Journal of *Marine Science and Engineering*

MDPI

Article

Description of the Joint Probability of Significant Wave Height and Mean Wave Period

Mingwen Zhao [1], Xiaodong Deng [2] and Jichao Wang [1,*]

[1] College of Science, China University of Petroleum, Qingdao 266580, China
[2] East China Sea Forecasting and Hazard Mitigation Center, MNR, Shanghai 200136, China
* Correspondence: wangjc@upc.edu.cn; Tel.: +86-1333-503-9103

Abstract: The bivariate probability distribution of significant wave heights and mean wave periods has an indispensable guiding role in the implementation of offshore engineering, which has attracted great attention. This work gives a new bivariate method to describe the bivariate distribution of significant wave height and mean wave period at the NanJi, BeiShuang, and XiaoMaiDao stations from 2018 to 2020. A mixed lognormal distribution is used for univariate probability analysis of wave data, and the method of connecting two mixed lognormal distributions with copula functions is applied to construct bivariate distribution. The results show that compared with Weibull and lognormal distributions, the mixed lognormal distribution shows good performance in fitting marginal distributions. In the bivariate probability analysis, the conditional model overestimates the probability of lower wave heights, and the bivariate function model has a poor fitting effect in the region with larger periods. In contrast, the copula model based on mixed lognormal distribution is more suited to describe the joint distribution of significant wave height and mean wave period.

Keywords: copula function; joint distribution; marginal distribution; mixed lognormal distribution; EM algorithm

Citation: Zhao, M.; Deng, X.; Wang, J. Description of the Joint Probability of Significant Wave Height and Mean Wave Period. *J. Mar. Sci. Eng.* **2022**, *10*, 1971. https://doi.org/10.3390/jmse10121971

Academic Editor: Christos Stefanakos

Received: 14 October 2022
Accepted: 8 December 2022
Published: 11 December 2022

1. Introduction

Some distributions are adopted to model the important wave parameters [1–5], such as significant wave heights and mean wave periods, to better understand the extremely complex marine environment, which is considered to be crucial for coastal engineering applications and the safety of offshore structures.

In practice, these wave parameters are correlated, so it is appropriate for the joint distributions to be used for statistical analysis. On the other hand, the modeled results may have a relatively large bias if we just utilize the univariate probabilistic models in the statistical analyses, so it is one-sided to study one of them alone. To provide better assistance in offshore operations and the construction of drilling platforms, studying the bivariate probability distribution of significant wave heights and mean wave periods has received growing attention recently [6–8].

Several parametric approaches have been pointed out to simulate the correlation between these two wave parameters, in which the bivariate function model has been widely employed [9–11]. Ochi [12] pointed out that a bivariate lognormal function can be used to simulate the bivariate probability of wave heights and periods. In addition, a two-dimensional Weibull model was proposed by Kimura [13] to provide a description of the statistical characteristics of wave heights and periods and discuss the influence of shape factors and related parameters on the fitting effect of the model. These bivariate methods mentioned above are very simple and easy to implement, but the problem is that the requirement for the dataset is relatively high [14]. On the other hand, the bivariate function models capture the joint behavior of two wave parameters as a whole. In general, different wave parameters should be fitted with different distributions. The conditional model can

select different probability distributions for the marginal distributions to fit [15–17]. For datasets with low correlation coefficients, the conditional model can add weight to the parts that need attention. However, the disadvantage is that it is difficult to determine the optimal expressions of the joint functions [18].

The copula function, proposed by Sklar [19], was used for the stock market and financial risk assessment. With continuous exploration, it has emerged in the joint analysis of ocean and water resources variables [20,21]. As a flexible statistical approach, copula allows any type of distribution function and has an indispensable role in the joint modeling of bivariate variables. The bivariate distribution of two ocean parameters, it can be decomposed into the marginal distributions of two parameters and a copula function, and the copula function acts as a bridge between the marginal distributions and the joint distribution. Gaussian copula, as a common copula function, is widely applied in hydrologic analysis [22–25]. In addition, Archimedean copula is also applied to establish the bivariate distribution due to its good mathematical properties. Compared with conditional modeling, Dong et al. [26] found that Clayton Copula showed good performance in bivariate probability analysis of group height and length. Iturrizaga and Zavoni [27] proposed to establish the bivariate distribution of wave heights and periods through copulas to achieve structural reliability research. Kim et al. [28] pointed out that Frank and Gaussian copula functions are most suitable for frequency analysis of wave heights and periods on the Korean Peninsula. At the same time, copula functions also show outstanding performance in the modeling of multivariate variables [29,30]; it should be noted that asymmetric copulas have certain advantages in the multivariable modeling process, which can more adequately simulate the asymmetric correlation between variables [31]. For extreme events, three extreme copulas were used by Mazas and Hamm [32] to model the bivariate distributions of wave heights and sea levels, and the return periods were discussed. Li et al. [33] showed that Gumbel–Hougaard copula could well fit the joint characteristics of extreme waves and surges.

In previous work, simple parametric probabilistic methods have often been used to fit marginal distributions, such as the Weibull distribution, Forristall distribution, lognormal distribution, and Gamma distribution [17,26,34]. Owing to their simple forms, it is difficult to describe the marginal distributions adequately when the probability distributions show some special features. Therefore, based on the advantage that copulas allow any type of marginal distribution function, a mixed lognormal distribution is proposed to fit the marginal distributions, which are combined with three common Archimedean copula functions to establish the bivariate distributions.

The rest of this paper is arranged as follows. Section 2 introduces the approaches to fitting univariate distribution and gives the construction of mixed lognormal distribution in detail. At the same time, it also presents the methods of describing joint distribution. Section 3 gives the research area and provides a statistical analysis of data. In Section 4, we discuss the fitting of marginal distribution and analyze the joint probability of significant wave heights and wave periods. In addition, the findings are summarized in Section 5.

2. Methodology

2.1. Univariate Distribution Methods

2.1.1. Common Distribution

In this work, we assume that x and y represent the significant wave height (H_s) and mean wave period (T_z), respectively. Weibull and lognormal distributions, as two common distributions, have been obtained as the widespread application in statistical analysis of wave parameters [18,35]. Its probability density functions are defined as follows:

$$f(x,\alpha,\beta) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} exp\left\{-\left(\frac{x}{\beta}\right)^{\alpha}\right\},\ x>0 \tag{1}$$

$$f(y,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma y}\exp\left\{-\frac{[\ln(y)-\mu]^2}{2\sigma^2}\right\},\ x>0 \tag{2}$$

where α and β are the shape and scale parameters, and μ and σ are the expectation and standard deviation of $\ln x$, respectively.

2.1.2. Mixed Lognormal Distribution

It is assumed that the distribution of the ocean parameter, z, can be represented by a mixed lognormal distribution, which is composed of k parts. The specific form is as follows:

$$f(z|\xi) = \sum_{j=1}^{k} \alpha_j f_j(z|\theta_j) \tag{3}$$

where $\xi = (\alpha_1, \cdots, \alpha_k, \theta_1, \cdots, \theta_k)$ indicates a series of parameters of the mixed lognormal distribution; α_j denotes the scale factor satisfying $\sum_{j=1}^{k} \alpha_j = 1$. θ_j is the parameter of the jth lognormal distribution. In general, k is required to be greater than or equal to 2. When $k = 1$, the mixed lognormal distribution becomes a general lognormal distribution.

