Title
The Effectiveness of Copulas for Modeling Compound Climate Extreme Events in Boulder County, Colorado

Author
Agrawal, Surabhi

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

The Effectiveness of Copulas for Modeling Compound Climate Extreme Events in Boulder County, Colorado

A thesis submitted in partial satisfaction of the requirements for the degree Master of Science in Statistics

by

Surabhi Agrawal

2022

ABSTRACT OF THE THESIS

The Effectiveness of Copulas

for Modeling Compound Climate Extreme Events

in Boulder County, Colorado

by

Surabhi Agrawal

Master of Science in Statistics

University of California, Los Angeles, 2022

Professor Karen McKinnon, Chair

This paper investigates the effectiveness of copula models for understanding, estimating, and predicting compound climate extreme events. It focuses on the bivariate temperature-humidity, temperature-wind speed, and wind speed-humidity distributions within the Boulder County, Colorado region. Climate model simulation data is bootstrapped to investigate the variability of the choice of copula families and accuracy of extreme event probability predictions given different lengths and internal variability of climate data. This showed that longer data records

have lower bias and variance than shorter data records in estimating the true probability of a compound extreme event. Fitting the ideal copula models to daily summary data from the region revealed that although there has been a slight increase in the frequency of the compound extreme events, this increase is within the expected range of sampling variability.

The thesis of Surabhi Agrawal is approved.

Rick Schoenberg

Mark S. Handcock

Karen McKinnon, Committee Chair

University of California, Los Angeles

2022

Table of Contents

List of Figures and Tables

1. Introduction

A large number of climate extremes have been observed in recent decades. As anthropogenic climate change continues, the frequency of hot climate extremes is expected to increase, posing a threat to agriculture, ecosystems, and natural resource availability (Coumou and Rahmstorf 2012; Rummukainen 2012). Some of the most negative impacts of climate change observed have been caused by multiple co-occurring climate extremes, such as combined extreme precipitation-temperature, hot-dry, hot-humid, and precipitation-wind events. These simultaneous events, termed "compound extremes," have a higher probability of impacting ecosystems than each extreme individually (Sedlmeier et al. 2016). Since dependencies often exist between the simultaneous extremes (Leonard et al. 2014; Martius et al. 2016), it is not enough to simply analyze each extreme individually. Separate analyses of the extremes often lead to underestimation or overestimation of the compound extreme probabilities. For example, Singh et al (2021) found that a warm-dry compound extreme was estimated to have a 100-year return period under independence assumptions but found to be 60 years when the joint dependency structures were modeled using copulas.

We are particularly motivated by the recent winter fires in Boulder County, Colorado. In December 2021, the Marshall fire swept through Boulder County, Colorado. This was the most destructive fire in Colorado's history. This fire was alarming for two reasons: 1) it occurred in the winter, and 2) it moved into densely populated areas. Normally, the ground is too moist from the snow for fires to spread. Wildfires in the American West generally occur in the forests and wildlands. Temperature, humidity, and wind extremes created the perfect conditions for such a

devastating fire to occur. In the months preceding the Marshall fire, Colorado experienced a strong drought. The fall was also unusually warm. This combination of low precipitation and higher temperatures resulted in drier weather. Additionally, the unusually high wind speeds, up to 105 mph, spread the unusually timed wildfire to populated regions (Chuck 2022; "Colorado" 2021). We will focus on the resulting bivariate compound extremes: hot-dry, hot-windy, and dry-windy events.

In this study, we will be focusing on understanding how well copula models perform to understand, estimate, and predict compound extreme events within the climate context.

Copulas are commonly used to model bivariate compound extremes because they allow the marginal distributions to be separated from the dependence structure (Nelsen 2007). Copulas are very commonly used in modeling climate extremes. Copulas are functions that couple a joint probability distribution to the marginal distributions, allowing us to separate the marginal distribution of each variable from the dependence structure between variables. Copulas are a special type of multivariate cumulative distribution functions for which the univariate marginal distributions are uniformly distributed between 0 and 1 (Hao 2018a). This allows copulas to encode strong assumptions about the dependence structure of the multivariate distribution through the choice of a copula family. Although copulas have previously been used to model the impact of hot, windy, and dry compound extremes (Tavakol et al. 2020), there is little work on the influence of sampling on these results, which is the focus of this study.

The **meta-Gaussian model** is a specific type of copula model that can represent a full range of

association and allows for flexible marginal distributions (Kelly and Krzysztofowicz 1997), which is essential to accurately modeling compound extremes. Under the meta-Gaussian model, the variables are independently transformed to the normal variate using the normal quantile transformation (Kelly and Krzysztofowicz 1997). The resulting multivariate normal distribution is used to conduct joint and conditional analyses of the variables in consideration. A benefit of the meta-Gaussian model is that it has an explicit form, even in high dimensions (Hao 2018b). A drawback is that the meta-Gaussian model might not accurately classify dependence in the extreme tails of the distribution (Wang et al. 2014).

It is essential that the statistical methods we use to model compound climate events are flexible enough to accurately estimate the extreme tails values of the joint distributions (AghaKouchak et al. 2014; Zscheischler et al. 2018). Copulas often are not flexible enough to accurately model higher dimensional compound extremes (Aas et al. 2009). A copula family must be chosen in order to use a copula modeling approach. This choice of family creates strong assumptions about the tail dependence structure of the distribution. Consequently, some models will choose different families for different regions to account for spatial variation in the dependence structure (Ribeiro et al. 2019), although these results can challenge interpretability because we expect there to be spatial structure in dependence structures. This can make copula models difficult to interpret while nonparametric models can allow for greater flexibility (Cooley et al. 2019).

Quantile regression and Markov chain models are also commonly used to model compound climate extremes, as we now review.

Quantile regression is commonly used because it is flexible and semi-parametric (Koenker and Bassett 1978). This method gives a nonparametric framework to model the way extreme percentiles of a conditional distribution change compared to the center. Unlike copulas, quantile regression doesn't require a set of assumptions to be made about the entire distribution. Although quantile regression is successful in detecting the interactions between variables but still has challenges performing accurately at high quantiles under limited sample sizes (Friederichs and Hense 2007).

Under anthropogenic climate change, the frequency, intensity, spatial extent, duration, and timing of extreme weather is changing, both individually and in combination with each other. However, there is very limited knowledge about the dynamical behavior of climate extremes. **Markov chain** analyses have been used to understand the dynamical nature of climate time series, both estimating past events and predicting future events. Markov chain analyses can also be a useful model validation tool for other analyses of compound extremes (Sedlmeier et al. 2016).

Our understanding of climate trends is strongly influenced by internal variability. Internal variability is the naturally occurring climate variability that results from the climate system processes. These can include the interactions between the land, oceans, and atmosphere. Different realizations of internal climate variability would result in different sets of observed data. (Deser et al. 2012). Additionally, within climate research, we often only have access to 40-50 years of climate data records. So, it is also important to consider the impact of limited data availability on the accuracy of copula models of climate extremes.

The aim of this study is to examine the implications of the assumptions made when choosing copula families for modeling compound climate extremes in terms of accuracy and interpretability. We intercompare the performance of the various copula models on temperature, humidity, and wind speed for climate model simulations and daily observations from Boulder County, Colorado to answer three questions. First, how well, in terms of bias and variance, can copula models estimate the true probability of a compound extreme even given 50-100 year long data records? Second, how much does the choice of copula family vary given different lengths and observations of data? Third, how, if at all, have the bivariate temperature-wind speed, temperature-humidity, and wind speed-humidity changed over recent decades, alongside global climate change? We address the first two questions by generating many potential data records of varying lengths by bootstrapping climate model simulation data from the gridbox closest to Boulder, Colorado. We address the third question by fitting the copula models to two different time periods (January 1983-December 2002 and January 2003-July 2022) of daily summary observational data from Broomfield, Colorado.

The rest of the study is organized as follows. The datasets are described in Section 2 and methodology are described in Section 3. The results are presented in Section 4, and summarized and discussed in Section 5.

2. Data

2.1 Community Earth System Model Version 2

In order to gain a preliminary understanding of the properties of the various copula families and the impact of data record length on copula fit, we use simulated climate model data from the Community Earth System Model Version 2 (CESM2), a fully coupled, open-source, comprehensive global climate model which provides simulations of past, present, and future climates on Earth. CESM2 includes coupled simulations of ocean, atmosphere, land, sea-ice, land-ice, river, and waves (Danabasoglu et al. 2020). The stationary climate state data comes from the pre-industrial control simulation, which provides a full stationary record assuming pre-1850 climate conditions. The data includes 2,000 simulated years. We restrict the model simulations to data points from the extended winter (November - March). We removed January, February, and March of the first year and November and December of the last year from the data to only consider fully simulated seasons.

