundings rather than having to be modeled using some version of a Markov model that takes into account the preceding soundings.

3.3 Variation by Radiosonde Station

Determining how the inferred parameters vary by radiosonde station involved repeating the process of creating kernel density estimates for each of the parameters and producing scatterplots between all pairs of parameters, but this time generating separate graphs for each of the four radiosonde stations. To compare the results, it was helpful to overlay the graphs associated with each station to look for major differences. Because it is extremely subjective to come up with quantitative metrics to determine whether there are significant discrepancies in the inferred gravity wave parameters among the four stations, a visual form of hypothesis testing was used to perform this analysis.

Given the close geographical proximity of the four stations under consideration, there was no a priori reason to believe that the gravity wave parameters inferred from the stations would be significantly different from one another. Therefore, the null hypothesis was that the inferred gravity wave parameters were identically distributed across all four stations. If there turned out to be major structural difference between the kernel density estimates and scatterplots associated with each of the stations, the null hypothesis would be rejected and the gravity wave parameters inferred from the stations would have to be modeled separately.

The kernel density estimate associated with one of the seven gravity wave parameters is reproduced below. In the graph, each of the four colored curves corresponds to one of the four stations (the specific station corresponding to each color is unimportant). The graphs corresponding to the six other gravity wave parameters share the same structure and can be found in Figure A.3.

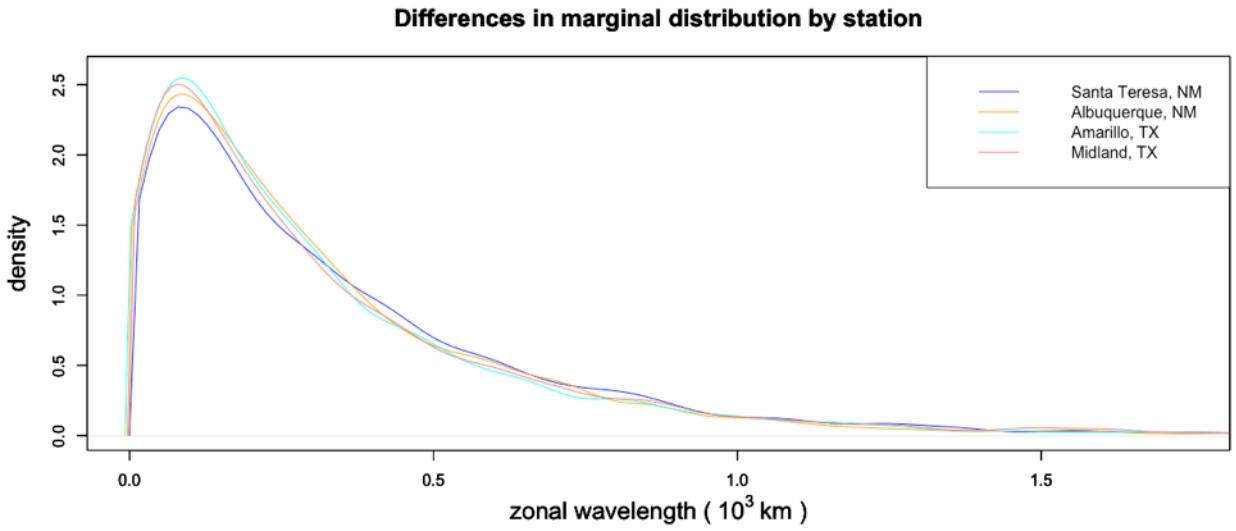


Figure 11: Overlaid kernel density estimates corresponding to each of the four stations for zonal wavelength. The overall shapes of the curves strongly coincide with one another and any minor deviations are attributed to statistical fluctuations.

Taken together, the graph in Figure 11 and those in Figure A.3 illustrate that there are no major differences in the gravity wave parameters that are inferred from each station. This is indicated by the fact that there are no major structural differences between the overall shapes of the kernel density estimates associated with each station. Instead, the minor discrepancies are attributed to statistical fluctuations. The 21 scatterplots, which have been omitted for brevity, corroborate this strong consistency across the four stations, as indicated by four overlapping scatterplots and Spearman's rank correlation coefficients that are within several hundredths of each other.

Together, the seven kernel density estimates and the 21 scatterplots led to the failure to reject the null hypothesis, which meant that the inferred gravity wave parameters were considered to be identically distributed across all four stations. This, in turn, allowed the inferred gravity wave parameters to be aggregated across the four stations.

It is important to note that a more quantitative technique could, in theory, be used to compare the empirical distributions across the four radiosonde stations. For example, the Kolmogorov-Smirnov test evaluates the hypothesis that two samples are drawn from the same distribution. The issue with this approach is that this null hypothesis that both samples are drawn from the same distribution is a very high standard to meet, which means that two empirical distributions under consideration need to be extremely similar to yield a p-value over 0.05 that would lead to the failure to reject the null hypothesis.

To illustrate this strict criterion, the p-values associated with the Kolmogorov-Smirnov tests that compare the blue kernel density estimate in Figure 11 to the other three kernel density estimates are 0.007, 0.002 and 0.025. This means that even though the blue curve is closely aligned with the other three curves, it can be asserted with roughly 97.5% confidence that the distribution of zonal wavelength parameters inferred in Santa Theresa, NM (the station corresponding to the blue curve) is different than those inferred in the other three locations. Consequently, using the Kolmogorov-Smirnov test to quantitatively compare the empirical distributions would force the gravity wave parameters to be modeled separately for the different stations, even though there exists a strong consistency in the gravity wave parameters inferred across the four stations. Therefore, it makes sense to use the less quantitative approach of visually comparing the parameters inferred at each station, since doing so allows for best judgement to be used instead of having to rely on fixed criteria that may not be suitable to specific situations. (This same argument against the use of a rigid metric such as the Kolmogorov-Smirnov test applies to the following discussions on the variation in the inferred gravity wave parameters by year, time of day and month.)

3.4 Variation by Year

The steps involved here are analogous to those discussed above, except that the data is categorized by year rather than by station.

Just as with regard to the four stations, there is no a priori reason to believe that the gravity wave parameters inferred across the years between 1998 and 2008 would be significantly

different from one another. (Although the literature discusses the possible influence of annual phenomena such as Quasi-Biennial Oscillation and El Niño – Southern Oscillation, the impact of these factors isn't significant enough to expect major differences in interannual gravity wave activity.) Therefore, the null hypothesis was that the inferred gravity wave parameters were identically distributed across all years between 1998 and 2008. If there turned out to be major structural differences between the kernel density estimates and scatterplots associated with each year, the null hypothesis would be rejected. In this case, more research would need to be done to identify which factors are responsible for the discrepancies so that the gravity wave activity in a given year could be modeled using data from previous years that shared similar atmospheric characteristics.

The kernel density estimates associated with each year are reproduced below for one of the gravity wave parameters. In the graph, each of the 11 colored curves corresponds to one of the 11 years between 1998 and 2009 (the specific year corresponding to each color is unimportant). Again, the graphs corresponding to the six other gravity wave parameters share the same structure and can be found in Figure A.4.

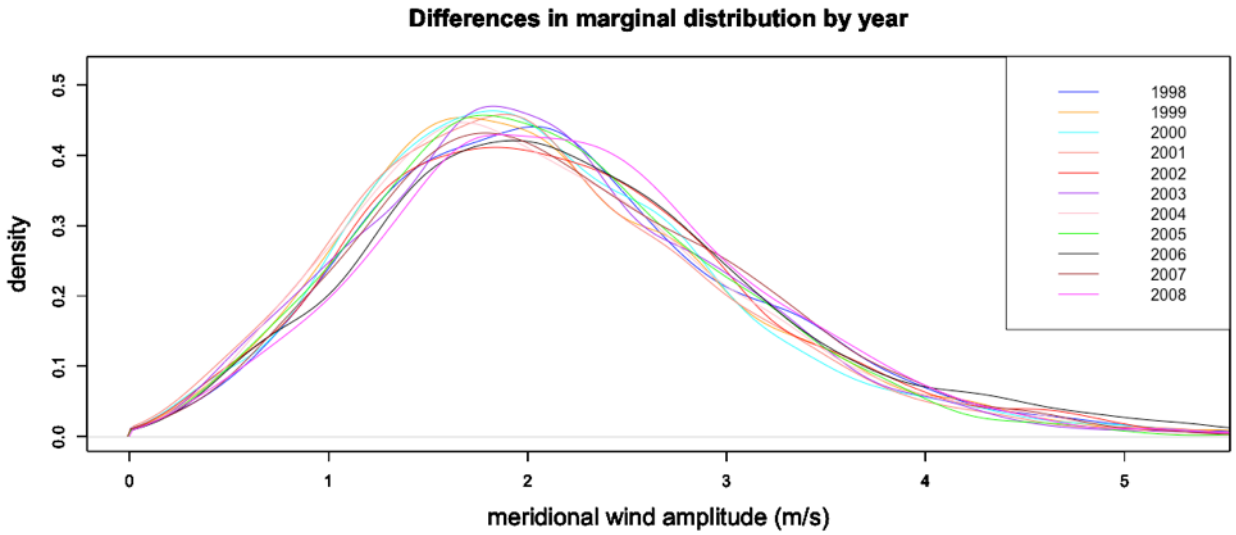


Figure 12: Overlaid kernel density estimates corresponding to each year between 1998 and 2008 for meridional wind amplitude. The overall shapes of the curves strongly coincide with one another and any minor deviations are attributed to statistical fluctuations.

As with the stations, the graphs in Figure 12 and Figure A.4 collectively illustrate that there are no major differences in the gravity wave parameters that are inferred during each year. This is indicated by the fact that there are no major structural differences between the overall shapes of the kernel density estimates associated with each year. Instead, the minor discrepancies are attributed to statistical fluctuations. The 21 scatterplots corroborate this strong consistency across the 11 years, but have been omitted for brevity.

Taken together, the seven kernel density estimates and the 21 scatterplots led to the failure

to reject the null hypothesis, which meant that the inferred gravity wave parameters were considered to be identically distributed across all years. This, in turn, allowed the inferred gravity wave parameters to be aggregated across the 11 years under consideration.

3.5 Variation by Time of Day

The steps involved here are analogous to those discussed above, except that the data is categorized by time of day rather than by station or year. The two times under consideration are midnight and noon, since the vast majority of the soundings are launched at these times. Again, there's no reason to believe that the gravity wave parameter inferred at these two times would be significantly different from each other, so the null hypothesis is that the inferred gravity wave parameters are identically distributed across all times of day.

The kernel density estimates associated with one of the seven gravity wave parameter are reproduced below. In the graphs, the two colored curves correspond to gravity wave parameter inferred at noon and at midnight (the specific time of day corresponding to each color is unimportant). Again, the graphs corresponding to the six other gravity wave parameters share the same structure and can be found in Figure A.5.

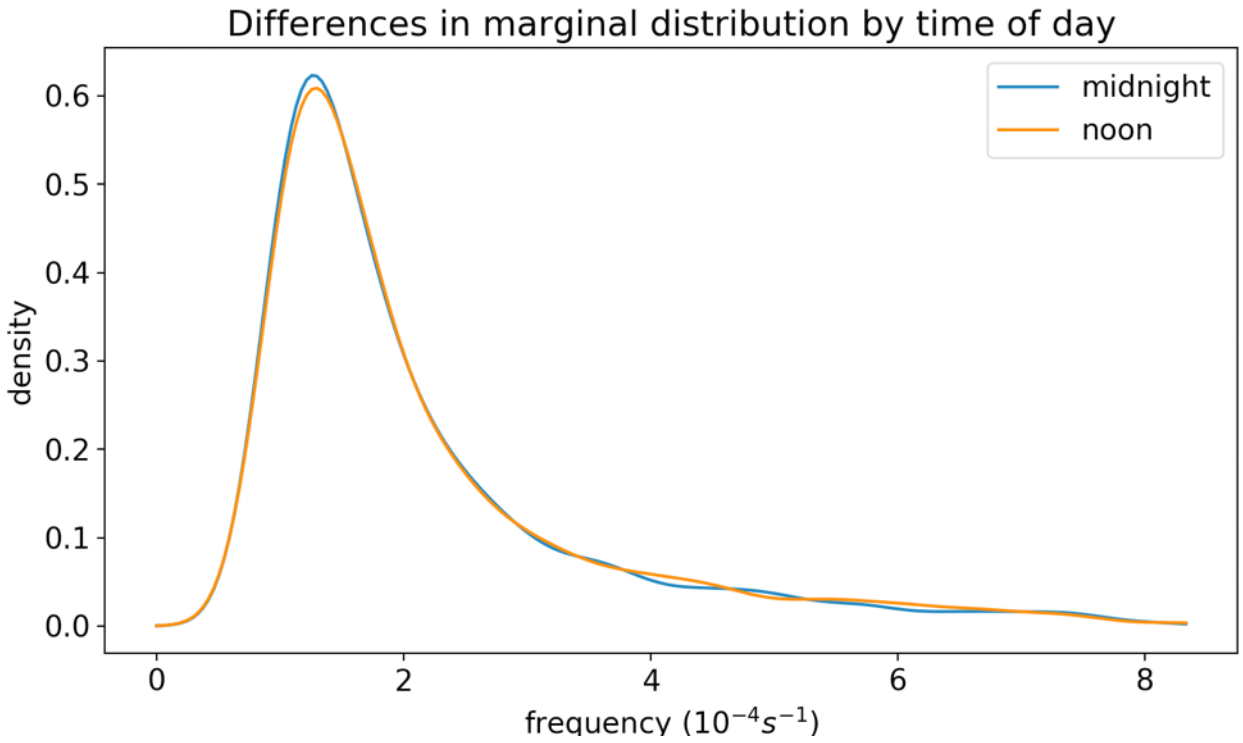


Figure 13: Overlaid kernel density estimates corresponding to noon and midnight for frequency. The overall shapes of the curves strongly coincide with each other and any minor deviations are attributed to statistical fluctuations.

Just as with the stations and years, the graphs in Figure 13 and Figure A.5 collectively il-

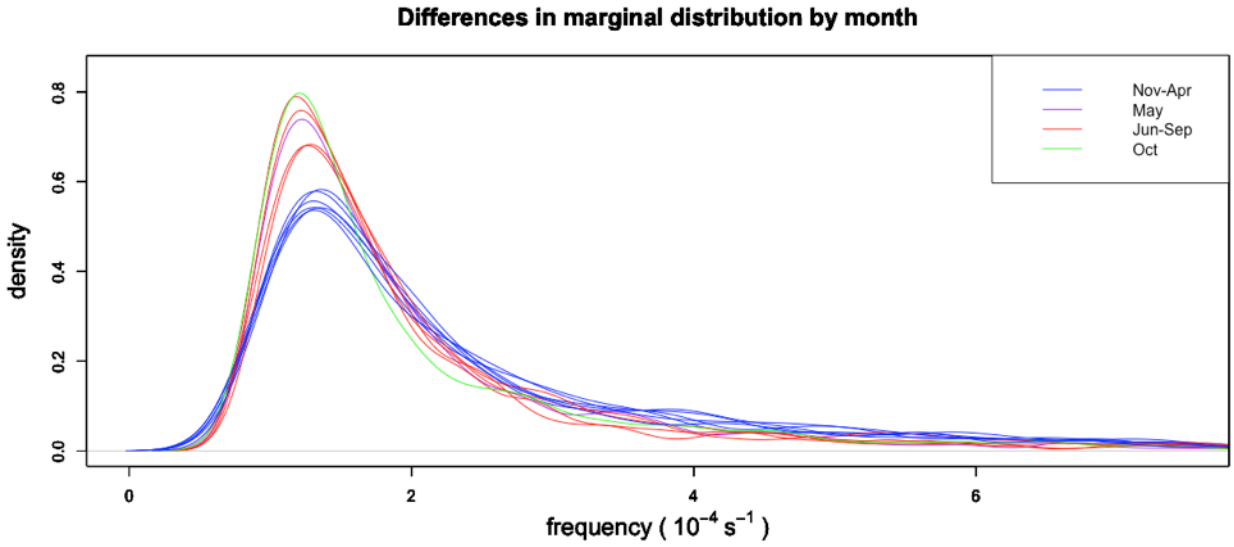
lustrate that there are no major differences in the gravity wave parameters that are inferred at different times of day. This is indicated by the fact that there are no major structural differences between the overall shapes of the kernel density estimates associated with each time of day. Instead, the minor discrepancies are attributed to statistical fluctuations. The 21 scatterplots corroborate this strong consistency across the two times of day but have been left out for succinctness.

Taken together, the seven kernel density estimates and the 21 scatterplots led to the failure to reject the null hypothesis, which meant that the inferred gravity wave parameters were considered to be identically distributed across all times of day. This, in turn, allowed the inferred gravity wave parameters to be aggregated across all times of day.

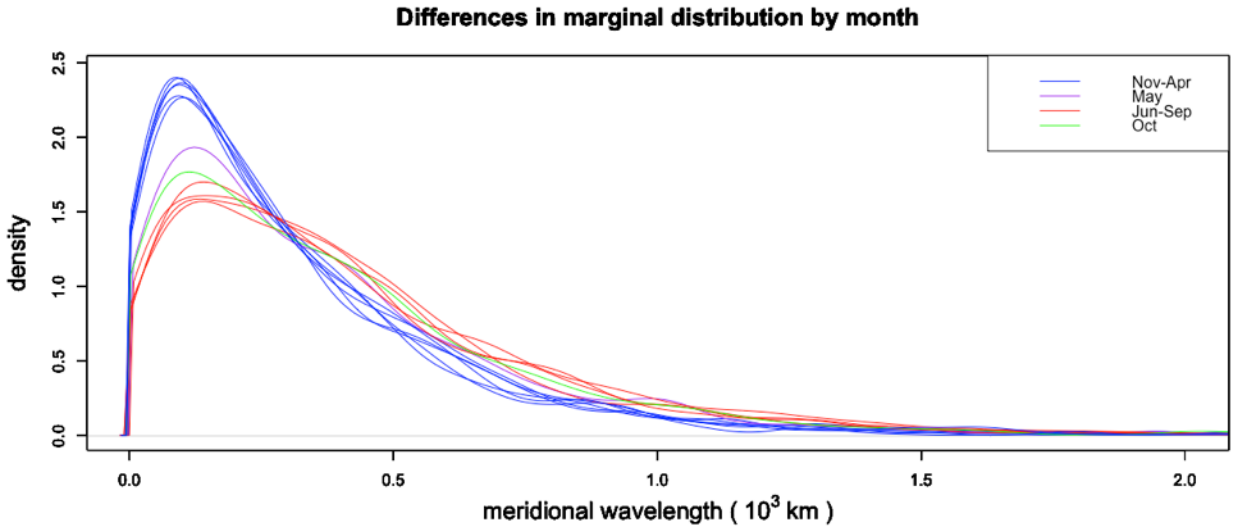
3.6 Variation by Month

The steps involved here are analogous to those discussed above, except that the data is categorized by month rather than by station, year or time of day. Just as with regard to the other categories, there is no a priori reason to believe that the gravity wave parameters inferred across the months of the year would be significantly different from each other. Therefore, the null hypothesis was that the inferred gravity wave parameters were identically distributed across all months of the year.

The kernel density estimates associated with two of the seven gravity wave parameter are reproduced below. In both graphs, each of the 12 colored curves corresponds to one of the months of the year (in this case, the specific months corresponding to each color are important). The graphs corresponding to the five other gravity wave parameters can be found in Figure A.6.



(a)



(b)

Figure 14: Overlaid kernel density estimates corresponding to each month of the year for frequency and meridional wavelength. In each graph, the blue curves correspond to the months between November and April (inclusive), the purple curve corresponds to the month of May, the red months correspond to the months between June and September (inclusive) and the green curve corresponds to the month of October.

Unlike in the graphs that were categorized by station, year and time of day, Figure 14 shows that there do exist discrepancies in the underlying structure of the kernel density estimates associated with different months of the year. Ideally there would be some quantitative metrics that could be used to justify grouping certain months of the year together. However, as discussed previously, these methods are limited in their usefulness since the decision on

where to draw the cutoff between distributions that are similar enough and distributions that differ too much is inherently arbitrary. As a result, the kernel density estimates were relied on to visually cluster the months of the year.

Figure A.6(a) illustrates that the kernel density estimates associated with the warmer months between June and September are skewed to the right compared to the kernel density estimators associated with the other months of the year. Figure 14(a) illustrates that the kernel density estimates associated with the colder months between November and April fall into one cluster, while the kernel density estimates associated with the other months of the year fall into another. Figure A.6(b) illustrates the same phenomenon as Figure 14(a) except for one of the warmer months that falls into the cluster associated with the colder months. Figure 14(b) further reinforces this trend, except in this case the month of May falls almost halfway between the cluster associated with the colder months and the cluster associated with the rest of the months. Figure A.6(c) illustrates a similar pattern, but this time the clusters overlap to some degree and the intermediate months of May and October are more closely related to the colder months. Unlike the other graphs, Figure A.6(d) doesn't display any distinct clusters. Finally, Figure A.6(e) illustrates a more amorphous clustering pattern between the colder and warmer months, where the month of October is clustered with the warmer months, while the month of May lies somewhere in between the two clusters.

In this case, the 21 scatterplots don't provide much additional information. The scatterplots associated with each month are consistent with one another, and while the Spearman's rank correlation coefficient varies to some degree between the colder and warmer months, the difference is isn't large enough (i.e. average difference is 0.062) to conclude that the discrepancies are statistically significant and cannot be attributed to statistical fluctuation.

Despite the lack of additional insight from the scatterplots, the fact that the kernel density estimates for the different months tend to form clusters across most of the gravity wave parameters provided enough motivation to model the joint probability distributions of the seven gravity wave parameters separately for different months. As discussed above, the colder months between November and April tend to form one cluster, while the warmer months between June and September tend to form another cluster, so the logical design decision was to model the joint distributions separately for these two groups of months. Meanwhile, the months of May and October, which are neither particularly warm nor cold, do not fall consistently into the clusters associated with either the warmer or the colder months. Therefore, to reflect a lower degree of certainty about the distributions of the gravity wave parameters during these intermediary months, the logical design decision was to model the joint distribution associated with these months using the inferred gravity wave data from all 12 months of the year.

Although it's unclear why there are significant differences in the distribution of gravity wave parameters at different times of year, there are several phenomena that could potentially account for the disparities. As discussed in the introduction to atmospheric gravity waves, two leading sources of atmospheric gravity waves are the jet stream and convection. In the Northern Hemisphere, the strength of the jet stream is related to the temperature difference

between the warm tropical air mass to the south and the cold arctic air mass to the north. This temperature difference reaches its maximum during the winter months, which could explain why the gravity wave activity during these colder months is different than it is during the warmer months. Moreover, convection tends to become more potent as the temperature increases [14], which offers another explanation as to why the gravity wave activity during the warmer months is different than it is during the colder months.

4 Modeling Joint Probability Distribution

As explained in the preceding section, a total of three models were built: one for the colder months of November through April, one for the warmer months of June through September and one for the intermediary months of May and October. For simplicity, all the figures and values presented in this section are from the model built for May and October that incorporates the data from all 12 months.

R was used exclusively to construct this joint probability distribution because its wide variety of packages and emphasis on machine learning and statistical analysis offered much more functionality and convenience than Python.

4.1 General Strategy

A high-quality joint probability distribution is one that accomplishes two things: first, it must accurately model the marginal distributions of all the variables and second, it must account for the correlation between all of the variables. To reflect these two distinct components and to avoid tackling everything at once, the approach to building the model was split into two steps. First, each of the seven gravity wave parameters was modeled independently of the others, and then the model was iteratively improved upon by gradually incorporating the correlation structures between the different parameters. Beyond facilitating the modeling process, by naively modeling the parameters independently of one another, it was possible to establish a baseline AIC value that could subsequently be decreased by taking into account the correlation. The most beneficial aspect of using the AIC to evaluate the model along the way was that this one-number summary could gauge the degree to which a given modification to the model improved or worsened the model. Specifically, a model that better matched the underlying structure of the raw data would be associated with a decreased AIC value, and vice versa.

4.2 Modeling Parameters Independently

The first step to modeling the marginal distributions of the seven gravity wave parameters was to explore the shape of the kernel density estimates of each parameter, as presented in Figure 5. The curve associated with vertical wavelength appears symmetric with tails that resemble those of a normal distribution and a flat, plateau-like peak on top. Meanwhile, the curves associated with the other six parameters all have a similar shape where they peak at

a low value and feature heavy right tails.

To facilitate the modeling process, the values of zonal wavelength and meridional wavelength, which are typically in the hundreds of kilometers, were divided by 1,000. Without this scaling step, the parameters fitted to the data would have been very large, which would have led to issues in the iterative algorithms used in R. Moreover, it is more convenient to work with and interpret parameters that are in the single digits as opposed to the hundreds or thousands. Next, the values of frequency were shifted downwards such that the minimum frequency value was just greater than zero. This shift made it possible to fit distributions that have a support between zero and infinity without having to shift them upwards to align with the raw frequency data.

The next step was to come up with a list of probability distributions that could be used to fit the raw data. Most distributions have been developed for particular applications, which means they are natural choices in certain situations. The binomial distribution, for example, is designed to model the number of successes and failures, while the exponential distribution is designed to model the waiting time until a certain event. Unfortunately, the true underlying distributions of atmospheric gravity wave parameters such as frequency and wavelength have yet to be characterized by atmospheric scientists, which means there aren't any distributions that are natural choices for modeling them.

Therefore, it was necessary to rely on identifying probability density functions (PDFs) that empirically seem to match the shape of the kernel density estimates. This manual approach is standard practice in statistics when there is no a priori reason to believe that a given variable is distributed according to a particular distribution. Because all the gravity wave parameters can only take on positive values, the set of potential distributions was limited to those with a support from zero to infinity. To model the vertical wavelength, the most logical choice was to use the generalized normal distribution, which shares the same characteristics as the kernel density estimate associated with the vertical wavelength. Both of these are symmetric and feature a flat, plateau-like peak and tails that resemble those of a normal distribution. (Although the support of the generalized normal distribution includes all real numbers, it is suitable for this application since the density corresponding to all negative values is essentially zero.) Meanwhile, there is no single distribution that stands out as the best one to use to model the other six gravity wave parameters. Fortunately, there are many distributions that are right-skewed and have a support of all positive values. These include the gamma, Weibull, lognormal, Gumbel, Burr, Dagum, inverse gamma, Rayleigh and generalized Rayleigh distributions.

Now that the candidate distributions for each gravity wave parameter had been specified, the last step was to determine which distribution best fits the raw data and, finally, to compute the parameters associated with the distribution that optimizes the fit. These two tasks were combined into one by solving for the maximum likelihood estimate (MLE) of each candidate distribution and then calculating the AIC corresponding to each MLE and selecting the distribution that corresponds to the lowest AIC. This procedure is outlined below in Figure 15 for the meridional wind amplitude.

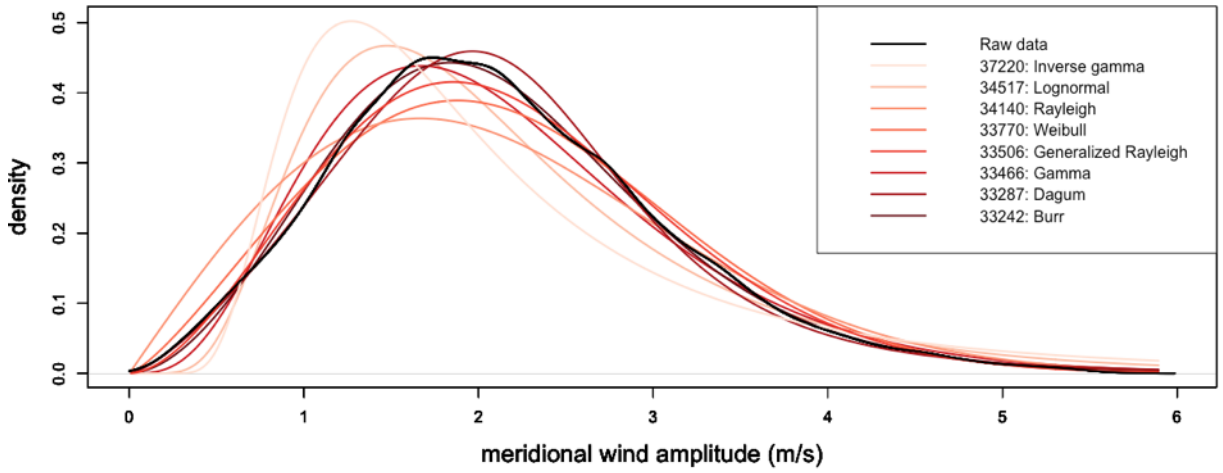
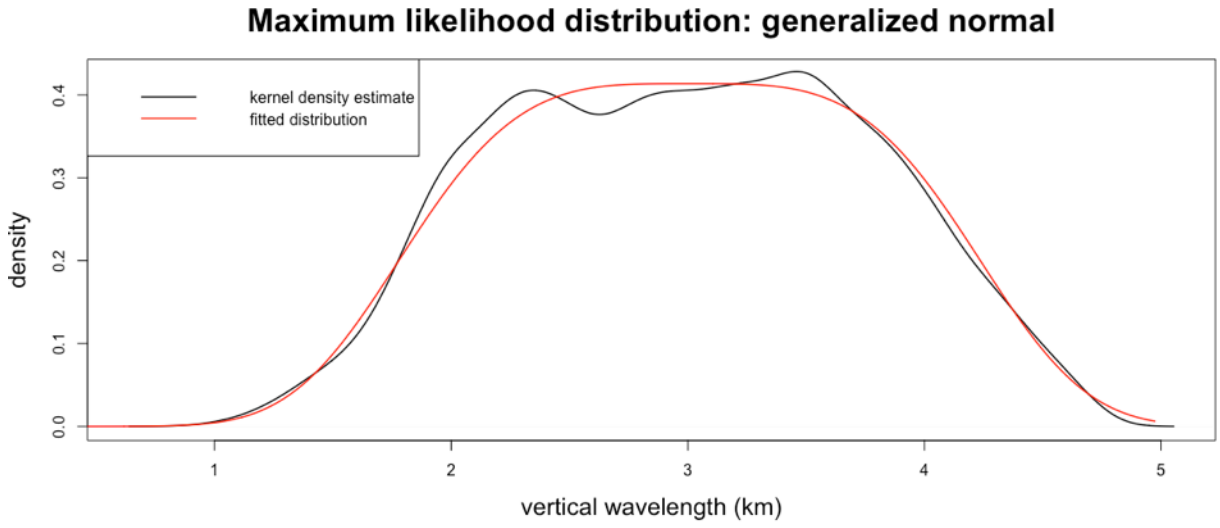


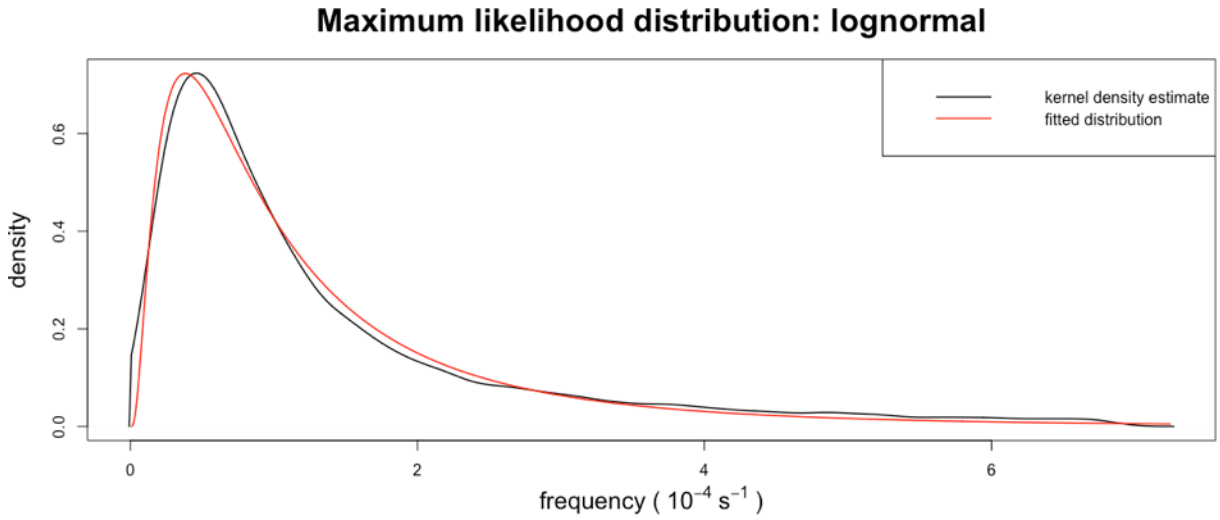
Figure 15: Maximum likelihood estimate and corresponding AIC value for each candidate distribution for the meridional wind amplitude. The black curve represents the kernel density estimate of the raw data, while the red curves represent the PDFs corresponding to the MLEs of the eight candidate distributions. The darker the curve, the more the PDF aligns with the raw data and the lower the AIC. Ultimately, the Burr distribution was used to model the marginal distribution of the meridional wind amplitude because its MLE had the lowest AIC.

As illustrated in Figure 15, the AIC was adept at evaluating the relative quality of the fits associated with each candidate distribution, as a reduction in the AIC corresponded to a visual improvement to the fitting of the raw data.

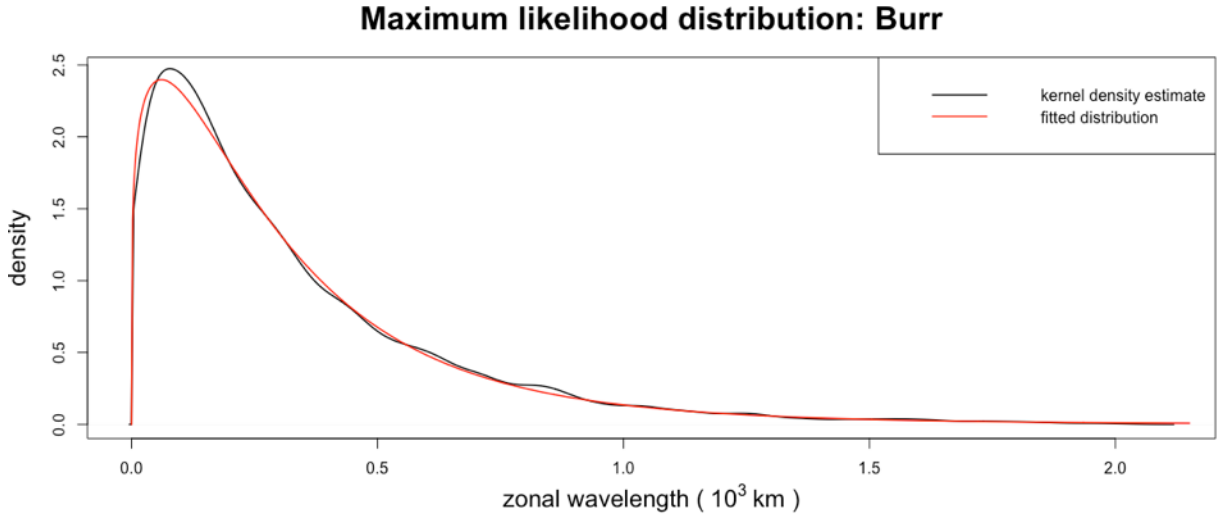
This process of calculating the MLE for each candidate distribution was then repeated for the other six gravity wave parameters. The distributions that corresponded to the lowest AIC for each parameter are presented below in Figure 16.


Maximum likelihood distribution: generalized normal
kernel density estimate
fitted distribution
density
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
vertical wavelength (km)

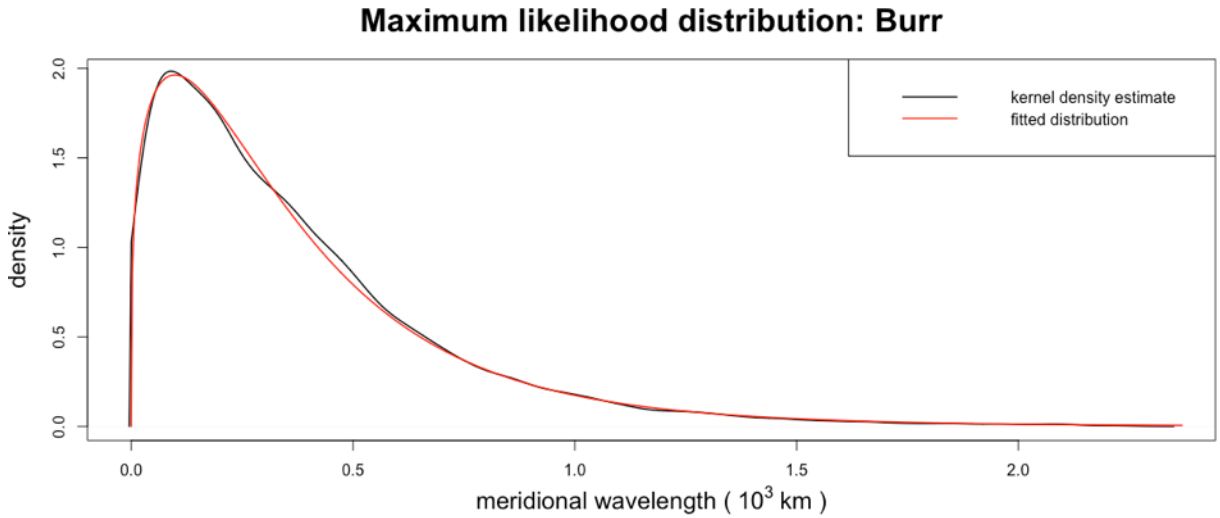
(a)


Maximum likelihood distribution: lognormal
kernel density estimate
fitted distribution
density
0.0
0.2
0.4
0.6
0
2
4
6
frequency (10^-4 s^-1)

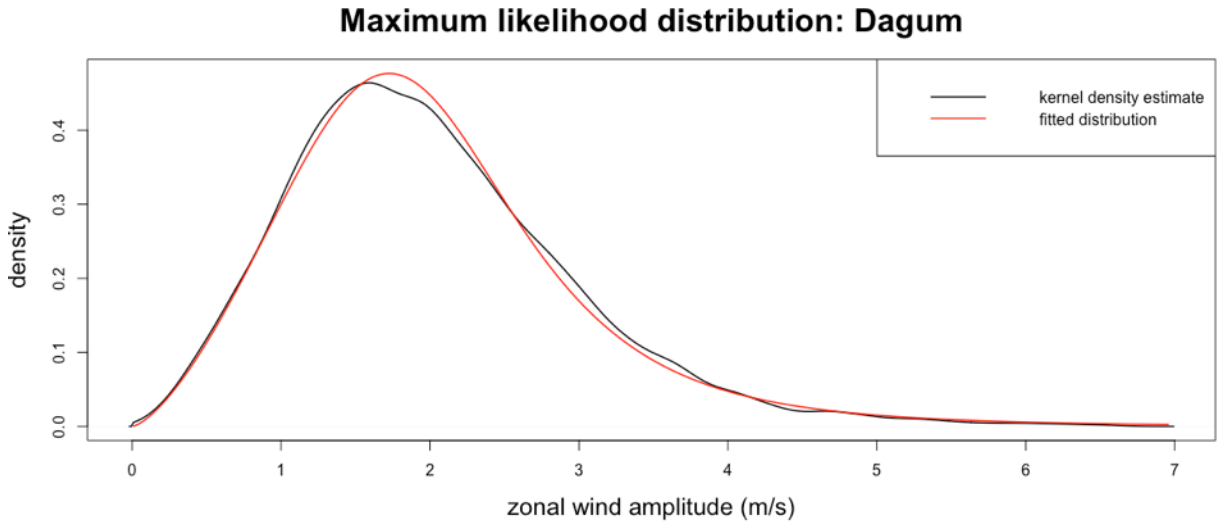
(b)


Maximum likelihood distribution: Burr
kernel density estimate
fitted distribution
density
0.0
0.5
1.0
1.5
2.0
2.5
zonal wavelength (10^3 km)

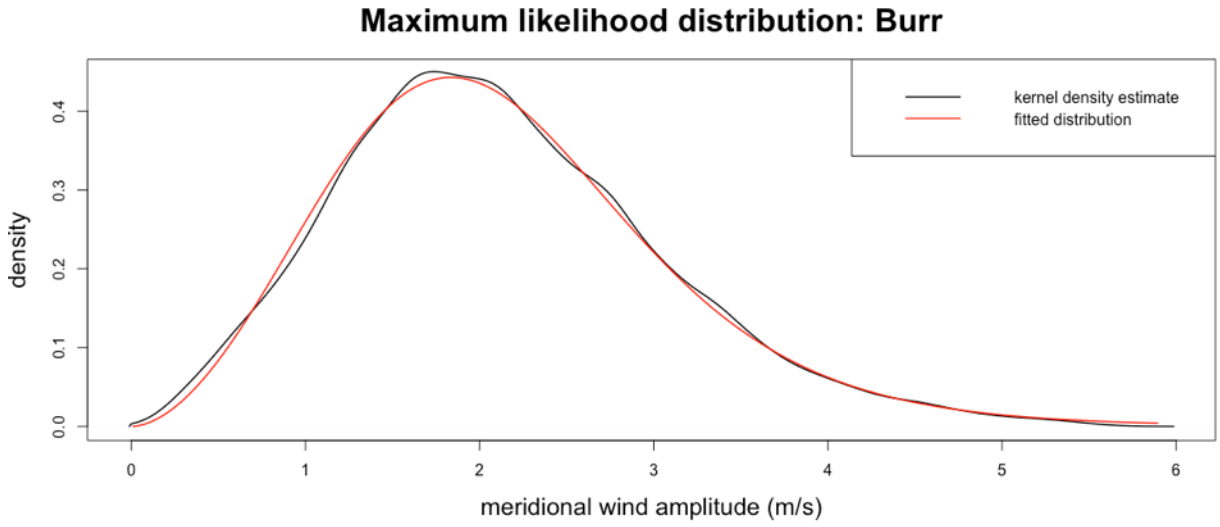
(c)


Maximum likelihood distribution: Burr
kernel density estimate
fitted distribution
density
0.0
0.5
1.0
1.5
2.0
meridional wavelength (10^3 km)

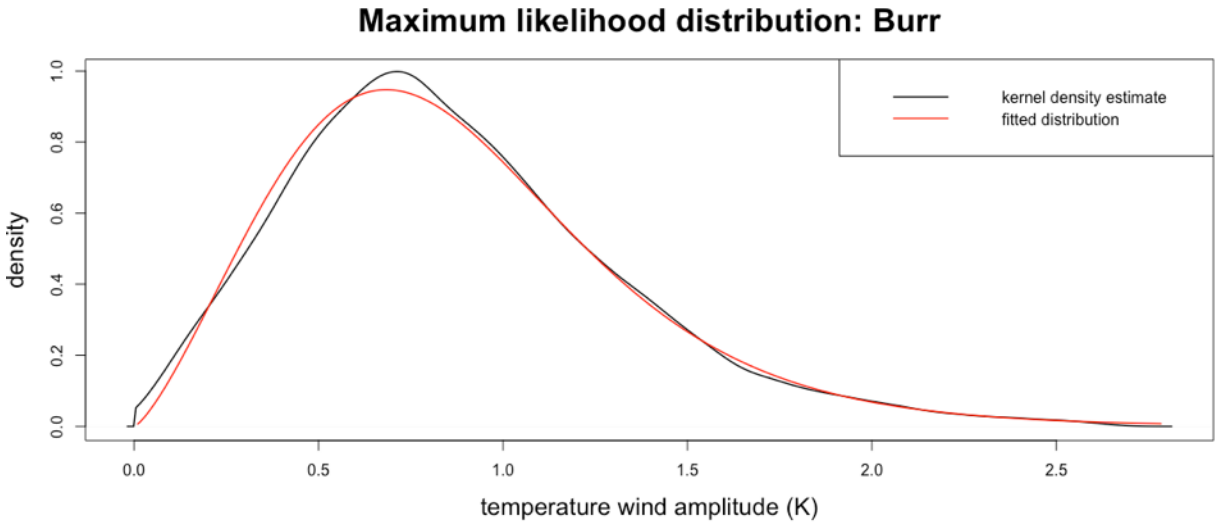
(d)


Maximum likelihood distribution: Dagum
kernel density estimate
fitted distribution
density
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
6
7
zonal wind amplitude (m/s)

(e)


Maximum likelihood distribution: Burr
kernel density estimate
fitted distribution
density
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
6
meridional wind amplitude (m/s)

(f)



(g)

Figure 16: MLEs for the marginal distributions of each of the seven gravity wave parameters. In each graph, the distribution corresponding to the MLE that minimizes the AIC is displayed in the title. The black curve represents the kernel density estimate of the raw data, while the red curve depicts the PDF associated with the MLE.