Evaluating the value of k is a primary consideration when applying the mixed lognormal distribution to the probability analysis of ocean parameters. Bayesian Information Criterion (BIC) is a practical method to select the number of mixture components. It penalizes the model more when the amount of data is large and prefers to choose the simple model with fewer parameters, which can effectively prevent the occurrence of overfitting problems. Therefore, the BIC is used to determine the best value of k for the mixed lognormal distribution in the following form:

$$BIC = -2L(\xi) + N_\xi \log(N) \tag{4}$$

in which N denotes the sample size and N_ξ represents the total amount of parameters of the mixed lognormal distribution; $L(\xi)$ is the log-likelihood function, with the specific form as follows:

$$L(\xi) = \sum_{i=1}^{N} \log f(x_i|\xi) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{k} \alpha_j f_j(x_i|\theta_j)\right) \tag{5}$$

Usually, the smaller the BIC value, the better. Therefore, we determine the optimal k value by minimizing the BIC.

In view of the parameter complexity of the mixed lognormal distribution, it is difficult to be solved by maximum likelihood estimation (MLE). In the present work, we introduce the Expectation-Maximization (EM) algorithm, which is a method of maximum posterior probability estimation. Each iteration of the EM algorithm needs to go through two parts: the E step and the M step. The main task of the E step is to obtain the expectation of the likelihood function through the samples and the proposed model, which is usually called the Q function, expressed as:

$$Q(\xi) = \sum_{i=1}^{N}\sum_{j=1}^{k} \frac{\alpha_j f_j(x_i|\theta_j)}{\sum_{s=1}^{k} \alpha_s f_s(x_i|\theta_s)} \ln\left(f_j(x_i|\theta_j)\right) + \sum_{i=1}^{N}\sum_{j=1}^{k} \frac{\alpha_j f_j(x_i|\theta_j)}{\sum_{s=1}^{k} \alpha_s f_s(x_i|\theta_s)} \ln \alpha_j \tag{6}$$

The main task of the M step is to calculate ξ when the function $Q(\xi)$ reaches its maximum value through continuous iteration, the specific steps are as follows:

$$\xi^{(r+1)} = argmax\ Q\left(\xi^{(r)}\right) \tag{7}$$

The former part of Equation (6) is maximized by Gibbs' inequality, and the proportion coefficient of the mixed lognormal distribution is obtained as follows:

$$\alpha_j^{(r+1)} = \frac{1}{N}\sum_{i=1}^{N} \frac{\alpha_j^{(r)} f_j(x_i|\theta_j^{(r)})}{\sum_{s=1}^{k} \alpha_s^{(r)} f_s(x_i|\theta_s^{(r)})} \tag{8}$$

The value of θ is obtained by taking the derivative of the second half:

$$\frac{\partial Q(\xi)}{\partial \theta} = 0 \tag{9}$$

It should be noted that the EM algorithm can only guarantee that the parameter estimation sequence converges to the stable point but not to the maximum point. Therefore, in the application, the selection of the initial value becomes very important. The common method is to select several different initial values randomly for iteration and then compare the estimated values to choose the best one. In addition, to ensure the accuracy of the solution, the following stopping criteria are selected:

$$\left|\xi^{(r+1)} - \xi^{(r)}\right| < \varepsilon \ (\varepsilon > 0) \tag{10}$$

By replacing the probability density function $f_j(z|\theta_j)$ in Equation (3) with the distribution in Equation (2), the specific forms of the mixed lognormal distribution can be obtained, which will be used to fit H_s and T_z.

2.2. Joint Distribution Models

2.2.1. Conditional Model

The conditional model is a method to obtain joint probability density based on the total probability theorem. It needs to know the density function of H_s and the density function of T_z conditional on H_s. The formula is as follows:

$$f(x, y) = f(x) \times f(y|x) \tag{11}$$

where $f(x, y)$ is the bivariate probability function of two ocean parameters, $f(x)$ is the marginal probability of H_s, $f(y|x)$ is the conditional probability of T_z. In this study, Weibull distribution in Equation (1) and lognormal distribution in Equation (2) are selected as the specific forms of $f(x)$ and $f(y|x)$, respectively. The parameters of $f(x)$ can be obtained by MLE, and the parameters of $f(y|x)$ can be calculated as follows:

$$f(y|x) = \frac{1}{y\sqrt{2\pi}\sigma_y(x)} \exp\left\{-\frac{\left[\ln(y) - \mu_y(x)\right]^2}{2\sigma_y(x)^2}\right\} y > 0 \tag{12}$$

$$\begin{cases} \mu_y(x) = A_1 + A_2 x^{A_3} \\ \sigma_y(x) = B_1 + B_2 exp(B_3 x) \end{cases} \tag{13}$$

where $\mu_y(x)$ and $\sigma_y(x)$ are the expectation and standard deviation of $\ln(y)$. By dividing H_s into several intervals, we can get the $\mu_y(x)$ and $\sigma_y(x)$ of corresponding intervals. In this way, we can get several arrays of $(x, \mu_y(x))$ and $(x, \sigma_y(x))$, and obtain the coefficients A_i and B_i of Equation (13) through the nonlinear fitting.

2.2.2. Bivariate Function Model

A bivariate lognormal distribution was pointed out by Ochi [12] to describe the joint characters of wave heights and periods and performed well in the case of small wave heights, with the specific formula as follows:

$$f(x, y) = \frac{1}{2\pi x y \sigma_x \sigma_y \sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(\ln x - \mu_x)^2}{\sigma_x^2} - \frac{2\rho(\ln x - \mu_x)(\ln y - \mu_y)}{\sigma_x \sigma_y} + \frac{(\ln y - \mu_y)^2}{\sigma_y^2}\right]\right\} \tag{14}$$

where μ_x, σ_x are the expectation and standard deviation of $\ln x$, μ_y, σ_y are the expectation and standard deviation of $\ln y$. ρ is the correlation coefficient of x and y.

2.2.3. Copula Model

A copula function, which describes the correlation between variables, connects the joint distribution and the marginal distributions. The bivariate distribution of two ocean parameters x and y can be obtained by the following formula:

$$F(x,y) = C(G(x), H(y)) \tag{15}$$

where $F(x,y)$ is the bivariate distribution, $G(x)$ and $H(y)$ are the marginal distributions of x and y, respectively and $C(\cdot,\cdot)$ is a copula function. By taking the derivative of Equation (15), the probability density function (PDF) $f(x,y)$ of variables x, y can be obtained:

$$f(x,y) = c(G(x), H(y))\cdot g(x)\cdot h(y) \tag{16}$$

where $g(x)\cdot$ and $h(y)\cdot$ are the PDF of marginal distributions $G(x)$ and $H(y)$, respectively. In addition, $c(\cdot,\cdot)$ can be given by the following Equation (16).

$$c(G(x), H(y)) = \frac{\partial C(G(x), H(y))}{\partial G(x)\partial H(y)} \tag{17}$$

Archimedean copula, as a common copula family, has been widely used in marine and coastal engineering because of its easy construction and calculation and some good properties. Therefore, Gumbel, Clayton, and Frank copulas are used in the bivariate statistical analysis of H_s and T_z. Table 1 presents their PDF and generator functions, in which u and v are distribution functions, θ is a parameter that represents the correlation between variables.