We focus on wind, humidity, and temperature model outputs from a gridbox close to Boulder, Colorado. We use the temperature at a reference height of 2m above ground (TREFT), humidity measurements from the same reference height (QREFT), and wind speed (U10). The bivariate distributions from the 1999 simulated extended winters are shown in the top row of Figure 2.1.

2.2 Global Summary of the Day

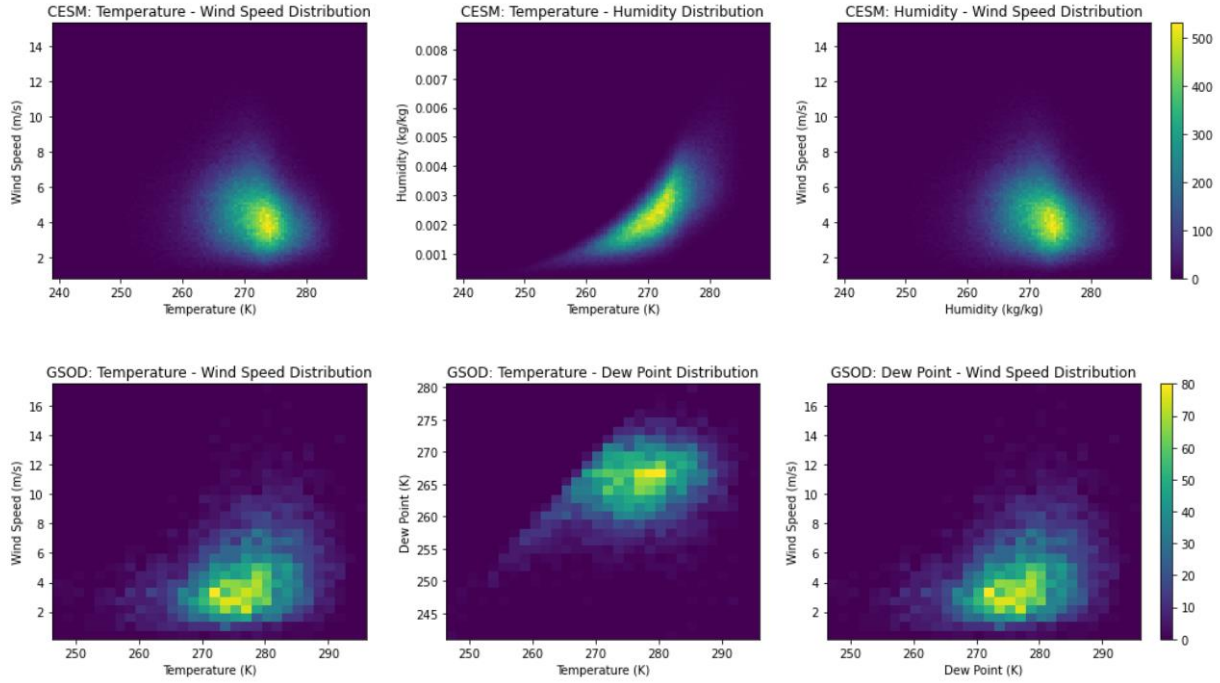
We further use data from the Global Summary of the Day (GSOD) obtained through the National Oceanic and Atmospheric Administration's (NOAA) National Centers for Environmental

Information (NCEI) website to investigate whether there has been a change in the temperature, humidity, and wind speed bivariate distributions over time. The GSOD data summarizes the hourly measurements from the Integrated Surface Database (ISD). The ISD is a global database containing hourly surface observations from the 1950s to present, which integrates wind, temperature, dew point, cloud, sea level pressure, and other climate variables from over a 100 data sources.

We pulled daily-average temperature, wind speed, and dew point observations from the Broomfield Jefferson Airport weather station in Boulder County, Colorado. These were the weather stations that had the longest continuous data record in the Boulder County region, containing continuous daily observations from January 1983 to July 2022. We filtered the data to only consider the extended winters (November - March). The GSOD data does not contain measurements of specific humidity. Instead, we use dew point, which is the temperature to which the air must be cooled to achieve a 100% relative humidity. We also converted the temperature and dew point observations from Fahrenheit to Kelvins and wind speed observations from knots to meters per second to align with our analysis of the CESM data. The bivariate distributions from the data are shown in the bottom row of Figure 2.1.

We split the data into two time periods, January 1983 to December 2002 and January 2003 to July 2022, to understand how these patterns may have changed over time, particularly in terms of climate change.

Figure 2.1. The bivariate density plots of temperature-humidity, temperature-wind speed, and wind speed-humidity from daily observations in the 1999 simulated extended winters in the CESM2 dataset are shown in the top row. The bivariate density plots of temperature-dew point, temperature-wind speed, and wind speed-dew point from the 1983-2022 GSOD dataset are shown in the bottom row. The color bar shows the number of occurrences of the bivariate observation within the dataset. Yellow regions are more frequently observed than blue regions.



3. Methodology

3.1 Copulas

Copulas are multivariate cumulative distribution functions that allow us to separate the univariate marginal distributions from the dependence structure (Nelsen 2007). Consider two variables X_1 and X_2 with continuous marginal cumulative distribution functions F_1 and F_2. Then, the transformations $U_i = F(X_i)$ are uniformly distributed between 0 and 1. Define the inverse of F_1 and F_2 as $F_i^{-1}(u) = \min\{F_i(x) = u\}$. The copula of X_1 and X_2 is the cumulative distribution function of U_1 and U_2: $C(u_1,u_2) = Pr(U_1 <= u_1, U_2 <= u_2)$. By the construction of the U_i variables, $C(u_1,u_2) = Pr(X_1 <= F_1^{-1}(u_1), X_2 <= F_2^{-1}(u_2))$.

There are many options for copula families. This study focuses on Gaussian, t, Frank, Clayton, and Gumbel copulas.

The **Gaussian copula** is defined as:

$$C(u_1,u_2) = \Phi[\Phi^{-1}(u_1), \Phi^{-1}(u_2)]$$

where Φ^{-1} is the inverse cumulative distribution function of the standard Gaussian distribution, and Φ is the bivariate joint cumulative distribution function of a Gaussian with zero means and the covariance matrix $\sum$ of U_1 and U_2.

The **t copula** is defined as:

$$C(u_1,u_2) = t[t^{-1}(u_1), t^{-1}(u_2)]$$

where t^{-1} is the inverse student t function, and t is the cumulative distribution function of bivariate student t distribution with dependence structure defined by the 2x2 matrix $\sum$ of the linear correlation parameters of u_1 and u_2.

The **Frank copula** is defined as:

$$C(u_1,u_2) = -(1/\theta) \log[1 + ((\exp(-\theta u_1) - 1)(\exp(-\theta u_2) - 1))/(\exp(-\theta) - 1)]$$

The **Gumbel copula** is defined as:

$$C(u_1,u_2) = \exp[-((-\log(u_1))^{\theta} + (-\log(u_2))^{\theta})^{1/\theta}]$$

The **Clayton copula** is defined as:

$$C(u_1,u_2) = [\max\{u_1^{-\theta} + u_2^{-\theta} -1, 0\}]^{-1/\theta}$$

3.2 Metrics

3.2.1 Bivariate Exceedance Probability of 99.9% Thresholds

In order to understand the accuracy and interpretability of various copula families on compound climate extremes, we are interested in the tail behavior of the copula fits. We particularly focus on 99.9% extreme tails. We are interested in high temperature, low humidity, and high wind conditions due to their high fire risk, so we consider a 99.9% threshold for temperature, 0.1% threshold for humidity, and 99.9% threshold for wind speed. These thresholds are defined as the 99.9% (or 0.1%) quantile value in the datasets and are shown in Table 3.2.1.1. We are interested in the predicted probability of bivariate exceedance of the threshold values. For example, if we were interested in the temperature-humidity distribution for the CESM2 data, we look at the

probability a copula model predicts of observing temperature value of 284.6027 K or higher and humidity value of 0.0004 kg/kg or lower.

Table 3.2.1.1. The 99.9% (or 0.1%) threshold values from the simulated extended winters from the CESM2 and GSOD data. Note that we observe specific humidity in kg/kg for the CESM2 data and dew point in Kelvins for the GSOD data.

	Temperature (99.9%)	Humidity (0.1%)	Wind Speed (99.9%)
CESM2	284.6027 K	0.0004 kg/kg	11.3078 m/s
GSOD	292.2580 K	246.5485 K	13.8879 m/s

3.2.2 Bayesian Information Criterion

We use the Bayesian Information Criterion (BIC) to cross-compare the fits of various copula families. We calculate the BIC values of the estimated copula probability values for each of the bivariate climate event observations. A lower BIC value indicates a better model fit. This approach has been used in the Multivariate Copula Analysis Toolbox (Sadegh et al. 2017).