Having modeled the seven gravity wave parameters independently of one another, the AIC corresponding to this naïve joint probability distribution model was computed. The joint probability density for a given sounding was computed by calculating the probability density for each gravity wave parameter and multiplying these values together. The log-likelihood was then calculated by summing the logs of these joint probability densities across all soundings. Meanwhile, the number of parameters in the model was simply equal to the total number of parameters in the seven MLEs corresponding to the seven gravity wave parameters. The AIC came out to 140,093.

4.3 Accounting for Correlation

4.3.1 Copula

Figure 9 and Figure A.1 illustrate that the vast majority of scatterplots display a highly linear correlation between pairs of variables. In fact, the only scatterplots that don't possess this feature are the ones plotting frequency against zonal wavelength and frequency against meridional wavelength. As a result of this linear relationship, the joint probability distribution of all the gravity wave parameters except for frequency were able to be constructed using a copula.

The copula is a powerful statistical technique that allows for the marginal distributions and correlation structure of a joint distribution to be modeled separately. It is particularly useful because for the vast majority of joint distributions that are comprised of different marginal

distributions, there are no built-in functions to generate the desired multivariate distributions. For example, while there exists a closed-form expression for the joint distribution of correlated random variables that have marginal normal distributions (i.e. the multivariate normal distribution), there is no equivalent joint distribution for correlated random variables that have marginal distributions that are, say, gamma and beta distributions or Weibull and lognormal distributions. The copula addresses this deficiency by providing a framework to construct joint distributions of random variables that have arbitrary marginal distributions.

The best way to demonstrate how a copula works is through an example. For simplicity, the example deals only with modeling the joint distribution of meridional wind amplitude and vertical wavelength, but the same method can easily be extended to all of the other parameters except for frequency.

The first step is to construct a scatterplot of meridional wind amplitude and vertical wavelength to confirm the existence of a linear relationship between the two, and then calculate the corresponding correlation coefficient.

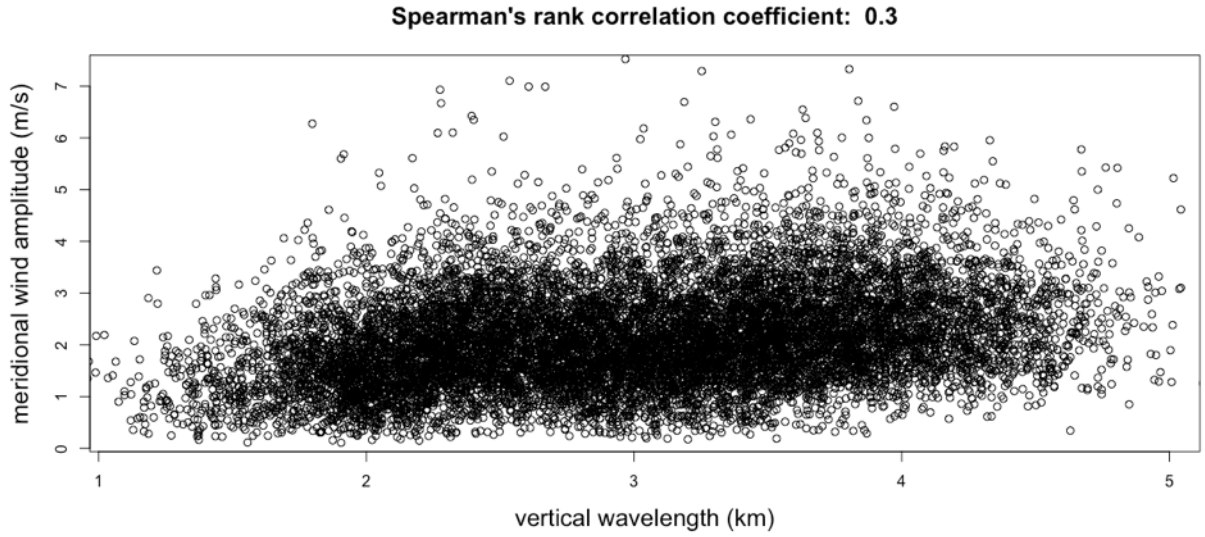


Figure 17: Scatterplot of vertical wavelength and meridional wind amplitude. The scatterplot indicates a clear linear correlation between the two parameters, and the corresponding Spearman's rank correlation coefficient is 0.30.

The next step is to construct a bivariate normal distribution with a mean vector of zeros and with a covariance matrix that uses the empirical correlation coefficient between the two parameters. In formal statistical notation, this joint distribution is given by:

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = N(\mu, \Sigma), \quad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

where Z_1, Z_2 are standard normal random variables, μ is the mean vector, Σ is the covariance matrix and ρ is the Spearman's rank correlation coefficient of 0.30 corresponding to the

relationship between meridional wind amplitude and vertical wavelength.

Generating 10,000 samples from this bivariate normal distribution yields the expected marginal distributions and correlation structure as illustrated in Figure 18.

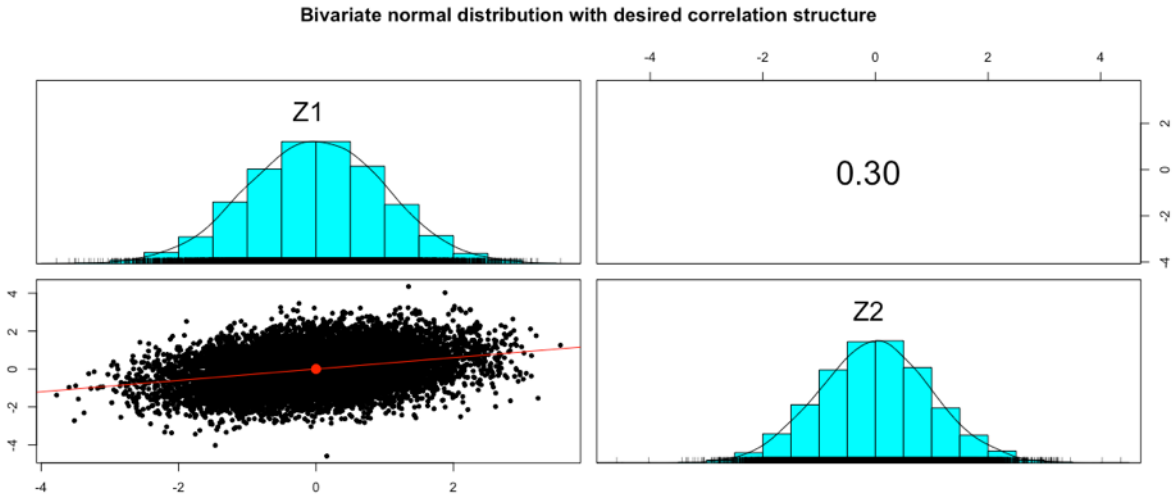


Figure 18: Visual overview of the bivariate normal distribution. The top-left and bottom-right panels illustrate that the marginal distributions of Z_1, Z_2 are standard normal, while the bottom-left and top-right panels indicate that the correlation between these two random variables is positive and equal to 0.30.

Probability theory states that plugging a random variable into its own distribution yields a standard uniform distribution. Therefore, the next step involves plugging each of the random variables Z_1, Z_2 into the cumulative distribution function (CDF) of the standard normal distribution, which transforms the standard normal-distributed Z_1, Z_2 random variables into standard uniform-distributed random variables, U_1, U_2 respectively. The key point is that even though the random variables U_1, U_2 are distributed according to standard uniform distributions, the correlation structure that was created by the multivariate normal distribution is not altered by plugging the normal random variables into their own CDF. This is illustrated below in Figure 19.

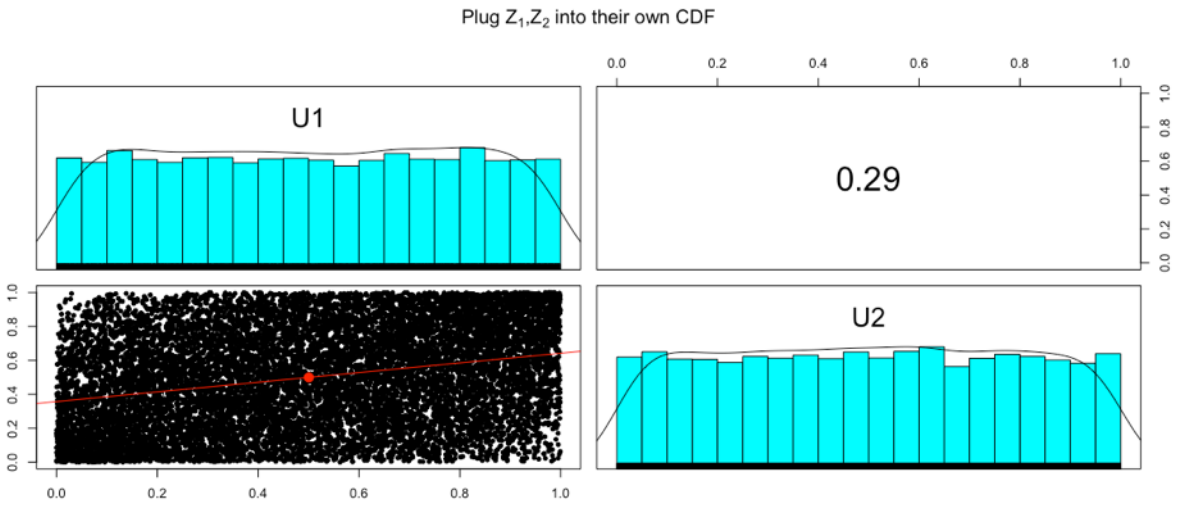


Figure 19: Visual overview of the distribution resulting from plugging Z_1, Z_2 into their own CDF to produce random variables U_1, U_2. The top-left and bottom-right panels illustrate that the marginal distribution of both U_1 and U_2 are standard uniform, while the bottom-left and top-right panels indicate that the correlation between these two random variables remains very close to the desired correlation of 0.30.

The final step exploits the fact that plugging a standard uniform distribution into the inverse CDF (F^{-1}) of an arbitrary probability distribution yields a random variable whose CDF is given by F. Therefore, the final step is to plug U_1 into the inverse CDF (also known as the quantile function) associated with the maximum likelihood estimate of the vertical wavelength distribution, and U_2 into the inverse CDF associated with the maximum likelihood estimate of the meridional wind amplitude distribution. Just as before, this operation leaves the correlation structure between the two random variables intact. As a result, this transformation produces a joint distribution that features the desired marginal distributions for each parameter and, most importantly, possesses the desired correlation structure between the two gravity wave parameters. This final joint distribution is illustrated below in Figure 20.

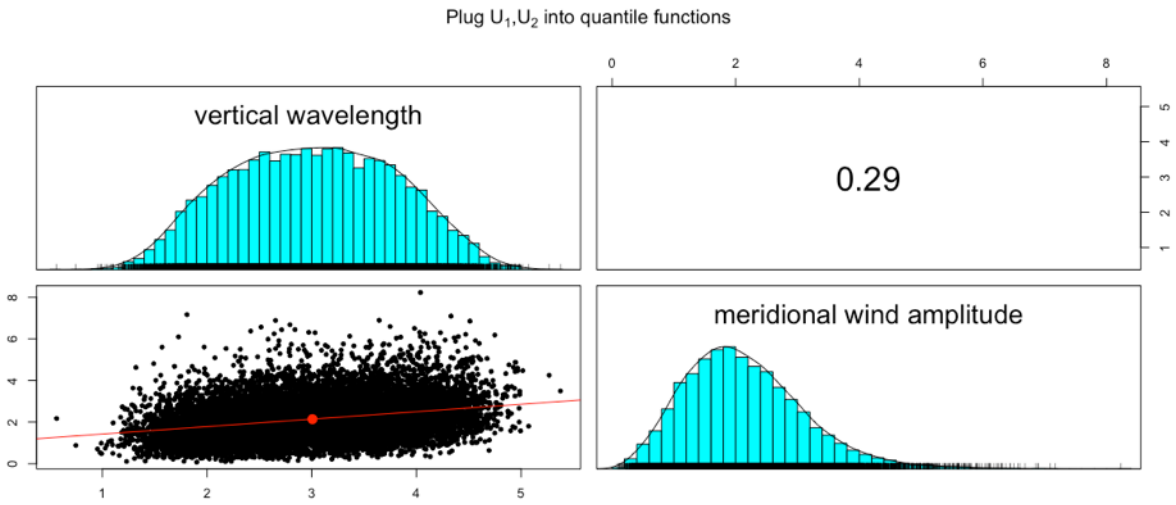


Figure 20: Visual overview of the joint distribution that results from plugging U_1 and U_2 into the quantile functions of the maximum likelihood estimates associated with the vertical wavelength and meridional wind amplitude, respectively. The top-left panel indicates that the vertical wavelength has a marginal generalized normal distribution, while the bottom-right panel indicates that the meridional wind amplitude has a marginal Burr distribution, as desired. Meanwhile, the bottom-left and top-right panels indicate that the correlation between these two variables is positive and is very close to the desired correlation of 0.30.

This explanation of the three primary steps involved in the copula method shows why there must exist a linear correlation between all pairs of random variables that are modeled using a copula. As discussed above, the way in which the correlation between two arbitrary random variables is introduced into the model is by setting the elements in the covariance matrix of the multivariate normal distribution that correspond to these random variables equal to the empirical correlation between these two random variables. However, because the multivariate normal distribution can only model linear correlation between random variables, the underlying correlation structure between two given random variables must be linear. This means that the highly nonlinear correlation between frequency and zonal wavelength and frequency and meridional wavelength displayed in Figure 9 could not be modeled using this copula method. As a result, a different approach was needed to incorporate frequency into the joint probability distribution.

As mentioned before, this overall framework can be applied to all the gravity wave parameters except for frequency to generate a six-dimensional joint probability distribution. Conveniently, the Copula package in R handles the entire process of creating the multivariate distribution and provides functionality to calculate the probability density and generate random samples from the multivariate distribution. The only input to the R functions are the marginal distributions associated with each parameter and the 6x6 correlation matrix specifying the correlation between all six parameters.

Because the marginal distributions had already been determined in the section on independent modeling, the only thing left to do was to optimize the correlation matrix to best capture the correlation structure of the raw data. While a reasonable first guess would be to populate the correlation matrix with the Spearman's rank correlation coefficients corresponding to each pair of parameters, there is no guarantee that these are the optimal values.

Therefore, a coordinate descent algorithm was run to find those correlation coefficients that minimize the AIC and thus best match the raw data. With a total of six parameters, this meant optimizing the values of the 15 correlation coefficients needed to completely specify the relationship between all the parameters. To run this coordinate descent algorithm, the values of the 15 correlation coefficients were initialized to the Spearman's rank correlation coefficient calculated from the data. Next, 14 of the 15 correlation values were held fixed while the remaining correlation value was adjusted in increments of 0.01 until the AIC associated with the multivariate distribution reached its minimum value. This individual step was then repeated for each of the other 14 correlation values. Finally, this process was repeated until none of the 15 parameters were further adjusted, signifying that the AIC had converged to a local minimum.

This coordinate descent algorithm, where one parameter is adjusted at a time while the others are held fixed, is standard practice in machine learning problems where there are multiple parameters to optimize and there is no closed-form objective function that could be differentiated to run a gradient descent algorithm. One unavoidable drawback of coordinate descent (or gradient descent, for that matter) is that it only guarantees convergence to a local minimum. This means that differences in the initializations of the parameters or the order in which they are adjusted can change the local minimum to which the algorithm converges. To assess the impact of these differences, the coordinate descent algorithm was run for various initializations and orderings of the parameters. While the final values of the parameters differed slightly based on the starting conditions, the maximum difference between the final AIC values was less than five, which indicated that neither the particular initialization nor the ordering of the parameters was of great consequence.

At this point, the joint probability distribution involved modeling all the gravity wave parameters except for frequency using a copula, and then modeling the frequency independently of the other six parameters. Overall, this step of applying the copula technique decreased the AIC from 140,093 (when all the parameters were modeled independently of each other) to 117,628. It's important to reemphasize that while the specific AIC values are meaningless, this significant reduction of 22,465 indicates that this modified model was much more consistent with the raw data.

4.3.2 Frequency Transformation

Now that the joint probability distribution had been completely specified for six of the gravity wave parameters, the next step was to take into account the correlation between frequency and the other six parameters. The first approach was to apply a transformation to the frequency values such that the correlation between these transformed frequency values

and the zonal/meridional wavelength would become linear. This, in turn, would justify the use of a seven-dimensional copula to model the joint probability distribution of all seven gravity wave parameters.

Given the strong inverse relationship between frequency and zonal/meridional wavelength as well as the fact that period is inversely proportional to frequency, the most logical transformation to apply to the frequency values was an inversion (i.e. $f(x) = \frac{1}{x}$). Having transformed frequency to period, the seven-dimensional copula could then be constucted. The steps involved in this process are outlined below:

1. Model the marginal distribution of the period values by applying the maximum likelihood estimation methodology described in Section 4.2. The MLE distribution that corresponded to the lowest AIC was the lognormal distribution, as shown in Figure 21.

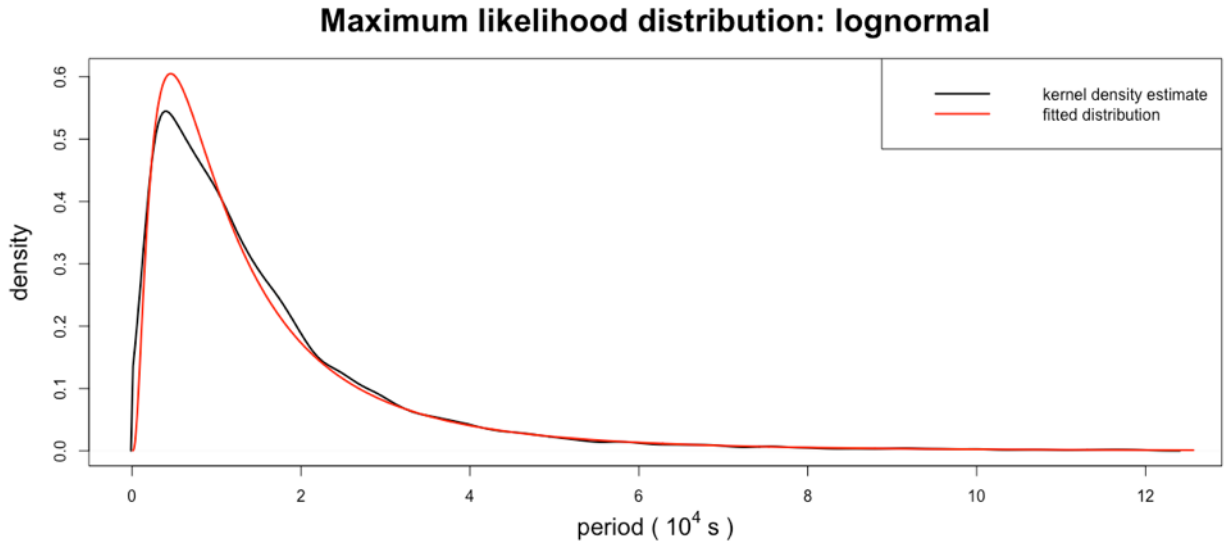
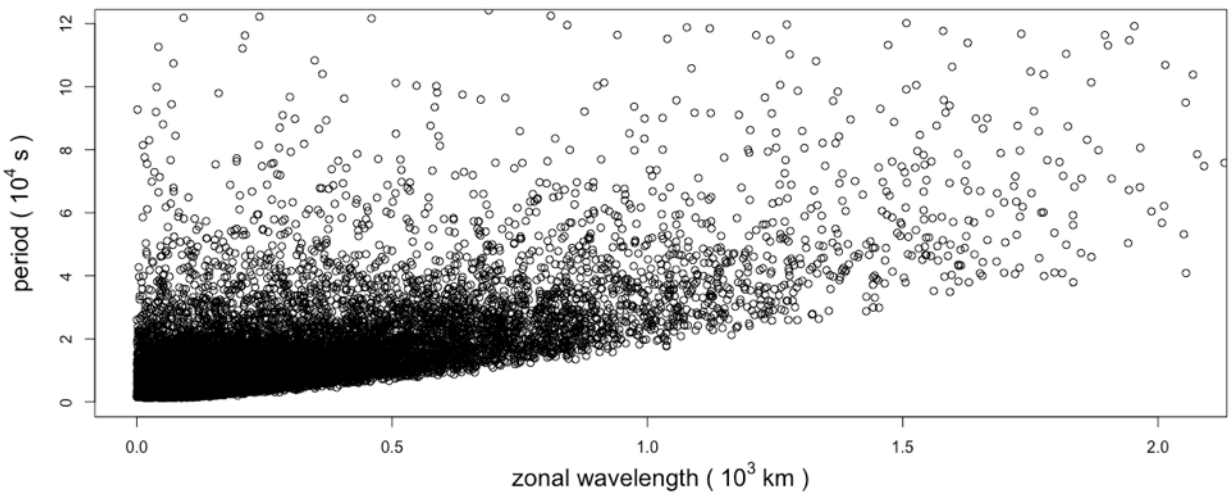
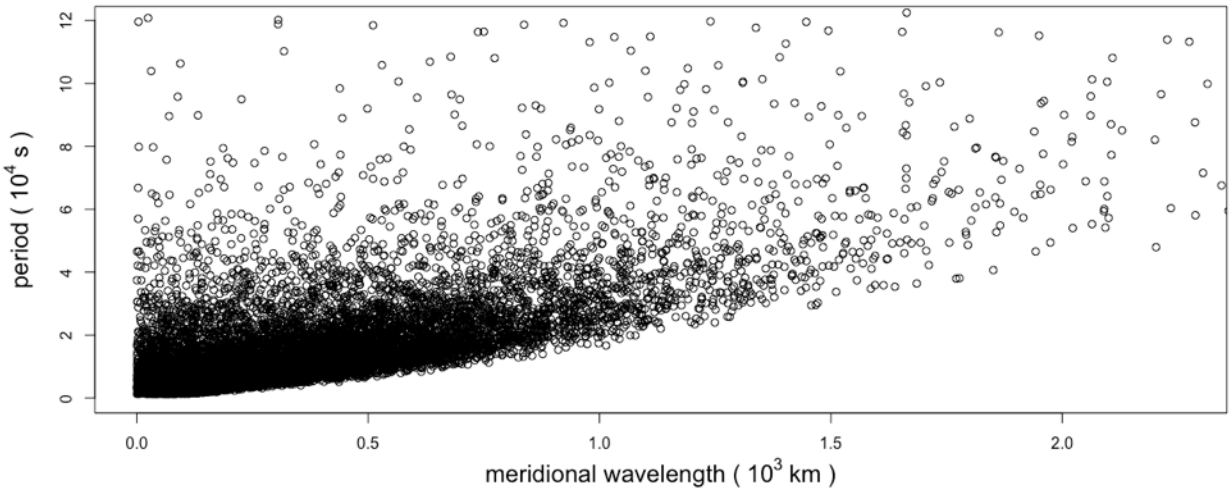


Figure 21: Maximum likelihood estimate associated with the marginal period distribution. The black curve represents the kernel density estimate of the raw data, while the red curve depicts the PDF associated with the MLE.

2. Confirm that the correlation between period and zonal/meridional wavelength is linear. This is done by inspecting the scatterplots displayed in Figure 22.



(a)



(b)

Figure 22: Scatterplots illustrating a linear relationship between period and zonal/meridional wavelength.

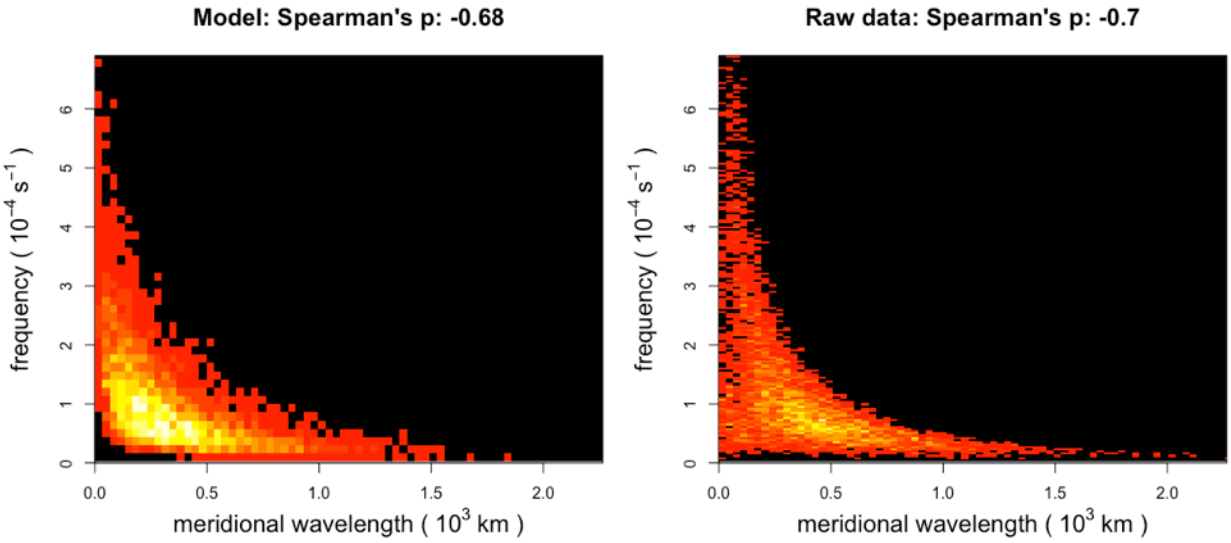
3. Construct a seven-dimensional copula and run the coordinate descent algorithm described in the previous section to optimize the 7x7 correlation matrix.

4. Invert the period values to transform them back to frequency.

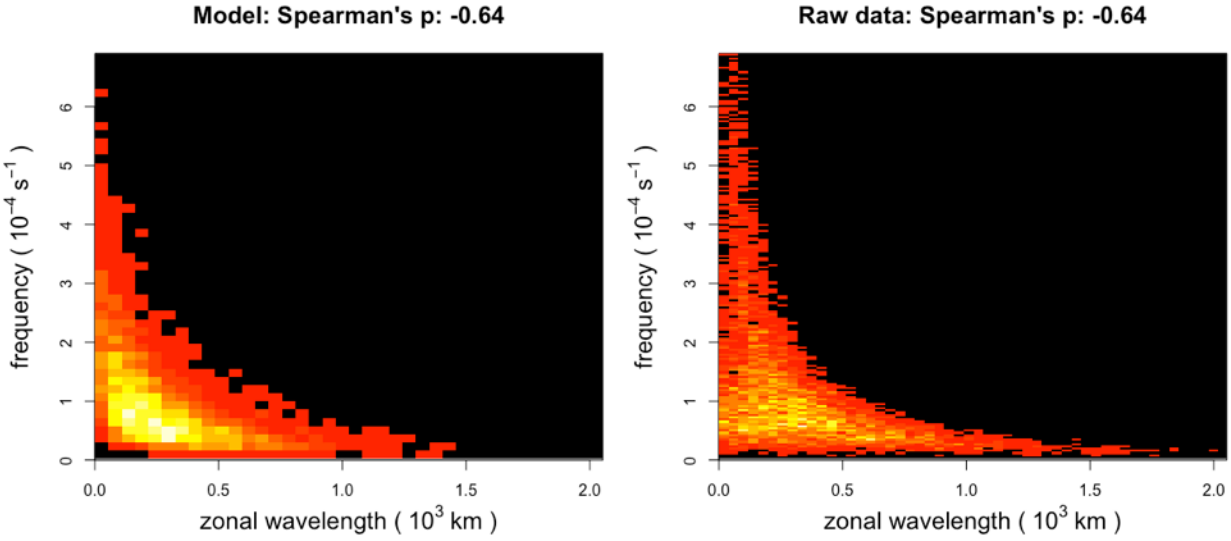
Modeling the joint probability distribution of the seven gravity wave parameters using this seven-dimensional copula reduced the AIC from 117,628 (the AIC after using a copula to model the six parameters) to 101,311. However, while this substantial reduction in the AIC

indicated that the model was significantly improved, there were still several major shortcomings to the model.

First, this seven-dimensional copula failed to completely capture the inverse relationship between frequency and zonal/meridional wavelength. This is illustrated in Figure 23, which compares the empirical correlation between these parameters to the corresponding correlation outputted by the model.



(a)



(b)

Figure 23: 2D histograms of zonal wavelength versus frequency and meridional wavelength versus frequency associated with the raw data (right) and the sample generated from the seven-dimensional copula model (left). The plots illustrate that the model failed to completely capture the correlation structure between frequency and zonal/meridional wavelength.

Second, while the transformation applied to the frequency values linearized the relationship with the zonal/meridional wavelength, it inadvertently detracted from the linear relationship with several of the other parameters. One example of this deviation from linearity is presented below in Figure 24.

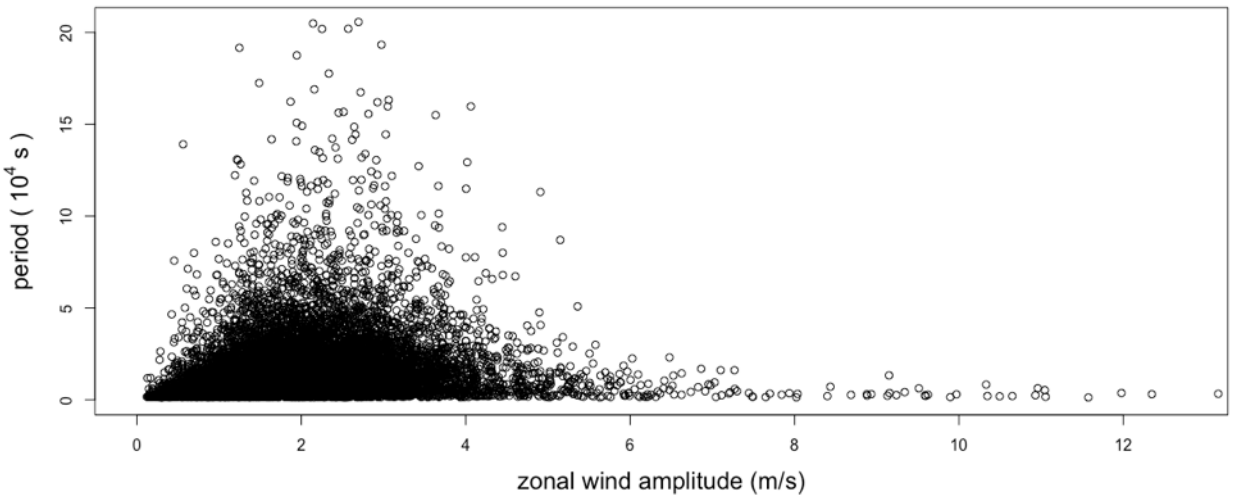


Figure 24: Scatterplot of zonal wind amplitude versus period. The fact that this scatterplot forms a triangular pattern illustrates that the relationship between these two parameters is not completely linear and thus cannot be thoroughly modeled using a copula.

Despite these drawbacks, this seven-dimensional copula was considered to be the new benchmark joint probability distribution model because it best modeled the data up to this point. However, this model was clearly not complete. The next step was to develop a more complex model that would better capture the correlation between frequency and zonal/meridional wavelength while preserving the linear relationship between frequency and the other four parameters.

4.3.3 Conditional Frequency Distribution

Before proceeding to discuss how the correlation between frequency and the other six parameters was ultimately accounted for in the model, it is important to discuss how the use of a copula modified the overall view of the data generating process. Before beginning the modeling, the data generating process was imagined as a one-step process where all seven gravity wave parameters are generated simultaneously from some joint probability distribution. However, once it became clear that frequency would need to be modeled separately, the data-generating process was envisioned as a two-step procedure: first the values of all the gravity wave parameters except for frequency are generated using a copula, and then the frequency value is generated based on the values of the parameters that had just been generated. Therefore, the final step was to model the conditional probability distribution of frequency given the six other gravity wave parameters.

Ideally, the correlation between frequency and each of the other six parameters could be incorporated into the model. However, it proved to be extremely difficult, if not impossible, to model the nonlinear correlation between frequency and zonal/meridional wavelength together with the linear correlation between frequency and the other four parameters. Fortunately,

these linear correlations were relatively weak, as the Spearman's rank correlation coefficient between frequency and vertical wavelength, zonal wind amplitude, meridional wind amplitude and temperature amplitude was -0.06, -0.24, -0.2 and 0.04, respectively. Therefore, it seemed like a reasonable simplification to treat frequency as independent of these four parameters and focus on capturing the strong inverse relationship between frequency and the zonal and meridional wavelength.

The steps involved in modeling the conditional distribution of frequency given zonal and meridional wavelength are outlined below:

1. Split the zonal wavelength values into a certain number of intervals, where each interval contains the same number of data points. For example, a total of 10 intervals would mean placing all zonal wavelength values from the minimum value to the 0.1 quantile in one interval, all the data between the 0.1 quantile and the 0.2 quantile in the second interval, etc.

2. For each interval, extract the frequency values that correspond to zonal wavelength values that lie within the given interval.

3. Calculate the maximum likelihood estimate for the frequency values corresponding to each interval using the same procedure discussed in the section on independent modeling. (This entails determining the maximum likelihood estimate associated with each candidate distribution and then selecting the candidate distribution that minimizes the AIC.)

4. Model the frequency by observing which interval a given zonal wavelength value falls into and then predicting the corresponding frequency value using the associated MLE distribution.

5. Use cross-validation to determine the optimal number of intervals. There are two competing factors at play when it comes to the optimal number of intervals: on the one hand, the more intervals are created, the more the model can capture the structure of the data. On the other hand, as the number of intervals increases, the risk of overfitting grows, which means the model may not generalize to new data. Taken to an extreme, if the number of intervals were to equal the total number of zonal wavelength data points (such that each zonal wavelength value would be placed into its own interval), the model would perfectly predict each frequency value (i.e. the maximum likelihood estimate would be a Dirac delta function), but the model would fail to generalize to new data. The steps involved in this 10-fold cross-validation procedure are outlined below [1] [2]:

[1] Because the following procedure is independent of the different choices for the number of intervals, this entire task was parallelized such that multiple interval values could be tested simultaneously.

[2] The question of how to set k in k-fold cross-validation is notoriously difficult to answer quantitatively. According to many sources, the use of 10 folds is a solid rule of thumb in a situation like this where the dataset consists of ∼12,500 radiosonde soundings.

(a) Create a list containing every integer between 10 and 200 in increments of 5. These numbers (denoted by n) represent the number of intervals to test. The ultimate goal is to determine which value of n is the best choice, which involves repeating the following steps for each value of n:

   i. Randomly split the seven-dimensional dataset containing all the gravity wave parameters into 10 equally sized chunks.
   ii. Train the model on 9 of the 10 chunks of data. This involves determining the MLE of $f(frequency \mid zonal\ wavelength)$ for each of the n intervals.
   iii. Test the model that was just created by calculating the joint log likelihood associated with the last of the 10 chunks of data. (To clarify, the likelihood of a single data point is calculated by identifying the interval that the zonal wavelength value falls into and then calculating the density of the frequency value using the MLE distribution corresponding to that interval.)
   iv. Repeat this process 10 times by holding out a different chunk of data, training the model on the remaining 9 chunks and testing the model on this held-out chunk.
   v. Sum the log likelihoods from each of these 10 cross-validation folds.
   vi. Add up all the parameters included in the model (this number is directly proportional to the total number of intervals)
   vii. Compute the AIC using the total log likelihood and the total number of model parameters.

(b) Create a plot of AIC as a function of the number of intervals and fit a 10^{th} degree polynomial regression curve through the data. (A 10^{th} degree polynomial was empirically determined to optimally fit the data and avoid both underfitting or overfitting. It's important to note that the selection of a different degree polynomial would have little, if any, impact on the minimum value of the curve.)

(c) Determine which value for the number of intervals corresponds to the lowest AIC value, and declare that value to be the optimal number of intervals.

As illustrated below in Figure 25, the regression curve was strictly convex. This makes sense because initially, as the number of intervals increased, the penalty associated with an increase in the number of parameters was overcome by a larger increase in the joint log likelihood. However, at a certain point, the penalty incurred from the increase in the number of parameters coupled with the onset of overfitting, where the model built on the training set did not generalize to the data held out in the testing set, led to a rise in the AIC.