Table 1. The structure of three copula functions.

Copulas	Bivariate Density Function	Generator
Gumbel	$\frac{exp\left\{-\left[(-\ln u)^\theta+(-\ln v)^\theta\right]^{\frac{1}{\theta}}\right\}\times(\ln u\times\ln v)^{\theta-1}}{u\times v\times\left[(-\ln u)^\theta+(-\ln v)^\theta\right]^{2-\frac{1}{\theta}}}$ $\times\left\{\left[(-\ln u)^\theta+(-\ln v)^\theta\right]^{1/\theta}+\theta-1\right\}$	$(-\ln t)^\theta$
Clayton	$(1+\theta)\times(u\times v)^{-\theta-1}\times\left(u^{-\theta}+v^{-\theta}-1\right)^{-2-1/\theta}$	$\frac{1}{\theta}\left(t^{-\theta}-1\right)$
Frank	$\frac{\theta\times exp[\theta\times(1+u+v)]\times(\exp(\theta)-1)}{(\exp(\theta)-\exp(\theta+\theta\times u)+\exp(\theta\times u+\theta\times v)-\exp(\theta+\theta\times v))^2}$	$-\ln\frac{e^{-\theta t}-1}{e^{-\theta}-1}$

2.2.4. Goodness of Fit

To assess the fitting ability of these models to the joint sample, the squared Euclidean distance (D^2) was introduced [18]. Specifically, the bivariate space is divided into m parts in H_s direction and n parts in T_p direction. D^2 can be expressed as:

$$D^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij} - q_{ij}\right)^2 \tag{18}$$

where p_{ij} is the probability obtained from the original data satisfying $x_i < x \leq x_{i+1}$ $(i = 1, 2, \cdots m)$, $y_j < y \leq y_{j+1}$ $(j = 1, 2, \cdots n)$. Similarly, q_{ij} is the probability obtained from the model satisfying the above conditions. Generally speaking, a lower value of D^2 indicates a better fitting effect of the model. It should be noted that D^2 can only compare the goodness of fit of different models on the same data set because the number of samples will affect the size of D^2 value. However, this does not affect our ability to compare the performance of several models in fitting the joint distribution of significant wave height and mean wave period. At the same time, the size of the grid also affects the D^2 value, but it has no effect on the final conclusion. In this study, the size of the bins is 0.3 m × 0.6 s. Of

course, the grid division in this study is not arbitrary, and it is consistent with the division of significant wave height and mean wave period in the process of univariate analysis.

3. Study Area and Data Analysis

The wave data are from the National Marine Data Center (http://mds.nmdis.org.cn/, accessed on 1 June 2022.). Two stations in the East China Sea, namely NanJi (NJ) station and BeiShuang (BS) station, are selected as the research objects, as shown in Figure 1. Their specific coordinates are 27.5° N 121.1° E and 26.7° N 120.3° E, respectively. The East China Sea is an important transportation hub for China's maritime interactions with various countries in the Pacific region. Here, the cold and warm currents converge, and the seawater exchanges smoothly, which is one of the important fishing grounds in China. Therefore, the analysis of wave characteristics and parameters, especially the joint distribution of H_s and T_z, is of positive significance for understanding the wave characteristics in the East China Sea and has important practical value for the design of the offshore structure, prevention of marine disasters, and navigation. In addition, the XaiMaiDao (XMD) station located in the Yellow Sea is selected to further verify the applicability of the methods proposed in this study, with a specific coordinate of 36.0° N 120.4° E. All data are from 2018 to 2020, and the sampling frequency is one hour. It is inevitable that there will be a small amount of missing observation data. Except for the missing part, the rest will be used for simulation experiments.

Figure 2 displays the scatter plots and histograms of H_s and T_z of the three stations. For the NJ and BS stations, the wave heights are mainly in the range of 0.5–2 m, and the wave periods are mainly in the range of 4–8 s. As for the XMD station, most of the wave heights are in the interval of 0–1 m, and most of the wave periods are in the interval of 3–7 s. In addition, Table 2 shows the statistical information of H_s and T_z. It can be found that the data of the XMD station is quite different from that of the other two stations. Therefore, it is feasible to use the XMD station to further illustrate the applicability of the proposed method. As can be seen from the skewness, the probability distribution curves of the two wave parameters are skewed to the right; on the other hand, kurtosis indicates that the probability distribution curves are steep. When fitting the probability density distributions of H_s and T_z, we should pay attention to these characteristics of the distribution curves.

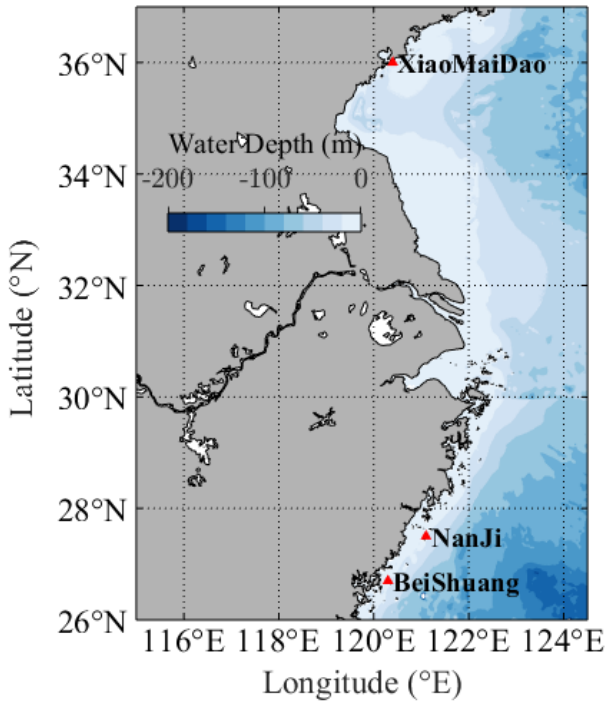

Figure 1. Map of NanJi, BeiShuang, and XiaoMaiDao stations.

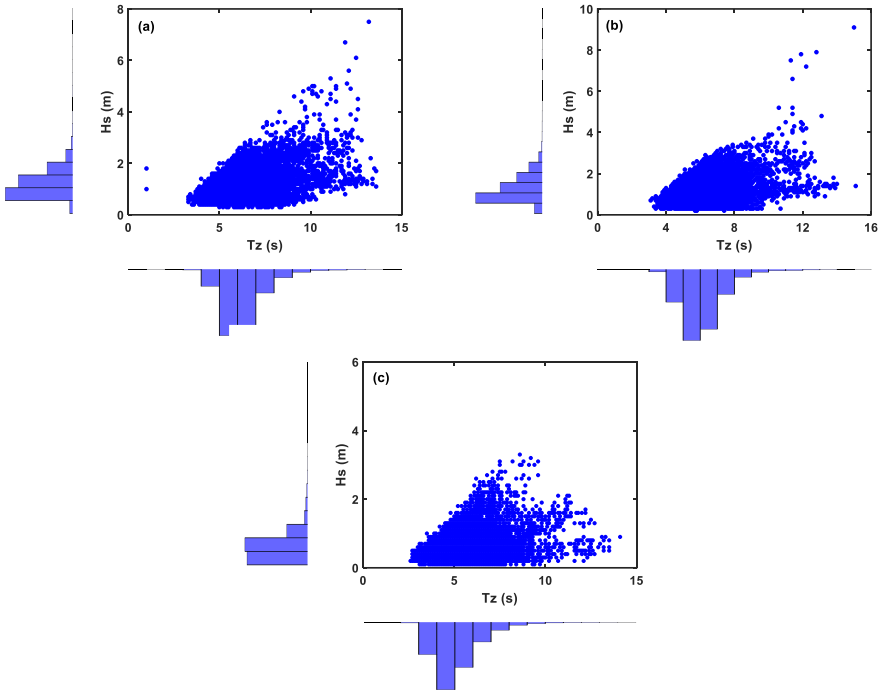

Figure 2. Scatter plots and histograms of H_s and T_z: (**a**) NJ station; (**b**) BS station; (**c**) XMD station.