The BIC is defined as BIC = k*ln(n) - 2ln(L_hat) where k is the number of parameters, n is the sample size, and L_hat is the maximized likelihood of the model on the data.

We use a modified version of the BIC rewritten in terms of the residual sum of squares (RSS) under the assumption that the model errors are independently, identically, and normally distributed. The model errors are defined as follows:

Let X and Y be two climate variables. Consider the bivariate climate event (X = x, Y = y). We estimate the “empirical” probability of bivariate exceedance of the event as the proportion of the

observations within our data record for which $X \geq x$ and $Y \geq y$. Then, we define the model errors as the absolute difference between the predicted copula probability of bivariate exceedance of the event and the "empirical" probability.

The modified version of the BIC is $BIC = n*\ln(RSS/n) + k*\ln(n)$. It is important to note that we removed a constant of $n + n*\ln(2\pi)$ from this version of the BIC. Since this constant only depends on the sample size, which remains consistent across the different models we intercompare, this does not impact the model selection. Consequently, this version of the BIC is negative when RSS/n is less than 1. Due to the seasonality of climate data, we have violated the normality and independence assumptions. This is an important topic to further explore in future research.

3.3 Uncertainty Estimates and Limited Data

Since we often only have access to 40-50 years of climate data records, we want to understand the accuracy of copula models when applied to limited data records. We use block bootstrapping to investigate whether we can reproduce a parameter spread if a different realization had been observed, with block size of one winter to maintain the seasonality of the data.

We use the case resampling bootstrap scheme described below:

1. The dataset contains daily weather observations for *j* years. Randomly sample *i* years from the *j* years with replacement.
2. Acquire the daily temperature, humidity, and wind speed values from the summers corresponding to those years.

3. Fit the copulas and find the BIC and probability of bivariate exceedance of 99.9% threshold events.
4. Repeat Steps 1-3 N=1000 times.

We perform this scheme across two scenarios:

1. We estimate the copula probabilities using multiple, **shorter, quasi-independent datasets** derived from the full 1999 year simulation ($i < j$; $i = 30$, 50, or 100; $j = 1999$). These are mentioned as **30 from 30**, **50 from 50**, and **100 from 100** throughout this paper.
2. We estimate the copula probabilities using a single **shorter data record** ($i = j = 30$, 50, or 100). These are mentioned as **30 from 1999**, **50 from 1999**, and **100 from 1999** throughout this paper.

3.4 Coding Packages

The copula estimates were generated using the copulafit and copulacdf functions in the stats package in Matlab. The data processing and analysis was performed in R.

4. Results

Before applying our methodology to the GSOD data, we develop an understanding of the implications of copula family choice with regards to accuracy and interpretability by analyzing the CESM2 climate model simulation data. The CESM2 dataset provides us with a long record and does not have any climate change signal which means that the full record is stationary.

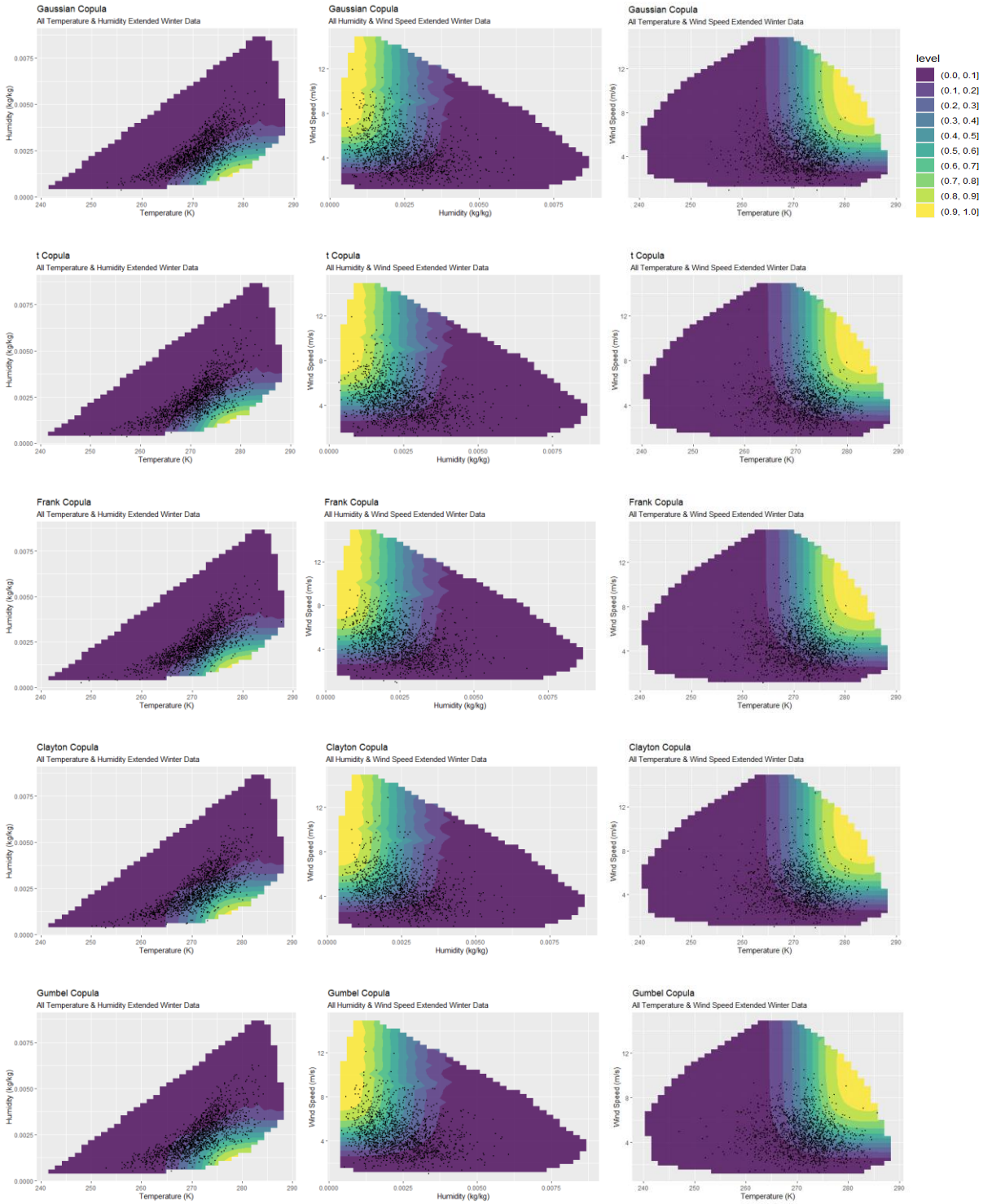
4.1 Full CESM2 Copula Fits

We begin by fitting all five copula families (Gaussian, t, Frank, Clayton, Gumbel) to the full 1999 extended winters available in the CESM2 dataset in all three bivariate variable combinations (temperature-humidity, wind speed-humidity, and wind-temperature) to gain a basic understanding of how the copulas fit on a long dataset.

In Figures 4.1.1, we compare the distribution of the predicted copula probabilities for the different variable combinations and copula families. These are contour plots of the predicted copula probability of bivariate exceedance of the realization of the two climate variables. Within each variable combination, there are broad similarities across the probability distributions for the five copula families.

In Table 4.1.1, we compare the probability of bivariate exceedance of the 99.9% threshold values for the fifteen copulas models. As expected, when provided with nearly 2,000 years of data, all the models predict a nearly 0.002% probability of observing a bivariate weather observation that

Figure 4.1.1. Contour plots of the predicted copula probability of the weather occurrences of the CESM2 data with scatter plot of a random subsample of the observations overlaid. The legend displays the predicted copula probabilities of bivariate exceedance of the climate events. For temperature and wind speed, this means the probability of observing the event or a higher value and for humidity this means observing the event or a lower value. Yellow regions have higher copula probabilities than purple regions.



of the 99.9% threshold or more extreme. This demonstrates that there is a positive correlation between the variables and that dependence structure of the variable combinations are important. This validates our choice to look at compound extremes rather than the extreme events individually.

In Table 4.1.2, we compare the BIC values for the fifteen copulas models. We use the BIC for model selection. The copula family with the lowest BIC value is considered the best model for each variable combination. When provided with nearly 2,000 years of data, the Gumbel copula is chosen as the best model for temperature-humidity. Either the Gumbel or Clayton copula would be good choices since the BICs are the most negative and very similar to each other. We will select the Gumbel copula since it is slightly more negative. For the wind speed-humidity case, all five of the copula families are reasonable choices since they are similar in value. We will select the Clayton copula since it is the most negative. Unlike the previous two variable combinations, for the wind speed-temperature case, the Gaussian copula is clearly the best choice. The BIC of the Gaussian copula is significantly more negative than those of the other four copula families.