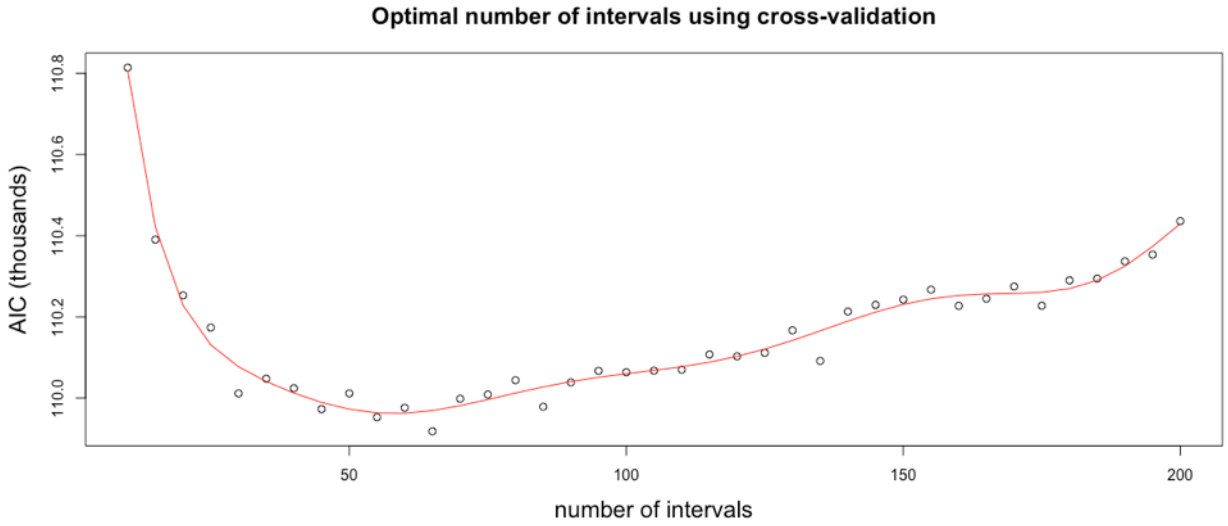


Figure 25: Plot of AIC as a function of the number of intervals along with a 10th degree polynomial regression curve. The number of intervals corresponding to the minimum AIC value (in this case 60) was chosen as the optimal number of intervals to use in the model.

Modeling frequency by conditioning on the zonal wavelength in this manner reduced the AIC from 117,628 (the AIC after using a copula to model the six parameters) to 107,651. This entire process was then repeated for modeling the conditional distribution of frequency given meridional wavelength. This corresponded to a reduction in the AIC from 107,651 to 104,891.

The next step was to take into account the information provided by both wavelengths by modeling the conditional distribution of frequency given both zonal and meridional wavelength. To accomplish this, it was helpful to compare side by side the 2D histogram of frequency versus meridional wavelength and the 2D histogram of frequency versus zonal wavelength, as illustrated below in Figure 26. These graphs indicate that the plot of frequency versus meridional wavelength matches up almost perfectly with the plot of frequency versus zonal wavelength (which makes sense due to the symmetrical relationship between zonal wavelength and meridional wavelength). Moreover, the graphs show a monotonically decreasing relationship between frequency and both wavelengths in the sense that as the value of zonal/meridional wavelength increases, the range of corresponding frequency values either shrinks toward zero or remains the same. This means that an increase in the zonal/meridional wavelength value is accompanied by an increase in confidence regarding the corresponding frequency value. For example, if the zonal/meridional wavelength is $0.1 \times 10^3\, km$, the corresponding frequency value is likely to fall anywhere between zero and $7.6 \times 10^{-4} s^{-1}$, whereas if the zonal/meridional wavelength is $1.0 \times 10^3\, km$, the corresponding frequency value is guaranteed to be less than $0.7 \times 10^{-4}\, s^{-1}$, and if the zonal/meridional wavelength is $2.0 \times 10^3\, km$, the corresponding frequency is guaranteed to be less than $0.3 \times 10^{-4}\ s^{-1}$.

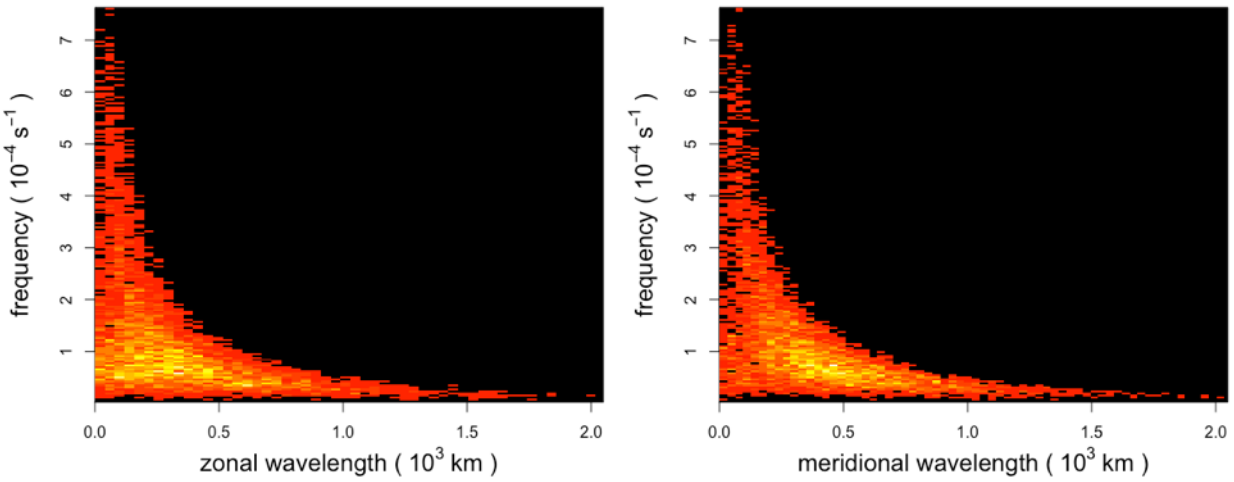


Figure 26: 2D histograms of frequency versus zonal wavelength and frequency versus meridional wavelength. The brighter the color at a given location, the higher the density of data points at that location. This fact that the two 2D histograms show such as strong resemblance indicates that the relationship between frequency and zonal wavelength is almost identical to the relationship between frequency and meridional wavelength. The graph also shows that as the value of the zonal/meridional wavelength increases, the range of possible frequency values decreases or remains the same.

These two pieces of information were then combined to come up with a simple method to model the conditional distribution of frequency given both zonal and meridional wavelength. The model works as follows: given a value of zonal wavelength and meridional wavelength, the frequency is modeled as the conditional probability given the larger of the two wavelength values. For example, if the zonal wavelength value is 0.5×10^3 km and the meridional wavelength value is 1.0×10^3 km, the frequency would be modeled as $f(frequency \,|\, meridional\, fwavelength = 1.0 \times 10^3\, km)$, which has already been determined by looking at which interval the meridional wavelength falls into and using the corresponding MLE distribution to model the frequency. If on the other hand, the zonal wavelength value is 1.0×10^3 km while the meridional wavelength is 0.5×10^3 km, the frequency would be modeled as $f(frequency \,|\, zonal\, wavelength = 1.0 \times 10^3\, km)$. Mathematically, this model is guaranteed to reduce the AIC because as the zonal/meridional wavelength value increases, the support of possible frequency values decreases, which means that the average density increases such that the area under the density curve remains one. This higher average density, in turn, boosts the log-likelihood, which, in turn, produces a lower AIC. Overall, this strategy for modeling the conditional distribution of frequency is particularly advantageous because it takes into account both wavelength values, yet ultimately results in the simpler task of conditioning frequency on just one of the two wavelength values.

The cross validation procedure outlined above was repeated to determine the optimal number of zonal and meridional wavelength intervals for this revised model that conditions frequency

on whichever of zonal and meridional wavelength is larger. As shown below in Figure 27, the optimal number of intervals again came out to 60. All in all, this revised model reduced the AIC from 104,891 (the AIC after modeling frequency as the conditional distribution given meridional wavelength) to 98,748.

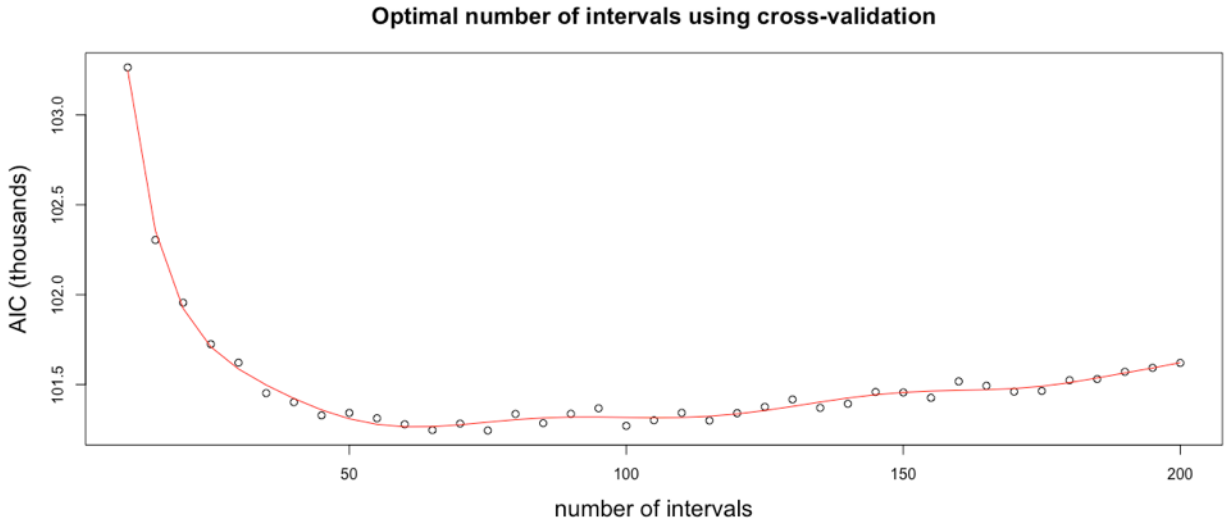
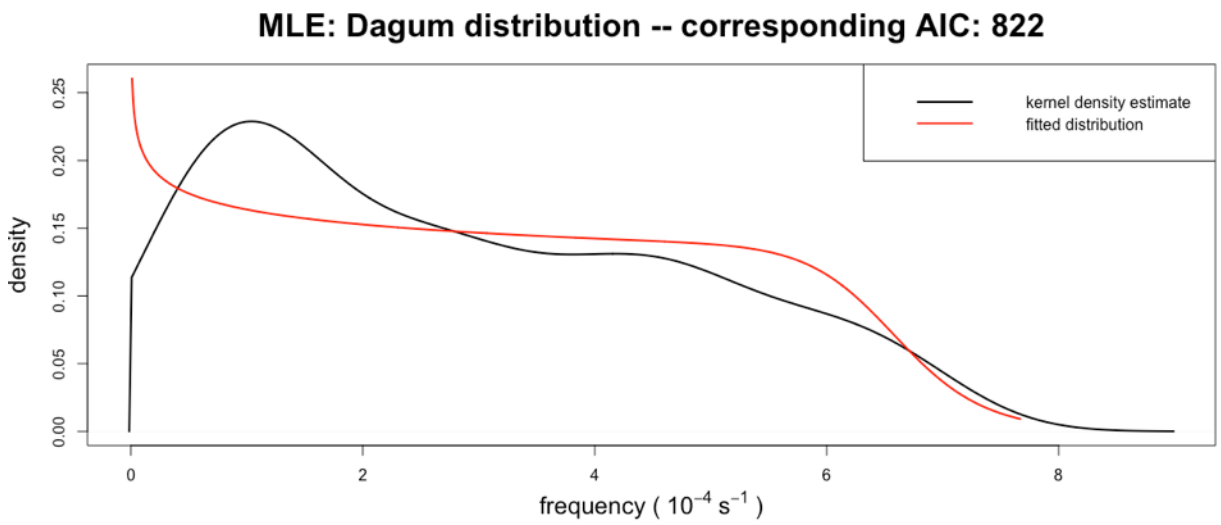


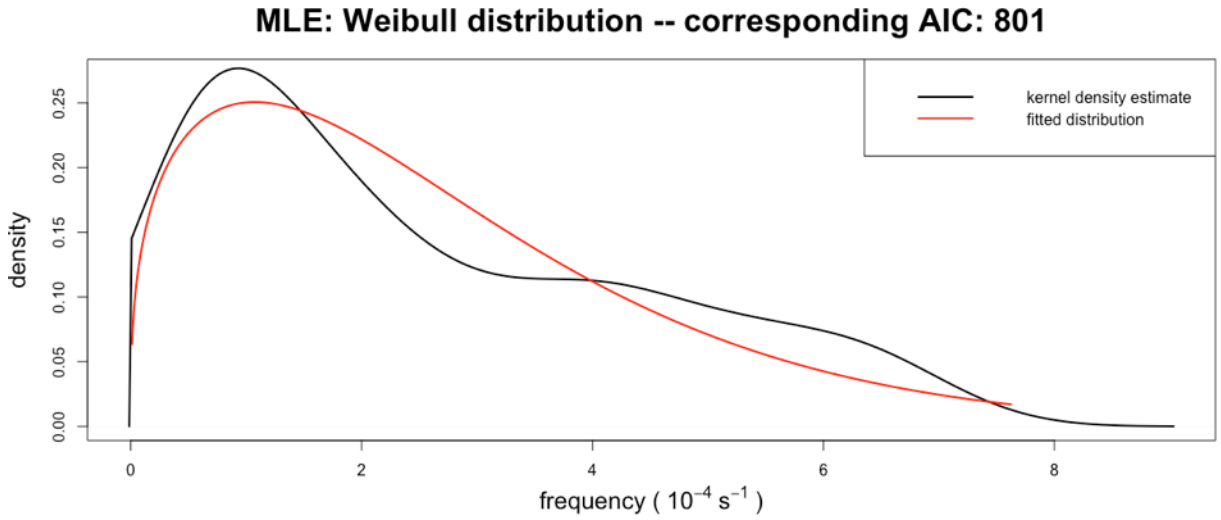
Figure 27: Plot of AIC as a function of the number of intervals along with a 10th degree polynomial regression curve. The number of intervals corresponding to the minimum AIC again turned out to be 60.

4.4 Improving the Maximum Likelihood Estimates

Now that the method for modeling the conditional frequency distribution had been fleshed out, the next step was to improve the quality of the MLE distributions corresponding to each zonal/meridional wavelength interval. When it came to modeling the marginal distributions of the seven gravity wave parameters, the kernel density estimates were extremely smooth (as illustrated in Figure 8), which meant it was sufficient to fit a single distribution to the data. However, because the number of frequency values contained in an interval is just one-sixtieth of the total number of frequency values, the kernel density estimates were much more jagged, which meant the data was much more difficult to model using a single distribution. Figure 28 shows two examples of how the MLE distributions fail to adequately model the conditional distribution of frequency.



(a)



(b)

Figure 28: Two examples comparing the kernel density estimate (represented in black) with the corresponding MLE distribution (represented in red). The top and bottom graphs model the frequency values that correspond to zonal wavelength values that fall in the interval between 0.071 and 0.078 and the interval between 0.057 and 0.064, respectively. Both graphs illustrate that the MLE distributions fail to capture all the structure in the raw data.

4.4.1 Mixture Models

The first approach to improving the quality of the MLE distributions was to supplement the list of candidate distributions (gamma, Weibull, lognormal, Gumbel, Burr, Dagum, inverse gamma, Rayleigh and generalized Rayleigh distributions) with various mixture models.

A K-component mixture model consists of a K-component categorical distribution and K sets of parameters. Each set of parameters specifies the parameters of the corresponding mixture component, with all the components belonging to the same parametric family of distributions (i.e. all gamma or Weibull distributions). This K-component mixture model then views the data generating process as a two-step procedure: first, select which of the K mixture components to sample from according to the K-component categorical distribution and then sample from the corresponding mixture component.

While the concept of a mixture model is relatively straightforward, the process of determining the maximum likelihood estimate of all the parameters in a mixture model is much more difficult than it is for a single distribution, where the MLE either has a closed-form expression or can be solved for using a simple iterative algorithm. This is the case because a mixture model assumes that each observed data point has a corresponding unobserved data point, or latent variable, that specifies the mixture component to which each data point belongs. This means that in order to construct an MLE for a mixture model, these latent variables must also be inferred.

Fortunately, an iterative method, known as the expectation-maximization (EM) algorithm exists to find the maximum likelihood estimates of the parameters in a mixture model. These parameters include the weights of the categorical distribution as well as the parameters corresponding to each mixture component. However, because the EM algorithm is both statistically and computationally intensive, it is difficult to manually implement for mixture models of desired distributions. Therefore, it was necessary to rely on the functionality provided by the MixR package in R, which performs maximum likelihood estimation for finite-dimensional mixture models for the normal, Weibull, gamma and lognormal families. Because the gravity wave parameters can only take on positive values, only the gamma, lognormal and Weibull mixture models were considered.

One issue with mixture models is that the number of components needs to specified beforehand. In order words, it is not possible to feed the data into a mixture model and get back the number of components that optimally fit the data. To get around this issue, the following approach was used:

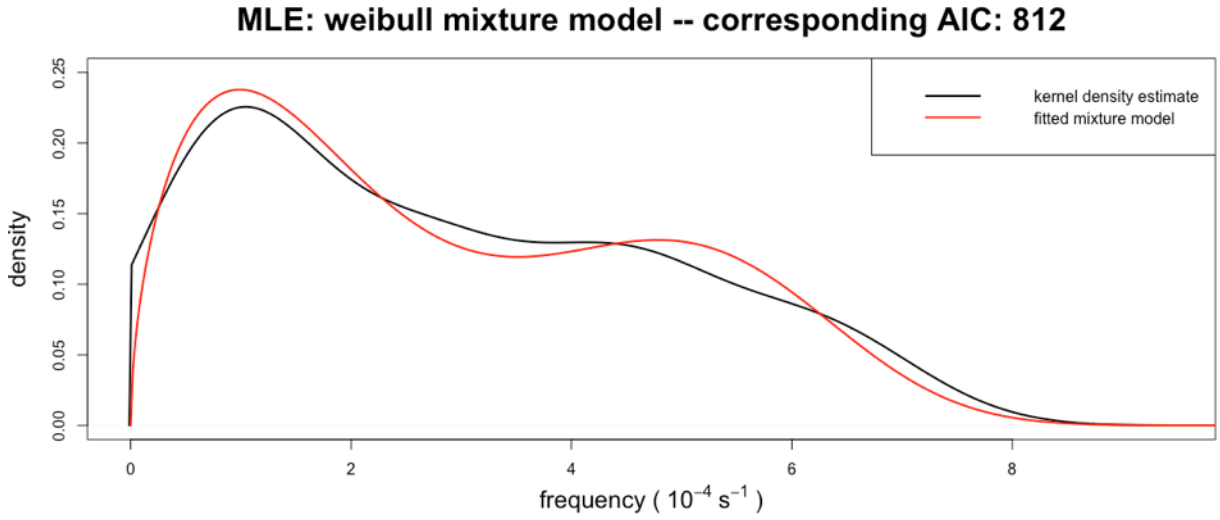
1. Begin with one mixture component, learn the corresponding mixture model and compute the corresponding AIC.

2. Continually increment the number of mixture components by one until the AIC increases for the first time. (Denote the number of components in this mixture model by j.)

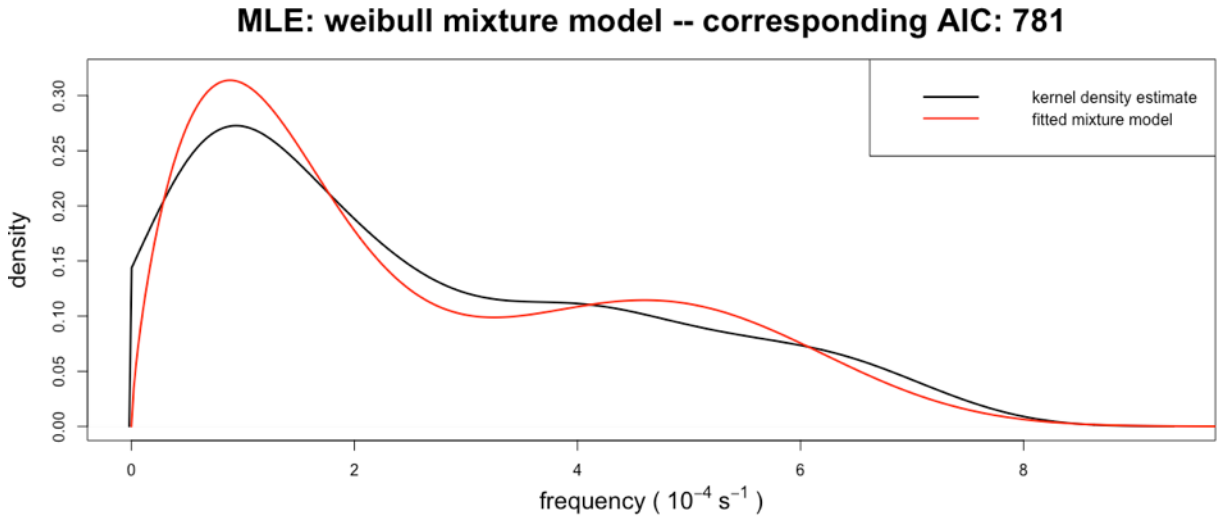
3. Declare the previous mixture model with $j - 1$ components to be the optimal one.

Adding these three mixture models to the list of candidate distributions caused many of the MLE distributions corresponding to the various intervals of zonal/meridional wavelength to

switch from single distributions to mixture models. Two such examples are illustrated below in Figure 29.



(a)



(b)

Figure 29: Two examples comparing the kernel density estimate (represented in black) with the corresponding MLE distribution (represented in red). These are the same examples shown in Figure 28, except now the maximum likelihood estimates are both Weibull mixture models as opposed to Dagum and Weibull distributions, respectively. It can clearly be seen from the red curves that the Weibull mixture models in the top and bottom graphs each have two components. The switch from an individual distribution to the mixture model corresponds to a respective reduction in the AIC of 10 and 20 for the two conditional frequency distributions.

Although it initially appeared that the addition of these mixture models would reduce the AIC, it actually increased it slightly from 98,748 to 98,866. This rise in the AIC means that the increase in the log-likelihood wasn't larger enough to overcome the penalty incurred from the dramatic increase in the number of parameters associated with the mixture models.

4.4.2 Polynomial Regression

The failure of the mixture model approach signaled that the complex conditional frequency distributions could not be adequately modeled using conventional probability distributions. This led to a shift in mentality away from parametric statistics and opened the door to more creative and unconventional approaches to modeling the conditional frequency distributions.

Before conceiving of a new approach, it was important to specify the two main requirements that any potential model would have to fulfill. First, the model needed to be able to calculate the probability density for each frequency value. This was necessary since it would enable the calculation of the log-likelihood and, as a result, the AIC associated with the model. Second, the model needed to be able to generate random samples from the distribution. This was essential because, as will be discussed later, the SCoPEx project team will utilize the final model by generating and analyzing random samples from the joint probability distribution.

Ultimately, a method was developed that uses a polynomial fit to model the probability density function. The steps involved in this procedure are outlined below:

1. Construct the adjusted kernel density estimate for the frequency values corresponding to a given zonal/meridional wavelength interval.

2. Perform a hundredth-degree polynomial regression on the adjusted kernel density estimate. The exact degree of the polynomial was arbitrary; the main specification was that the polynomial regression curve needed to fit the adjusted kernel density estimate so well that if plotted on the same graph, the two curves would be virtually indistinguishable from each other. It was empirically determined that a hundredth-degree polynomial regression curve was able to achieve this high level of accuracy for all the intervals.

3. Consider this polynomial fit to be the probability density function of the conditional frequency (denote PDF by $f(x)$). It's important to note that there is no danger of overfitting in this scenario because the kernel density estimate is a smoothing function and thus prevents any sharp spikes or dips in the PDF that are indicative of overfitting. Because fitting a kernel density estimate is immune to overfitting, the polynomial was not factored into the calculation of the total number of parameters contained in the model. This nonparametric modeling technique contrasts strongly with the parametric aspects of the model where each additional parameter increases the risk of overfitting and must therefore be penalized.

4. Locate the frequency value at which the polynomial curve drops below zero on the right side of the PDF (denote this value by X_{max}).

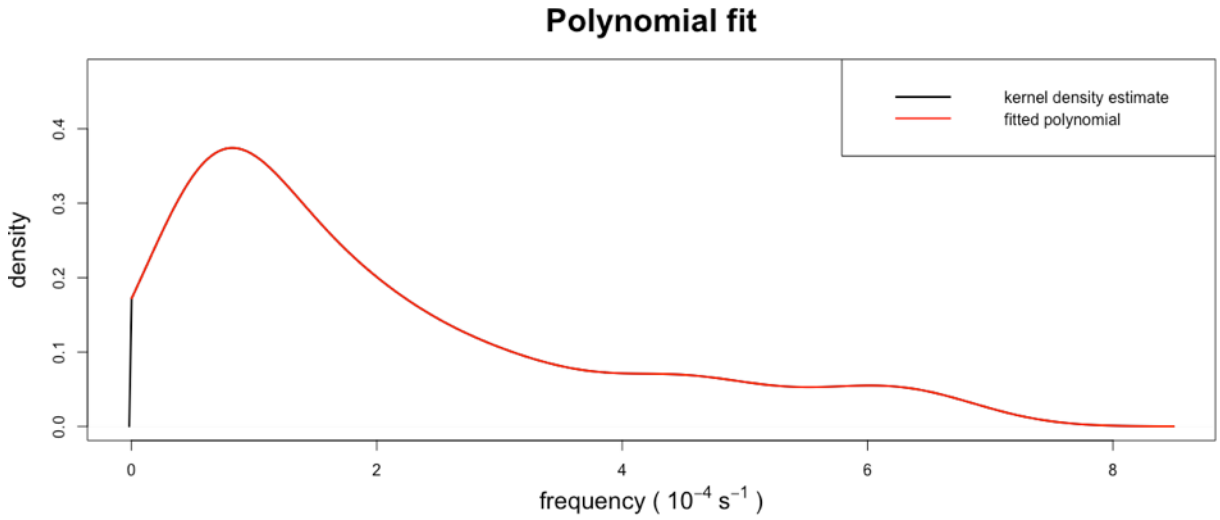
5. Locate the frequency value at which the polynomial curve drops below zero on the left side of the PDF (denote this value by X_{min}).

6. Calculate the positive area under f(x) by computing $F(X_{max}) = \int_{X_{min}}^{X_{max}} f(t)dt$. This function $F(X_{max})$ represents the cumulative distribution function evaluated at X_{max}. This value is usually within 0.0001 of 1, but is never exactly equal to 1.

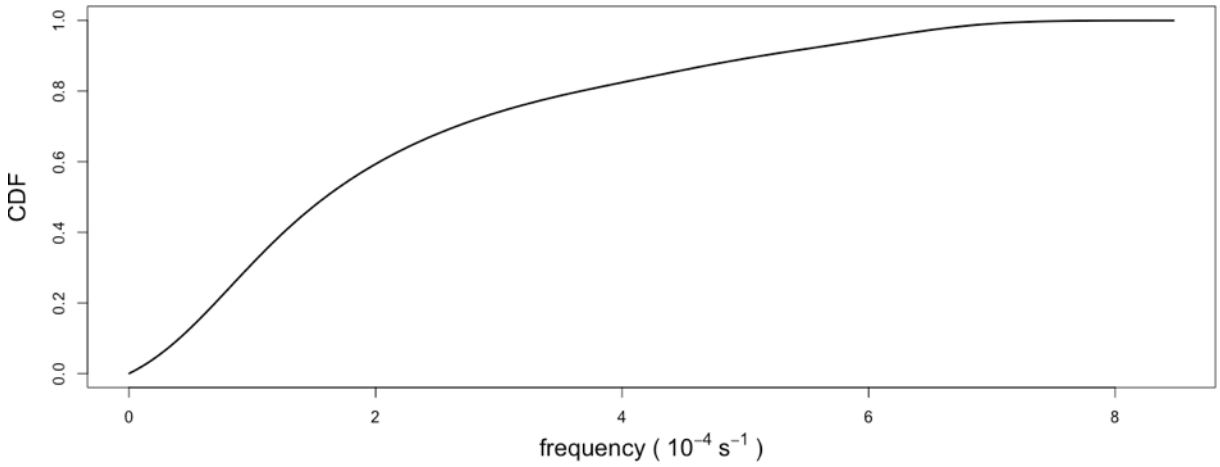
7. Divide $f(x)$ by this area to ensure that the positive area under the curve is exactly equal to one.

8. Revise the CDF, $F(x)$, by integrating this modified expression for the PDF from zero to x.

A typical polynomial PDF and corresponding CDF is reproduced below in Figure 30:



(a)



(b)

Figure 30: A polynomial regression PDF (a) and corresponding CDF (b). The top graph shows that the polynomial regression curve (red) overlaps almost perfectly with the kernel density estimate (black), while the bottom graph illustrates that integrating this PDF yields a valid CDF that is 0 for all frequency values less than 0 and approaches 1 as the frequency increases.

These polynomial estimations of the PDF and CDF of the conditional frequency distributions could then be used to calculate densities and generate random samples. Evaluating the density of a given frequency value could be achieved by simply plugging the frequency value into the PDF, $f(x)$. Meanwhile, generating a random sample from the conditional frequency distribution could be accomplished by generating a uniform random variable and, in theory,

plugging that random variable into the inverse cumulative distribution function, $F^{-1}(x)$. However, in practice, it was cumbersome to derive and work with the inverse of the CDF, which is a 101^{st} degree polynomial. Thus, an iterative algorithm was used instead to perform this calculation. The steps of this algorithm are detailed below:

1. Generate a random sample of size 1 (denoted by U) from the standard uniform distribution.
2. Label X_{min} and X_{max} as the lower and upper bounds, respectively, on the corresponding frequency value.
3. Initialize the guess of the frequency value (denoted by X) to the midpoint of X_{min} and X_{max}
4. Repeat the following process until convergence:
   (a) Compute $Y = F(X) - F(X_{min})$
   (b) Declare X to equal $F^{-1}(U)$ if $|Y - U| < 1 \times 10^{-4}$ (this value was chosen as the threshold because the frequency values do not need to be accurate beyond four decimal places).
   (c) Otherwise:
      i. Set the upper bound to X and set X equal to $\frac{X+Bound_{lower}}{2}$ if $Y > U$
      ii. Set the lower bound to X and set X equal to $\frac{X+Bound_{upper}}{2}$ if $Y < U$

Overall, this polynomial regression method reduced the AIC from 98,748 to 98,159. At first, this may not seem like a major improvement compared to the other modifications made to the model. However, it is important to note that the extent to which refining the conditional frequency modeling could reduce the AIC is much lower than it is for the other improvements that better accounted for the correlation between the gravity wave parameters.

Taking a step back, this polynomial regression method provides a concrete example of how the final model is selected. It was stated in the introduction that a model would be considered final only once it captured both the marginal and correlation structure of all the gravity wave parameters, and no more obvious steps could be taken to improve the model. The fact that the polynomial regression reduced the AIC by only 589 signifies that the model would not be significantly worse if a single probability distribution were used to model each conditional frequency interval. However, even though such a model would strongly capture the marginal distributions and correlation structures between the seven gravity wave parameters, it would not be considered complete because one obvious step to improve the model would be to enhance the degree to which the conditional frequency distributions align with the raw data. Therefore, this step would have to be done at some point before the model could be considered complete.

4.5 Comparison of Model to Raw Data

Now that as much of the correlation as possible had been incorporated into the model, the next step was to visualize how the model actually fits the data and identify areas for improvement. Up until this point, the AIC had been used to assess the relative quality of the models, but while a decreasing AIC suggests a better model, it provides no information about the absolute quality of the model.

In order to visually compare the model to the raw data, it was necessary to generate a random sample from the joint probability distribution. This was a two-step process.

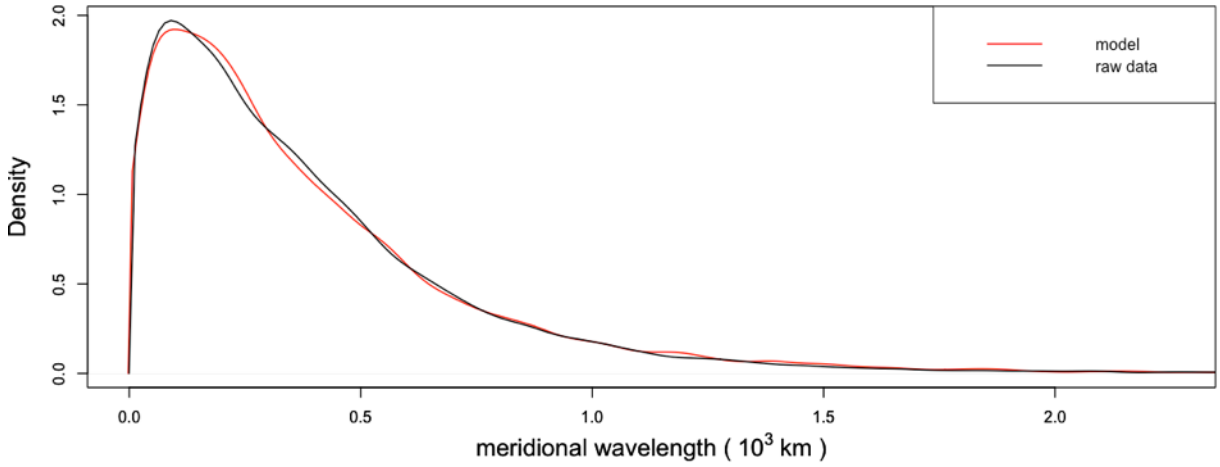
1. Generate a sample of size n from the copula, where n is the total number of radiosonde soundings used in the model. Carrying this out is trivial since, as mentioned previously, R provides functionality to generate random samples from a copula object. This step produces values for all the gravity wave parameters except for frequency.

2. Do the following for each of the n individual samples:

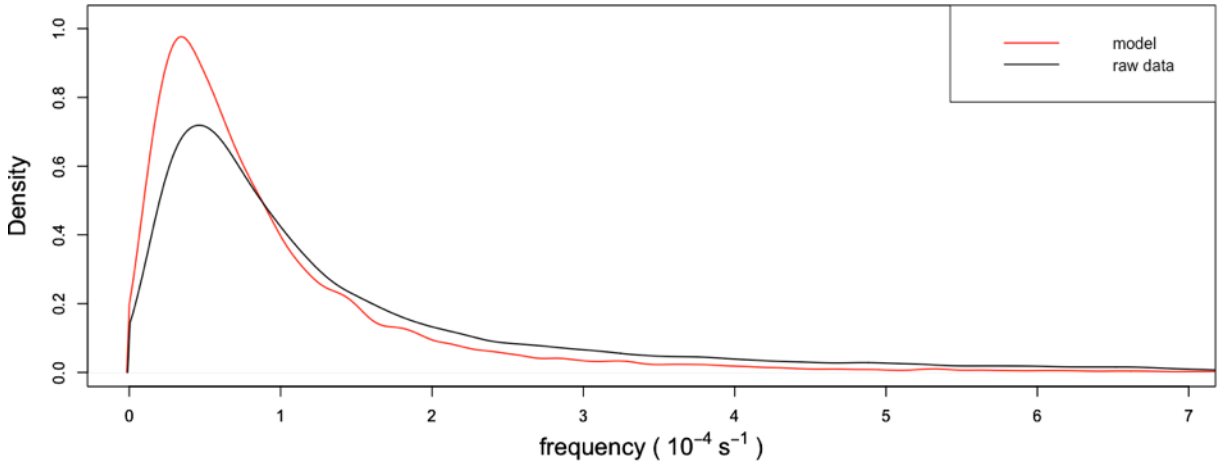
   (a) Identify the larger value among zonal wavelength and meridional wavelength.

   (b) Generate a sample of size one from the conditional frequency distribution corresponding to the interval that this larger wavelength value belongs to.

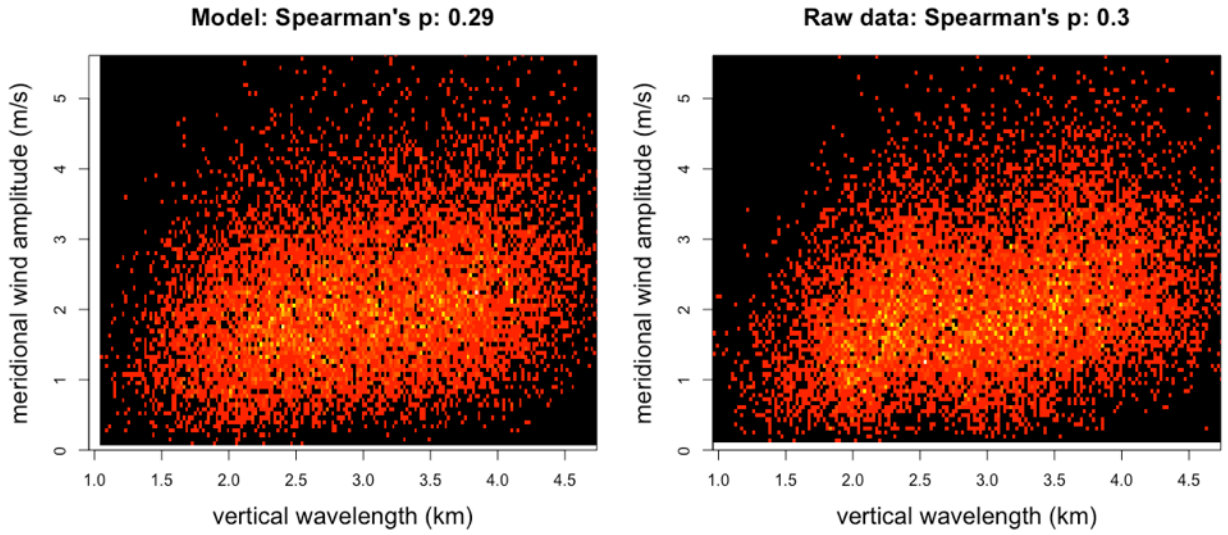
Two of the seven marginal distributions and three of the 21 two-dimensional histograms needed to fully characterize the joint probability distribution are displayed below in Figure 31. The rest of the graphs are presented in Figure A.7.


model
raw data
Density
0.0
0.5
1.0
1.5
2.0
meridional wavelength (10^3 km)

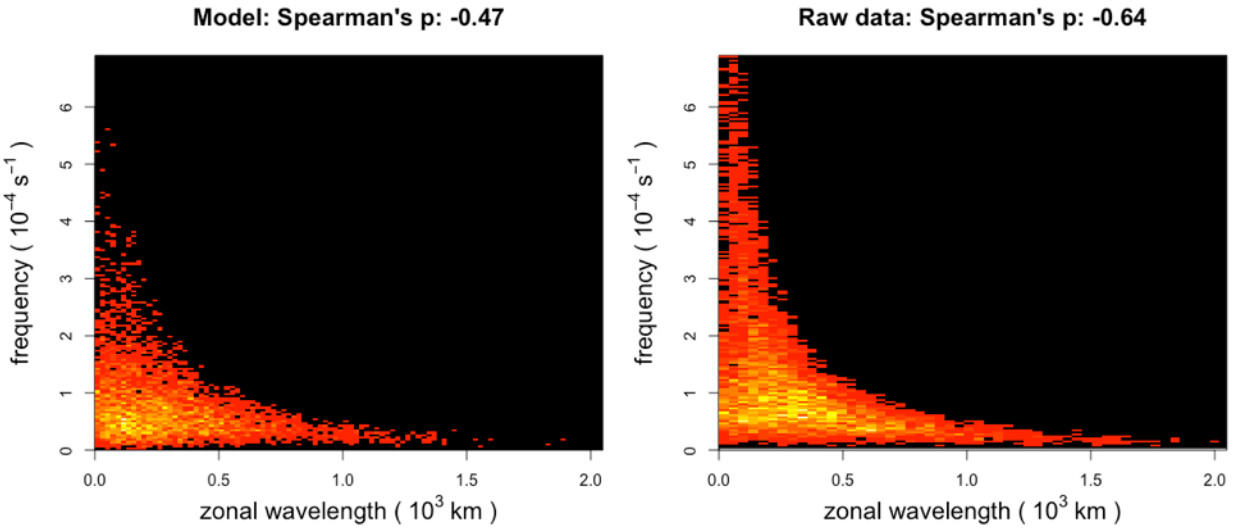
(a)


model
raw data
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7
frequency (10^{-4} s^{-1})

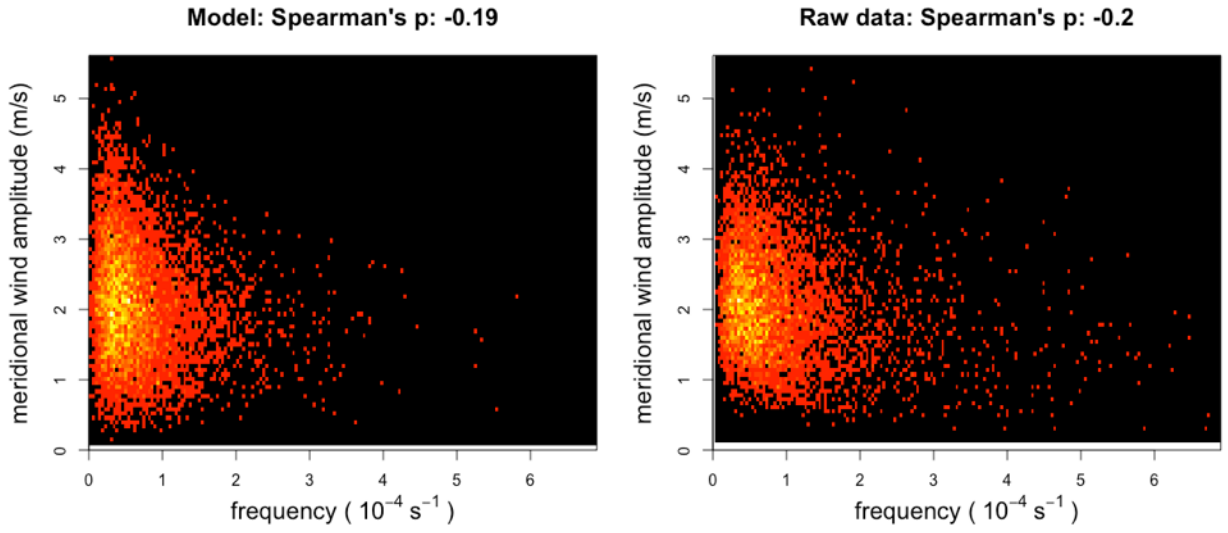
(b)


Model: Spearman's p: 0.29
Raw data: Spearman's p: 0.3
meridional wind amplitude (m/s)
vertical wavelength (km)
0
1
2
3
4
5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

(c)


Model: Spearman's p: -0.47
Raw data: Spearman's p: -0.64
frequency (10^{-4} s^{-1})
zonal wavelength (10^{3} km)
0
1
2
3
4
5
6
0.0
0.5
1.0
1.5
2.0

(d)



(e)

Figure 31: Figures (a) and (b) present the kernel density estimates for both the raw data (black curve) and the sample generated from the model (red curve) for meridional wavelength and frequency. Figures (c), (d) and (e) display side by side the 2D histograms associated with the raw data (right) and the sample generated from the model (left) for three pairs of gravity wave parameters.