Table 2. Statistics for ocean data.

Dataset	Statistic	Mean	Standard Deviation	Kurtosis	Skewness
NJ	H_s (m)	1.1169	0.4924	10.1386	1.6093
	T_z (s)	6.2277	1.2250	6.7266	1.4120
BS	H_s (m)	1.1468	0.5966	9.9534	1.4862
	T_z (s)	6.0671	1.2975	7.1869	1.4275
XMD	H_s (m)	0.5069	0.3435	9.9329	2.0887
	T_z (s)	5.0290	1.3177	7.4531	1.5932

4. Results and Discussion

4.1. Fitting the Marginal Distributions

First, we conduct an experimental analysis on NJ and BS stations. Before constructing the bivariate distribution of H_s and T_z, it is necessary to conduct probability analysis on each variable to determine their marginal distributions. Weibull distribution has been applied to fit the wave heights [36], and lognormal distribution is a better approach for probability analysis of wave periods [14]. Therefore, in order to adequately explain the advantages of the mixed lognormal distributions in estimating the probability of H_s and T_z, the fitting results are compared with Weibull and lognormal distributions, respectively.

The probability analysis of H_s and T_z from NJ and BS stations is carried out in this section, it is necessary to determine the parameter values of distributions before fitting the wave data. For Weibull and lognormal distributions, the parameters can be solved directly by MLE. On the other hand, the quantity of mixed components of the mixed lognormal distribution is given by *BIC* (Table 3), and the results of other parameters are obtained with the help of the EM algorithm. Figure 3 presents the probability density functions and frequency histograms of H_s from NJ and BS stations. Through intuitive comparison, we can find that the Weibull distribution is not enough to predict H_s in the middle region, and the mixed lognormal distribution performs well on the whole. As shown in Figure 3b, the prediction of H_s in the range of 1.2 to 1.8 by Weibull distribution is higher than the empirical value. At the same time, the empirical distributions and marginal distributions from the two methods are also plotted in Figure 3. There are obvious differences between

the Weibull distributions and the empirical distributions, especially in the middle area. In contrast, the curves of the mixed lognormal distributions are basically consistent with the curves of the empirical distributions. Therefore, the mixed lognormal distribution may be a new option to effectively fit H_s.

Table 3. Number of parts of mixed lognormal distribution based on experimental data.

Station	Variable	Number of Components	$BIC(\times 10^4)$
NJ	H_s (m)	3	2.3558
	T_z (s)	2	6.3125
BS	H_s (m)	4	2.2244
	T_z (s)	3	4.8030

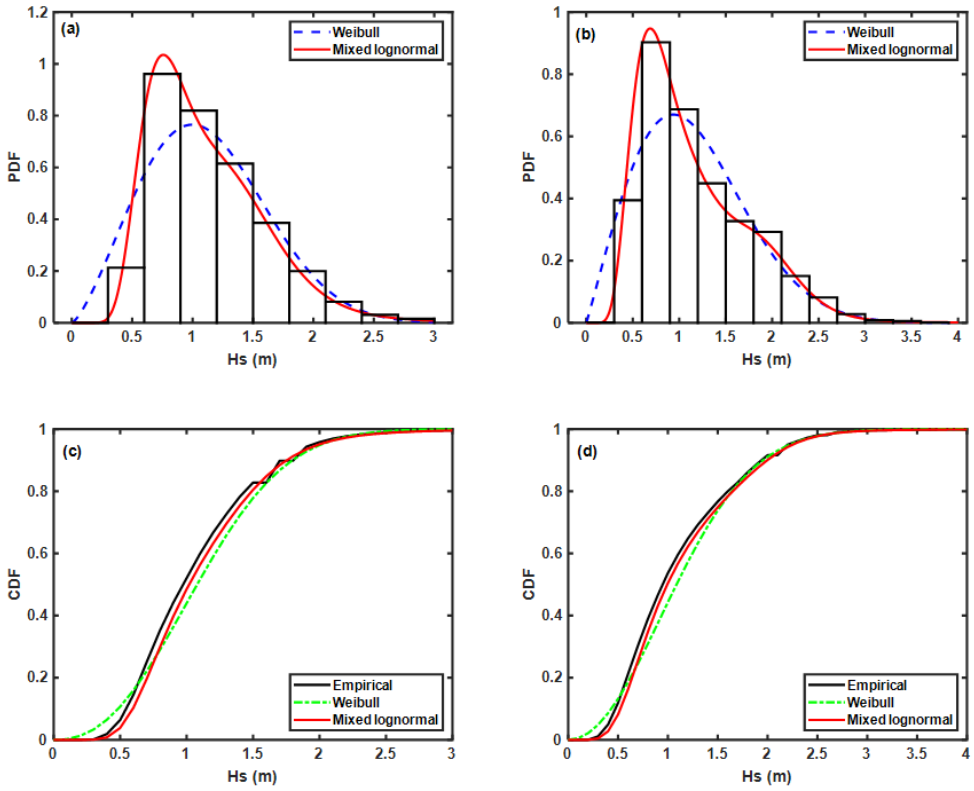

Figure 3. Probability density functions and Cumulative distributions of H_s : (**a**,**c**) NJ station; (**b**,**d**) BS station.

The lognormal distributions and mixed lognormal distributions are used to describe T_z of NJ and BS stations, and their probability density functions are given in Figure 4. According to Figure 4a, the lognormal distribution does not provide enough probability prediction in the range of 5 to 6 and overestimates the probability in the range of 7 to 8. By comparison, the mixed lognormal distribution can adequately fit the distribution characteristics of T_z. The empirical distributions generated from the two datasets are shown in Figure 4c,d, from which it can be concluded that the fitting accuracy of the mixed lognormal distribution is higher. In order to more specifically verify the performance of mixed lognormal distribution in describing marginal probability distribution, this paper uses the root mean square error ($RMSE$) as an evaluation index and defines it as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_e(x_i) - F_m(x_i))^2} \tag{19}$$

where $F_e(x)$ represents the empirical distribution and $F_m(x)$ the theoretical distribution. Table 4 presents the values of the $RMSE$ test. In general, the smaller the value of $RMSE$, the better the performance of this method. The results show that the mixed lognormal distribution is quite different from the other two methods and has a smaller $RMSE$ value, indicating that it produces better consistency in fitting field data.