Table 4.1.1. The predicted copula probabilities of bivariate exceedance of 99.9% threshold events on the full 1999 extended simulated winters. Regardless of copula family or variable choice, all fifteen models yield approximately 0.002% probability of bivariate exceedance of the 99.9% threshold events.

	Gaussian	t	Frank	Clayton	Gumbel
Temperature Humidity	0.002003	0.002003	0.002003	0.002002	0.002002
Wind Speed Humidity	0.001995	0.001995	0.002002	0.002002	0.001852
Wind Speed Temperature	0.002003	0.002003	0.002003	0.002002	0.002002

Table 4.1.2. The BIC values for the copula fits on the full 1999 extended simulated winters. The smallest BIC value for each variable combination is bolded. In this specific climate model simulation, the Gumbel copula is chosen as the best model for temperature-humidity, the Clayton model is chosen as the best model for wind speed-humidity, and the Gaussian model is chosen as the best model for wind speed-temperature.

	Gaussian	t	Frank	Clayton	Gumbel
Temperature Humidity	-526,205	-534,575	-545,163	-624,180	**-627,661**
Wind Speed Humidity	-670,521	-686,794	-681,842	**-688,280**	-678,553
Wind Speed Temperature	**-2,968,461**	-1,045,135	-935,532	-896,348	-876,021

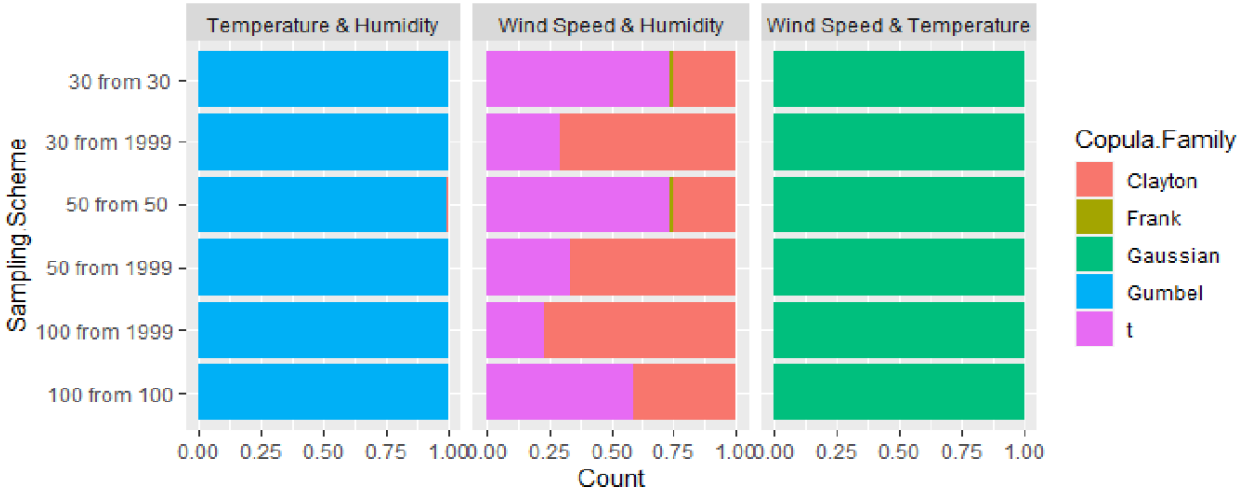
4.2 Uncertainty Estimates

The analysis of the 1999 years of data from the climate model simulation data in Section 4.1 provides insight into the performance of the copula models on one instance of potential data. However, our understanding of climate trends is strongly influenced by the specific realizations of internal climate variability we observe. Additionally, we often only have access to limited data records. As discussed in Section 3.3, we use a case resampling block bootstrap scheme on the climate model simulation data to investigate whether we can reproduce a parameter spread if a different realization of the data had been observed and how accurately the copula models predict extreme tail behavior with limited data availability.

We compare the performance of the Gaussian, t, Frank, Clayton, and Gumbel copulas on the BIC and probability of bivariate exceedance of the 99.9% threshold events using the full 1999-year dataset, shorter quasi-independent datasets derived from the full 1999 year simulation, and single shorter records.

In this analysis, the probability of bivariate exceedance of the 99.9% threshold events from the full 1999-year dataset are used as the "true" probabilities of the compound extreme events. These can be referred to in Table 4.1.1.

Figure 4.2.1. A visualization of the number of times the copula family was chosen by having the minimum BIC value for each of the bootstrap schemes and variable combinations.



The BIC is used as a criterion for selecting the best copula family for each sampling scheme and variable combinations. Figure 4.2.1 shows the number of times (out of 1000 repetitions) each copula family had the lowest BIC, for each of the sampling schemes and variable combinations. The Gumbel distribution was almost always chosen as the best copula family for the temperature-humidity variable combination. The Gaussian copula was always chosen as the best copula family for the wind speed-temperature variable combination. The wind speed-humidity variable combination had more variety in copula family choice. The t copula was most frequently chosen as the best for shorter data records (30 years sampled from 30 years with replacement, 50 years sampled from 50 years with replacement, 100 years sampled from 100 years with replacement). However, the Clayton copula was most frequently chosen as the best for the shorter quasi-independent datasets 30 years sampled from 1999 years with replacement,

50 years sampled from 1999 years with replacement, 100 years sampled from 1999 years with replacement).

We use the copula family most often chosen as the best model for the sampling scheme and variable combination for all further analysis. So, we use the Gumbel copula for all temperature-humidity analyses. We use the Gaussian copula for all wind speed-temperature analyses. We use the Clayton copula for all wind speed-humidity analyses.

This model selection aligns with the analysis of the entire 1999-year CESM2 dataset in Section 4.1. In both sections, Gumbel is chosen as the best copula family for temperature and humidity and Gaussian distribution is chosen as the best copula family for temperature and wind speed. The analysis of the entire 1999-year CESM2 dataset chose Clayton as the best copula family, which aligns with the way Clayton is chosen as the best copula family. Overall, we conclude that although the choice of copula family can vary extensively with variable choice, the ideal copula family remains consistent regardless of data record length.

In order to understand how well the "true" probability of bivariate exceedance of the 99.9% threshold events can be estimated given shorter data records, for each of the three variable combinations, we compare the distributions of the estimated copula probability of bivariate exceedance of the 99.9% threshold events across the six sampling schemes. Figures 4.2.2.a-c show histograms of these distributions.

Under these bootstrap schemes, there are three trends we expect to observe:

1) The shorter data record sampling schemes only sample from the first 30, 50, or 100 years as opposed to the "true" probability which sampled from all 1999 years. Any trends in the beginning of the dataset would be exacerbated by this sampling scheme. So, we expect the mean of the probability of bivariate exceedance of the 99.9% threshold events from the shorter data record copula fits to differ from the "true" probability of bivariate exceedance of the 99.9% threshold events.
2) The shorter quasi-independent datasets take samples of size 30, 50, or 100 years from the full 1999-year dataset. Thus, we expect the means of the probability of bivariate exceedance of the 99.9% threshold events from the shorter quasi-independent datasets to be close to that of the "true" probability of bivariate exceedance of the 99.9% threshold events.
3) For both types of sampling schemes, we expect the distributions to get narrower as the number of years sampled increases.

In all three bivariate variable combinations, the difference between the mean estimated probability and "true" probability of bivariate exceedance of the 99.9% threshold events is larger for the shorter data records (shown in the top rows of Figures 4.2.2.a-c) than for the shorter quasi-independent datasets (shown in the bottom rows of Figures 4.2.2.a-c). The difference is approximately 10 times larger for shorter data records than shorter quasi-independent datasets when looking at the wind speed-temperature distribution. The difference is approximately 100 times larger for shorter data records than shorter quasi-independent datasets when looking at the wind speed-humidity distribution. The difference is only slightly larger for shorter data records than shorter quasi-independent datasets when looking at the temperature-humidity distribution.

The standard deviations of the estimated probabilities of 99.9% decrease as sample size increases regardless of sampling scheme, for all three variable combinations. This can also be seen by looking from left to right in the Figures 4.2.2.a-c.

Figure 4.2.2.a Histograms of the estimated probability of exceedance of the 99.9% threshold wind speed-temperature event (temperature 284.6027 K or higher and wind speed 11.3078 m/s or higher) for each of the six sampling schemes (30 from 30, 50 from 50, 100 from 100, 30 from 1999, 50 from 1999, 100 from 1999). The red line shows the "true" probability. The blue line shows the mean estimated probability of bivariate exceedance of the 99.9% threshold event across the 1000 repetitions of the sampling scheme. All results are shown from the ideal copula family selected based on the most negative BIC.