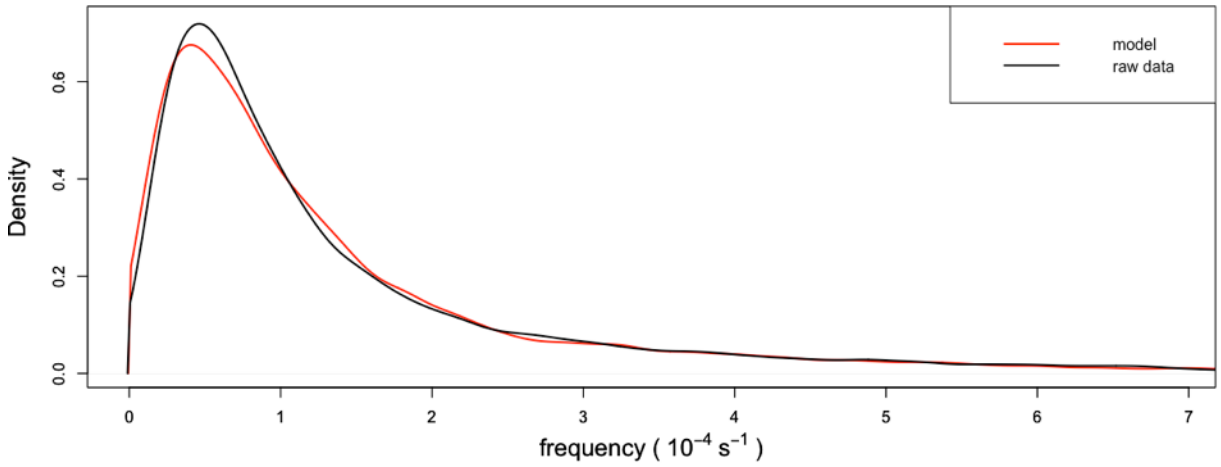
The graphs presented in Figure 31 and in Figure A.7 reveal several key features about the model built thus far. First, the kernel density estimates show that the marginal distributions generated from the model are very consistent with the raw data for all the parameters except for frequency. Second, the 2D histograms show that the correlation plots generated from the model are very consistent with the raw data for all pairs of parameters excluding frequency, as indicated by comparable histograms and Spearman's rank correlation coefficients that are within several hundredths of each other. A difference of several hundredths indicates a high degree of consistency between the data and the model. Third, the kernel density estimate of frequency as well as the 2D histograms of frequency and zonal/meridional wavelength display a scarcity of large frequency values. This means that the marginal frequency distribution associated with the model puts too much weight on smaller frequency values and too little weight on larger frequency value. Finally, the model correlation between frequency and zonal/meridional wind amplitude strongly concur with the empirical correlation plots (as indicated by Spearman's rank correlation coefficients that are within 0.03 of each other), while the model correlation between frequency and temperature amplitude/vertical wavelength moderately concur with the empirical correlation plots (as indicated by Spearman's rank correlation coefficients that are within 0.18 of each other).

Overall, there were three main takeaways from the graphs presented in Figure 31. First, the copula was extremely effective at modeling both the marginal distributions and the correlation structures between all the gravity wave parameters except for frequency. Second,

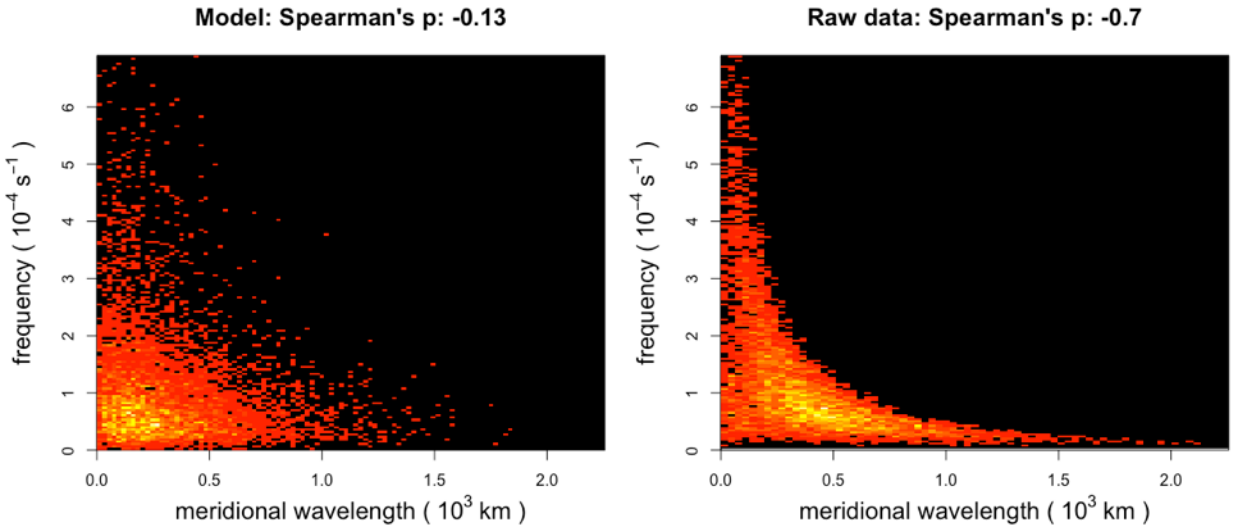
even though frequency was modeled independently of vertical wavelength, zonal wind amplitude, meridional wind amplitude and temperature amplitude, the correlation structure in the model between these pairs of parameters shares many similarities with the empirical correlation plots. This implies that by conditioning on the values of zonal/meridional wavelength, the frequency values are indirectly influenced by these other parameters. This make sense given that the correlation structures between the zonal/meridional wavelength and these other parameters had been modeled using the copula. Finally, it was clear that the conditional frequency distribution needed to be further revised to better model its marginal distribution as well as its correlation with zonal and meridional wavelengths.

4.6 Revision of Conditional Frequency Distribution

One easy way to reduce the discrepancy in the marginal frequency distribution between the model and the raw data is to condition the frequency on either zonal wavelength or meridional wavelength, but not both. However, as discussed above, this simpler model corresponds to a higher AIC. This means that while this model enhances the degree to which the marginal frequency distribution aligns with the empirical kernel density estimate, it leads to a deterioration in the degree to which the model correlation structure matches up with the empirical correlation structure. (Most notably, the model correlation between frequency and the wavelength parameter that frequency isn't conditioned on deviates significantly from the empirical correlation structure.) The improved marginal frequency distribution, but worsened correlation structures are illustrated below in Figure 32:



(a)



(b)

Figure 32: Effect of conditioning frequency on just zonal wavelength. Figure (a) shows that this improves the extent to which the model marginal frequency distribution (red curve) aligns with the empirical marginal frequency distribution (black curve). Figure (b), however, illustrates that the model correlation between frequency and meridional wavelength (left) no longer matches up well with the corresponding empirical correlation between the two parameters (right).

The next approach was to condition frequency on both zonal wavelength and meridional wavelength simultaneously, rather than condition on each separately and use the conditional distribution corresponding to the larger wavelength value. The steps involved in this procedure were similar to those used to condition frequency on just one of the wavelength values

at a time:

1. Split the zonal wavelength values and meridional wavelength values into the same number of intervals, where each interval contains the same number of data points. For example, a total of 10 intervals would mean placing all zonal/meridional wavelength values from the minimum value to the 0.1 quantile in one interval, all the data between the 0.1 quantile and the 0.2 quantile in the second interval, etc.

2. Visualize the data as a three-dimensional surface plot, where the x and y axis represent meridional and zonal wavelength and the z axis represents frequency, as illustrated in Figure 33:

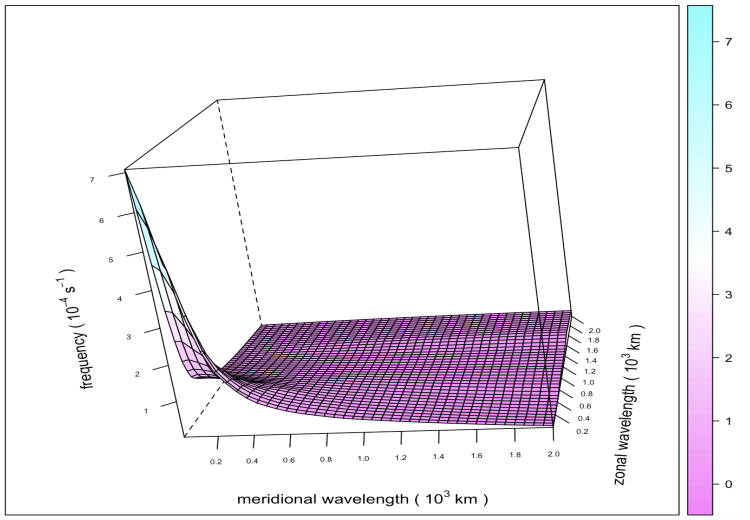


Figure 33: 3D surface plot where the x and y axis represent meridional and zonal wavelength, respectively, and the z axis represents frequency.

3. For each two-dimensional region in the x-y plane that is defined by the intersection of a zonal wavelength and a meridional wavelength interval, extract those frequency values that correspond to zonal/meridional wavelength values that lie within their respective intervals. Denoting the number of zonal/meridional intervals by n, there are a total of n^2 regions in the x-y plane.

4. Apply the polynomial regression method to the frequency values corresponding to each region to approximate the conditional frequency distribution given zonal and meridional wavelength.

5. Model the frequency by observing which region a given zonal wavelength and meridional wavelength value fall into and then predicting the corresponding frequency value using the associated polynomial probability density function.

6. Use 10-fold cross-validation to determine the optimal number of intervals. The same risk of overfitting exists in this situation, so cross-validation is necessary to prevent

this from occurring. The steps involved in this 10-fold cross-validation procedure are practically identical to those outlined previously. The only difference is that the number of intervals to test included all integers from 5 to 20 rather than every integer between 10 and 200 in increments of 5. These lower numbers were used because the total number of regions is equal to the square of the number of intervals.

Figure 34 shows the cross-validation plot that approximates the AIC as a function of the number of intervals and selects the number of intervals corresponding to the lowest AIC value [3]. As expected, the regression curve is convex and thus reaches a global minimum.

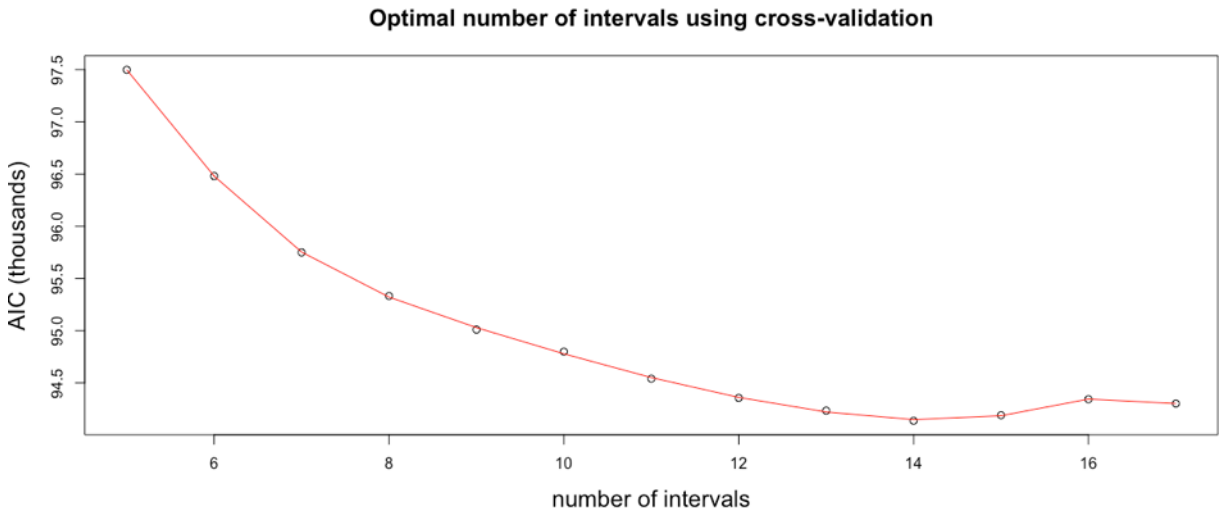


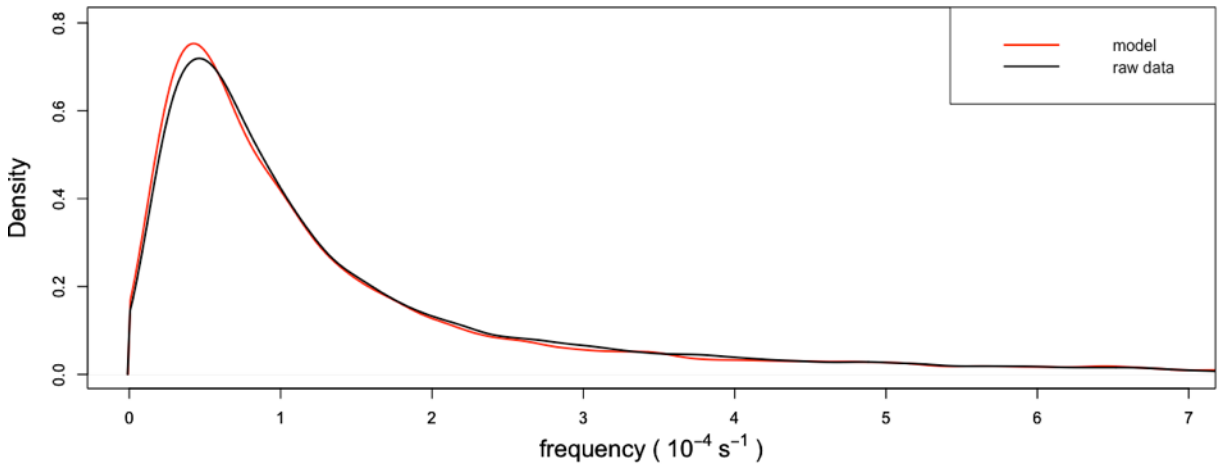
Figure 34: Plot of AIC as a function of the number of intervals along with a 10th order polynomial regression curve. The number of intervals corresponding to the minimum AIC turned out to be 14.

In theory, this approach of modeling the conditional distribution of frequency given zonal and meridional wavelength could be extended to the other four gravity wave parameters. However, the amount of data necessary to split the frequency values into six-dimensional regions and learn distributions on the frequency values in each region is inordinate, which means this cannot be implemented in practice. (Even conditioning on one more gravity wave parameter in addition to zonal/meridional wavelength would lead to issues with data sparsity, as the total number of regions would be equal to n^3 where n is the number of intervals.)

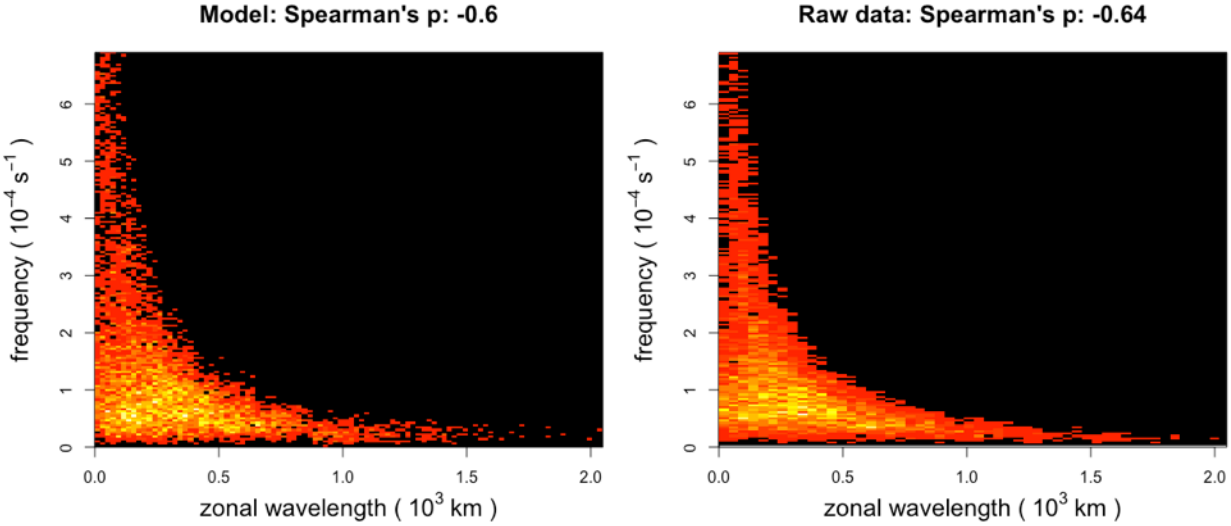
This new way of calculating the conditional frequency distribution lowered the AIC from 98,159 (the AIC after implementing the polynomial regression PDF) to 93,145. As usual, this sizable drop in the AIC indicates that this alteration led to a significant improvement in

[3]Although all integers between 5 and 20 were tested, only the integers that are less than 18 can be found in the graph. This is the case because starting with 18 intervals, there were not enough data points in each of the 324+ regions to learn a distribution on the frequency values. As a result, these high number of intervals were disregarded entirely.

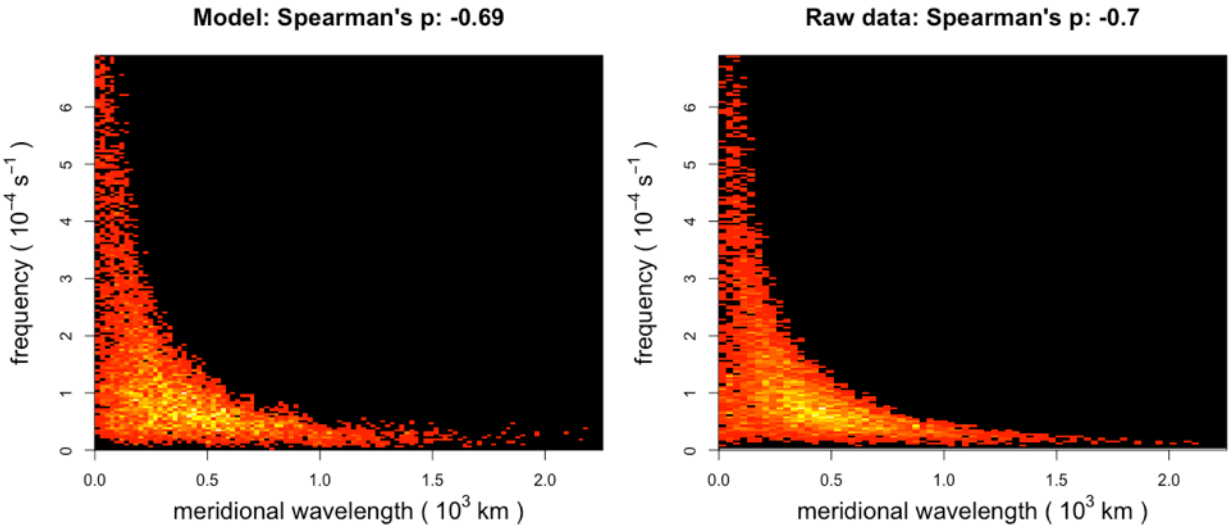
the model. To confirm this, it is helpful to recompare those aspects of the model involving frequency to the raw data, as shown below in Figure 35.



(a)



(b)



(c)

Figure 35: The top graph displays the kernel density estimates of frequency for both the raw data (black curve) and the sample generated from the model (red curve). The bottom two graphs display side by side the 2D histograms of zonal wavelength versus frequency and meridional wavelength versus frequency associated with the raw data (right) and the sample generated from the modified model (left).

Based on the graphs presented in Figure 35, it is clear that the marginal frequency distribution is aligned much more closely with the raw data and is no longer biased toward smaller frequency values. Moreover, the correlation structure between frequency and zonal and meridional wavelength better matches the raw data as indicated by the 2D histograms that are more alike and closer Spearman's rank correlation coefficients. Finally, the correlation structure between frequency and the other gravity wave parameters aren't significantly altered as indicated by the 2D histograms found in Figure A.8.

4.7 Model Conclusion

Having revised the conditional frequency distribution, the joint probability distribution now accurately models all of the marginal distribution as well as the correlation structures. At this point there were no obvious modifications that could be made to the model to make it better match up with the raw data. Moreover, the final AIC value of 93,145 is significantly lower than the AIC value of 101,311 associated with the seven-dimensional copula model discussed in Section 4.3.2. Therefore, this model was considered to be complete and the best possible joint probability distribution on the seven atmospheric gravity wave parameters.

As mentioned briefly before, the deliverable is a function that generates a sample of a given size from this joint probability distribution. The SCoPEx will then analyze these samples to gain insight into the gravity wave activity in the lower stratosphere over New Mexico.

Beyond being easy to deal with, sampling from the joint probability distribution is the only feasible way to interface with the model because the model itself is incredibly unwieldy and difficult to work with directly.

5 Future Work

While the model presented in this report is complete from a statistical standpoint, the reliability of the model can be improved by increasing the amount of radiosonde data from which it is constructed. While it may seem tempting to acquire this data from other radiosonde stations that are further away from the experimental site in eastern New Mexico, doing so requires a rigorous justification as to why the gravity wave activity in a different region should theoretically be similar to that over eastern New Mexico. This requires a deep understanding of atmospheric science and is thus outside the scope of the project. Instead, the most sensible course of action is to supplement the radiosonde data from 1998 to 2008 with the radiosonde data from 2009 to the present that is associated with the four radiosonde stations used in the model. As mentioned in Section 2.1, this involves running a Fortran program that converts the encoded BUFR files into plain text. Therefore, the next step is to figure out how to run this Fortran program or to develop some alternative method to decode the BUFR files.

6 Applications

Exploring the gravity wave parameters inferred from different radiosonde stations across the United States indicates that the overall structure of the gravity wave parameters is consistent across a wide range of locations. This is important because the model makes certain assumptions about the data that must hold in order for it to accurately capture the structure of the data. For example, the parameters are assumed to be independent of year and time of day, and the correlation between all pairs of parameters except for frequency are assumed to be linear so that a copula can be applied.

As a result of this consistency, the process of inferring the gravity wave parameters from raw radiosonde data and then constructing a joint probability distribution on the seven parameters can be carried out for many geographical locations other than eastern New Mexico. This can provide insight into the gravity wave activity in other specific regions and could help atmospheric scientists better understand the phenomenon as a whole. One such example is a NASA-funded project that the Keutsch lab is working on where it is exploring how strong summertime convective storms over North America can alter the chemistry of the stratosphere. Because the experiments will be taking place over Kansas, the lab will build a model using radiosonde data from nearby stations in the Great Plains to account for the influence of atmospheric gravity waves on the dynamics in the stratosphere.

Another application involves comparing the joint probability distribution that this model outputs to the gravity wave parameters that are generated by high-resolution computer simulations for given atmospheric conditions in specific regions. Performing this comparison

could shed light on both the reliability of the hodograph method at inferring gravity wave parameters and the ability of the simulations to accurately model gravity wave characteristics.

Finally, this joint probability distribution can serve its original purpose of being incorporated into the SCoPEx team's atmospheric transport and chemistry model, where it will account for the influence of atmospheric gravity waves on the movement of aerosol particles in the lower stratosphere.

7 Budget

This project was completed entirely in Python and R, and the only input to the model was publicly accessible radiosonde data. Therefore, the total cost of the project was $0.

Item	Cost
Python	Free
R	Free
US High Vertical Resolution Radiosonde Data	Publicly available through SPARC (Stratosphere-troposphere Processes And their Role in Climate)
Literature on atmospheric gravity waves	Free through Harvard Hollis

Table 2: Budget items

8 Acknowledgements

There are several people in particular to whom I'm incredibly grateful for their help in making this project possible.

First, I would like to thank my advisor, John Dykema, for his guidance through the project. With his deep expertise in atmospheric sciences, John helped me navigate the complexities of atmospheric gravity waves and offered valuable advice throughout the design process.

I would also like to thank my faculty advisor, Prof. Frank Keutsch, for giving me the opportunity to work on such a meaningful project. It's been extremely rewarding to play a role, however small, in working towards a solution to climate change.

Next I'd like to thank my section leader, Michelle Rosen, for putting in the effort to learn a significant amount of statistics so that she could provide valuable feedback throughout the project. I'd also like to thank her for constantly challenging me to better justify my design decisions and for always being available to discuss my progress reports and rough drafts.

Finally, I'd like to thank my family and friends for their constant support and encouragement throughout the entire project.

9 References

[1] L. Wang and M. A. Geller, "Morphology of gravity-wave energy as observed from 4 years (1998–2001) of high vertical resolution u.s. radiosonde data," *Journal of Geophysical Research: Atmospheres*, vol. 108, no. 16, 2003.

[2] J. A. Dykema, D. W. Keith, J. G. Anderson, and D. Weisenstein, "Stratospheric controlled perturbation experiment: a small-scale experiment to improve understanding of the risks of solar geoengineering," *Philosophical Transactions of the Royal Society*, vol. 372, no. 2031, 2014.

[3] J. S. Sawyer, "Quasi-periodic wind variations with height in the lower stratosphere," *Quarterly Journal of the Royal Meteorological Society*, vol. 87, no. 371, 1961.

[4] R. Plougonven and F. Zhang, "Internal gravity waves from atmospheric jets and fronts," *Reviews of Geophysics*, vol. 52, no. 1, pp. 33–76, 2014.

[5] S. D. Zhang, F. Yi, C. M. Huang, and K. M. Huang, "High vertical resolution analyses of gravity waves and turbulence at a midlatitude station," *Journal of Geophysical Research: Atmospheres*, vol. 117, no. D2, 2012.

[6] S. D. Zhang and F. Yi, "A statistical study of gravity waves from radiosonde observations at wuhan (30 n, 114 e) china," *Annales Geophysicae*, vol. 23, no. 3, 2005.

[7] S. D. Zhang and F. Yi, "Latitudinal and seasonal variations of inertial gravity wave activity in the lower atmosphere over central china," *Journal of Geophysical Research: Atmospheres*, vol. 112, no. D5, 2007.

[8] F. Zhang, S. Wang, and R. Plougonven, "Uncertainties in using the hodograph method to retrieve gravity wave characteristics from individual soundings," *Geophysical Research Letters*, vol. 31, no. 11, 2004.

[9] L. Wasserman, *All of Nonparametric Statistics*. International series of monographs on physics, Springer Science+Business Media, Inc., 2006.

[10] A. Fitzgibbon, M. Pilu, and R. B. Fisher, "Direct least square fitting of ellipses," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 5, pp. 476–480, 1999.

[11] R. A. Vincent and M. Joan Alexander, "Gravity waves in the tropical lower stratosphere: An observational study of seasonal and interannual variability," *Journal of Geophysical Research: Atmospheres*, vol. 105, no. D14, pp. 17971–17982, 2000.

[12] T. P. Lane, M. J. Reeder, B. R. Morton, and T. L. Clark, "Observations and numerical modelling of mountain waves over the southern alps of new zealand," *Quarterly Journal of the Royal Meteorological Society*, vol. 126, no. 569, pp. 2765–2788, 2000.

[13] G. Simpson, "Good methods for density plots of non-negative variables in r?," 2013. Accessed on 2019-1-12.

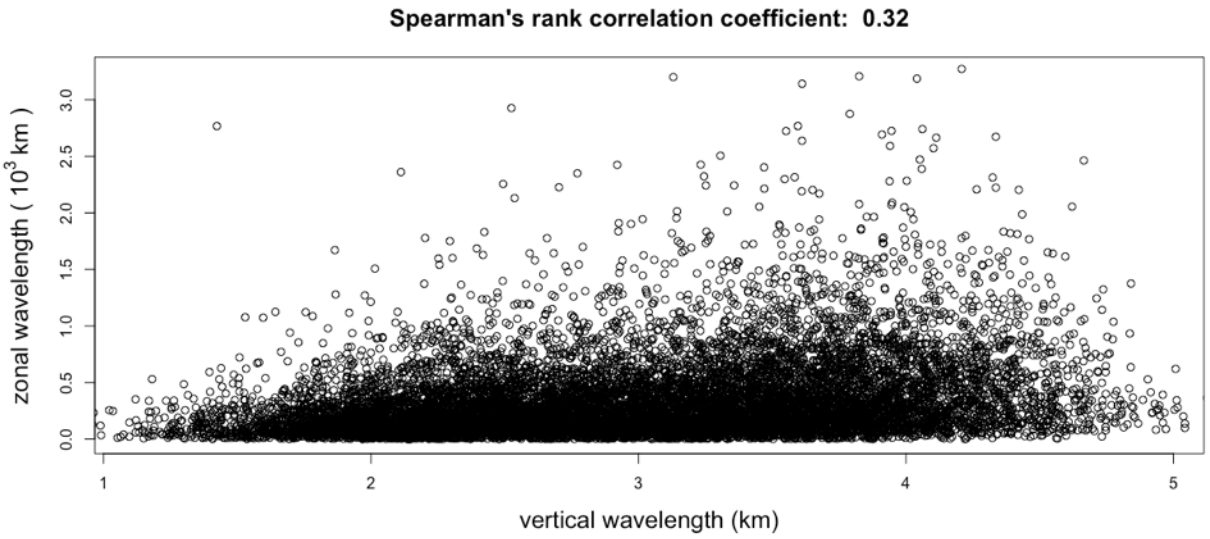
[14] A. D. Del Genio, M.-S. Yao, and J. Jonas, "Will moist convection be stronger in a warmer climate?," *Geophysical Research Letters*, vol. 34, no. 16, 2007.

10 Link to Code

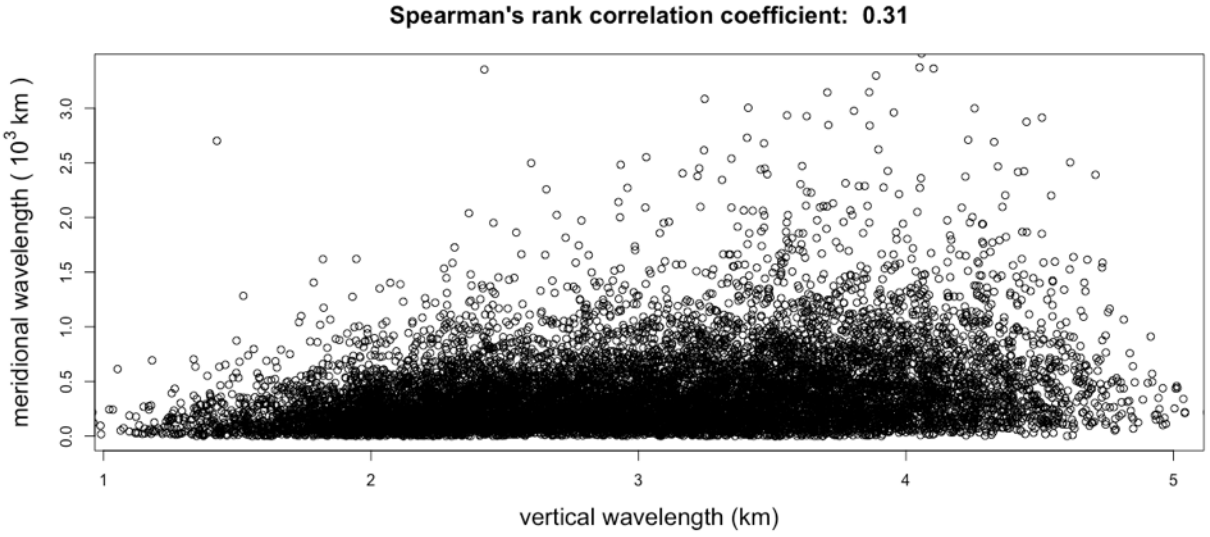
All the code needed to build the model presented in this report can be found here.

(https://github.com/williamfried/Joint-Probability-Distribution-of-Atmospheric-Gravity-Wave-Parameters)

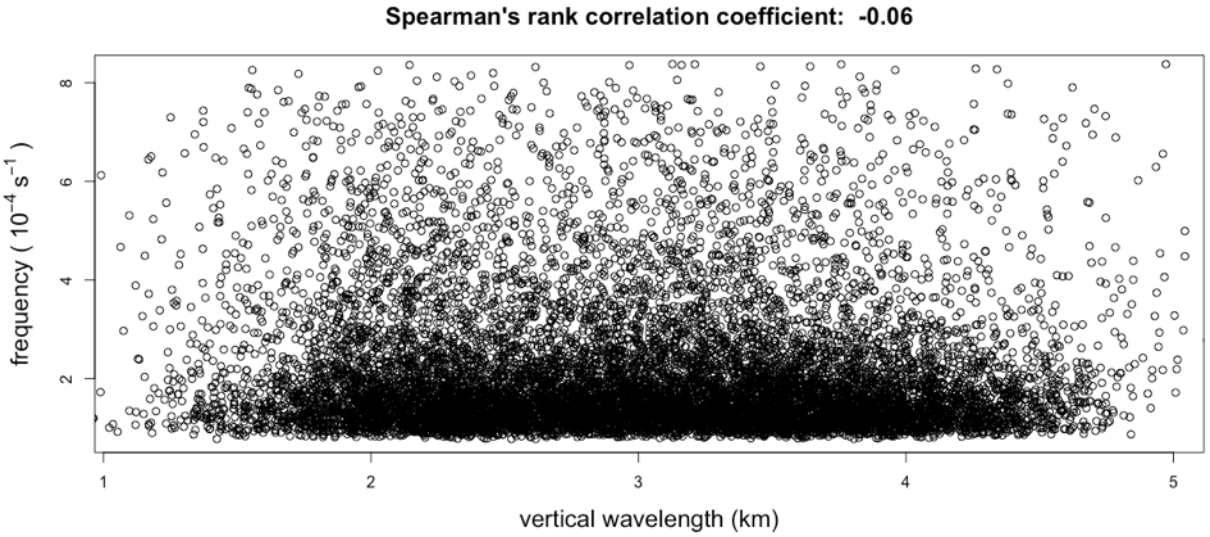
A Appendix



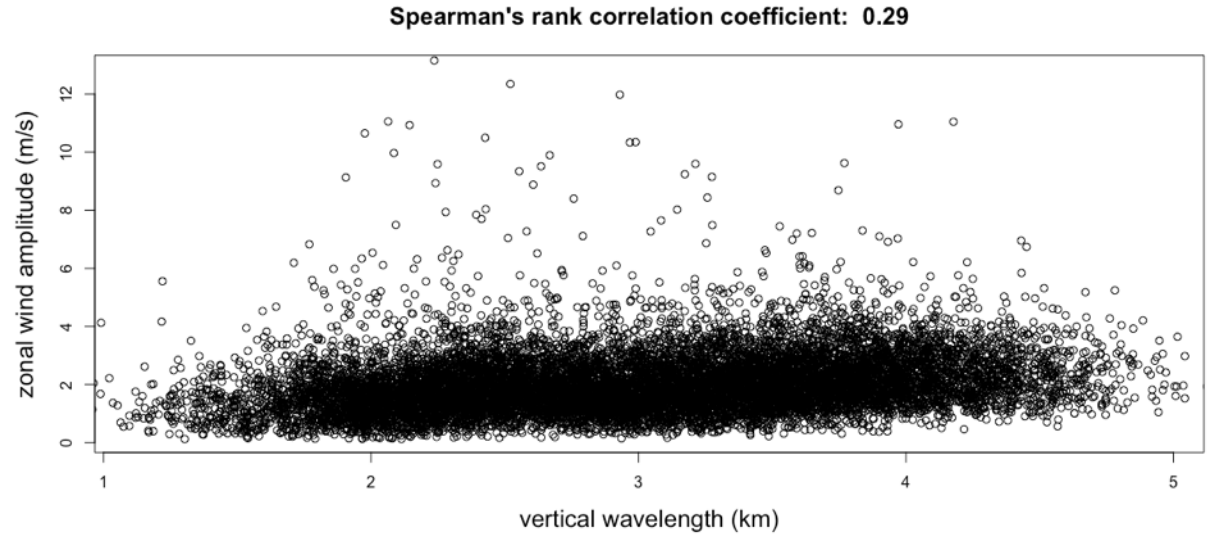
(a)



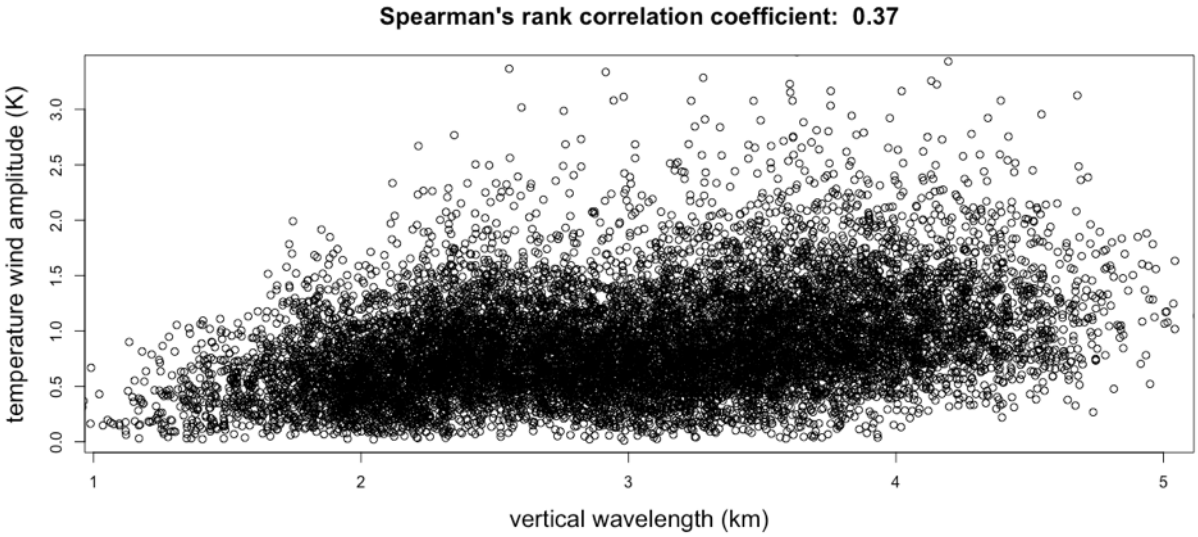
(b)


Spearman's rank correlation coefficient: -0.06
frequency (10^{-4} s^{-1})
vertical wavelength (km)