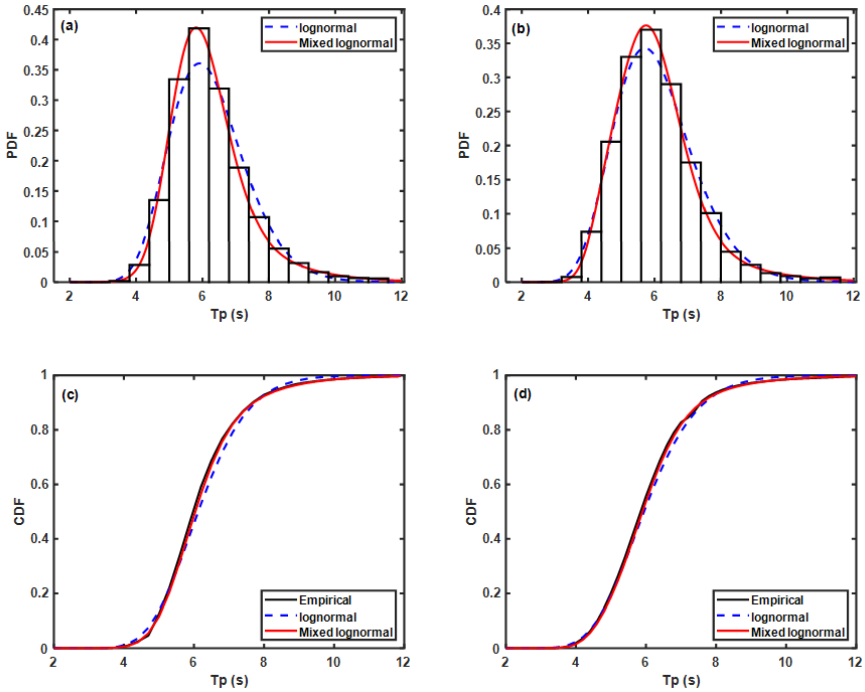

Figure 4. Probability density functions and Cumulative distributions of T_z : (**a**,**c**) NJ station; (**b**,**d**) BS station.

Table 4. Evaluation indexes of methods for experimental data.

	Distribution	*RMSE*
H_s for NJ	Weibull	0.0578
	Mixed lognormal	0.0373
T_z for NJ	lognormal	0.0381
	Mixed lognormal	0.0172
H_s for BS	Weibull	0.0570
	Mixed lognormal	0.0323
T_z for BS	lognormal	0.0272
	Mixed lognormal	0.0144

The mixed lognormal distribution, as a flexible statistical method, performs well in univariate analysis of H_s and T_z. When the marginal distribution shows some special features, such as heavy tail and saddle shape. Because of its simplicity, it is difficult to capture these characteristics for Weibull distribution and lognormal distribution, in which case the mixed lognormal distribution will show a greater advantage. At the same time, the mixed lognormal distribution may be a new option in the extreme analysis of ocean parameters.

4.2. Fitting the Bivariate Distributions

It is not easy to construct the joint distributions of H_s and T_z in the mixed sea state [14]. The main purpose of this section is to apply the conditional model, bivariate function model, and copula model to establish the bivariate distributions of wave data from NJ and

BS stations under the total sea state. In order to assess the fitting ability of these models, the squared Euclidean distance is introduced.

4.2.1. Bivariate Distributions with the Conditional Model and Bivariate Function Model

When using the conditional model to establish the bivariate distribution, it is important to note that the interval of dividing H_s may affect the final fitting performance [24]. In this present work, the interval of H_s is selected as 0.25 m, and the parameter values of the conditional model can be obtained by nonlinear fitting. For the bivariate function model, the bivariate lognormal model has been proven to be useful for describing the joint behavior of H_s and T_z. The parameters of the bivariate lognormal model can be obtained by MLE based on experimental data.

The contour plots obtained based on the conditional model and the bivariate lognormal model are shown in Figures 5 and 6. To intuitively assess the applicability of the above two models, the empirical distributions obtained from the experimental data are also drawn in the contour plots. As can be summarized from Figures 5a and 6a, the conditional model overestimates the probability of lower wave heights and larger periods for NJ and BS stations. In addition, the conditional model fits the contours of the empirical distributions poorly. Figure 6b presents the contour plot of the bivariate lognormal model at BS station, and we can find that this model underestimates the probability of larger wave heights. The conditional model and the bivariate lognormal model have larger the squared Euclidean distance, which further verifies that the two methods have a poor ability to fit joint samples, as shown in Table 5.

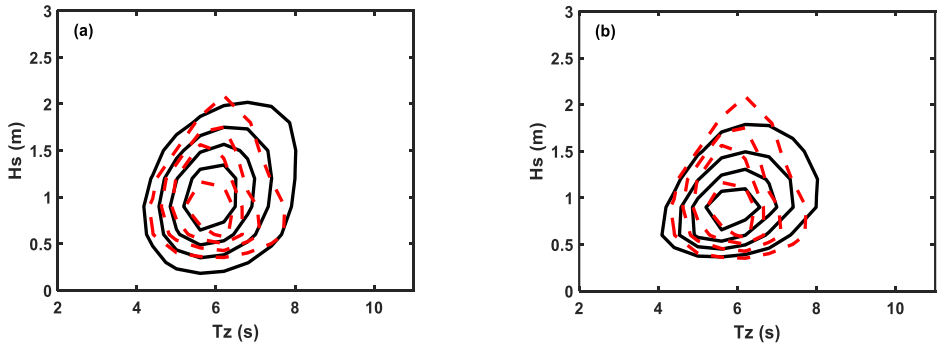

Figure 5. (**a**) Contour plot of conditional model for the NJ station; (**b**) Contour plot of bivariate lognormal model for the NJ station. Red lines are generated from raw data.

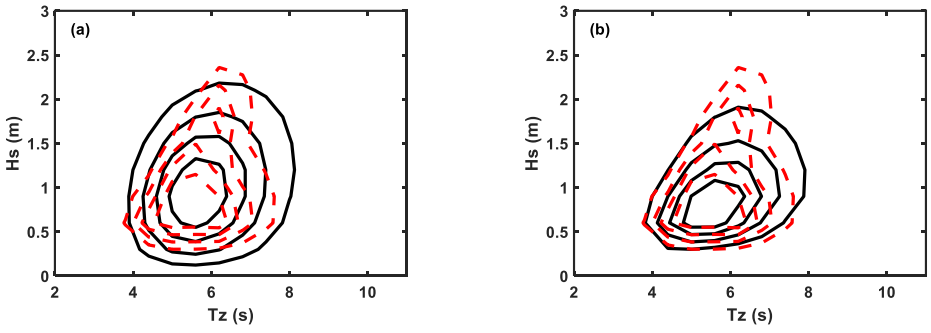

Figure 6. (**a**) Contour plot of conditional model for the BS station; (**b**) Contour plot of bivariate lognormal model for the BS station. Red lines are generated from raw data.

Table 5. The D^2 for each model.