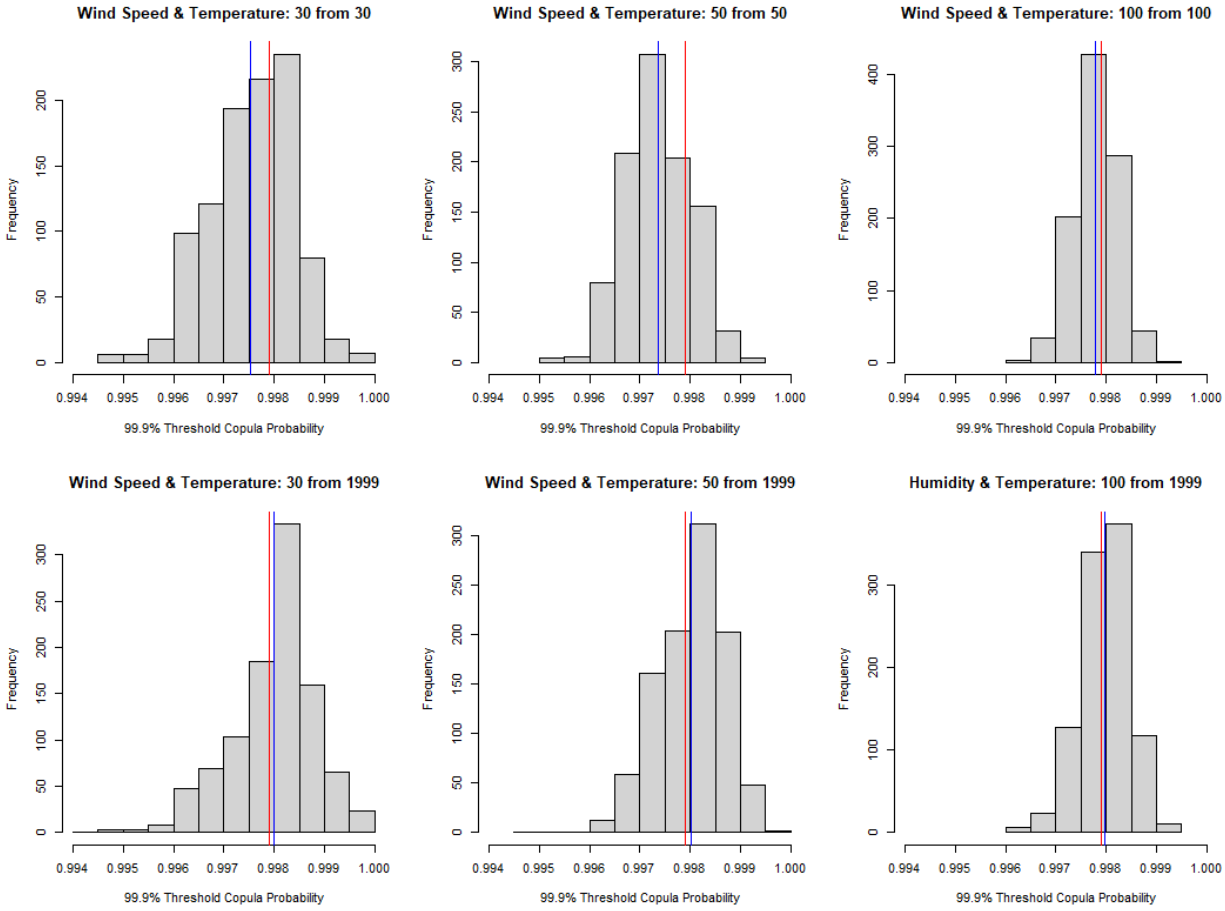


Figure 4.2.2.b Histograms of the estimated probability of exceedance of the 99.9% threshold humidity-wind speed event (humidity 0.0004 kg/kg or lower and wind speed 11.3078 m/s or higher) for each of the six sampling schemes (30 from 30, 50 from 50, 100 from 100, 30 from 1999, 50 from 1999, 100 from 1999). The red line shows the "true" probability. The blue line shows the mean estimated probability of bivariate exceedance of the 99.9% threshold event across the 1000 repetitions of the sampling scheme. All results are shown from the ideal copula family selected based on the most negative BIC.

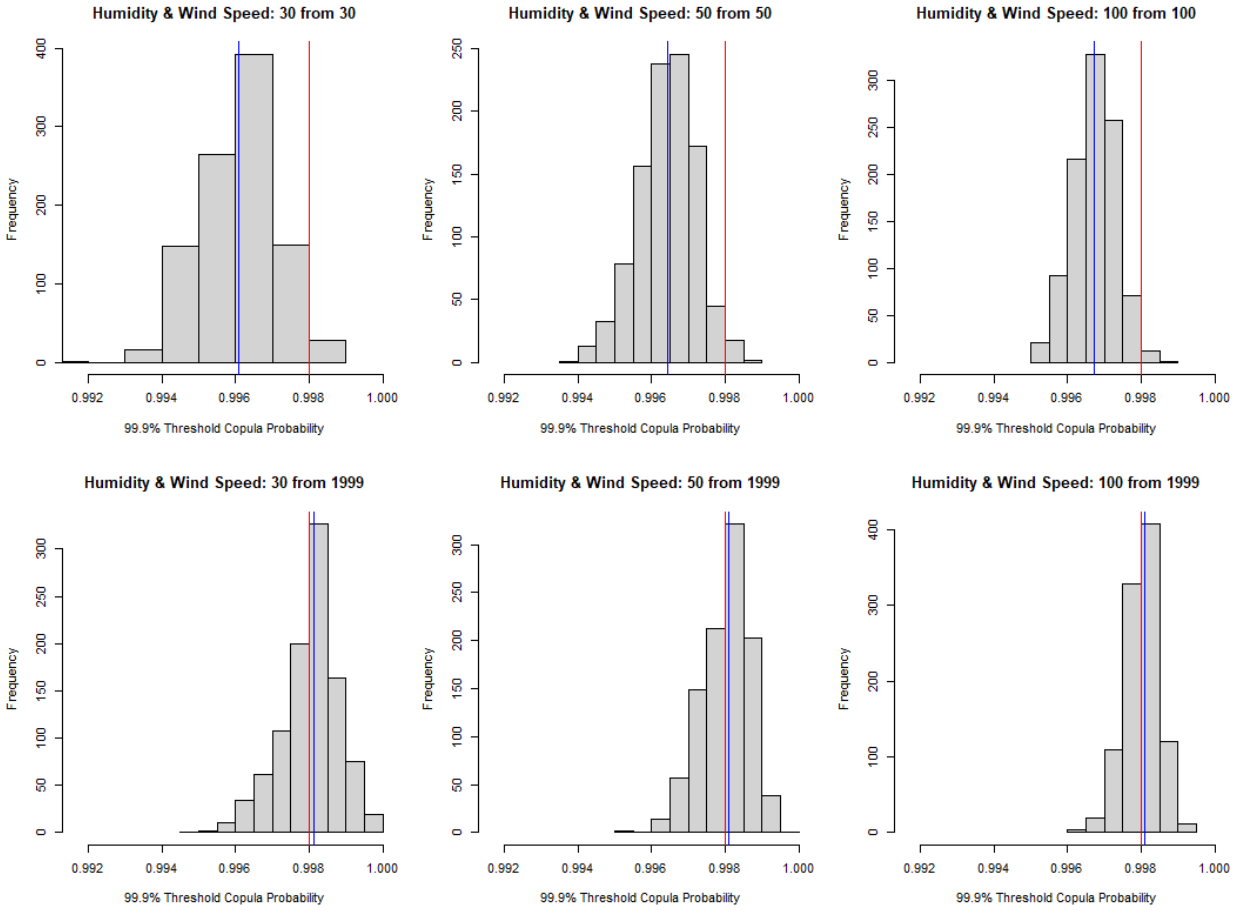


Figure 4.2.2.c Histograms of the estimated probability of exceedance of the 99.9% threshold humidity-temperature event (humidity 0.0004 kg/kg or lower and temperature 284.6027 K or higher) for each of the six sampling schemes (30 from 30, 50 from 50, 100 from 100, 30 from 1999, 50 from 1999, 100 from 1999). The red line shows the "true" probability. The blue line shows the mean estimated probability of bivariate exceedance of the 99.9% threshold event across the 1000 repetitions of the sampling scheme. All results are shown from the ideal copula family selected based on the most negative BIC.