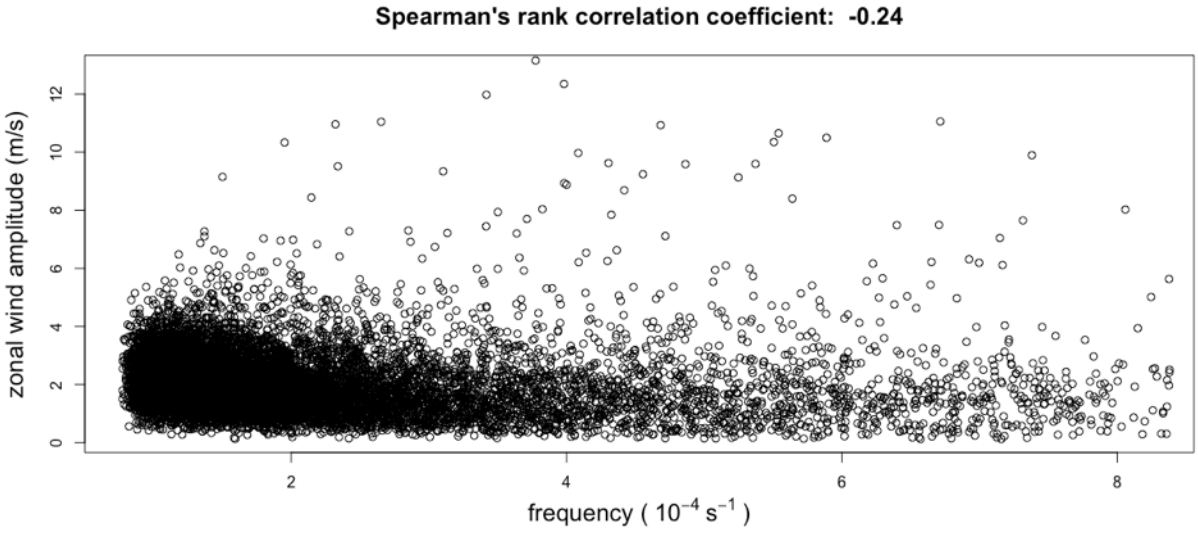
(c)


Spearman's rank correlation coefficient: 0.29
zonal wind amplitude (m/s)
vertical wavelength (km)

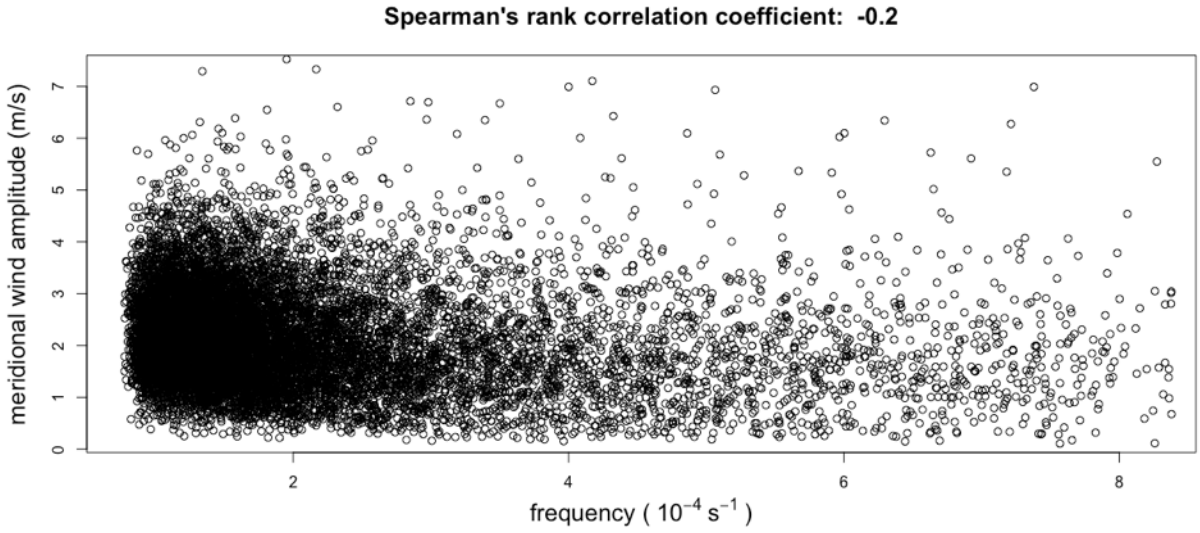
(d)


Spearman's rank correlation coefficient: 0.37
temperature wind amplitude (K)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1
2
3
4
5
vertical wavelength (km)

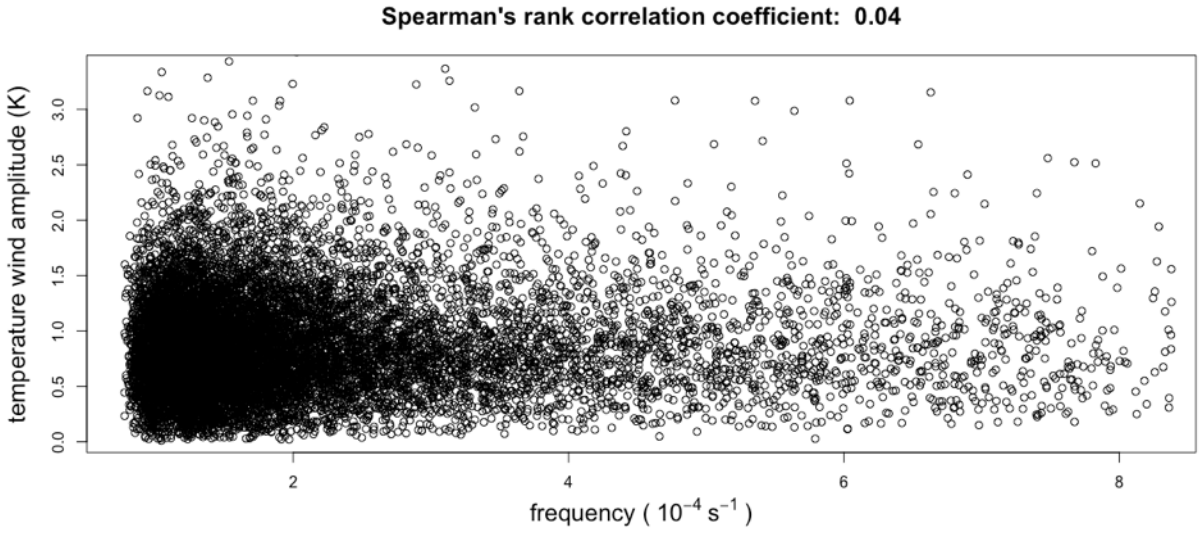
(e)


Spearman's rank correlation coefficient: -0.24
zonal wind amplitude (m/s)
0
2
4
6
8
10
12
2
4
6
8
frequency (10^{-4} s^{-1})

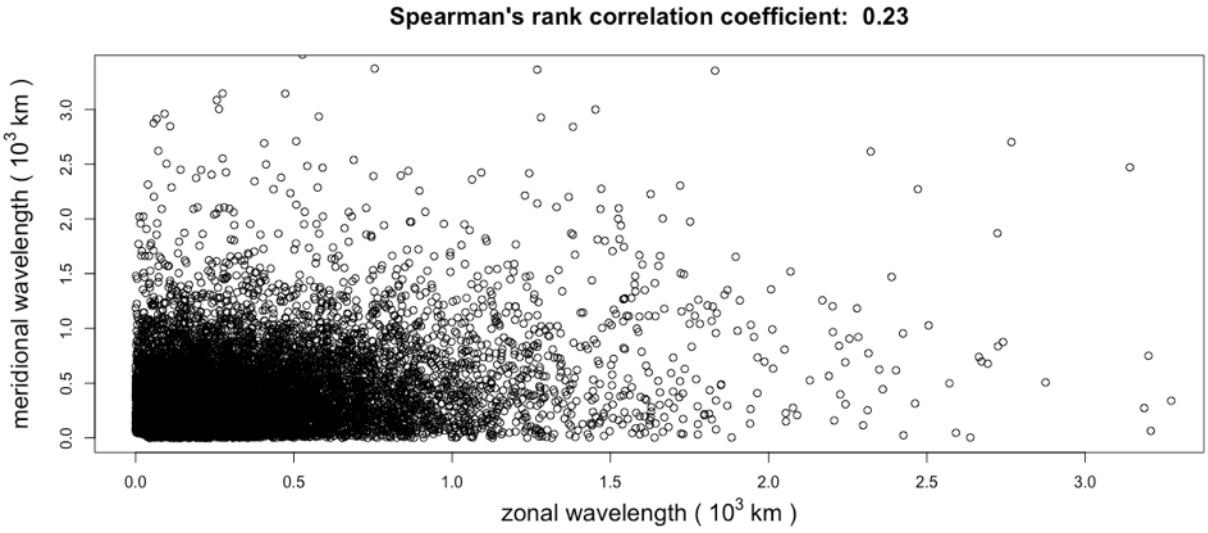
(f)


Spearman's rank correlation coefficient: -0.2
meridional wind amplitude (m/s)
0
1
2
3
4
5
6
7
2
4
6
8
frequency (10^{-4} s^{-1})

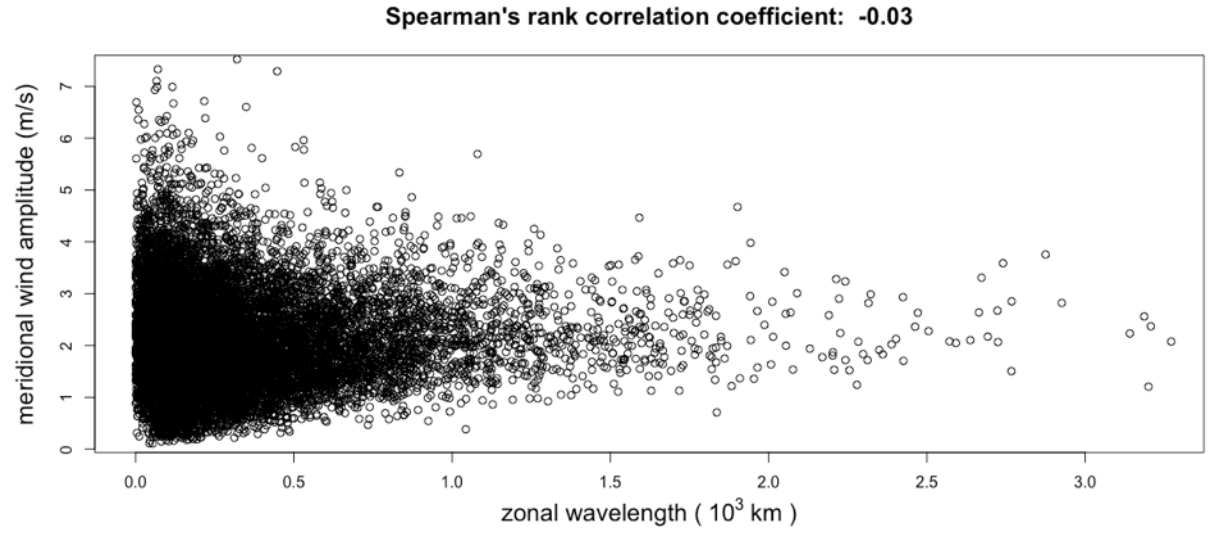
(g)


Spearman's rank correlation coefficient: 0.04
temperature wind amplitude (K)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
4
6
8
frequency (10^{-4} s^{-1})

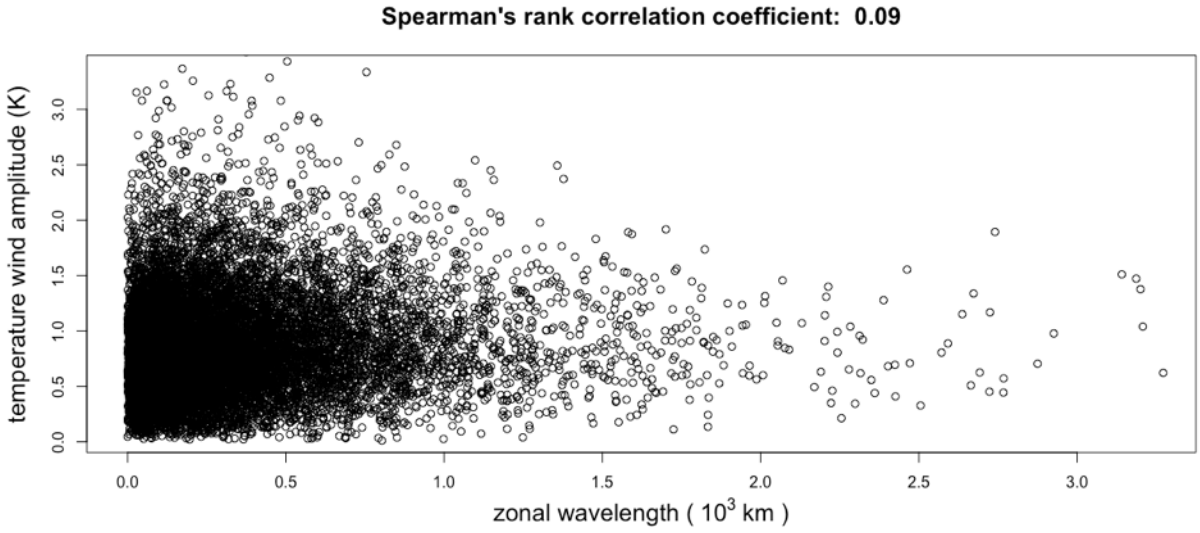
(h)


Spearman's rank correlation coefficient: 0.23
meridional wavelength (10^3 km)
zonal wavelength (10^3 km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

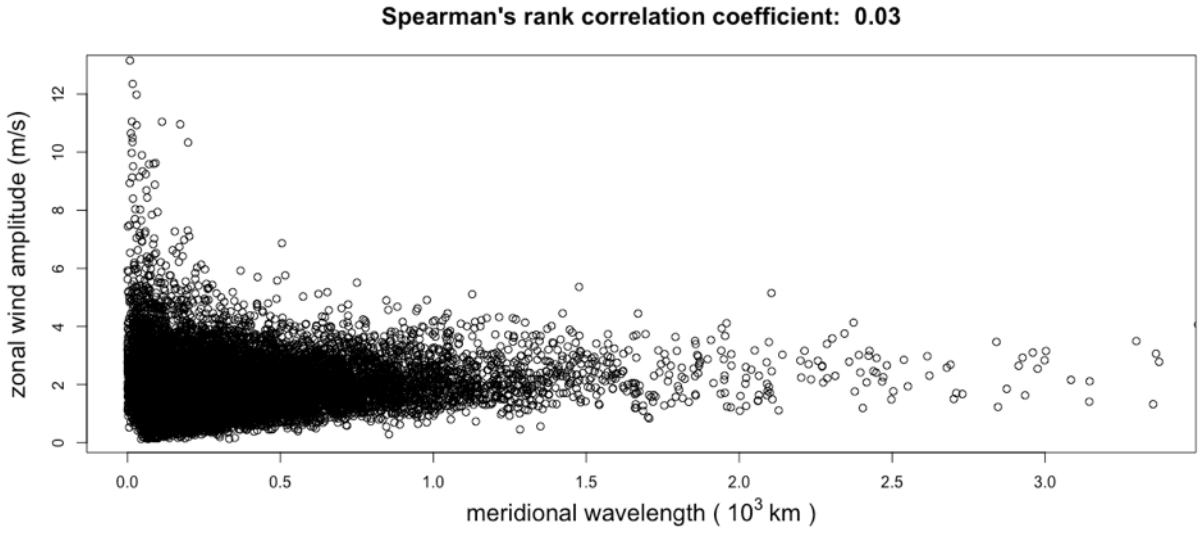
(i)


Spearman's rank correlation coefficient: -0.03
meridional wind amplitude (m/s)
zonal wavelength (10^3 km)
0
1
2
3
4
5
6
7
0.0
0.5
1.0
1.5
2.0
2.5
3.0

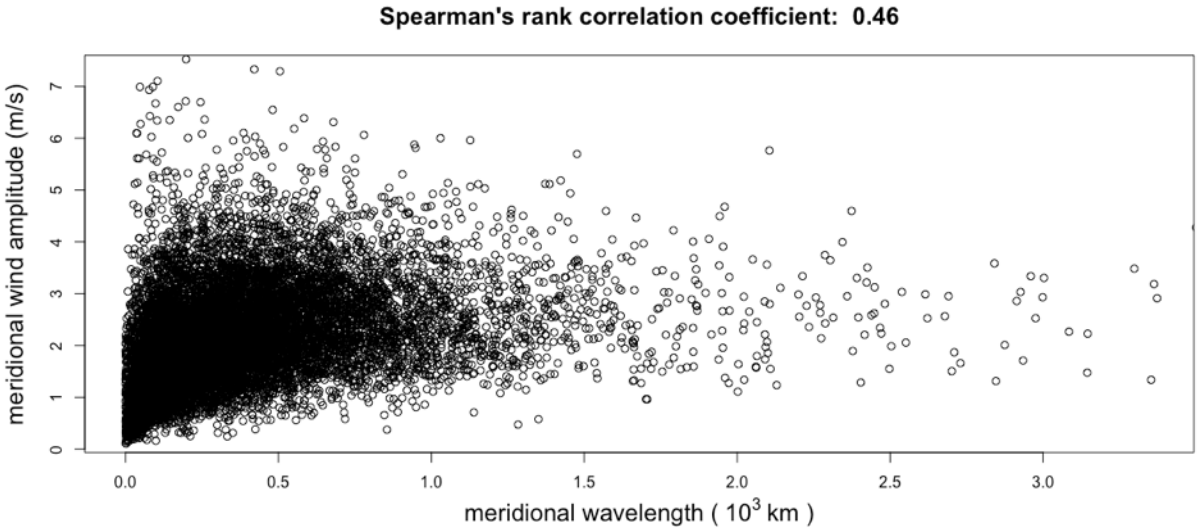
(j)


Spearman's rank correlation coefficient: 0.09
temperature wind amplitude (K)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
zonal wavelength (10^3 km)

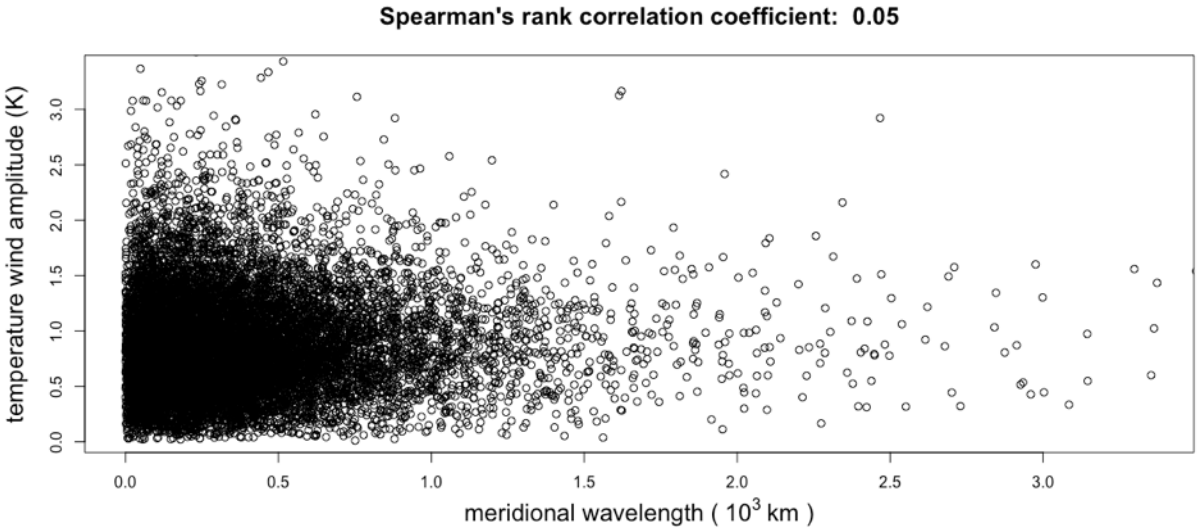
(k)


Spearman's rank correlation coefficient: 0.03
zonal wind amplitude (m/s)
0
2
4
6
8
10
12
meridional wavelength (10^3 km)

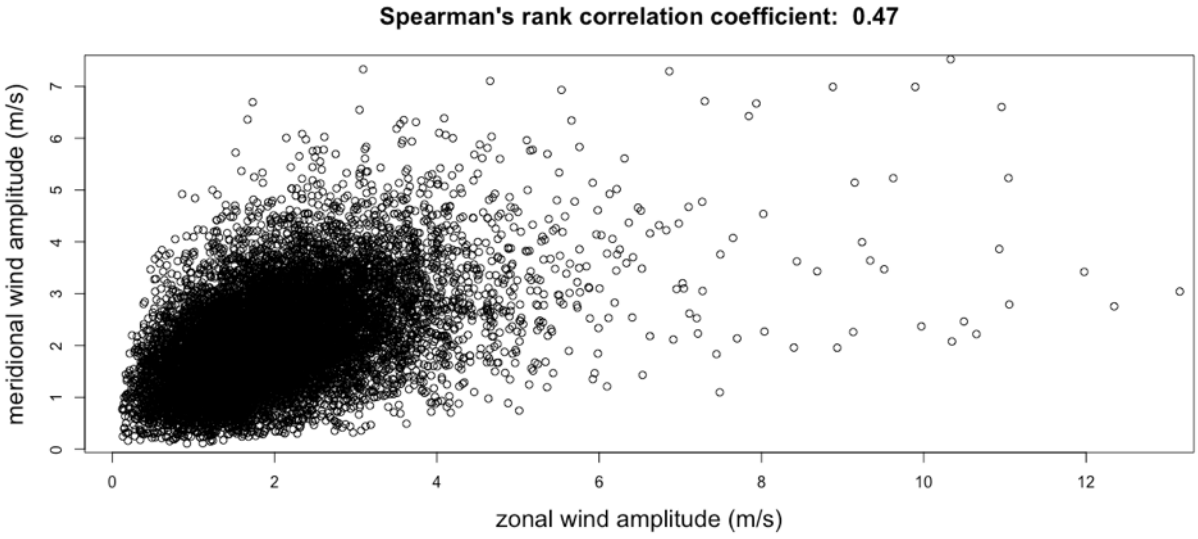
(l)


Spearman's rank correlation coefficient: 0.46
meridional wind amplitude (m/s)
meridional wavelength (10^3 km)

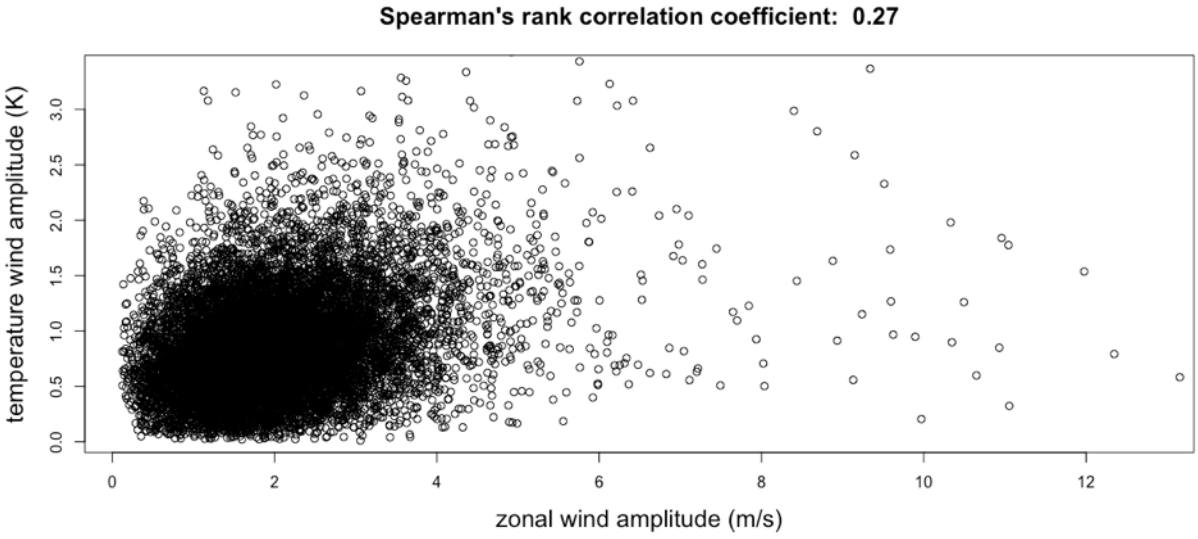
(m)


Spearman's rank correlation coefficient: 0.05
temperature wind amplitude (K)
meridional wavelength (10^3 km)

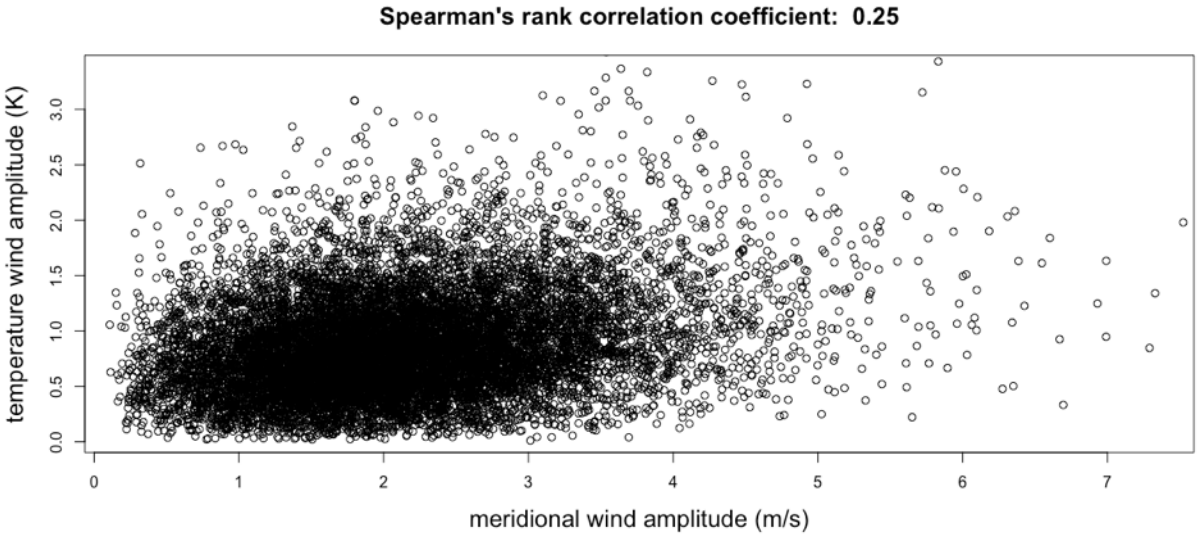
(n)


Spearman's rank correlation coefficient: 0.47
meridional wind amplitude (m/s)
zonal wind amplitude (m/s)

(o)

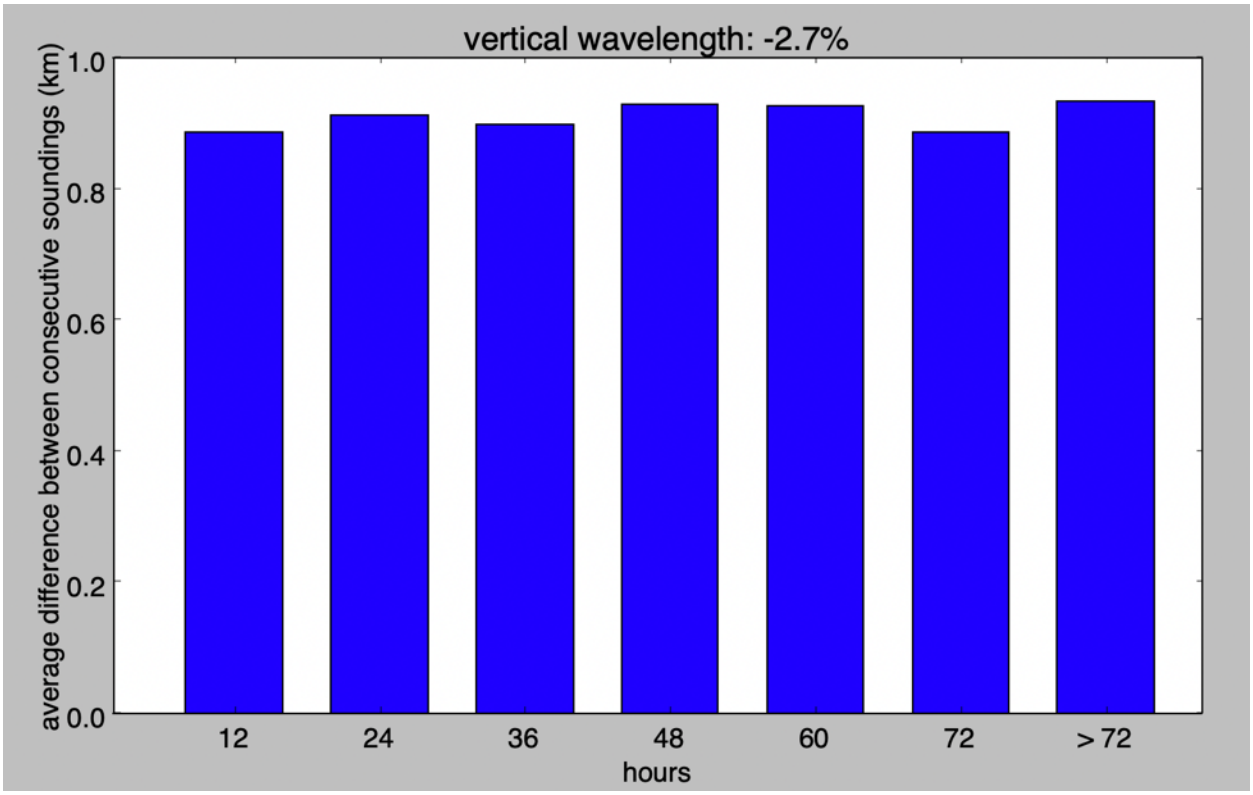

Spearman's rank correlation coefficient: 0.27
temperature wind amplitude (K)
zonal wind amplitude (m/s)

(p)

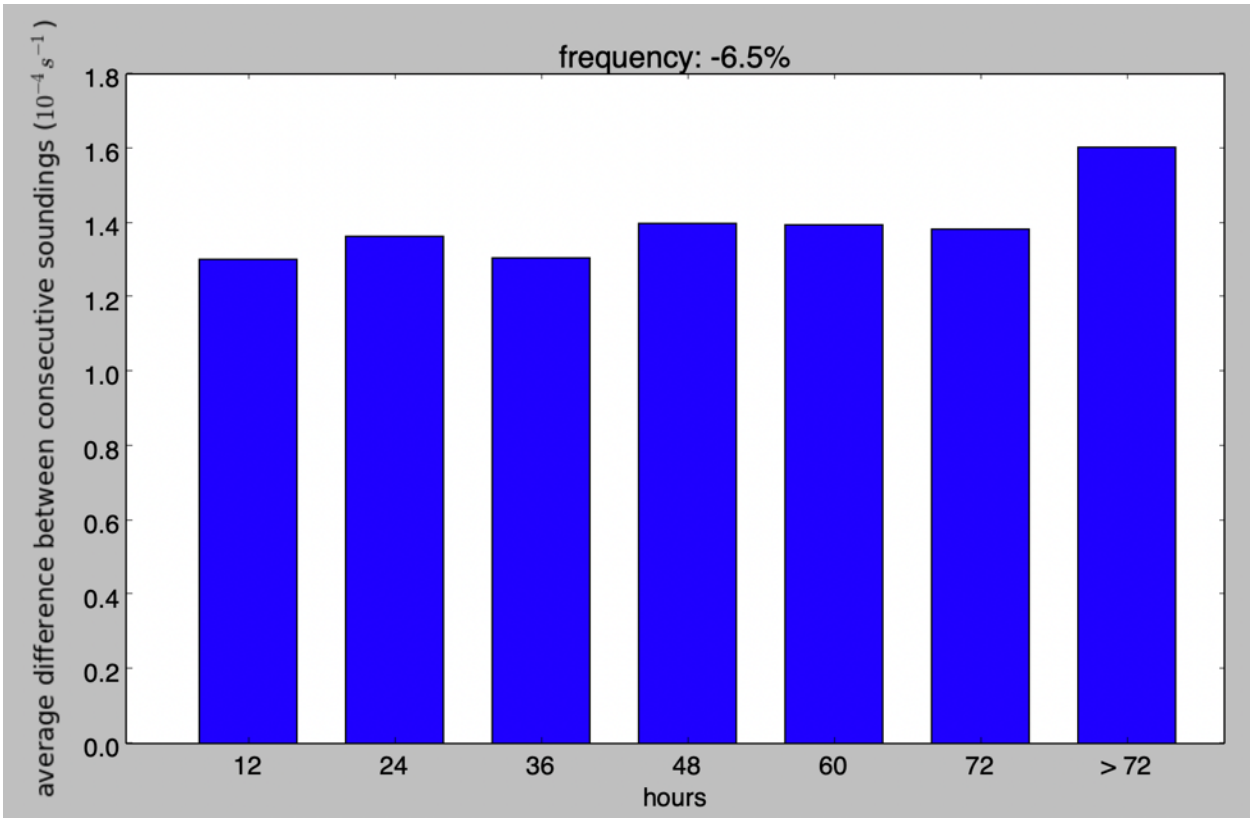


(q)

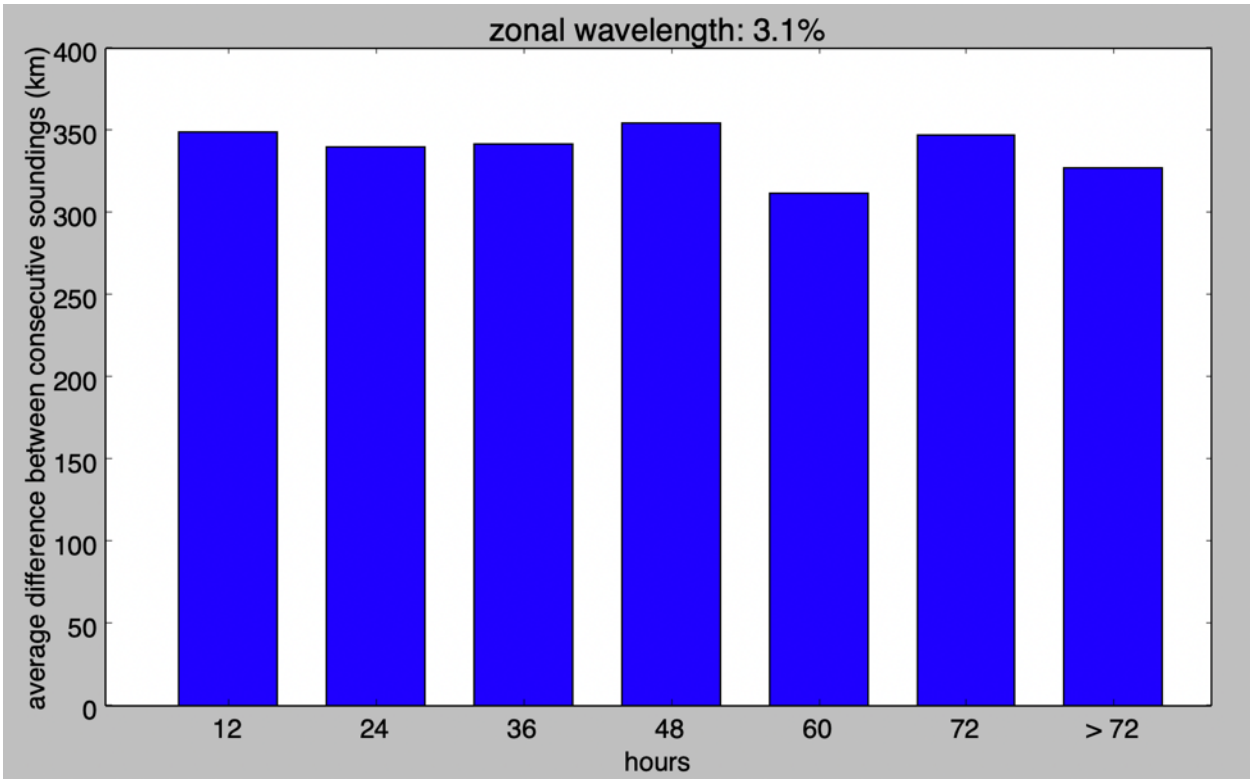
Figure A.1: Scatterplots of the pairs of gravity wave parameters that were not included in Figure 9. The Spearman's rank correlation coefficient is reported in the title of each scatterplot.


vertical wavelength: -2.7%
average difference between consecutive soundings (km)
1.0
0.8
0.6
0.4
0.2
0.0
12
24
36
48
60
72
> 72
hours

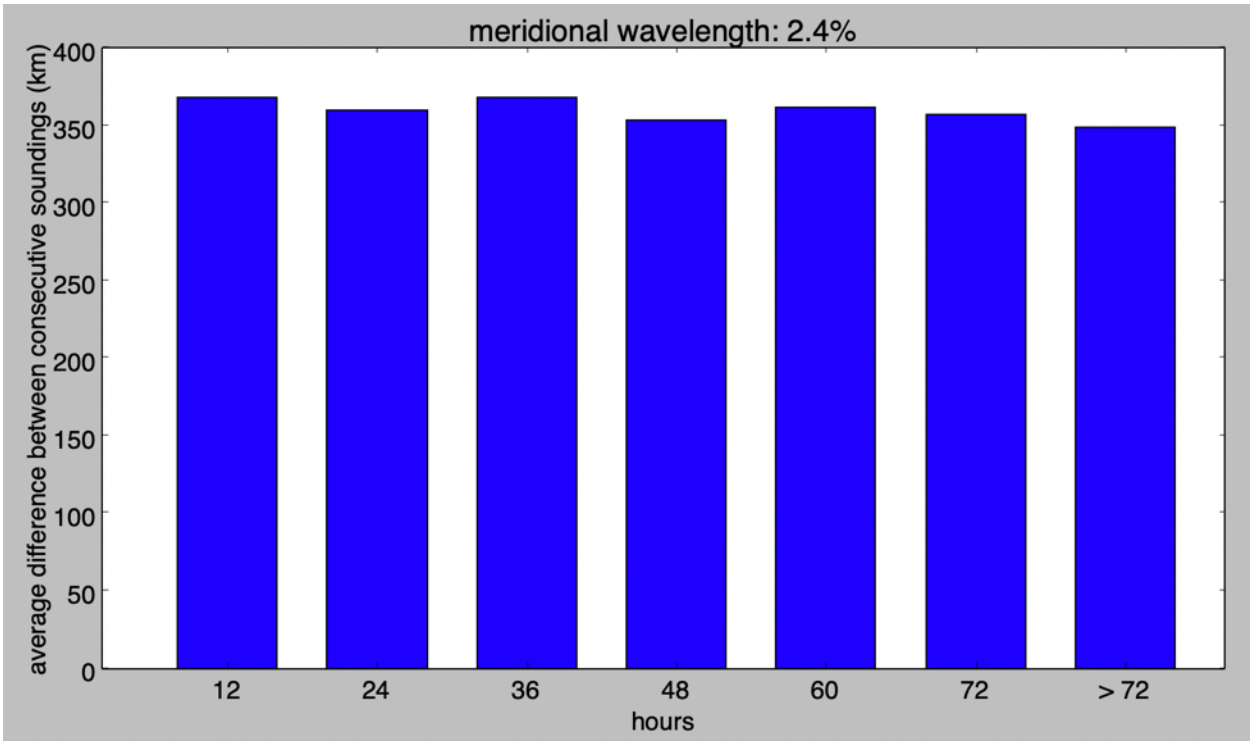
(a)


frequency: -6.5%
average difference between consecutive soundings ($10^{-4}\,s^{-1}$)
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
12
24
36
48
60
72
> 72
hours