Model	H_sMarginal	T_zMarginal	$D^2$for NJ	$D^2$for BS
Conditional model	Weibull	lognormal	0.1025	0.1211
Bivariate lognormal model	-	-	0.0961	0.0782
Gumbel copula model	Weibull	lognormal	0.1120	0.1158
	Mixed lognormal	Mixed lognormal	0.0436	0.0582
Clayton copula model	Weibull	lognormal	0.1278	0.1255
	Mixed lognormal	Mixed lognormal	0.0407	0.0600
Frank copula model	Weibull	lognormal	0.1238	0.1164
	Mixed lognormal	Mixed lognormal	0.0564	0.0655

4.2.2. Bivariate Distributions with the Copula Model

In contrast to conditional modeling and bivariate lognormal distribution, the copula method is different in that it has no requirement on the distribution form of the two connected variables. Therefore, in this section, two types of copula models are introduced: one is that Weibull distribution is chosen to fit H_s, and lognormal distribution is chosen to fit T_z, and then construct bivariate distribution by copula function. The other is that combine the mixed lognormal distributions proposed in Section 2 with the copula function to obtain a new copula model for constructing the bivariate distribution of H_s and T_z. In view of the variety of copula functions, the Gumbel, Clayton, and Frank copulas are selected for analysis and comparison. The parameters of copulas are solved by the $copulafit$ function in the MATLAB toolbox.

Figures 7 and 8 display the contour plots of the copula models connecting Weibull and lognormal distributions. For the NJ station, all three copula models overestimate the probability of a larger H_s and T_z, which may be caused by the inadequate fitting of Weibull and lognormal distributions to the marginal distributions. For the BS station, the Gumbel and Clayton copula models have poor fitting for the lower wave heights. Meanwhile, the Frank copula model underestimates the probability of smaller wave heights. From Table 5, we can see that the copula models cannot achieve satisfactory results by connecting the Weibull distribution and lognormal distribution to establish bivariate distributions.

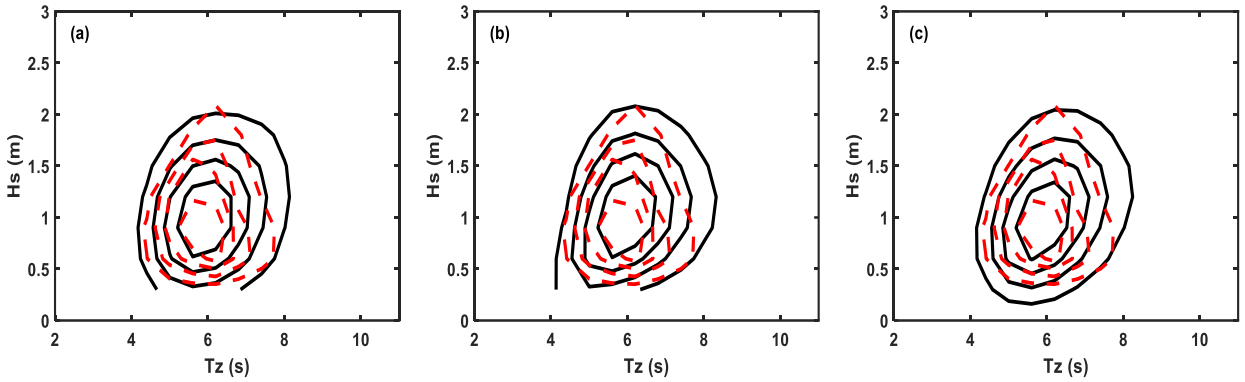

Figure 7. Contour plots of NJ station: (**a**) Gumbel copula model H_s-Weibull and T_z-lognormal; (**b**) Clayton copula model H_s-Weibull and T_z-lognormal; (**c**) Frank copula model H_s-Weibull and T_z-lognormal. Red lines are generated from raw data.

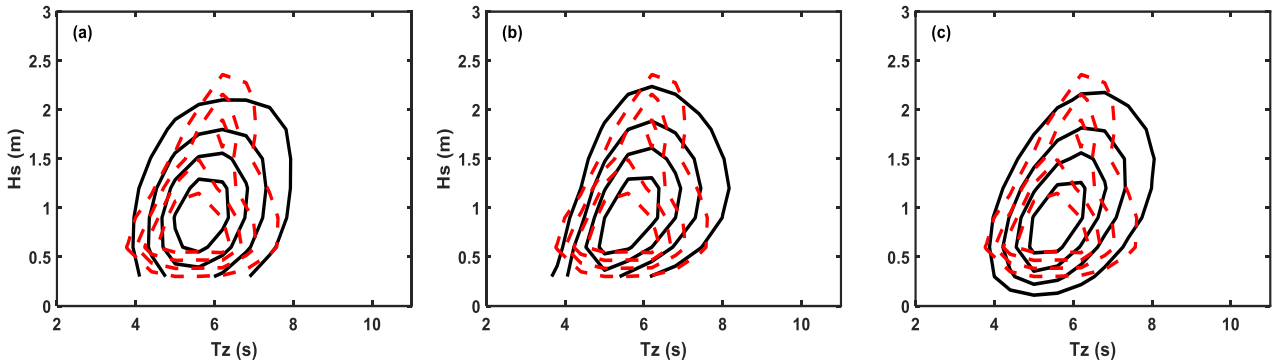

Figure 8. Contour plots of BS station: (**a**) Gumbel copula model H_s-Weibull and T_z-lognormal; (**b**) Clayton copula model H_s-Weibull and T_z-lognormal; (**c**) Frank copula model H_s-Weibull and T_z-lognormal. Red lines are generated from raw data.

Weibull and T_z-lognormal; (**b**) Clayton copula model H_s-Weibull and T_z-lognormal; (**c**) Frank copula model H_s-Weibull and T_z-lognormal. Red lines are generated from raw data.

The analysis in Section 4.1 shows that the mixed lognormal distributions are more suitable for fitting the probability distributions of H_s and T_z than the Weibull distribution and lognormal distribution. Combining two mixed lognormal distributions with copula function may be a good option to effectively simulate the bivariate distribution of H_s and T_z. The contour plots of copula models with mixed lognormal distributions as the marginal distributions are shown in Figures 9 and 10. From this figure, it can be found that the three copula functions based on the mixed lognormal distributions can better fit the curve of the empirical distributions.

From the contour plots, it is difficult to determine which of the three copula functions performs best. Table 5 gives the values of the squared Euclidean distance D^2 for NJ and BS stations. For the NJ station, the D^2 of Gumbel, Clayton, and Frank copulas are 0.0436, 0.0407, and 0.0564, respectively. According to the data analysis in Section 3, it is concluded that the wave height data accounts for the largest proportion in the range of 0.5–1.5 m. The Clayton copula fits the contour plot of the empirical distribution best in the above range, as shown in Figure 9, which may be the reason for the smallest value of D^2 for the Clayton copula. For the BS station, the D^2 of Gumbel, Clayton, and Frank copulas are 0.0582, 0.0600, and 0.0655, respectively. The Gumbel copula is optimal for the BS station.