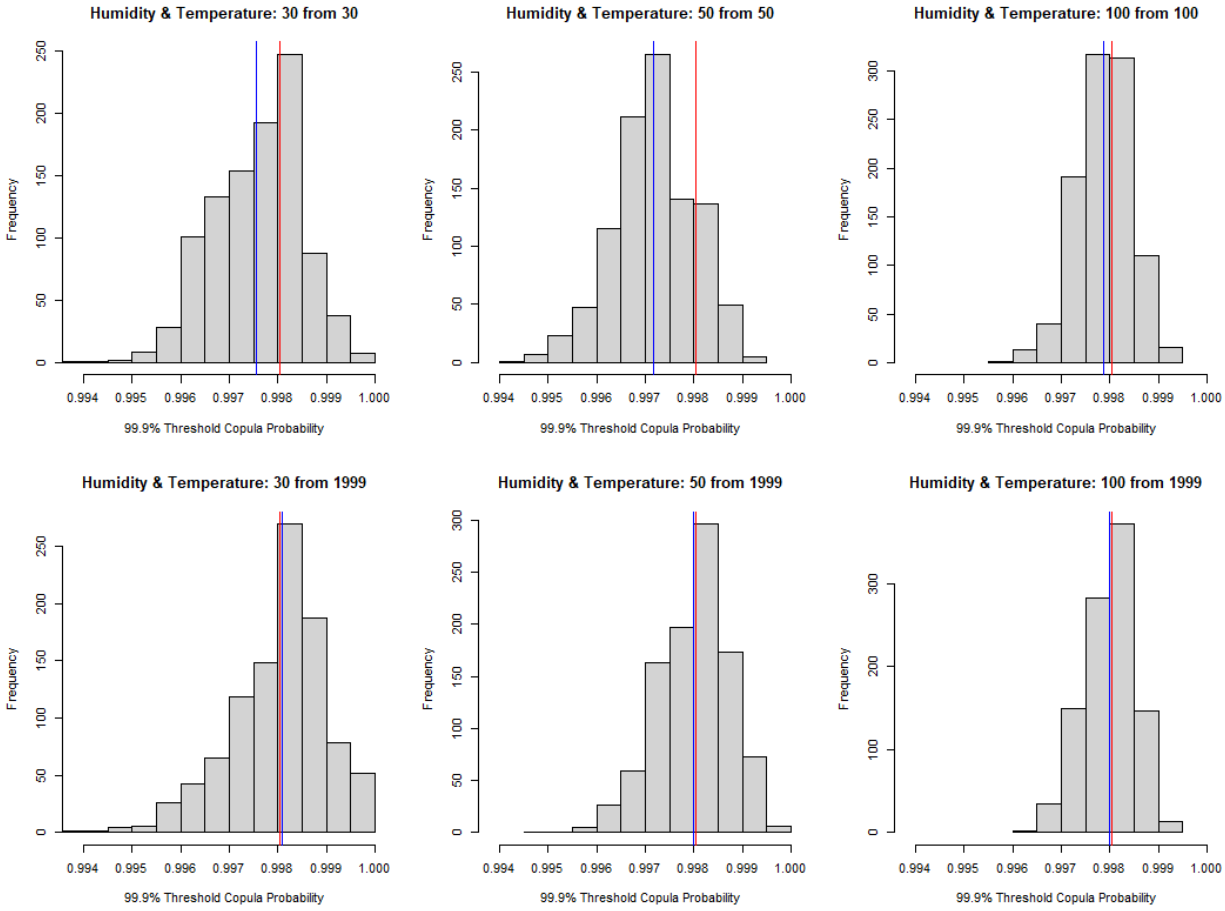


Table 4.2.1.a The 5-95% quantile range of the distribution of predicted probability of bivariate exceedance of a 99.9% threshold wind speed-temperature event (temperature 284.6 K or higher and wind speed 11.3 m/s or higher) for each of the six sampling schemes (30 from 30, 50 from 50, 100 from 100, 30 from 1999, 50 from 1999, 100 from 1999).

	Shorter Data Records	Shorter Quasi-Independent Datasets
30 years	0.994312 - 0.997798	0.996702 - 0.999117
50 years	0.995003 - 0.997624	0.996839 - 0.998941
100 years	0.995788 - 0.997689	0.997233 - 0.998742

Table 4.2.1.b The 5-95% quantile range of the distribution of predicted probability of bivariate exceedance of a 99.9% threshold humidity-wind speed event (humidity 0.0004 kg/kg or lower and wind speed 11.3078 m/s or higher) for each of the six sampling schemes (30 from 30, 50 from 50, 100 from 100, 30 from 1999, 50 from 1999, 100 from 1999).

	Shorter Data Records	Shorter Quasi-Independent Datasets
30 years	0.994312 - 0.997798	0.996702 - 0.999117
50 years	0.995003 - 0.997624	0.996839 - 0.998941
100 years	0.995788 - 0.997689	0.997233 - 0.998742

Table 4.2.1.c The 5-95% quantile range of the distribution of predicted probability of bivariate exceedance of a 99.9% threshold humidity-temperature event (humidity 0.0004 kg/kg or lower and temperature 284.6027 K or higher) for each of the six sampling schemes (30 from 30, 50 from 50, 100 from 100, 30 from 1999, 50 from 1999, 100 from 1999).

	Shorter Data Records	Shorter Quasi-Independent Datasets
30 years	0.996027 - 0.998894	0.996236 - 0.999556
50 years	0.995764 - 0.998541	0.996689 - 0.999071
100 years	0.996953 - 0.998740	0.997086 - 0.997233

Tables 4.2.1.a-c show the central 90% range of the distribution of the predicted probabilities of bivariate exceedance of a 99.9% threshold event. We are interested in the 90% coverage rate. By comparing these ranges to the "true" probabilities shows in Table 4.1.1, we find that the "true" probability falls in the 90% range for shorter quasi-independent dataset sampling schemes but not for shorter data record sampling schemes for the wind speed-temperature and wind speed humidity variables combinations. However, the "true" probability falls in the 90% range for all sampling schemes for the temperature-humidity variable combination.

It is important to keep in mind that we only sampled from the first 30, 50, and 100 years of the 1999 year long CESM2 dataset in the shorter data record sampling schemes. In order to understand if this was due to any specific trends in the beginning of the CESM2 dataset, or a feature of the variable selection, we compare the coverage rate for all the possible 30-year time periods. We were unable to repeat this analysis for the 50 year and 100-year sample sizes due to computational limitations.

We split the 1999-year dataset into sixty-six different 30-year datasets. These datasets consist of the first 30 years, second 30 years, and so on. We repeat the analysis above for each of these 66 datasets in the 30 from 30 bootstrap scheme. In each time duration, we generate 1000 different 30 from 30 bootstrapped samples, fit the ideal copula to each of these, calculate the estimated probability of bivariate exceedance of a 99.9% threshold event, and then construct a central 90% range.

These 90% intervals are shown in Figure 4.2.3 and the coverage rates are displayed in Table 4.2.2. Regardless of variable choice, the coverage rate for a 30-year dataset is approximately 80%. So, overall, copula models adequately capture the true probability value of a 99.9% extreme event even with limited sample sizes of 30 years, although there is still noticeable undercoverage. Since copula models have more accurate performance when trained with more data, we have strong reason to believe that this coverage rate would be just as high or even higher for 50- or 100-year long datasets.

Figure 4.2.3. Coverage plots of the central 90% range of the predicted probabilities of bivariate exceedance of a 99.9% threshold event for 1000 iterations of the 30 from 30 sampling scheme for each of the 66 different 30 year time periods for all three variable combinations. The vertical lines show the 90% range, beginning at the 5% quantile value and ending at the 95% quantile value of the predicted probability of bivariate exceedance of a 99.9% threshold event. These lines are black if the interval includes the "true" probability of a 99.9% event and blue if they do not. The dashed red line represents the "true" probability of the bivariate exceedance of a 99.9% threshold event.