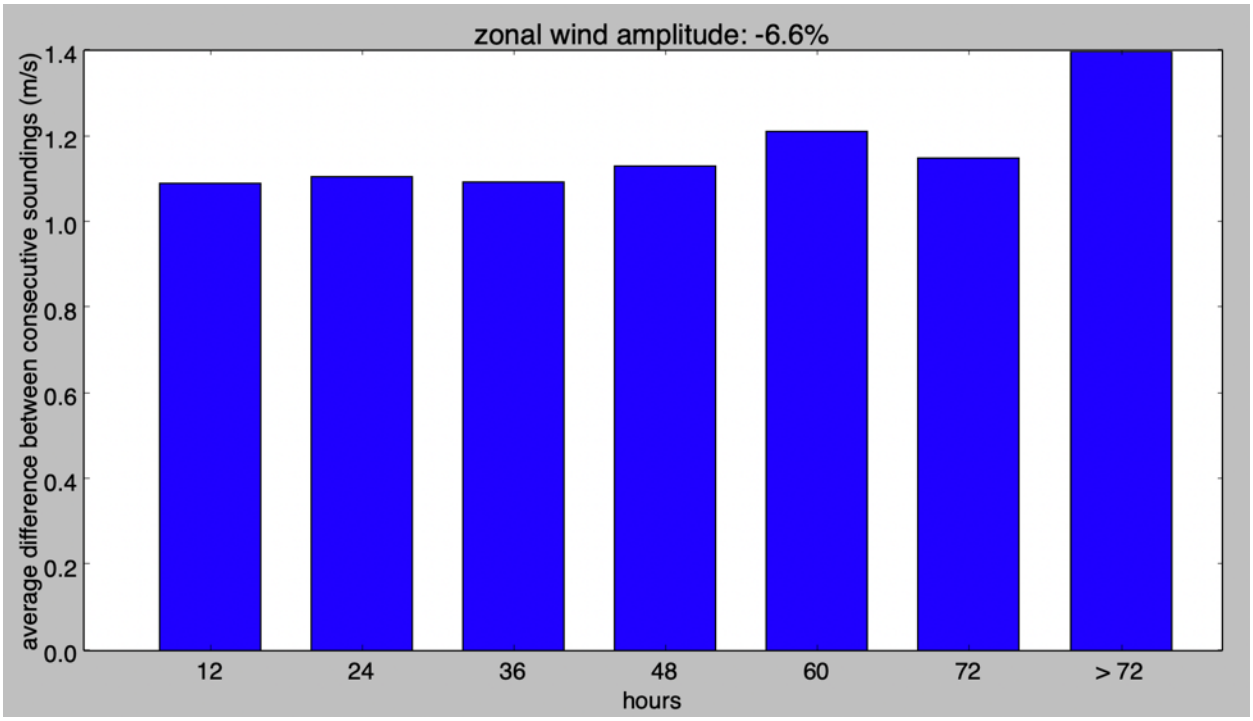
(b)


zonal wavelength: 3.1%
average difference between consecutive soundings (km)
0
50
100
150
200
250
300
350
400
12
24
36
48
60
72
> 72
hours

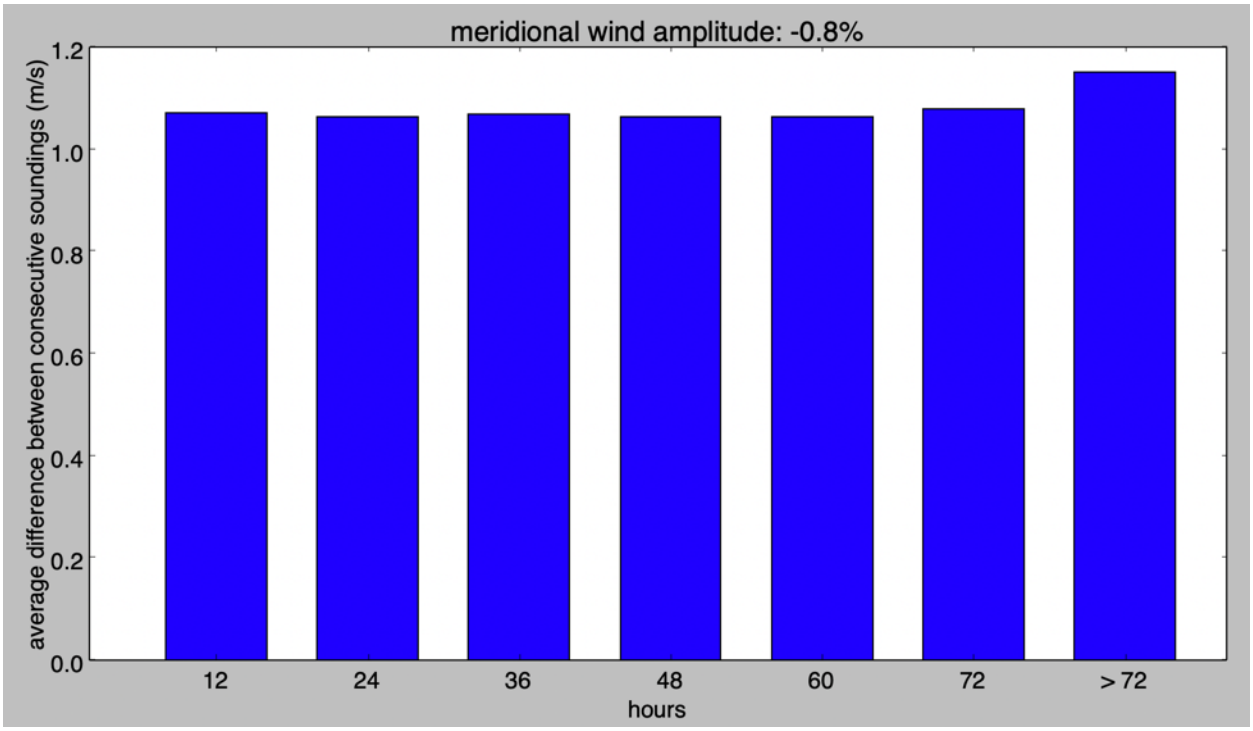
(c)


meridional wavelength: 2.4%
average difference between consecutive soundings (km)
0
50
100
150
200
250
300
350
400
12
24
36
48
60
72
> 72
hours

(d)

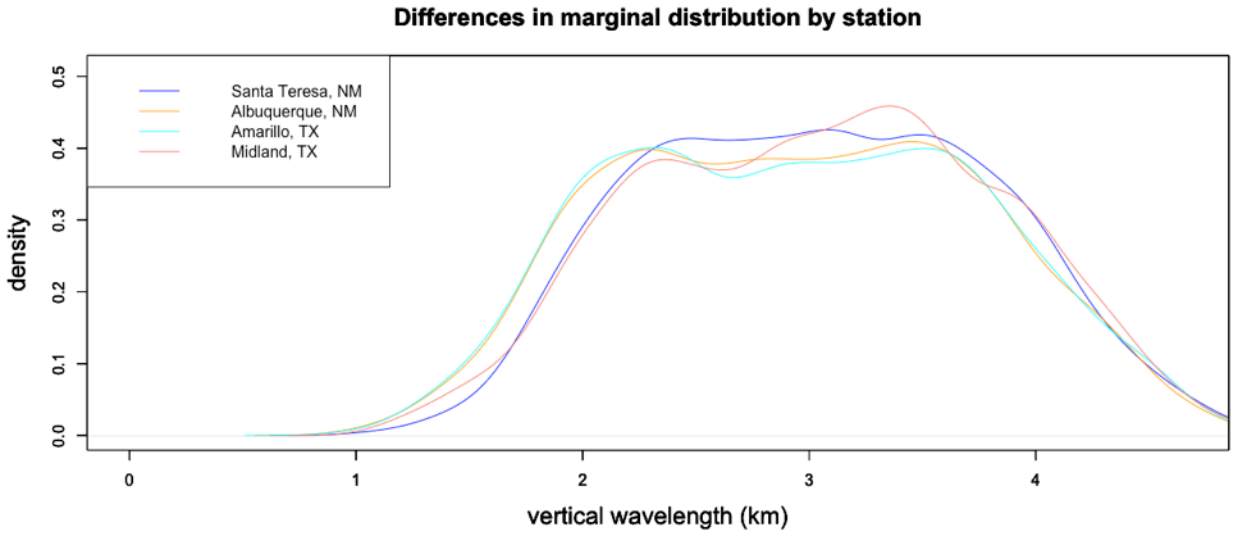

zonal wind amplitude: -6.6%
average difference between consecutive soundings (m/s)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
12
24
36
48
60
72
> 72
hours

(e)

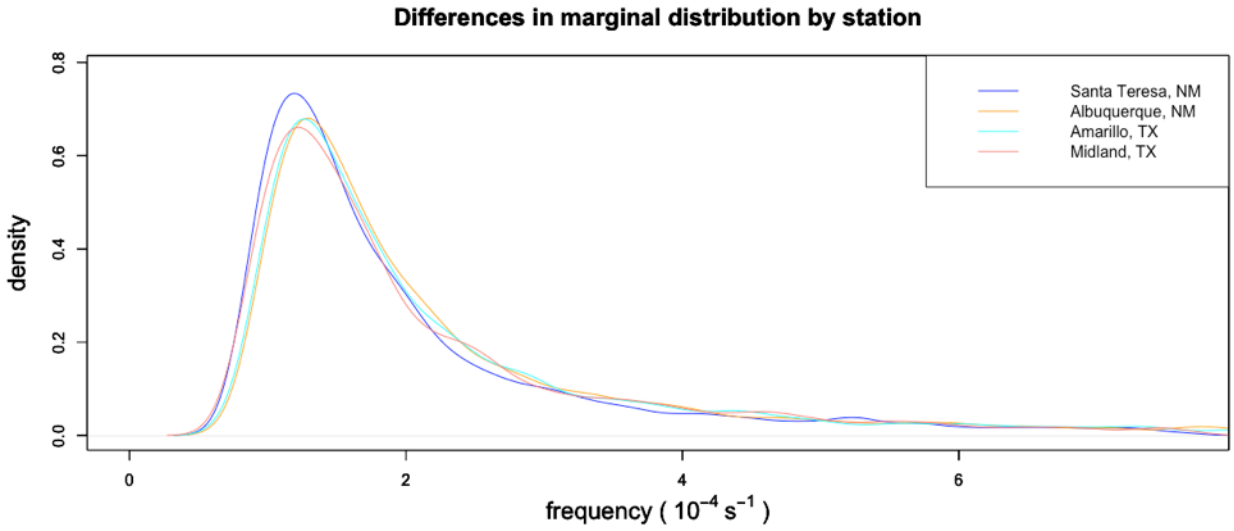


(f)

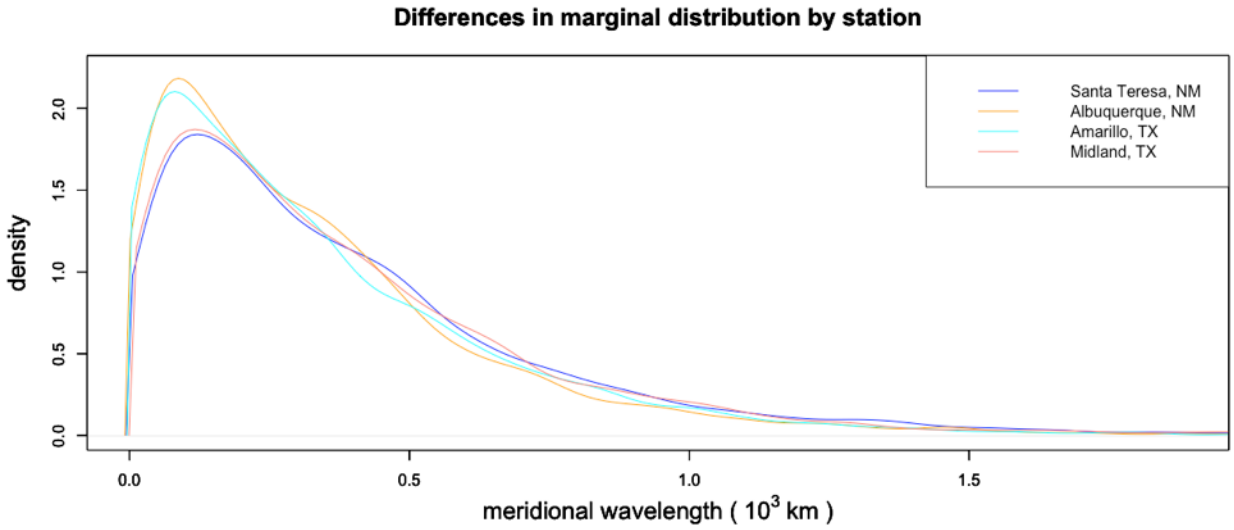
Figure A.2: Bar charts for the six gravity wave parameters that were not included in Figure 10 that illustrates how the average difference between consecutive soundings varies as a function of the number of hours between consecutive soundings. The percentage presented in the title of the graph indicates the percent difference between the average difference between consecutive soundings that take place 12 hours apart and the average difference between consecutive soundings across all soundings regardless of the number of hours between consecutive soundings.


Differences in marginal distribution by station
Santa Teresa, NM
Albuquerque, NM
Amarillo, TX
Midland, TX
density
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
vertical wavelength (km)

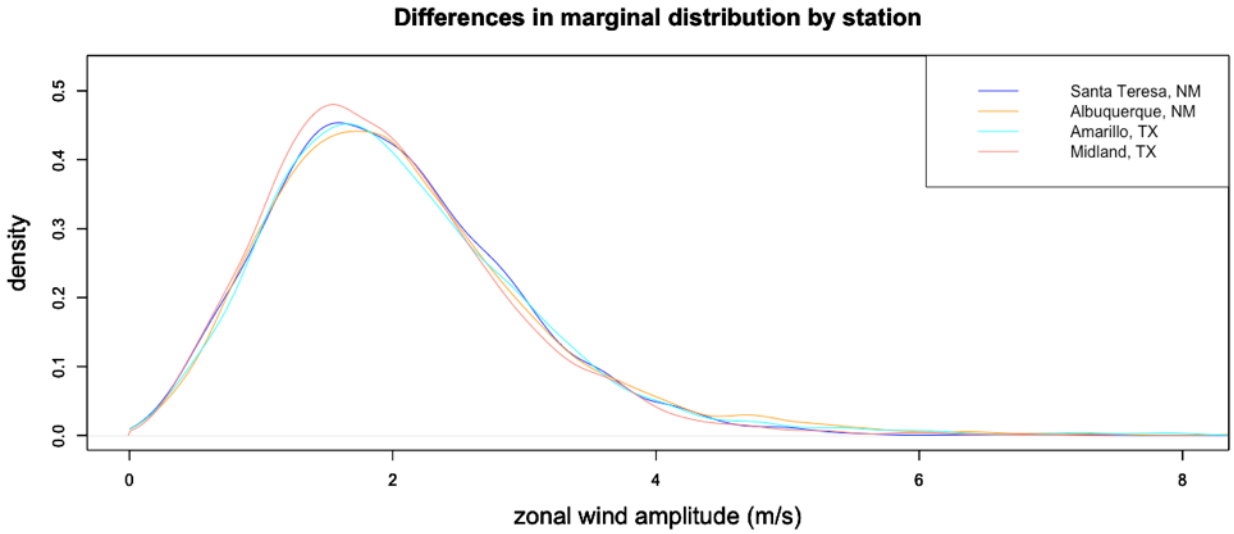
(a)


Differences in marginal distribution by station
Santa Teresa, NM
Albuquerque, NM
Amarillo, TX
Midland, TX
density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
frequency (10^-4 s^-1)

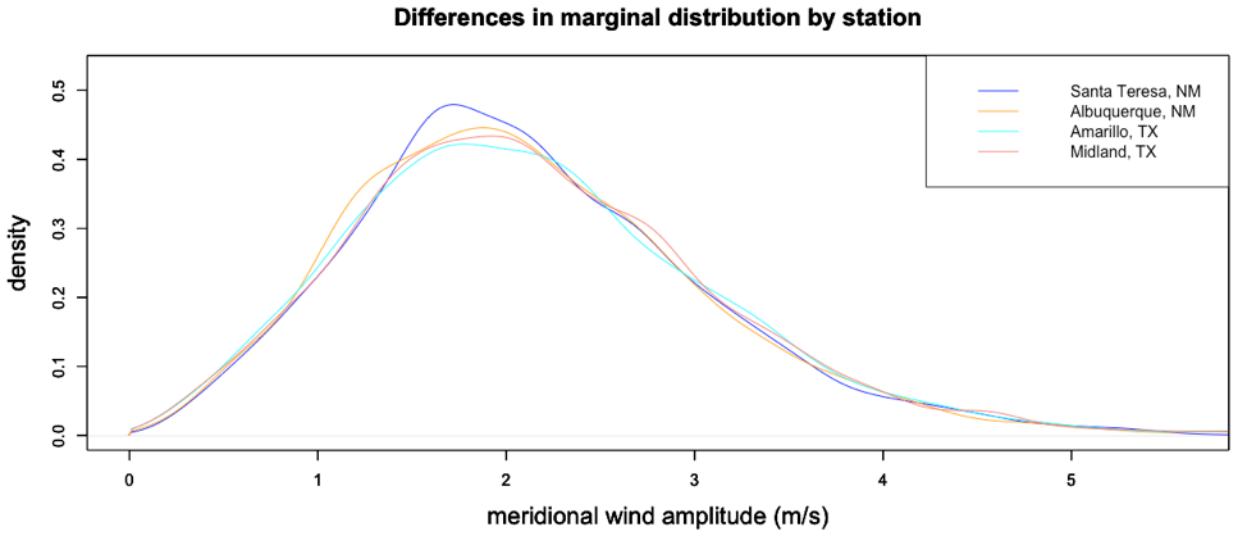
(b)


Differences in marginal distribution by station
density
meridional wavelength (10^3 km)
Santa Teresa, NM
Albuquerque, NM
Amarillo, TX
Midland, TX

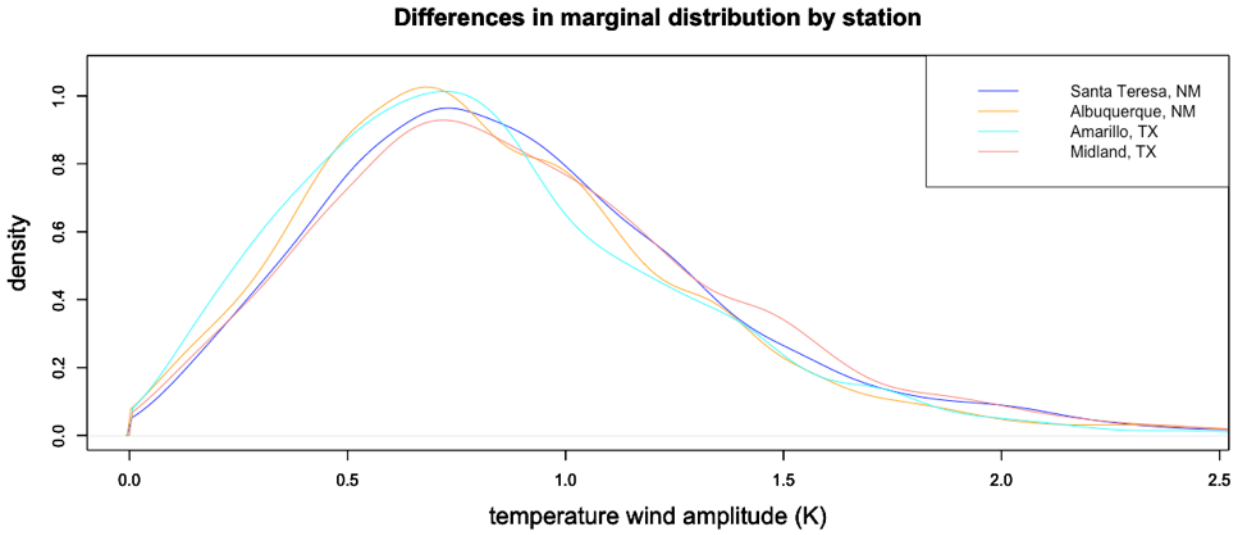
(c)


Differences in marginal distribution by station
density
zonal wind amplitude (m/s)
Santa Teresa, NM
Albuquerque, NM
Amarillo, TX
Midland, TX

(d)

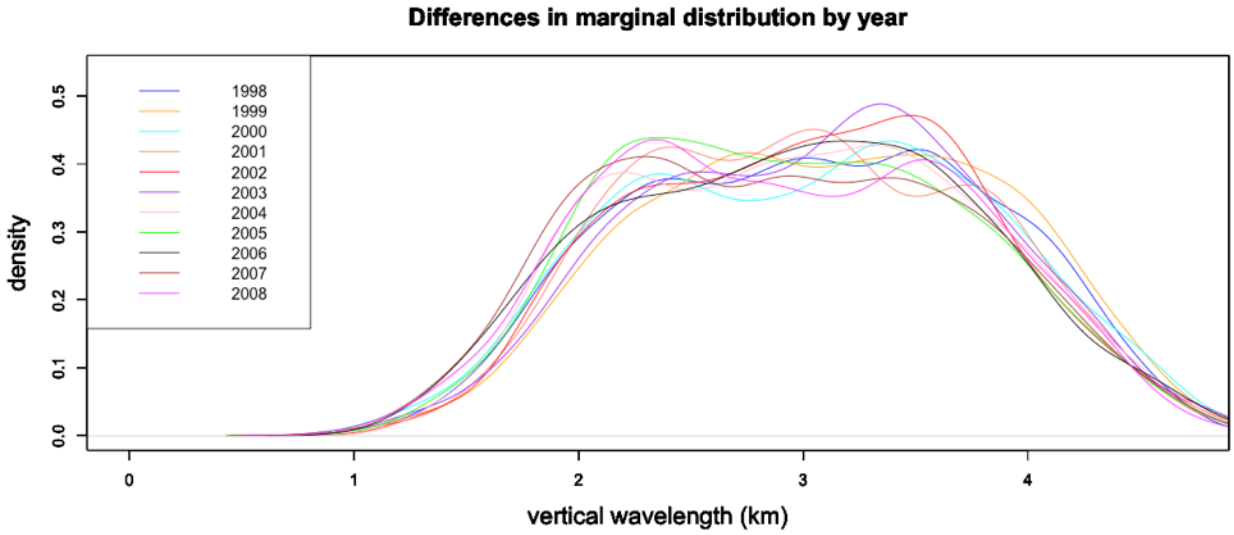


(e)

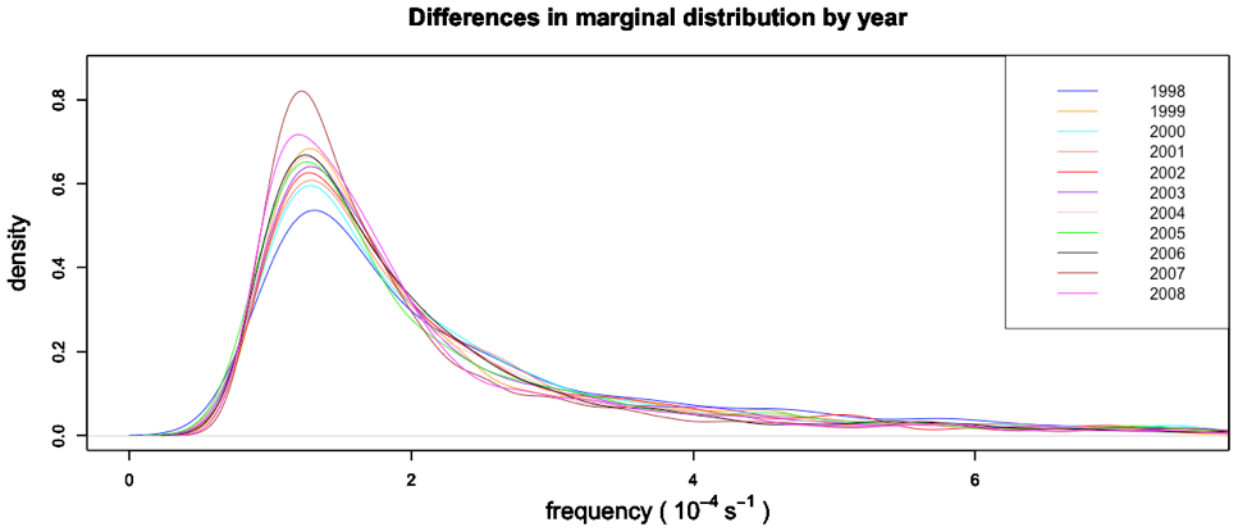


(f)

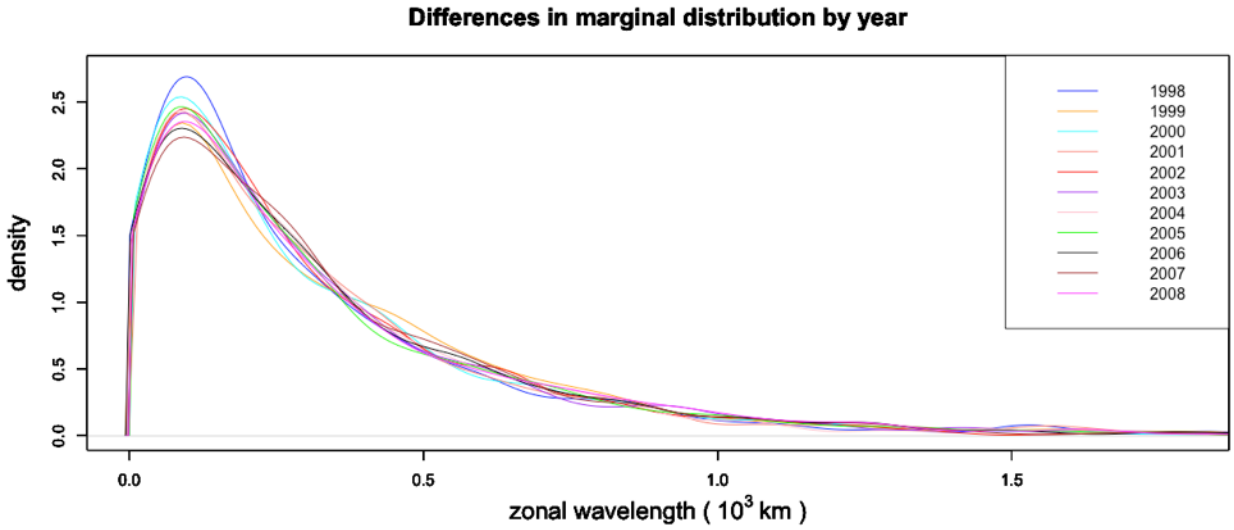
Figure A.3: Overlaid kernel density estimates corresponding to each of the four stations for all the gravity wave parameters that were not included in Figure 11. The overall shapes of the curves strongly coincide with each other and any minor deviations are attributed to statistical fluctuations.


Differences in marginal distribution by year
density
vertical wavelength (km)
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008

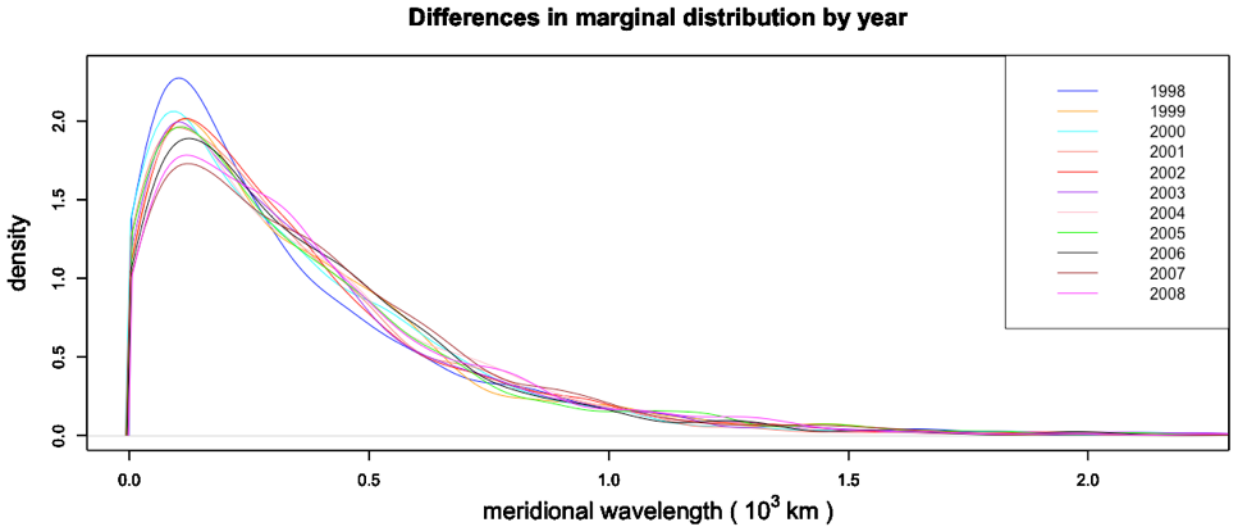
(a)


Differences in marginal distribution by year
density
frequency (10^{-4} s^{-1})
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008

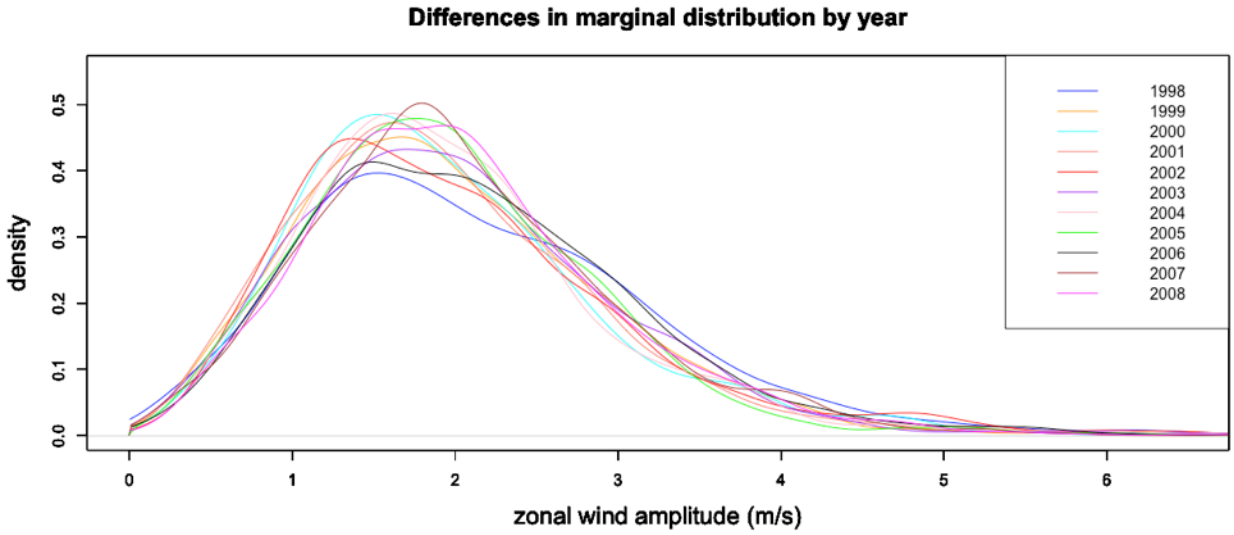
(b)


Differences in marginal distribution by year
density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.5
1.0
1.5
zonal wavelength (10^3 km)
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008

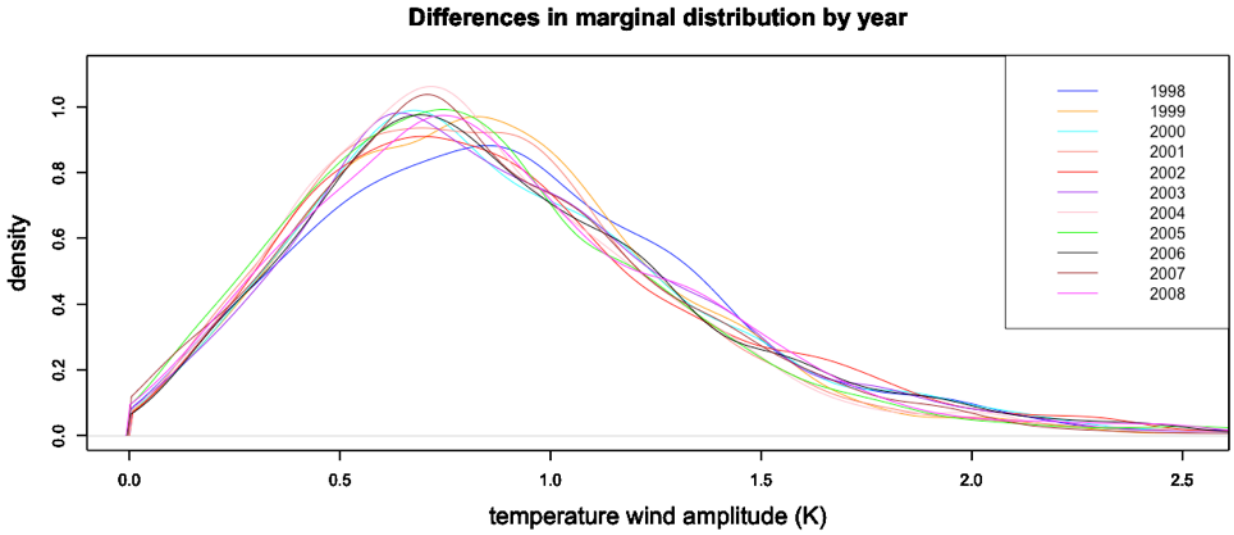
(c)


Differences in marginal distribution by year
density
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0
meridional wavelength (10^3 km)
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008

(d)

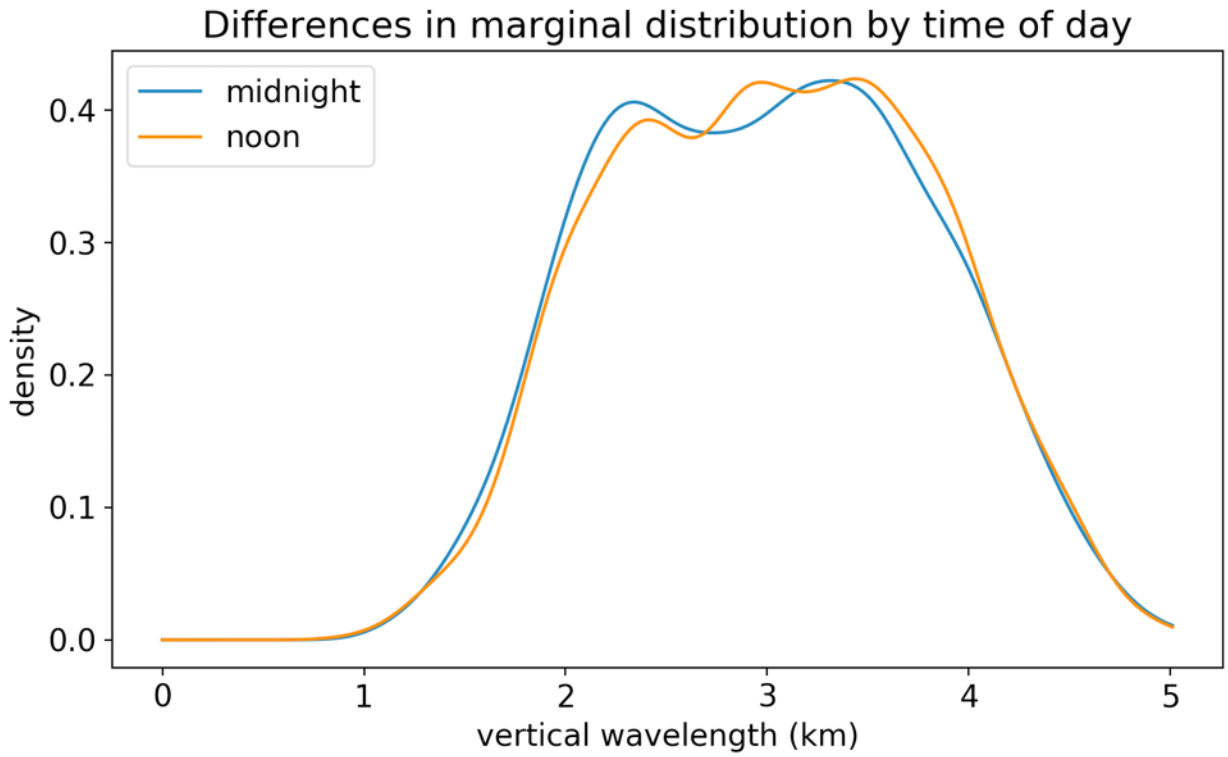


(e)

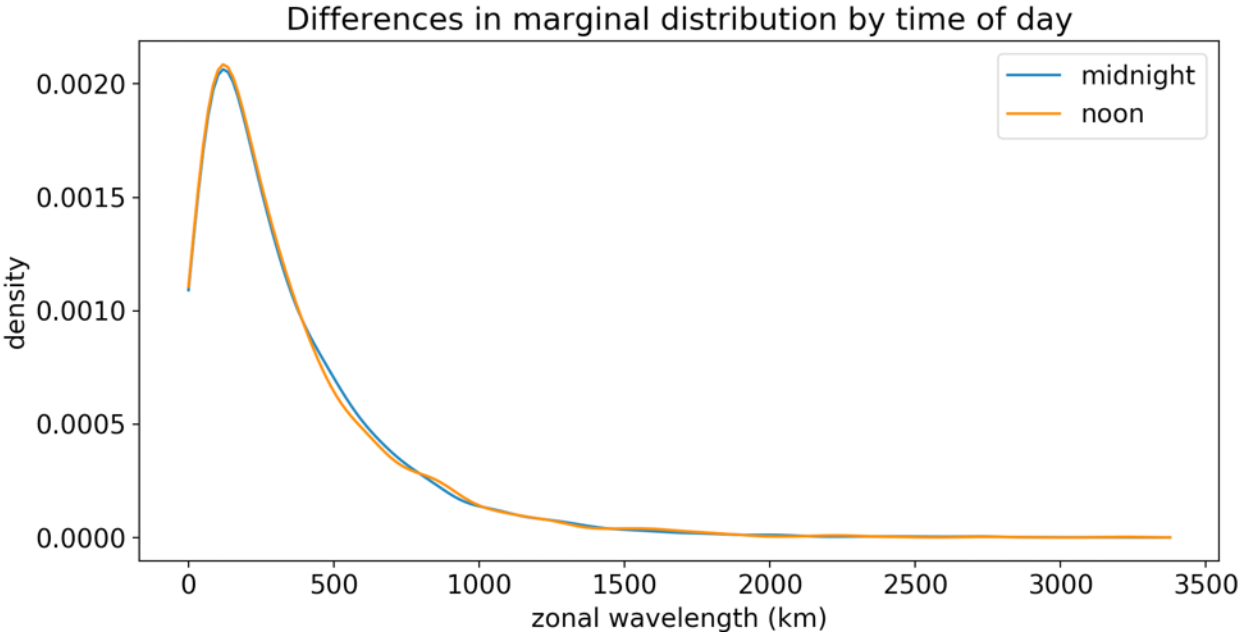


(f)

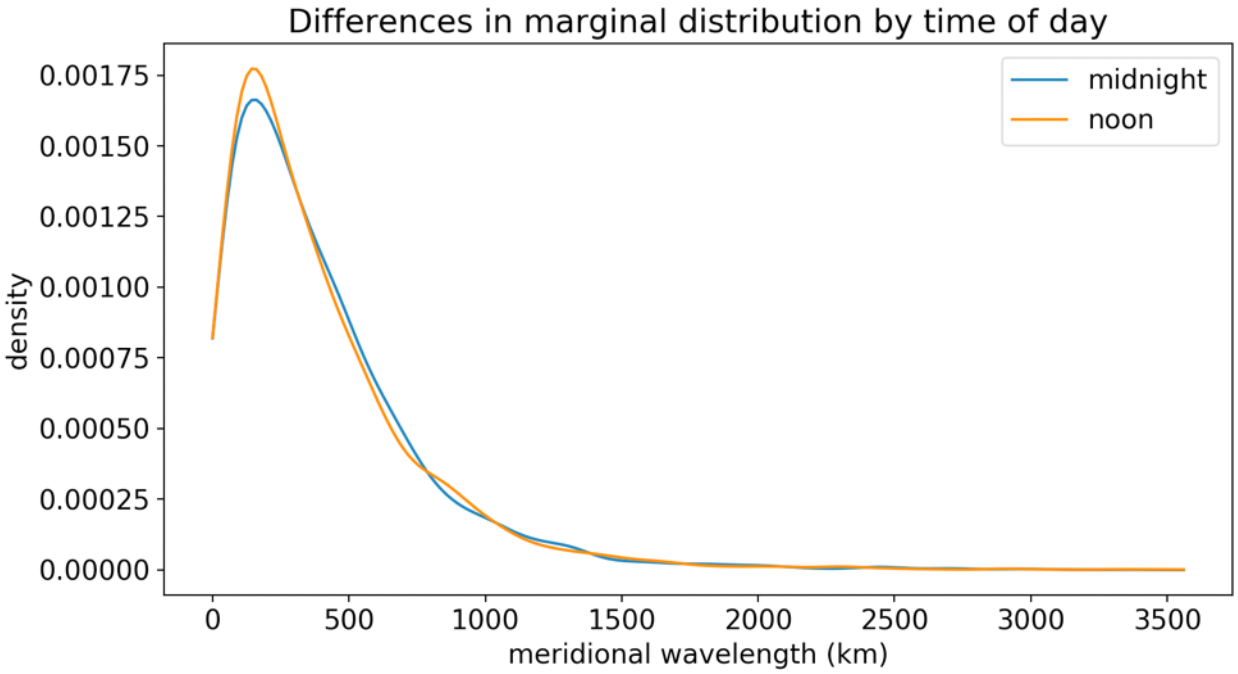
Figure A.4: Overlaid kernel density estimates corresponding to each year between 1998 and 2008 for all the gravity wave parameters that were not included in Figure 12. The overall shapes of the curves strongly coincide with each other and any minor deviations are attributed to statistical fluctuations.