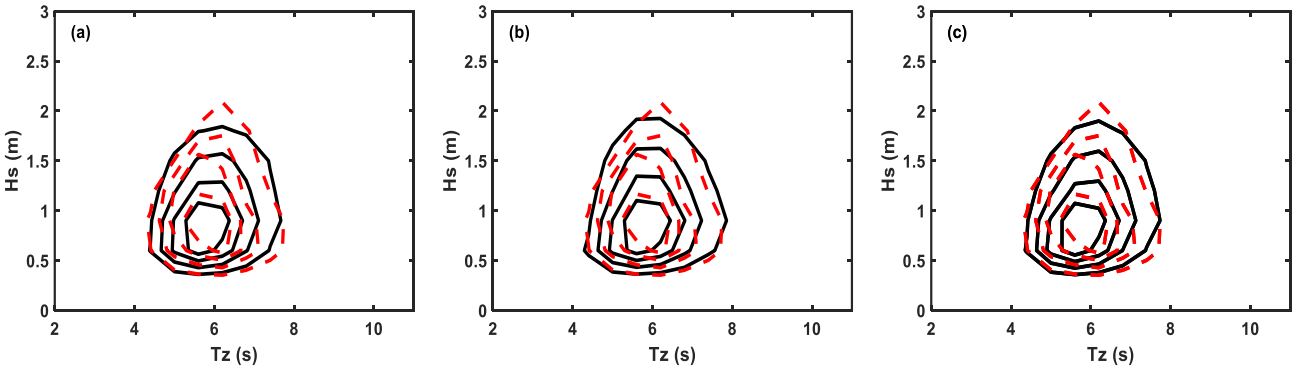

Figure 9. Contour plots of NJ station: (**a**) Gumbel copula model based on mixed lognormal distribution; (**b**) Clayton copula model based on mixed lognormal distribution; (**c**) Frank copula model based on mixed lognormal distribution. Red lines are generated from raw data.

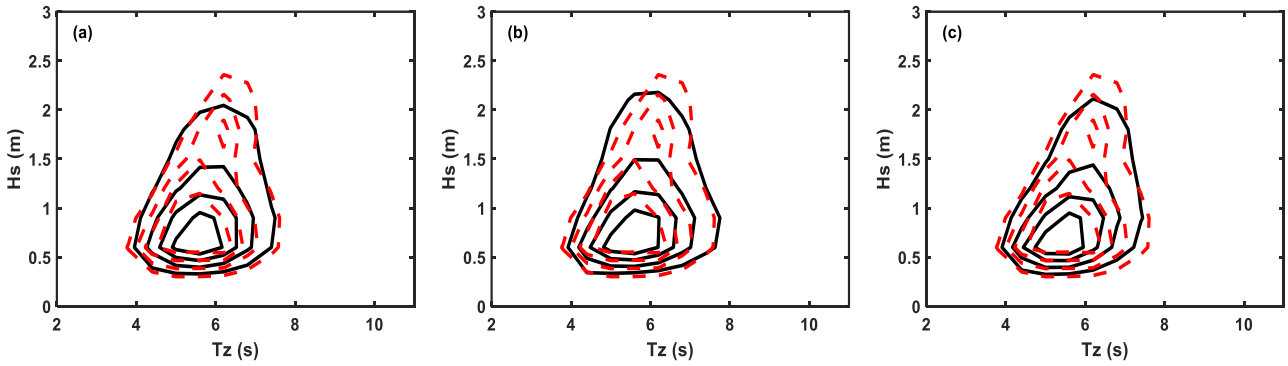

Figure 10. Contour plots of BS station: (**a**) Gumbel copula model based on mixed lognormal distribution; (**b**) Clayton copula model based on mixed lognormal distribution; (**c**) Frank copula model based on mixed lognormal distribution. Red lines are generated from raw data.

4.3. Verification

In view of the similarity of the data from the two stations in the East China Sea, it is insufficient to demonstrate the wide applicability of the proposed method. Therefore, the three bivariate models involved in this study, namely the conditional model, the bivariate lognormal model, and the copula model based on mixed lognormal distribution, are applied to the XMD station to further illustrate the wide applicability of the new copula method. Since the data of the XMD station is different from that of the other two stations, in order to make the contour plot size appropriate, we have made appropriate adjustments to the grid division. Although this affects the size of D^2, it does not affect the performance comparison of several methods. In addition, the parameters of all bivariate models are obtained by using the solution methods mentioned earlier.

The contour plots obtained based on the conditional model and the bivariate lognormal model are shown in Figure 11. From this figure, it can be found that the conditional model overestimates the probability of larger wave heights and periods. At the same time, the bivariate lognormal model also overestimates the probability of larger wave periods. The D^2 of conditional model and bivariate, lognormal model is 1.9359 and 0.8465, respectively, which further verifies that the two methods have poor ability to fit joint samples. Figure 12 displays the contour plots of the copula models based on mixed lognormal distribution. It can be found that the three copula models can better fit the curve of the empirical distribution. The D^2 of Gumbel, Clayton, and Frank copulas are 0.4314, 0.5111, and 0.4986, respectively. By comparison, the Clayton copula is optimal for the XMD station.

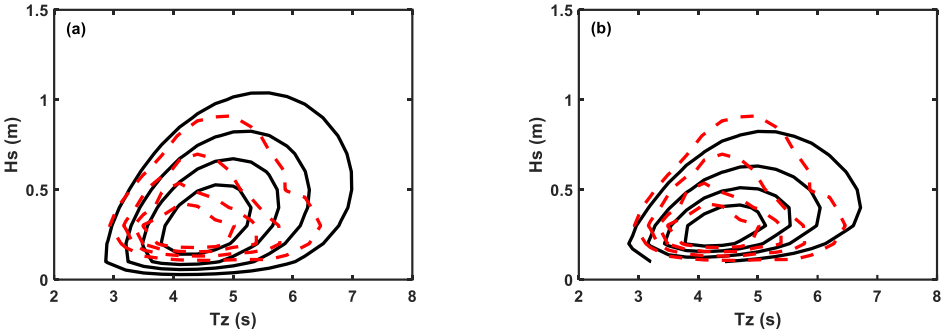

Figure 11. Contour plots of XMD station: (**a**) Conditional model; (**b**) Bivariate lognormal model. Red lines are generated from raw data.

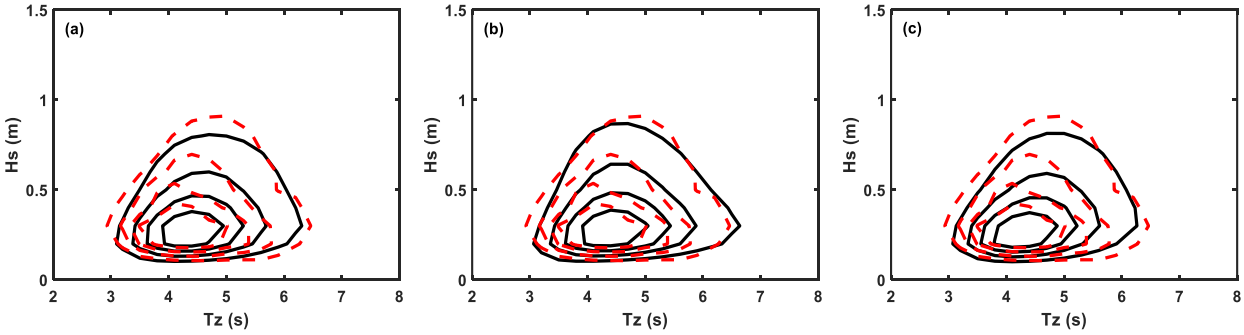

Figure 12. Contour plots of XMD station: (**a**) Gumbel copula model based on mixed lognormal distribution; (**b**) Clayton copula model based on mixed lognormal distribution; (**c**) Frank copula model based on mixed lognormal distribution. Red lines are generated from raw data.