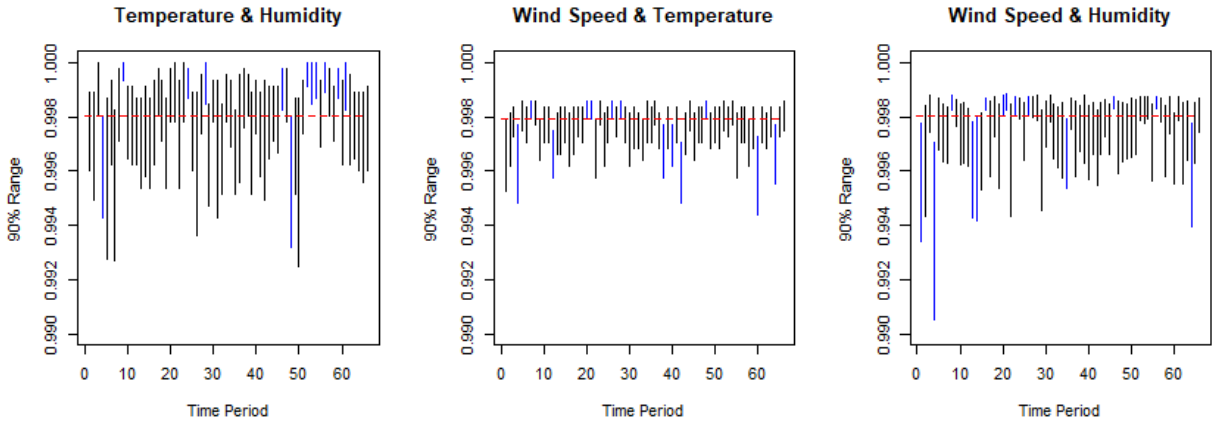


Table 4.2.2. The coverage rates of the central 90% intervals across the 66 different time durations for each of the variable combinations.

Temperature - Humidity	Wind Speed - Humidity	Wind Speed - Temperature
81.8%	78.8%	80.3%

4.3 GSOD Data Application

Now that we have developed an understanding of the performance of various copula families on various data records for a variety of bivariate variable combinations, we focus our attention on the GSOD data from Broomfield Jefferson, Colorado. We split the data into two sub-datasets: January 1983 - December 2002 and January 2003 - July 2022. We are interested in analyzing the ways the occurrence of compound extremes in wind speed-temperature, wind speed-humidity, and temperature-humidity have changed alongside climate change. We use dew point instead of specific humidity in this analysis. We fit all five copula family models to each of the bivariate variable combinations for both time periods. Figures 4.3.1a-f show contour plots of the predicted copula probability of bivariate exceedance of climate events for both time periods for the ideal copula family as selected in Section 4.2. There are broad similarities in the contour plots across the time periods. We see an apparent shift in probability of lower dew point values given high temperature observations from the earlier time period to the more recent time period.

The predicted probability of bivariate exceedance of 99% threshold values of temperature, dew point, and wind speed are shown in Table 3.2.1.1. Tables 4.3.1 and 4.3.2 show us that, regardless of the copula family or variable choice, the estimated probabilities of bivariate exceedance of a 99.9% threshold event were between 0.25 and 0.28% for the 1983-2002 data and between 0.2 and 0.23% for the 2003-2022 data. This shows that there has been a slight increase in occurrence of extreme high temperature-low dew point, high temperature-high wind speed, and high wind speed-low dew point events in recent decades. However, it is important to note that this slight difference seems to be within the expected variability as shown in the CESM2 analysis in Section 4.2.

Figure 4.3.1. Contour plots of the predicted copula probability of the weather occurrences of the GSOD data with scatter plot of a random subsample of the observations overlaid, for wind speed-temperature, wind speed-dew point, and temperature-dew point, for 1983-2002 and 2003-2022. The legend displays the predicted copula probabilities of bivariate exceedance of the bivariate climate event. For temperature and wind speed, this means the probability of observing the event or a higher value and for humidity this means observing the event or a lower value. Yellow regions have higher copula probabilities than purple regions. All plots are shown for the copula family selected as the ideal family for the variable combination within the 30 from 30 sampling scheme in Section 4.2.

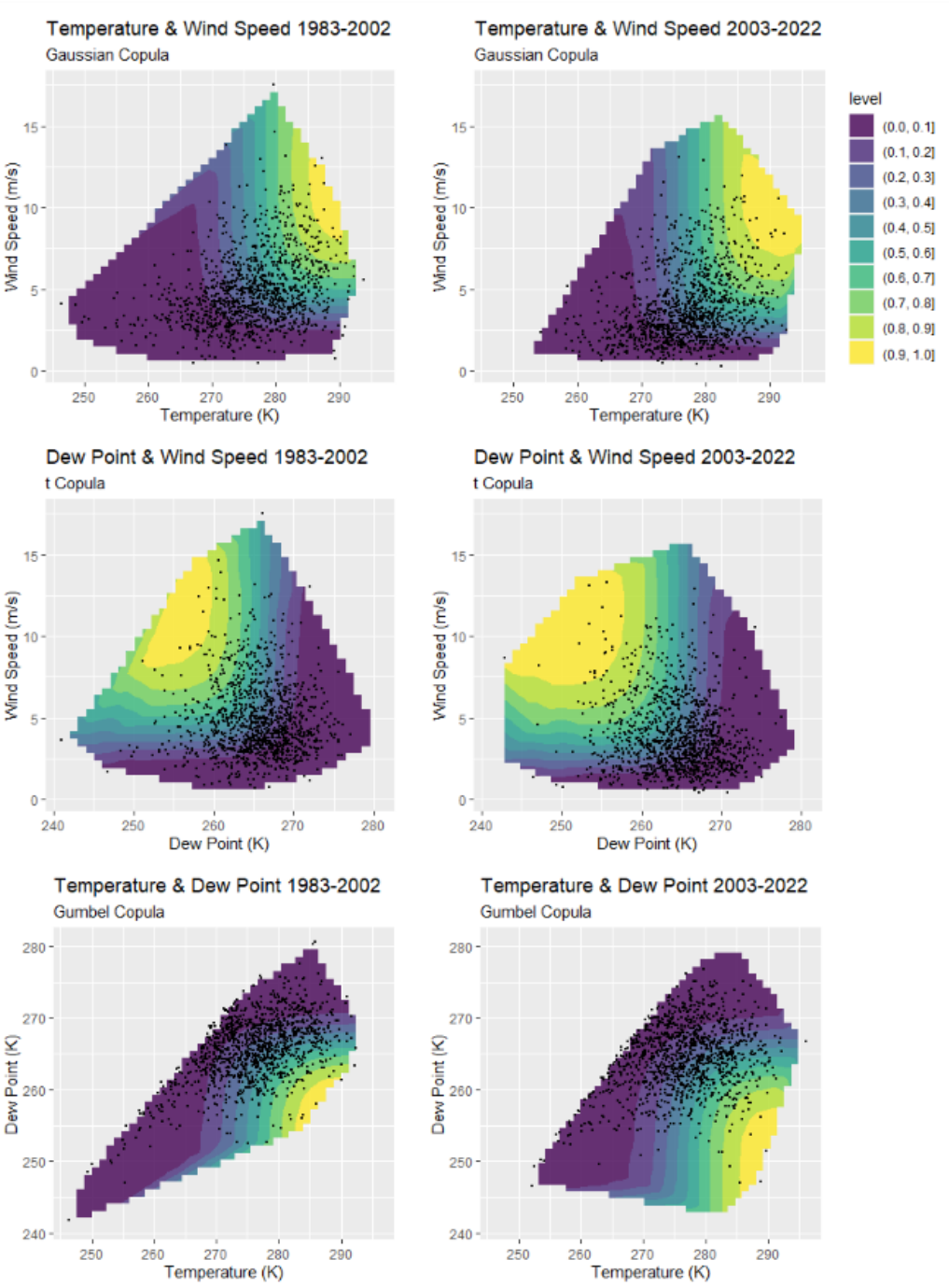


Table 4.3.1. The predicted bivariate exceedance probabilities of a 99.9% event for copula fits on the 1983-2002 GSOD data.

	Gaussian	t	Frank	Clayton	Gumbel
Temperature Dew Point	0.002586	0.002686	0.002503	0.002705	0.002353
Wind Speed Dew Point	0.002506	0.002806	0.002706	0.002505	0.002705
Wind Speed Temperature	0.002633	0.002633	0.002634	0.002634	0.002594

Table 4.3.2. The predicted bivariate exceedance probabilities of a 99.9% event for copula fits on the 2003-2022 GSOD data.

	Gaussian	t	Frank	Clayton	Gumbel
Temperature Dew Point	0.002343	0.002343	0.002362	0.002364	0.002098
Wind Speed Dew Point	0.002366	0.002362	0.002365	0.002364	0.002364
Wind Speed Temperature	0.002024	0.002018	0.002026	0.002027	0.001924

Tables 4.3.3 and 4.3.4 show the BIC values for the copula fits to the 1983-2002 and 2003-2022 GSOD data. In both time periods, the Gaussian family yields a drastically lower BIC value for the temperature-dew point analysis than any of the other copula families. The bootstrap analysis of the CESM2 data in Section 4.2 selected the Gumbel family as the ideal choice. This may be a feature of using dew point instead of specific humidity measurements. We also saw in the previous section that different instances of internal variability could yield different ideal copula families. It would be of interest to investigate this in future studies.