Differences in marginal distribution by time of day
midnight
noon
density
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
vertical wavelength (km)

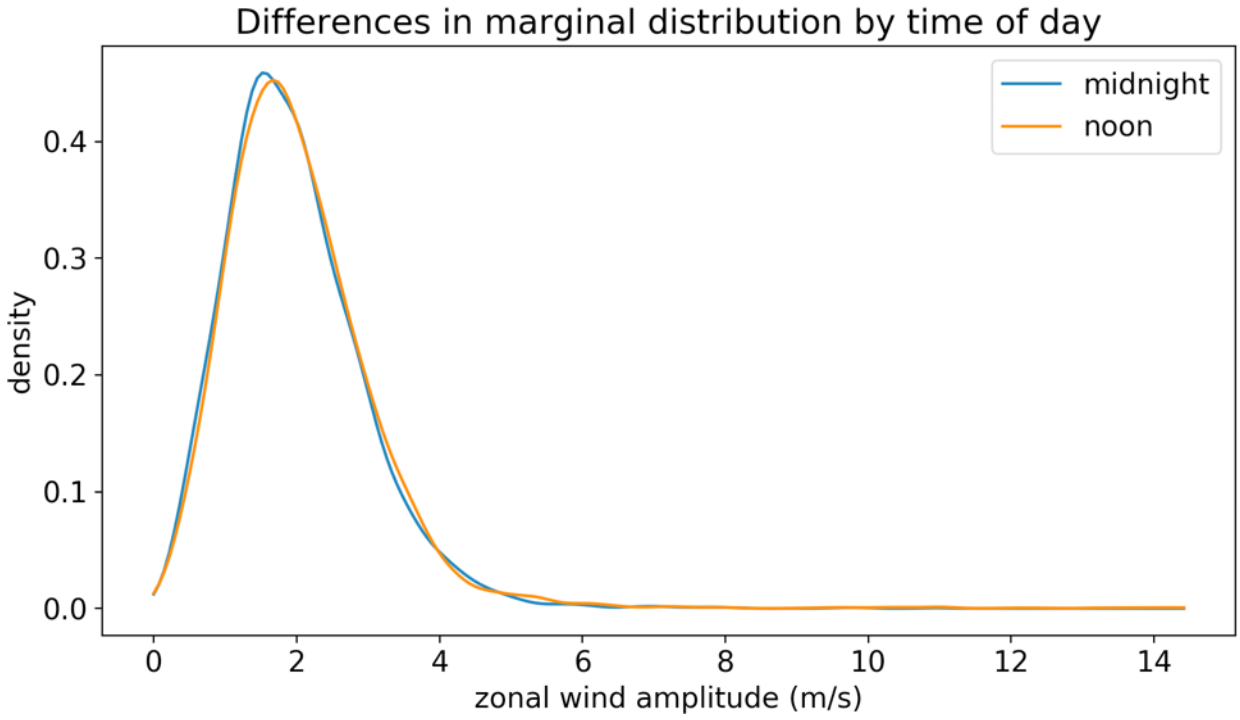
(a)


Differences in marginal distribution by time of day
midnight
noon
density
0.0000
0.0005
0.0010
0.0015
0.0020
0
500
1000
1500
2000
2500
3000
3500
zonal wavelength (km)

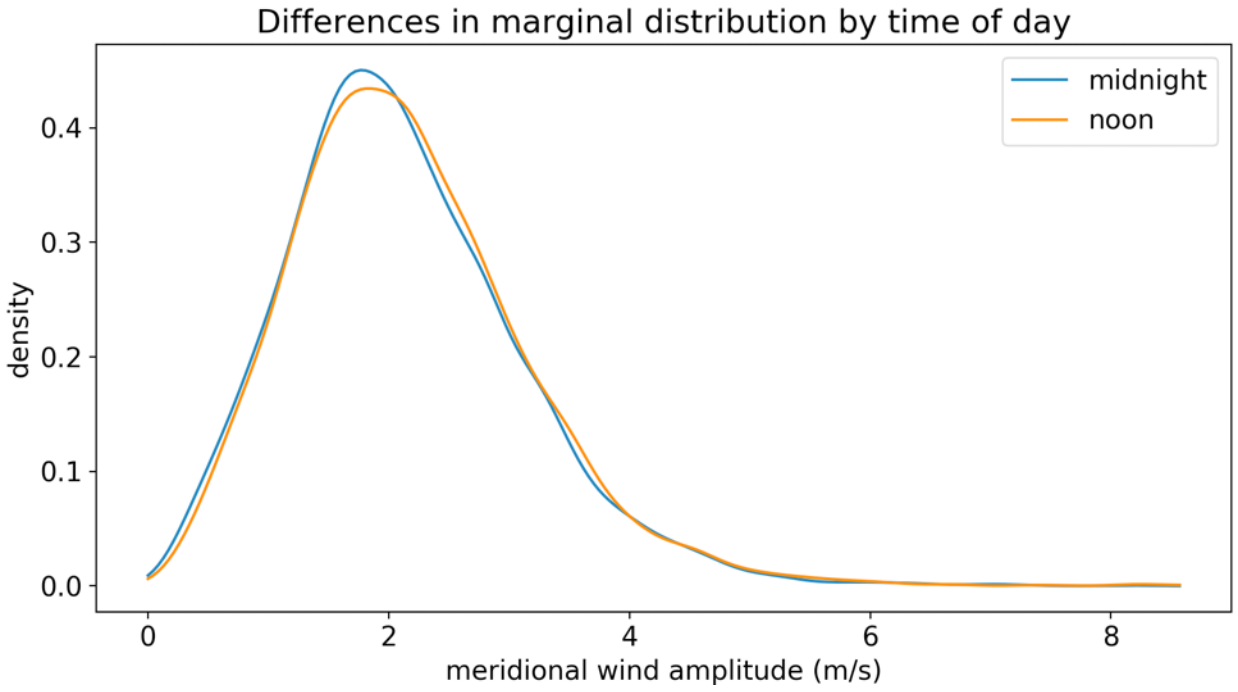
(b)


Differences in marginal distribution by time of day
midnight
noon
density
0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
0
500
1000
1500
2000
2500
3000
3500
meridional wavelength (km)

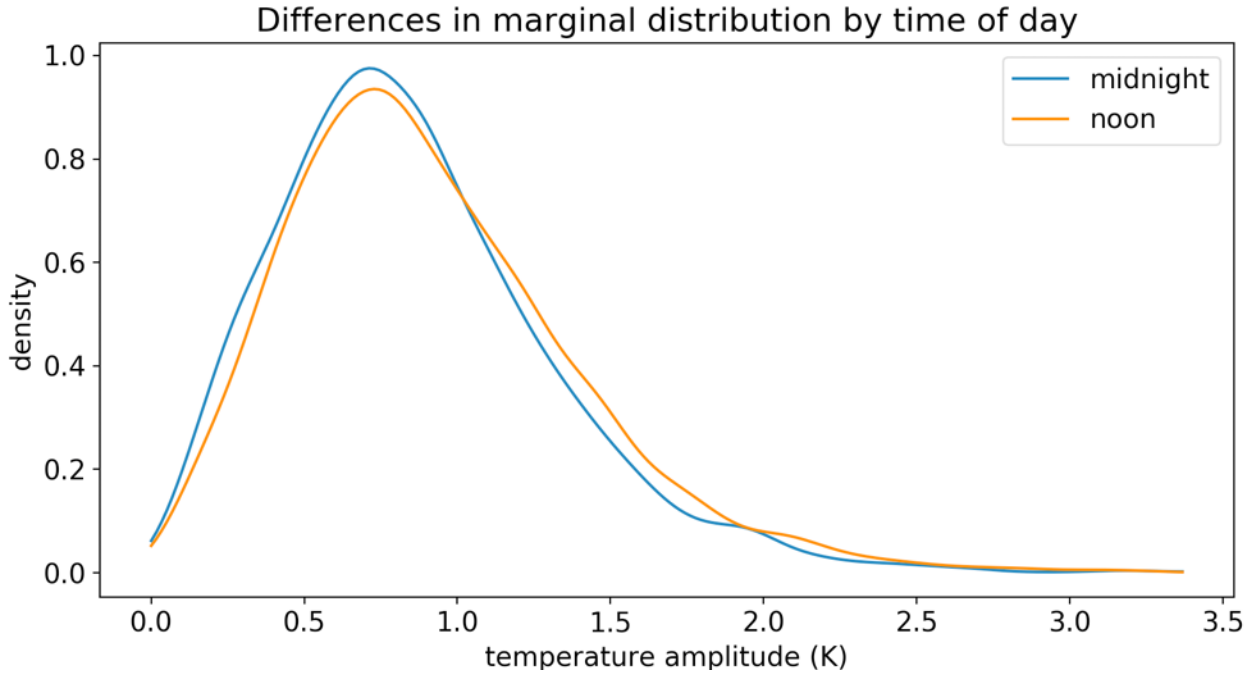
(c)


Differences in marginal distribution by time of day
midnight
noon
density
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
10
12
14
zonal wind amplitude (m/s)

(d)

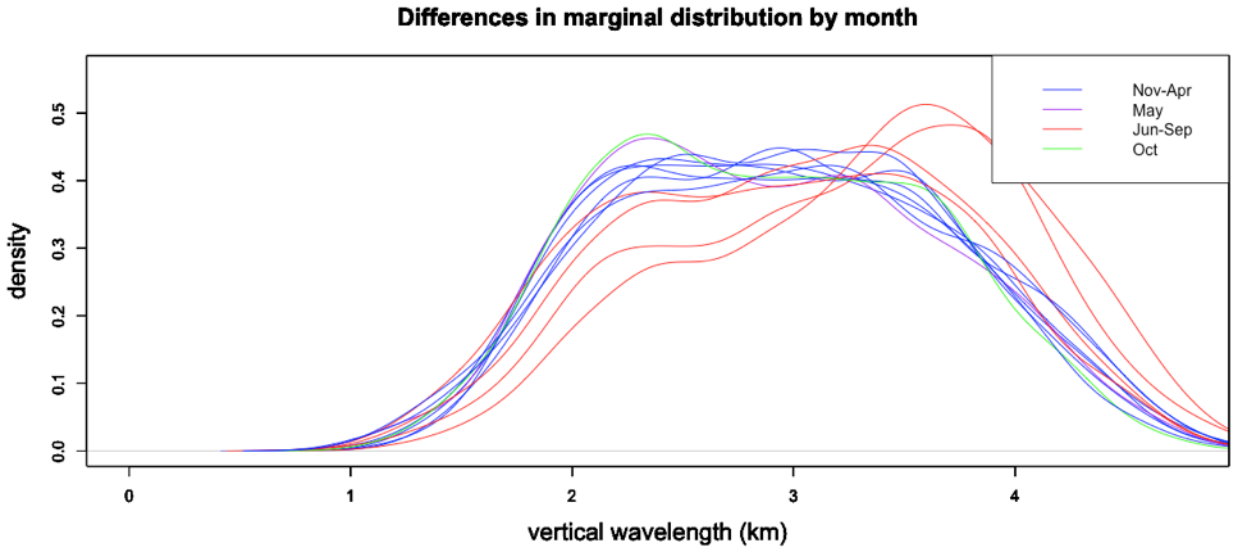


(e)

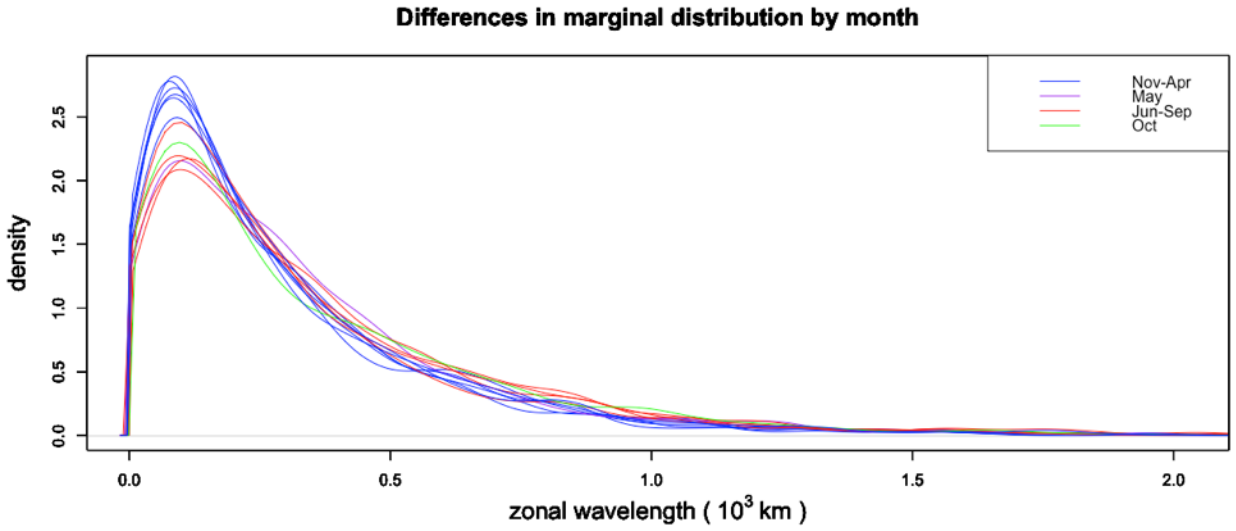


(f)

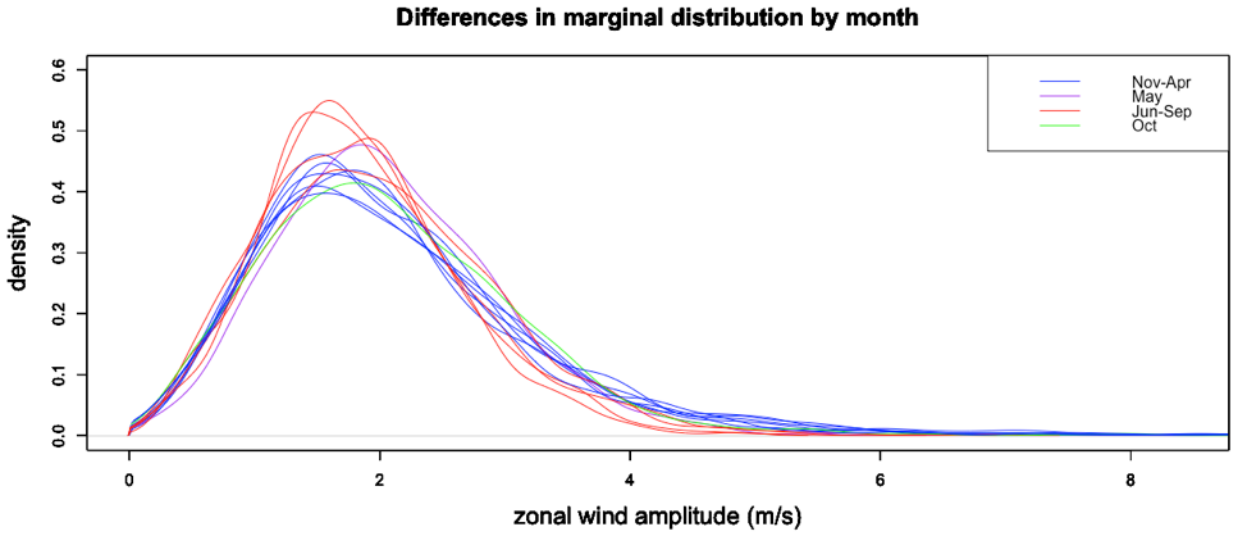
Figure A.5: Overlaid kernel density estimates corresponding to noon and midnight for all the gravity wave parameters that were not included in Figure 13. The overall shapes of the curves strongly coincide with each other and any minor deviations are attributed to statistical fluctuations.


Differences in marginal distribution by month
density
vertical wavelength (km)
Nov-Apr
May
Jun-Sep
Oct
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4

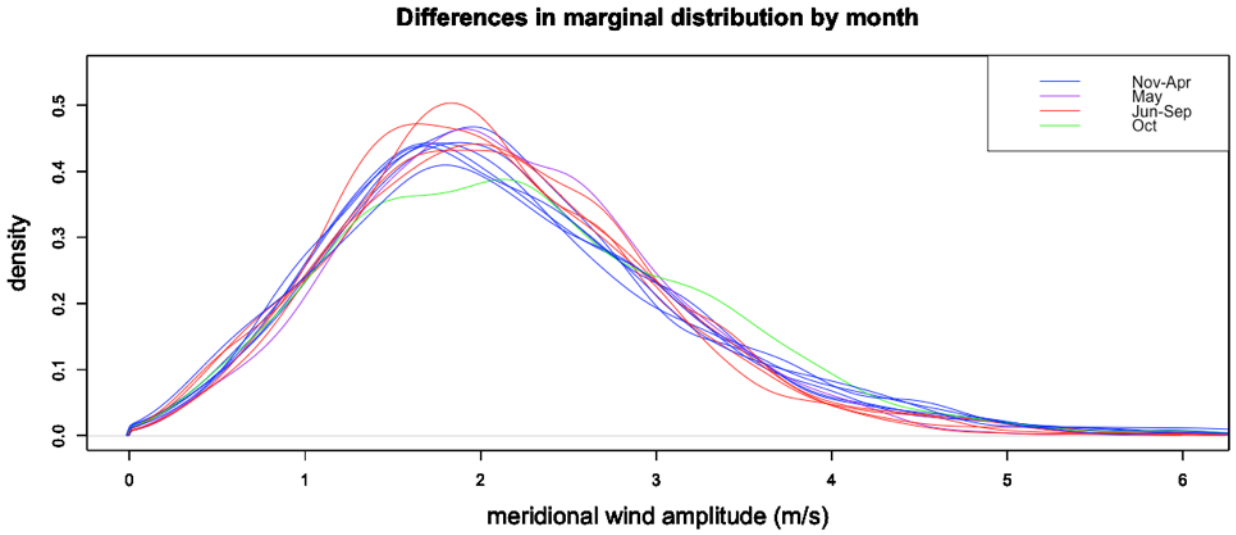
(a)


Differences in marginal distribution by month
density
zonal wavelength (10^3 km)
Nov-Apr
May
Jun-Sep
Oct
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.5
1.0
1.5
2.0

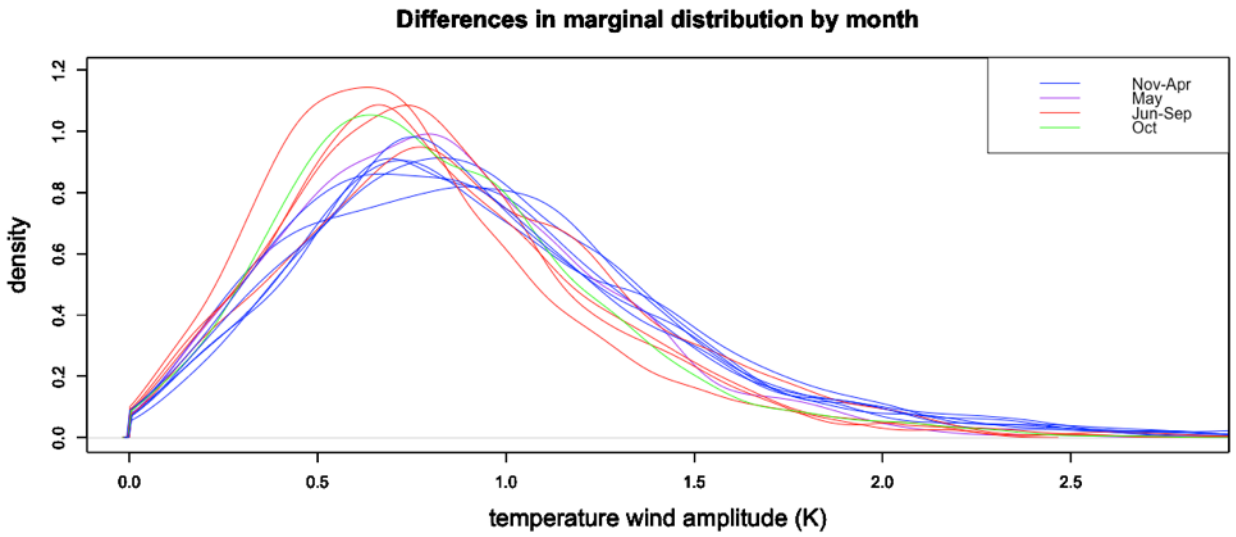
(b)


Differences in marginal distribution by month
density
zonal wind amplitude (m/s)
Nov-Apr
May
Jun-Sep
Oct
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8

(c)

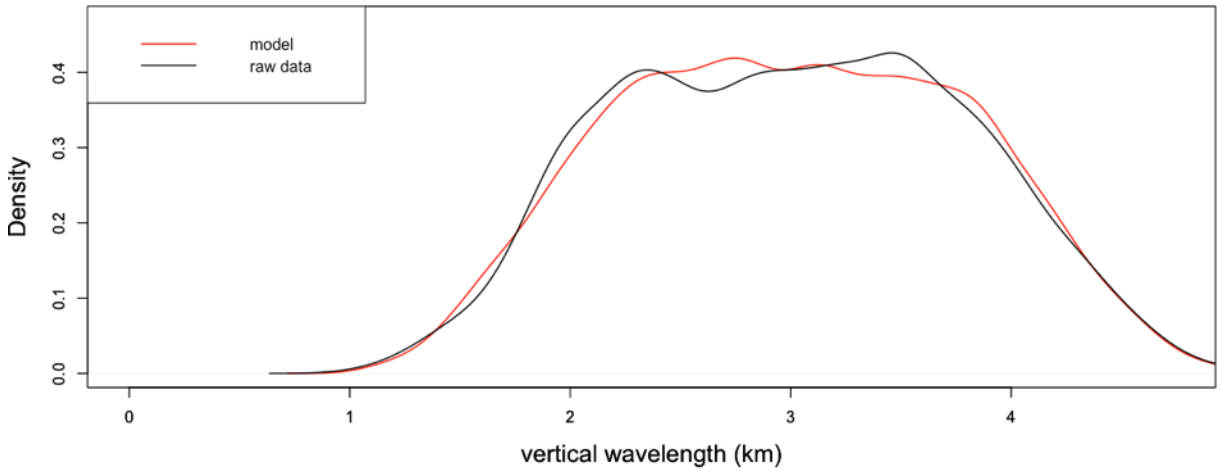

Differences in marginal distribution by month
density
meridional wind amplitude (m/s)
Nov-Apr
May
Jun-Sep
Oct
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6

(d)

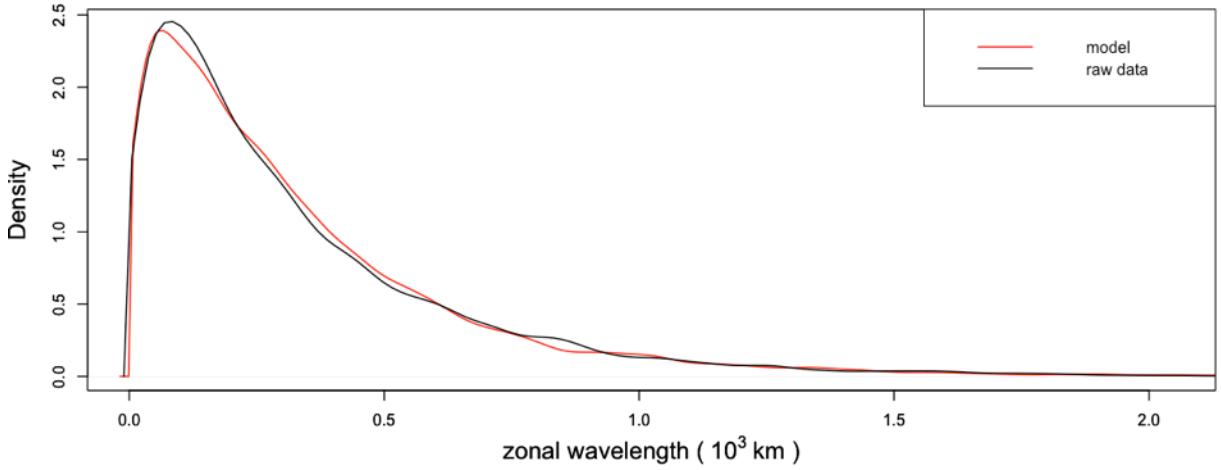


(e)

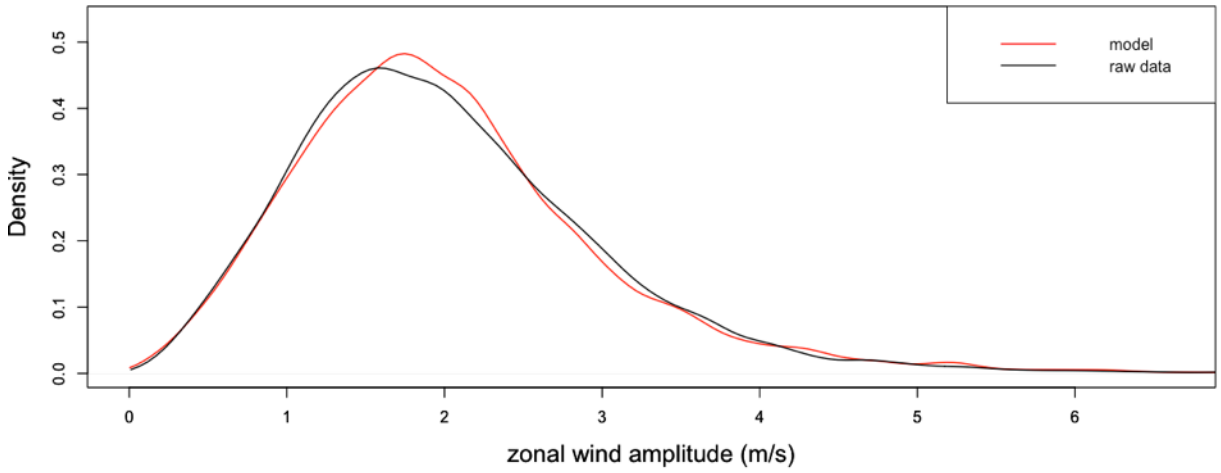
Figure A.6: Overlaid kernel density estimates corresponding to each month of the year for all the gravity wave parameters that were not included in Figure 14. In each graph, the blue curves correspond to the months between November and April (inclusive), the purple curve corresponds to the month of May, the red months correspond to the months between June and September (inclusive) and the green curve corresponds to the month of October.


model
raw data
Density
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
vertical wavelength (km)

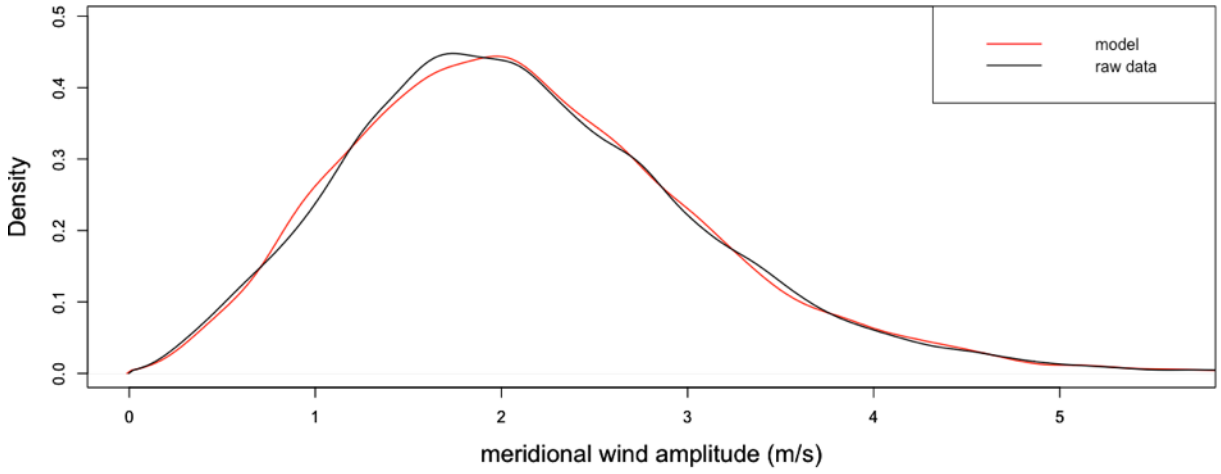
(a)


model
raw data
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.5
1.0
1.5
2.0
zonal wavelength (10^3 km)

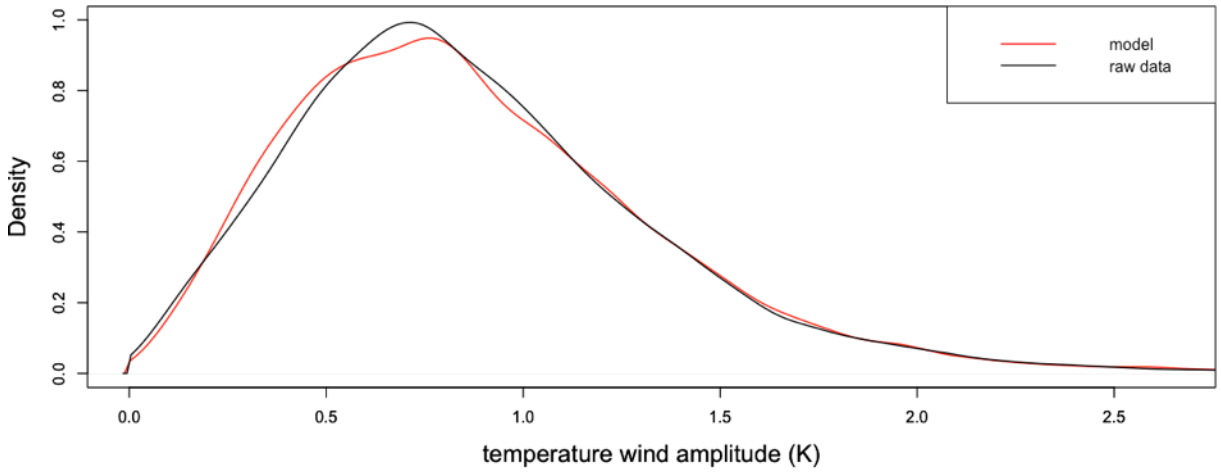
(b)


model
raw data
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6
zonal wind amplitude (m/s)

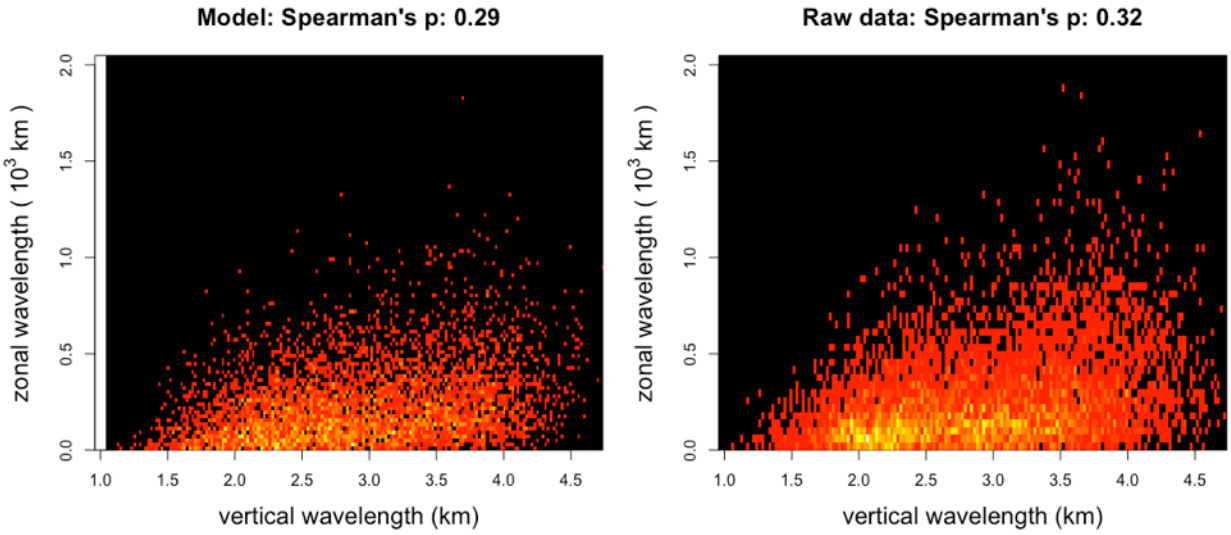
(c)


model
raw data
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
meridional wind amplitude (m/s)

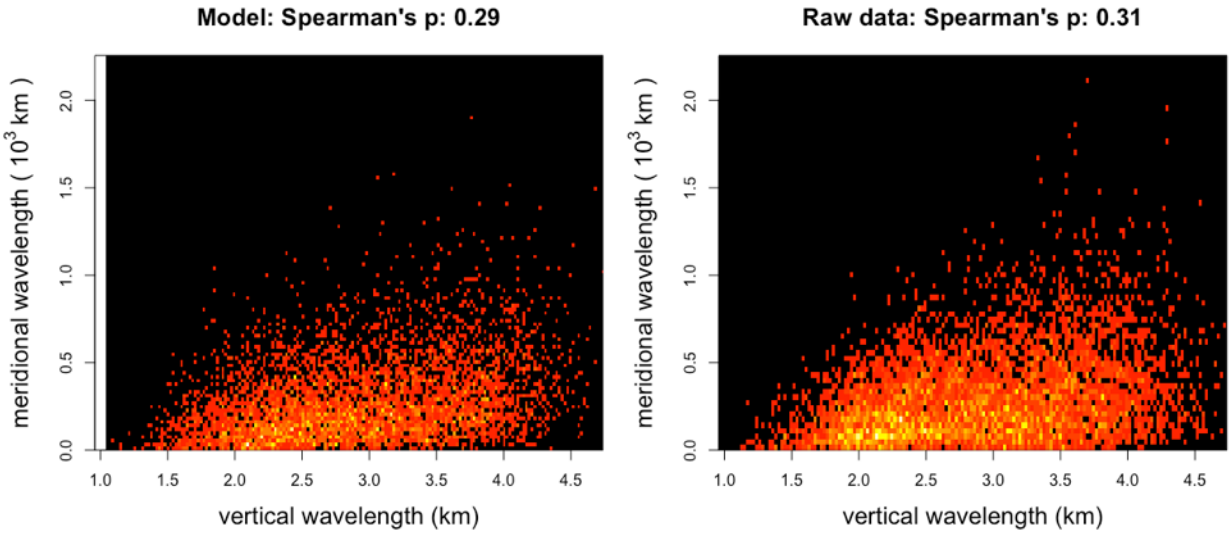
(d)


model
raw data
Density
temperature wind amplitude (K)

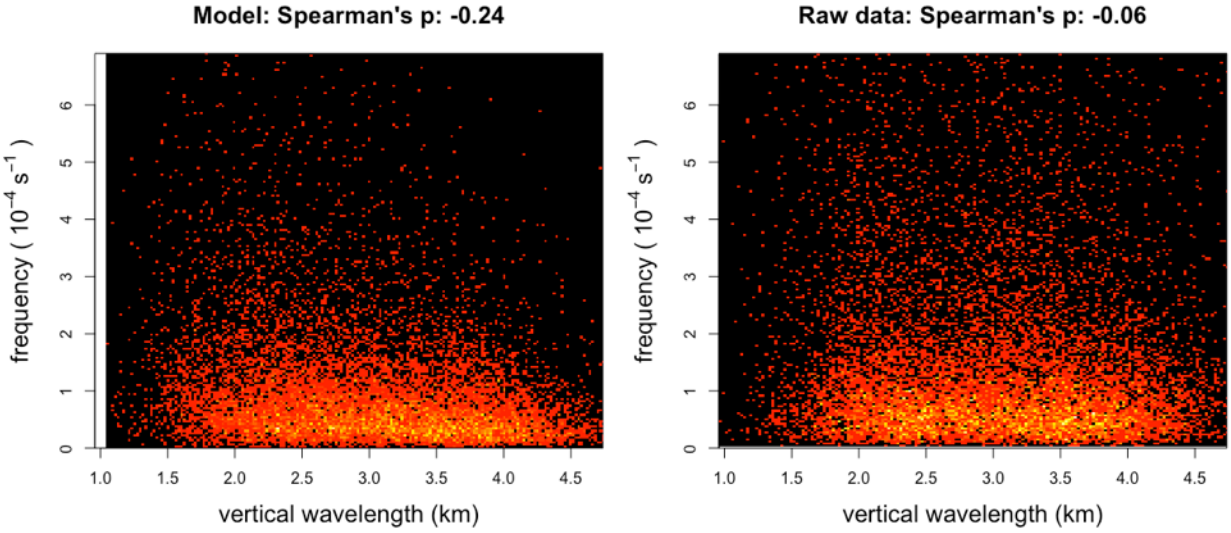
(e)


Model: Spearman's p: 0.29
Raw data: Spearman's p: 0.32
zonal wavelength (10^3 km)
vertical wavelength (km)

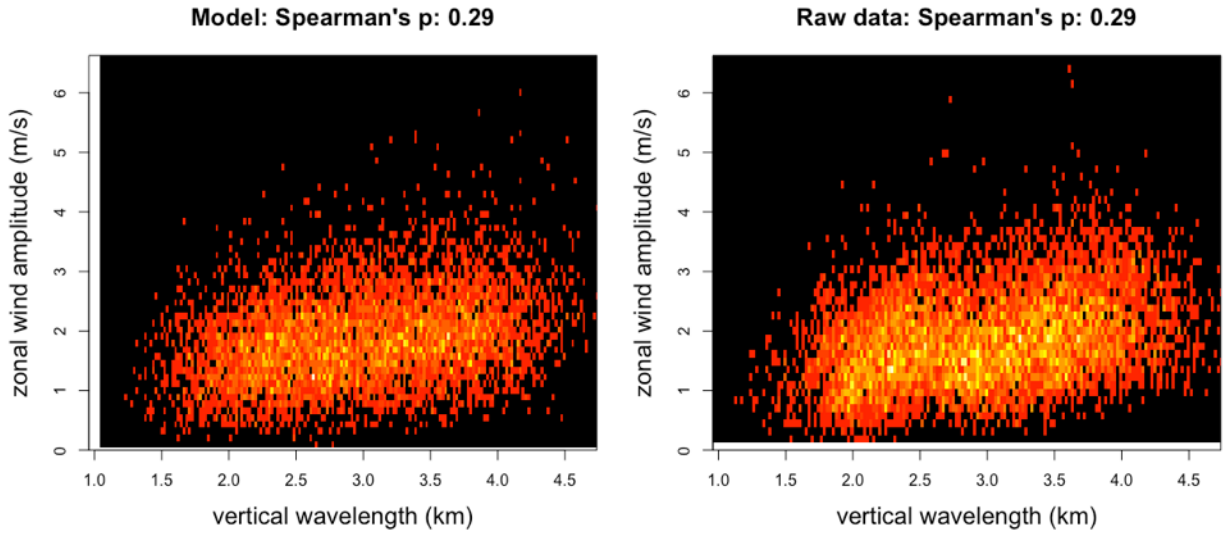
(f)


Model: Spearman's p: 0.29
Raw data: Spearman's p: 0.31
meridional wavelength (10^3 km)
vertical wavelength (km)
0.0
0.5
1.0
1.5
2.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

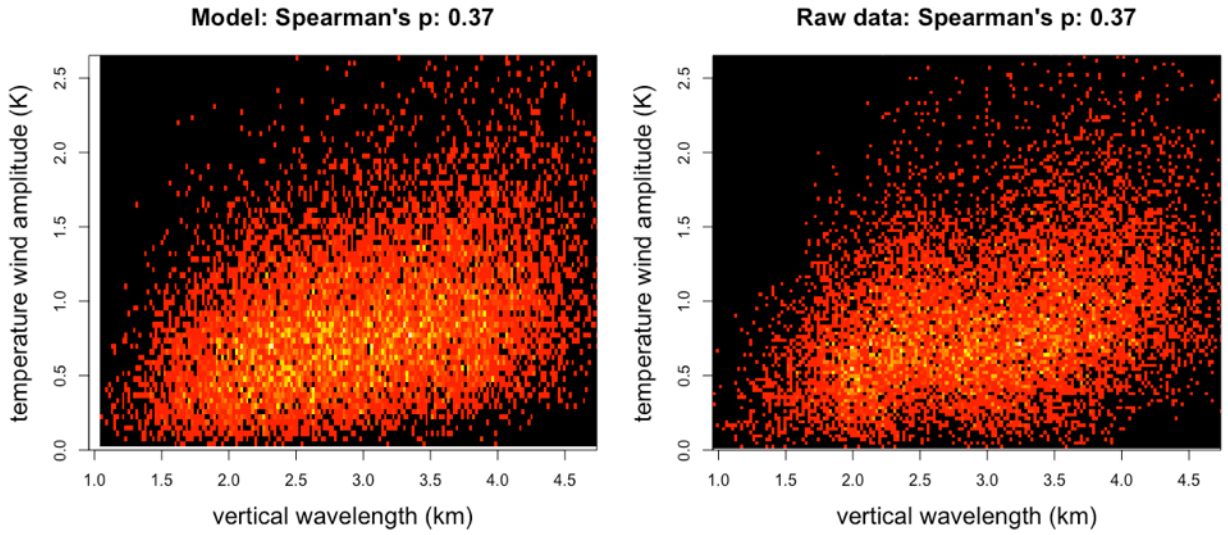
(g)


Model: Spearman's p: -0.24
Raw data: Spearman's p: -0.06
frequency (10^{-4} s^{-1})
vertical wavelength (km)
0
1
2
3
4
5
6
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

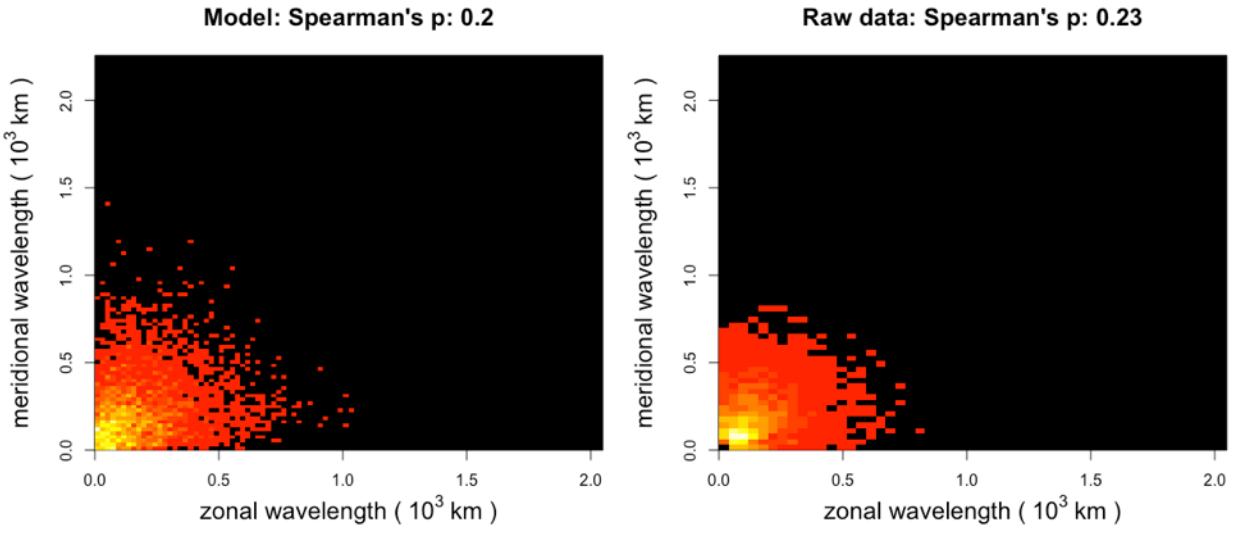
(h)


Model: Spearman's p: 0.29
Raw data: Spearman's p: 0.29
zonal wind amplitude (m/s)
vertical wavelength (km)
0
1
2
3
4
5
6
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

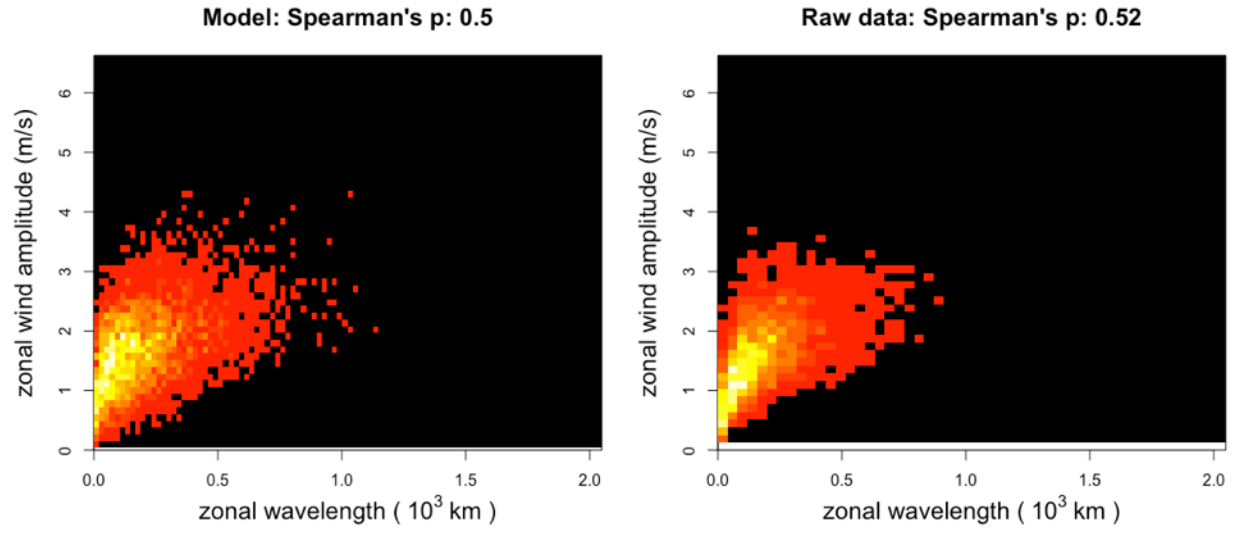
(i)


Model: Spearman's p: 0.37
Raw data: Spearman's p: 0.37
temperature wind amplitude (K)
vertical wavelength (km)
0.0
0.5
1.0
1.5
2.0
2.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

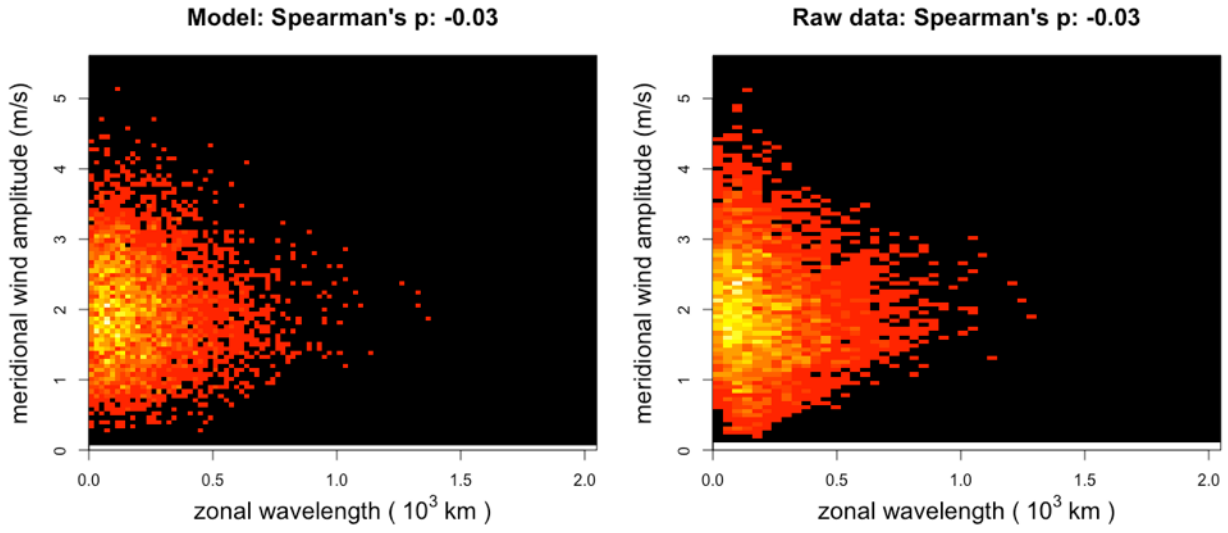
(j)


Model: Spearman's p: 0.2
Raw data: Spearman's p: 0.23
meridional wavelength (10^3 km)
zonal wavelength (10^3 km)
0.0
0.5
1.0
1.5
2.0

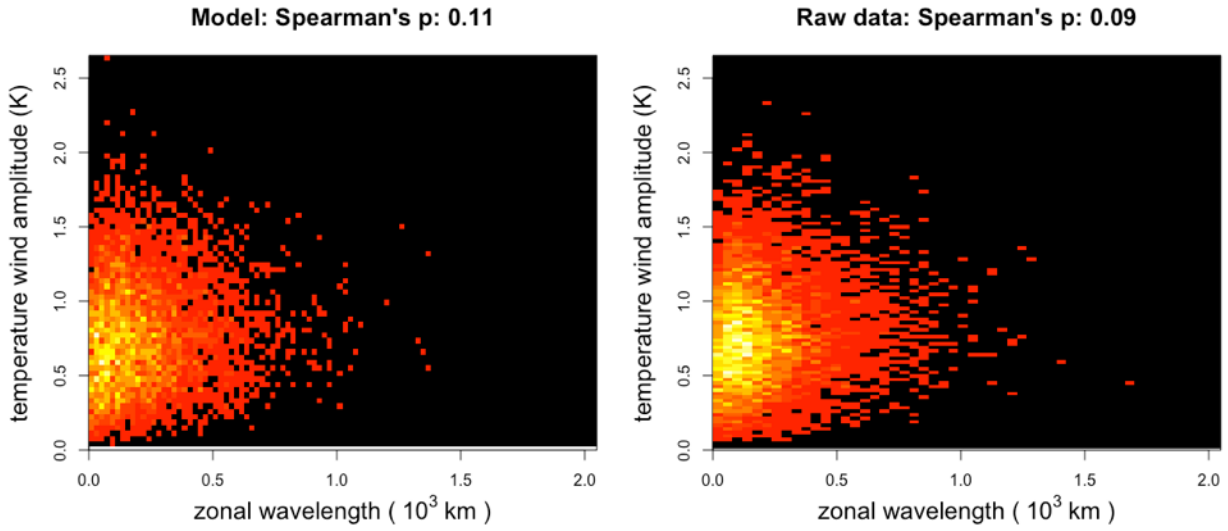
(k)


Model: Spearman's p: 0.5
Raw data: Spearman's p: 0.52
zonal wind amplitude (m/s)
zonal wavelength (10^3 km)
0
1
2
3
4
5
6
0.0
0.5
1.0
1.5
2.0

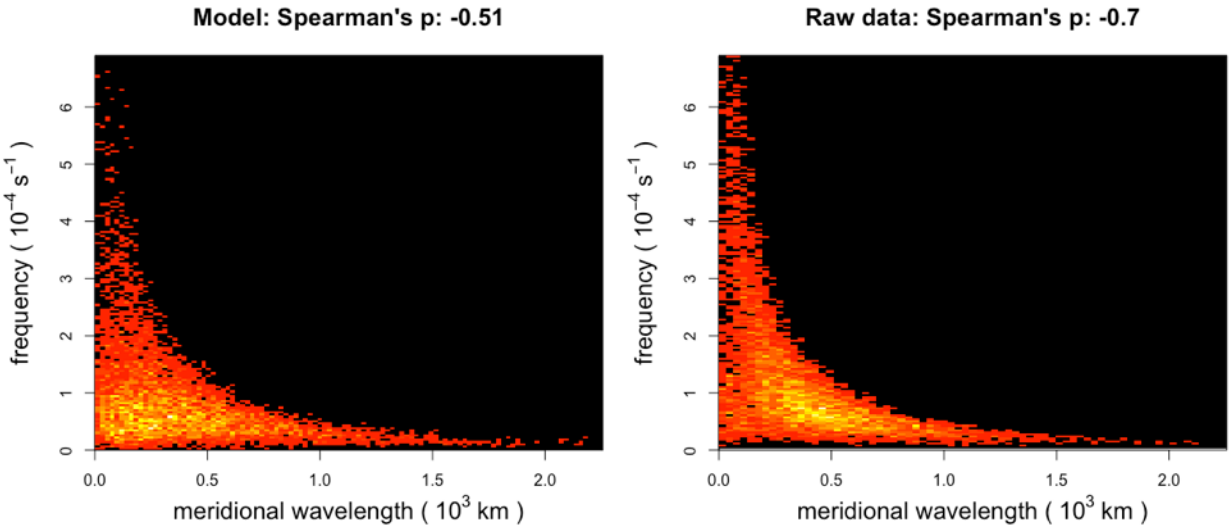
(l)


Model: Spearman's p: -0.03
Raw data: Spearman's p: -0.03
meridional wind amplitude (m/s)
zonal wavelength (10^3 km)
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5

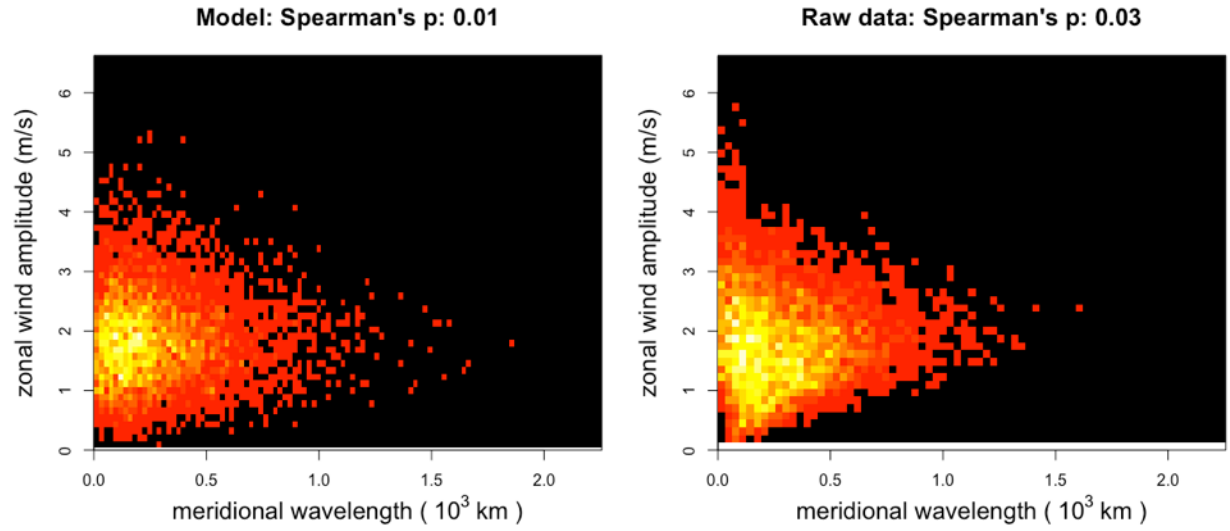
(m)


Model: Spearman's p: 0.11
Raw data: Spearman's p: 0.09
temperature wind amplitude (K)
zonal wavelength (10^3 km)
0.0
0.5
1.0
1.5
2.0
2.5

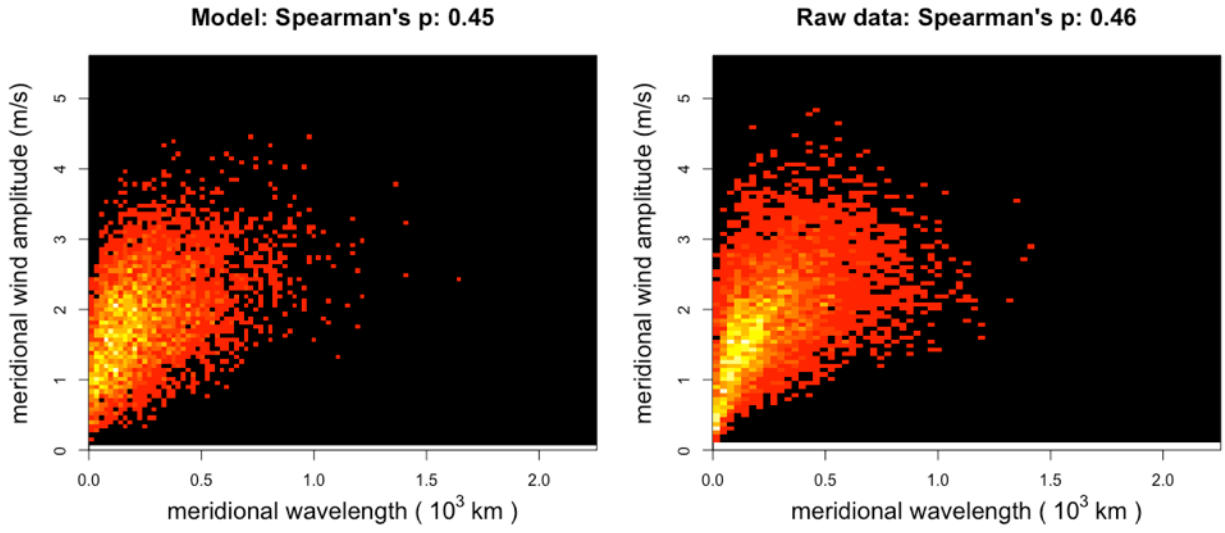
(n)


Model: Spearman's p: -0.51
Raw data: Spearman's p: -0.7
frequency (10^{-4} s^{-1})
meridional wavelength (10^3 km)
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
6

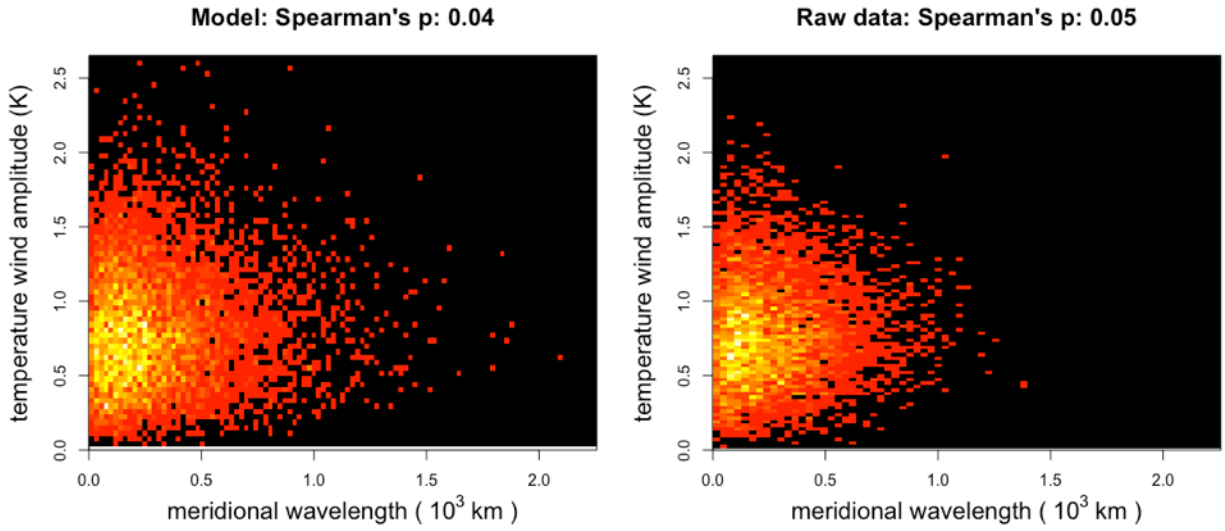
(o)


Model: Spearman's p: 0.01
Raw data: Spearman's p: 0.03
zonal wind amplitude (m/s)
meridional wavelength (10^3 km)
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
6

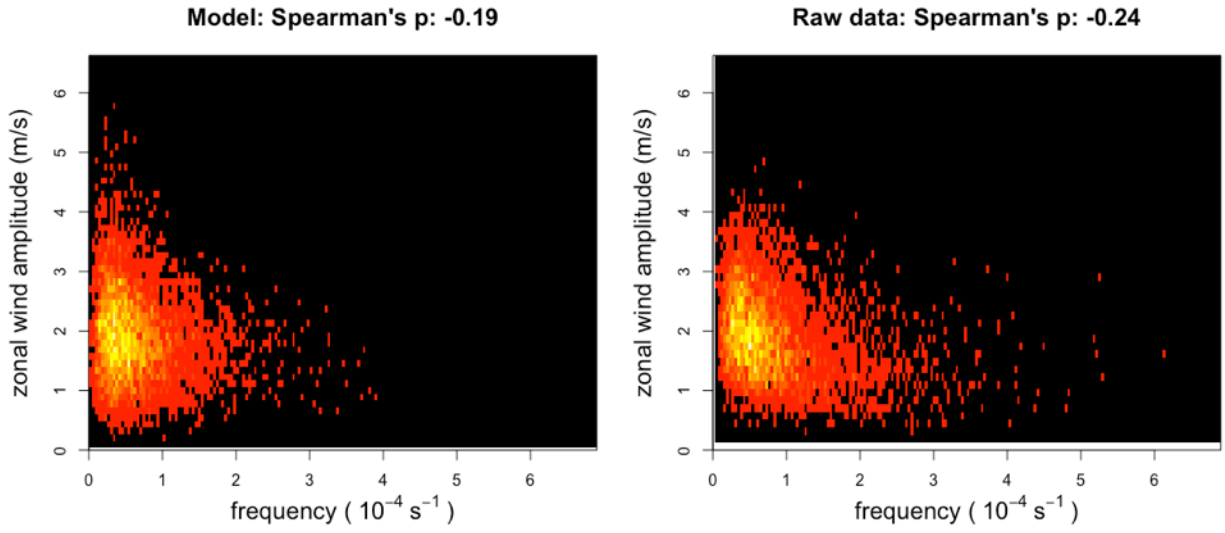
(p)


Model: Spearman's p: 0.45
Raw data: Spearman's p: 0.46
meridional wind amplitude (m/s)
meridional wavelength (10^3 km)
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5

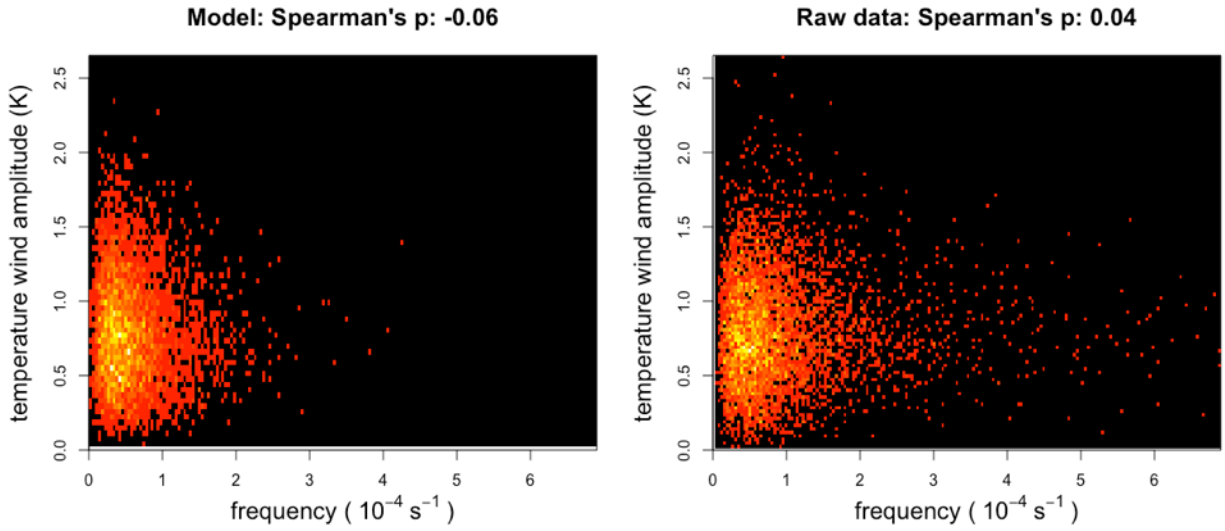
(q)


Model: Spearman's p: 0.04
Raw data: Spearman's p: 0.05
temperature wind amplitude (K)
meridional wavelength (10^3 km)
0.0
0.5
1.0
1.5
2.0
2.5

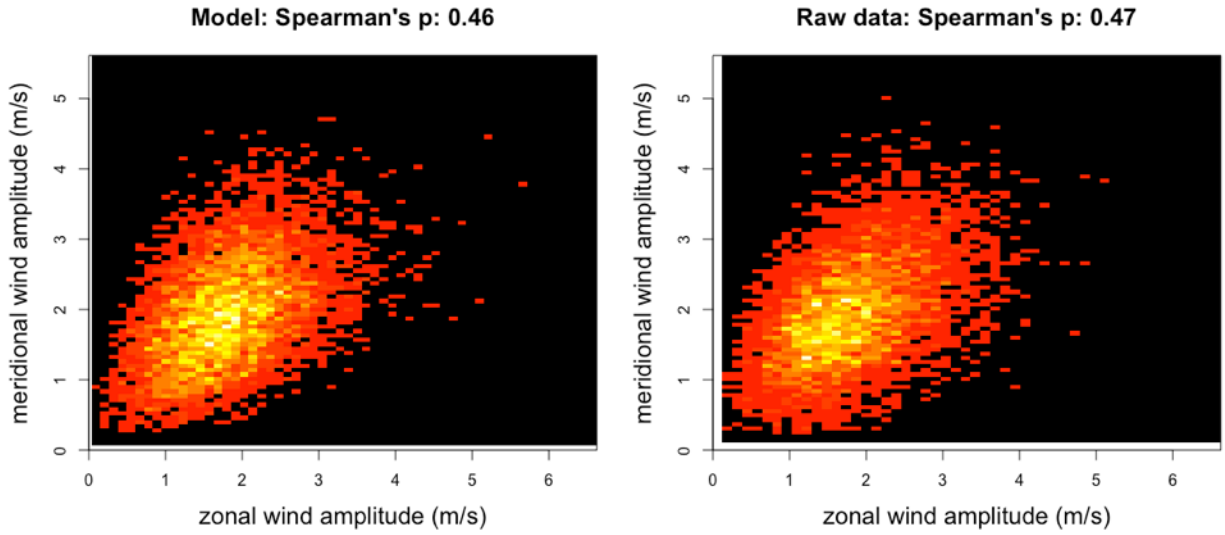
(r)


Model: Spearman's p: -0.19
Raw data: Spearman's p: -0.24
zonal wind amplitude (m/s)
frequency (10^{-4} s^{-1})

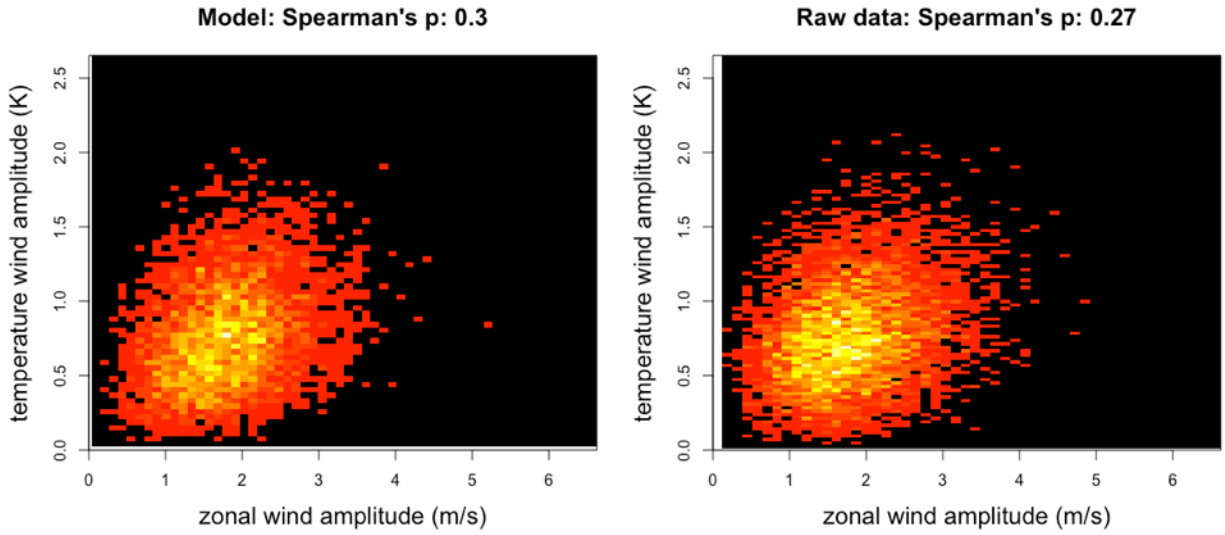
(s)


Model: Spearman's p: -0.06
Raw data: Spearman's p: 0.04
temperature wind amplitude (K)
frequency (10^{-4} s^{-1})

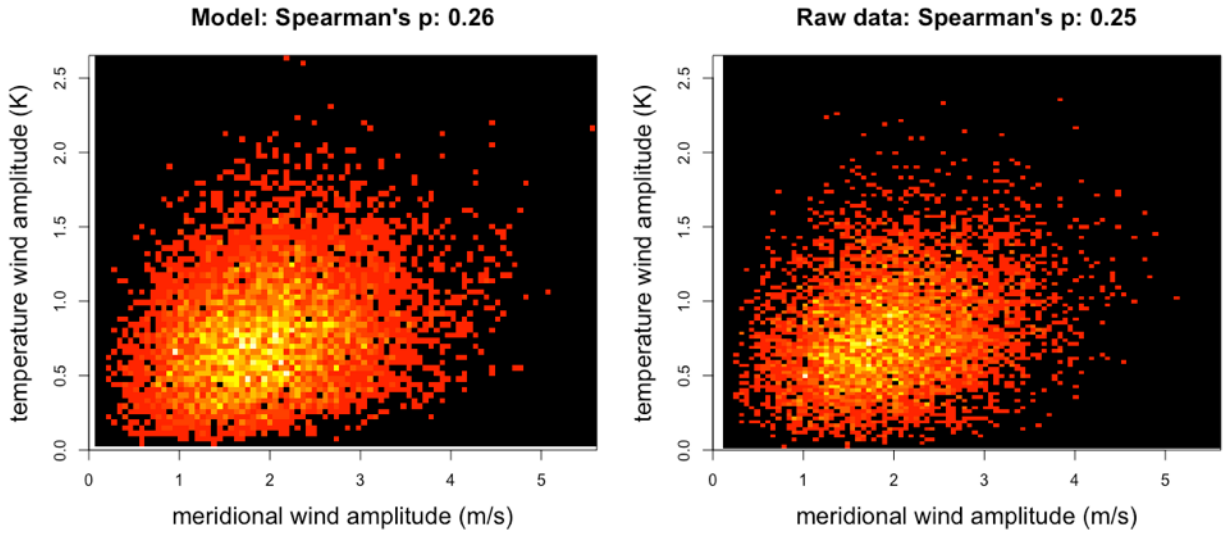
(t)


Model: Spearman's p: 0.46
Raw data: Spearman's p: 0.47
meridional wind amplitude (m/s)
zonal wind amplitude (m/s)
0
1
2
3
4
5
6

(u)

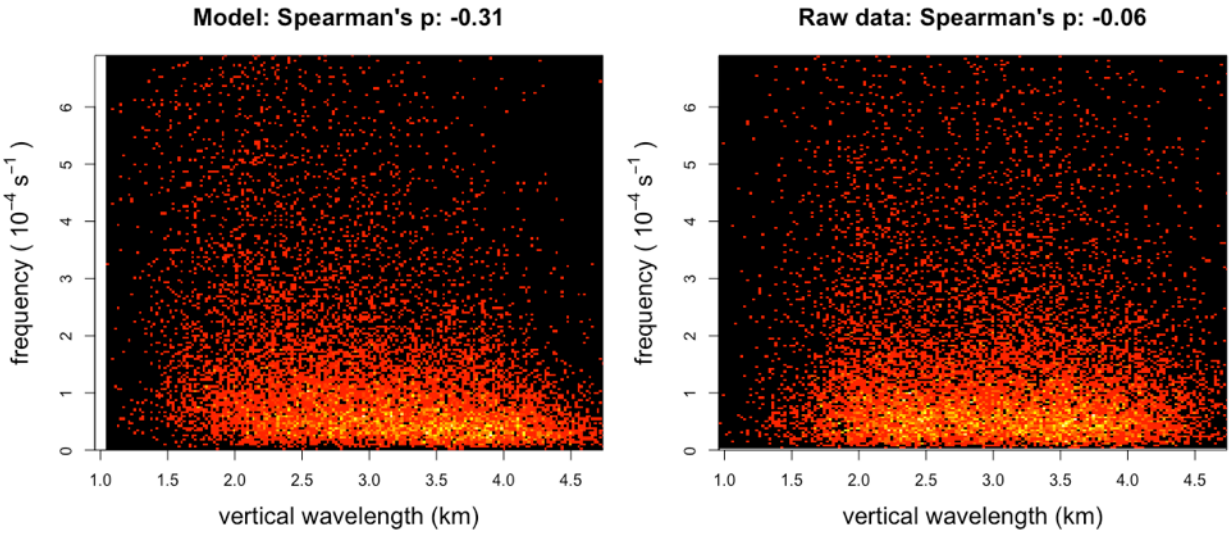

Model: Spearman's p: 0.3
Raw data: Spearman's p: 0.27
temperature wind amplitude (K)
zonal wind amplitude (m/s)
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3
4
5
6

(v)

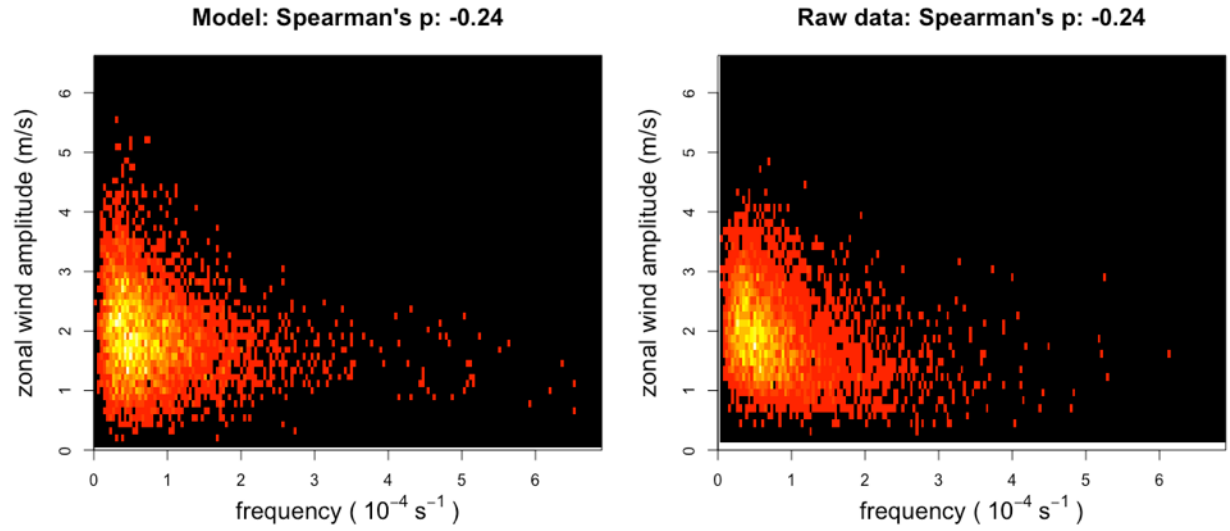


(w)

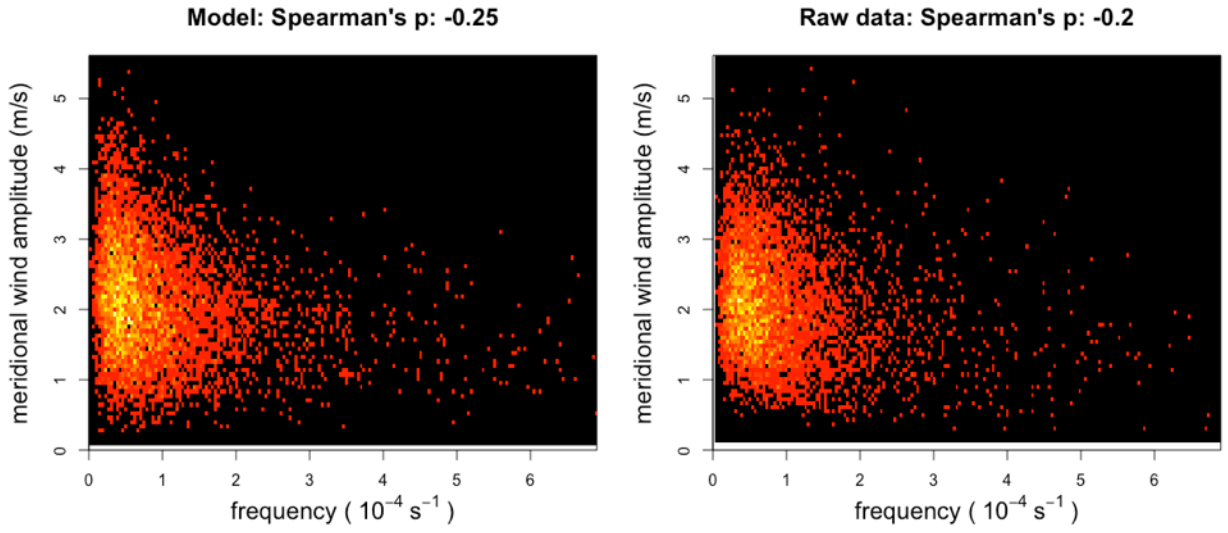
Figure A.7: Figures (a)-(e) present the kernel density estimates for both the raw data (black curve) and the sample generated from the model (red curve) for all the gravity wave parameters that were not included in Figure 31. Meanwhile, Figures (f)-(w) display side by side the 2D histograms associated with the raw data (right) and the sample generated from the model (left) for each pair of gravity wave parameters that was not included in Figure 31.


Model: Spearman's p: -0.31
Raw data: Spearman's p: -0.06
frequency (10^{-4} s^{-1})
vertical wavelength (km)
0
1
2
3
4
5
6
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

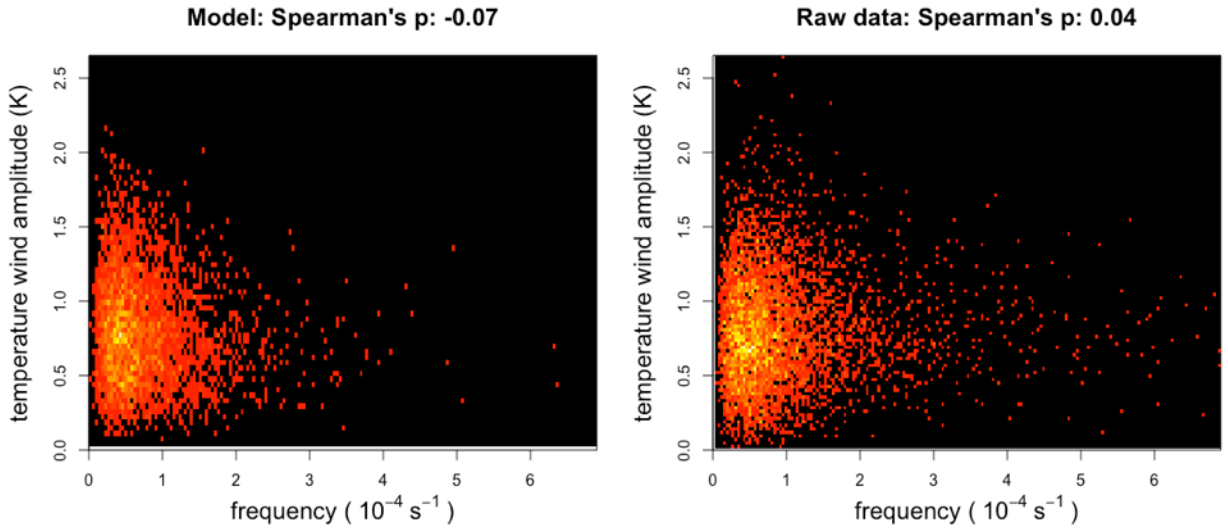
(a)


Model: Spearman's p: -0.24
Raw data: Spearman's p: -0.24
zonal wind amplitude (m/s)
frequency (10^{-4} s^{-1})
0
1
2
3
4
5
6

(b)



(c)



(d)

Figure A.8: The four graphs display side by side the 2D histograms associated with the raw data (right) and the sample generated from the model (left) for each pair of gravity wave parameters that was not included in Figure 35.