The copula families with the lowest BIC values for the wind speed-dew point and wind speed-temperature analyses also differ between the GSOD and CESM2 analyses. The GSOD copula fits yield Gumbel as the best family for wind speed-dew point and Clayton for wind-speed-temperature. Whereas the CESM2 copula fits yield t as the best family for wind speed-humidity and Gaussian for wind-speed-temperature. However, the BIC values across families are much closer in scale for these variable combinations than the temperature-dew point analysis. It is also of interest to continue to investigate what causes these differences.

Table 4.3.3. The copula BIC values for the 1983-2002 GSOD data. The model with the lowest BIC for each variable combination is bolded.

	Gaussian	t	Frank	Clayton	Gumbel
Temperature Dew Point	**-32427.160**	-8719.512	-7501.357	-7234.432	-6935.733
Wind Speed Dew Point	-6464.537	-6500.950	-6573.152	-6832.907	**-6855.224**
Wind Speed Temperature	-6836.546	-6813.762	-6922.635	**-7049.783**	-7006.004

Table 4.3.4. The copula BIC values for the 2003-2022 GSOD data. The model with the lowest BIC for each variable combination is bolded.

	Gaussian	t	Frank	Clayton	Gumbel
Temperature Dew Point	**-32864.042**	-8069.476	-7046.139	-6815.687	-6570.230
Wind Speed Dew Point	-6484.255	-6533.511	-6641.954	-6706.446	**-6707.877**
Wind Speed Temperature	-6535.572	-6537.206	-6667.172	**-6669.048**	-6591.004

5. Conclusion

The increase in the prevalence of hot climate extremes, alongside global climate change, poses a threat to the anthropogenic world. These effects are particularly felt in compound extremes, multiple co-occurring climate extremes, such as temperature-humidity, temperature-wind speed, and humidity-wind speed. By developing methods to accurately statistically model these events, we can better understand and potentially troubleshoot the impacts of compound extreme events.

In this study, we focused on the usage of copula models to predict temperature-humidity, temperature-wind speed, and humidity-wind speed extremes in Boulder County, Colorado. We intercompared the fit of the Gaussian, t, Clayton, Gumbel, and Frank copulas on data records including 30, 50, 100, and 1999 years of data. By using case resampling block bootstrap schemes on climate model simulation data, we found that longer data records have lower bias and variance in estimating the true probability of a compound extreme event than shorter data records. We also found that although the choice of copula family can vary extensively with variable choice, the ideal copula family remains consistent regardless of data record length. Afterwards, we fit the copula models to observational daily summary data from Broomfield Jefferson, Colorado on 1983-2002 and 2003-2022 data to understand the ways these climate extremes have changed over time. We used dew point as a measure of humidity instead of specific humidity due to data availability. We found that although there has been a slight increase in the frequency of high temperature-low dew point, high temperature-high wind speed, and low dew point-high wind speed events in the past two decades, this difference is within the expected envelope of sampling variability explored in the CESM2 analysis in Section 4.2

There are some suggestions for future directions of research. First, it is of interest to do a similar analysis after removing the season cycle from the data. It would be interesting to also consider the frequency of relative extreme events within each season, alongside this analysis of absolute climate extreme events. Second, our current analysis violated the linearity and normality assumptions of the BIC. It would be of interest to compare this analysis to one with another metric for model selection. Third, due to computational limitations, we analyzed the bivariate distributions of temperature-humidity, temperature-wind speed, and wind speed-humidity separately. It would be interesting to run similar analyses on the trivariate temperature-humidity-wind speed distribution and compare the results to those in this study. Fourth, there is high autocorrelation between climate variables, such as temperature and humidity. It would be of interest to investigate the impact of such autocorrelation and in future research. Lastly, our current work does not consider the non-stationarity caused by climate change. It would be of interest to examine the impact of climate change on wind speed, temperature, and humidity extremes in Boulder County, CO using non-stationary copula models.

Our understanding of compound extreme events and modeling techniques still have a far way to go. By continuing to understand the implications of the assumptions made when choosing copula families, we can ensure that we use these assumptions to our advantage when understanding climate extremes.

References

Aas, Kjersti et al. “Pair-Copula Constructions of Multiple Dependence.” Insurance, mathematics & economics 44.2 (2009): 182–198. Web.

AghaKouchak, Amir et al. “Global Warming and Changes in Risk of Concurrent Climate Extremes: Insights from the 2014 California Drought.” Geophysical research letters 41.24 (2014): 8847–8852. Web.

Chuck, Elizabeth. “How climate change primed Colorado for a rare December wildfire.” *Nbcnews.com*. NBC Universal. 01 Jan 2022. Web.

“Colorado Faces Winter Urban Firestorm.” *Earth Observatory*. NASA. 30 Dec. 2021. Web.

Cooley, Daniel et al. “A Nonparametric Method for Producing Isolines of Bivariate Exceedance Probabilities.” Extremes (Boston) 22.3 (2019): 373–390. Web.

Coumou, Dim, and Stefan Rahmstorf. “A Decade of Weather Extremes.” Nature climate change 2.7 (2012): 491–496. Web.

Danabasoglu, G et al. “The Community Earth System Model Version 2 (CESM2).” Journal of advances in modeling earth systems 12.2 (2020): n. pag. Web.

Deser, Clara et al. “Communication of the Role of Natural Variability in Future North American Climate.” Nature climate change 2.11 (2012): 775–779. Web.

Friederichs, P, and A Hense. “Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression.” Monthly weather review 135.6 (2007): 2365–2378. Web.

Hao, Zengchao et al. “A Multivariate Approach for Statistical Assessments of Compound Extremes.” Journal of hydrology (Amsterdam) 565 (2018): 87–94. Web.

Hao, Zengchao, Vijay Singh, and Fanghua Hao. “Compound Extremes in Hydroclimatology: A Review.” Water (Basel) 10.6 (2018): 718–. Web.

Kelly, K S, and R Krzysztofowicz. “A Bivariate Meta-Gaussian Density for Use in Hydrology.” Stochastic hydrology and hydraulics : research journal 11.1 (1997): 17–31. Web.

Koenker, Roger, and Gilbert Bassett. “Regression Quantiles.” Econometrica 46.1 (1978): 33–50. Web.

Leonard, Michael et al. “A Compound Event Framework for Understanding Extreme Impacts.” Wiley interdisciplinary reviews. Climate change 5.1 (2014): 113–128. Web.

Martius, Olivia, Stephan Pfahl, and Clément Chevalier. “A Global Quantification of Compound Precipitation and Wind Extremes.” Geophysical research letters 43.14 (2016): 7709–7717. Web.

National Centers For Environmental Information, Integrated Surface Database (ISD). www.ncdc.noaa.gov/isd (accessed June 23, 2022).

Nelsen, R. B. (2007), An introduction to copulas, Springer Science & Business Media.

Renard, B, and M Lang. “Use of a Gaussian Copula for Multivariate Extreme Value Analysis: Some Case Studies in Hydrology.” Advances in water resources 30.4 (2007): 897–912. Web.

Ribeiro, Andreia F.S et al. “Copula-Based Agricultural Drought Risk of Rainfed Cropping Systems.” Agricultural water management 223 (2019): 105689–. Web.

Rummukainen, Markku. “Changes in Climate and Weather Extremes in the 21st Century.” Wiley interdisciplinary reviews. Climate change 3.2 (2012): 115–129. Web.

Sadegh, Mojtaba, Elisa Ragno, and Amir AghaKouchak. “Multivariate Copula Analysis Toolbox (MvCAT): Describing Dependence and Underlying Uncertainty Using a Bayesian

Framework.” Water resources research 53.6 (2017): 5166–5183. Web.

Sedlmeier, Katrin et al. “Compound Extremes in a Changing Climate - A Markov Chain Approach.” *Nonlinear processes in geophysics* 23.6 (2016): 375–390. Web.

Singh, Harsimrenjit, Mohammad Reza Najafi, and Alex J Cannon. “Characterizing Non-Stationary Compound Extreme Events in a Changing Climate Based on Large-Ensemble Climate Simulations.” *Climate dynamics* 56.5-6 (2021): 1389–1405. Web.

Tavakol, Ameneh, Vahid Rahmani, and John Harrington Jr. “Probability of Compound Climate Extremes in a Changing Climate: A Copula-Based Study of Hot, Dry, and Windy Events in the Central United States.” Environmental research letters 15.10 (2020): 104058–. Web.

Wang, Zhuo, Jun Yan, and Xuebin Zhang. “Incorporating Spatial Dependence in Regional Frequency Analysis.” Water resources research 50.12 (2014): 9570–9585. Web.

Zscheischler, Jakob et al. “Future Climate Risk from Compound Events.” Nature climate change 8.6 (2018): 469–477. Web.