Obviously, the biggest problem of the conditional model and bivariate lognormal model is poor flexibility. They may perform well in some special cases, but the fitting effect is not satisfactory in most cases. After the analysis of univariate fitting, a copula method based on mixed lognormal distribution is proposed to establish the joint distribution of H_s and T_z, and satisfactory results are obtained. The characteristics of the copula function provide the basis for selecting the optimal distribution forms for the marginal distributions of two variables, which directly affects the performance of the constructed bivariate distribution. Of course, the choice of copula function type is also a problem, which reflects the coupling of two ocean parameters. We can choose the best one by comparing D^2 of several copula models.

5. Conclusions

In this work, two mixed lognormal distributions are connected by copula function to establish the bivariate distribution of H_s and T_z, and compared with the conditional model and bivariate function model. The squared Euclidean distance D^2 is used to verify the fitting performance of these models.

Since the copula function allows any type of marginal distribution, in order to obtain the optimal form of marginal distribution, the probability distributions of H_s and T_z are analyzed. The experimental results show that the Weibull distribution fits the probability distribution of H_s poorly, and the lognormal distribution underestimates the probability of T_z in the middle region. In contrast, the mixed lognormal distribution, as a flexible statistical method, provides satisfactory fitting results for both H_s and T_z. Although the solution of many parameters brings inevitable drawbacks to the mixed lognormal distribution, the EM algorithm can effectively solve this problem. In the analysis of bivariate probability, the conditional model and bivariate function model have poor fitting effect in the region with larger T_z. The copula model, which connects Weibull and lognormal distributions, performs poorly in predicting the probability of smaller H_s. In comparison, the copula model based on mixed lognormal distribution is more suited to simulate the joint distribution of H_s and T_z.

An accurate prediction of the joint distribution of H_s and T_z is of high practical value in the design of maritime structures and mitigation of marine disasters. The method of connecting mixed lognormal distributions by copula function may be a new option to effectively simulate the joint distribution.

Author Contributions: Conceptualization, J.W.; methodology, J.W.; data curation, J.W.; supervision, J.W.; project administration, J.W.; funding acquisition, J.W.; investigation, X.D.; resources, X.D.; software, X.D.; formal analysis, M.Z.; writing—original draft preparation, M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the National Natural Science Foundation of China [grant number 42176011, 41976201]; and the Shandong Provincial Natural Science Foundation [grant number ZR2020MD060].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets during the current study can be found on the National Marine Science Data Center.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chou, C.R.; Yan, S.S.; Fang, H.M. Wave height distributions around submerged structure in wave-current field. *Eng. Anal. Bound. Elem.* **1997**, *20*, 45–49. [CrossRef]
2. Rodríguez, G.; Soares, C.G.; Pacheco, M. Wave period distribution in mixed sea-states. *J. Offshore Mech. Arct. Eng.* **2004**, *126*, 105–112. [CrossRef]
3. Hou, Y.; Guo, P.; Song, G.; Song, J.; Yin, B.; Zhao, X. Statistical distribution of nonlinear random wave height. *Sci. China Ser. D* **2006**, *49*, 443–448. [CrossRef]
4. Martucci, G.; Carniel, S.; Chiggiato, J.; Sclavo, M.; Lionello, P.; Galati, M.B. Statistical trend analysis and extreme distribution of significant wave height from 1958 to 1999–an application to the Italian Seas. *Ocean Sci.* **2010**, *6*, 525–538. [CrossRef]
5. Gibson, R.; Christou, M.; Feld, G. The statistics of wave height and crest elevation during the December 2012 storm in the North Sea. *Ocean Dyn.* **2014**, *64*, 1305–1317. [CrossRef]
6. Baarholm, G.S.; Haver, S.; Økland, O.D. Combining contours of significant wave height and peak period with platform response distributions for predicting design response. *Mar. Struct.* **2010**, *23*, 147–163. [CrossRef]
7. Zhang, H.D.; Soares, C.G. Modified joint distribution of wave heights and periods. *China Ocean Eng.* **2016**, *30*, 359–374. [CrossRef]
8. Wang, Y.G. Prediction of height and period joint distributions for stochastic ocean waves. *China Ocean Eng.* **2017**, *31*, 291–298. [CrossRef]
9. Analysis and prediction of long term probability distributions of wave heights and periods. Technical Report of the National Maritime Institute London. Available online: https://trid.trb.org/view/423921 (accessed on 5 August 2022).
10. Yue, S. The Gumbel logistic model for representing a multivariate storm event. *Adv. Water Resour.* **2000**, *24*, 179–185. [CrossRef]
11. Antão, E.M.; Soares, C.G. Approximation of the joint probability density of wave steepness and height with a bivariate gamma distribution. *Ocean Eng.* **2016**, *126*, 402–410. [CrossRef]
12. Ochi, M.K. On long-term statistics for ocean and coastal waves. *Coast. Eng. Proc.* **1978**, *2*, 59–75. [CrossRef]
13. Kimura, A. Joint distribution of the wave heights and periods of random sea waves. *Coast. Eng. Jpn.* **1981**, *24*, 77–92. [CrossRef]
14. Huang, W.; Dong, S. Joint distribution of significant wave height and zero-up-crossing wave period using mixture copula method. *Ocean Eng.* **2021**, *219*, 108305. [CrossRef]
15. Myrhaug, D.; Kjeldsen, S.P. Parametric modelling of joint probability density distributions for steepness and asymmetry in deep water waves. *Appl. Ocean Res.* **1984**, *6*, 207–220. [CrossRef]
16. Haver, S. Wave climate off northern Norway. *Appl. Ocean Res.* **1985**, *7*, 85–92. [CrossRef]
17. Athanassoulis, G.A.; Skarsoulis, E.K.; Belibassakis, K.A. Bivariate distributions with given marginals with an application to wave climate description. *Appl. Ocean Res.* **1994**, *16*, 1–17. [CrossRef]
18. Lucas, C.; Soares, C.G. Bivariate distributions of significant wave height and mean wave period of combined sea states. *Ocean Eng.* **2015**, *106*, 341–353. [CrossRef]
19. Sklar, M. Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **1959**, *8*, 229–231.
20. Salvadori, G.; Durante, F.; Tomasicchio, G.R.; D'Alessandro, F. Practical guidelines for the multivariate assessment of the structural risk in coastal and off-shore engineering. *Coast. Eng.* **2015**, *95*, 77–83. [CrossRef]
21. Ward, P.J.; Couasnon, A.; Eilander, D.; Haigh, I.D.; Hendry, A.; Muis, S.; Veldkamp, T.I.E.; Winsemius, H.C.; Wahl, T. Dependence between high sea-level and high river discharge increases flood hazard in global deltas and estuaries. *Environ. Res. Lett.* **2018**, *13*, 84012. [CrossRef]
22. Wist, H.T.; Myrhaug, D.; Rue, H. Statistical properties of successive wave heights and successive wave periods. *Appl. Ocean Res.* **2004**, *26*, 114–136. [CrossRef]
23. Antão, E.M.; Soares, C.G. Approximation of bivariate probability density of individual wave steepness and height with copulas. *Coast. Eng.* **2014**, *89*, 45–52. [CrossRef]
24. Vanem, E. Joint statistical models for significant wave height and wave period in a changing climate. *Mar. Struct.* **2016**, *49*, 180–205. [CrossRef]
25. Jane, R.; Valle, L.D.; Simmonds, D.; Raby, A. A copula-based approach for the estimation of wave height records through spatial correlation. *Coast. Eng.* **2016**, *117*, 1–18. [CrossRef]
26. Dong, S.; Wang, N.; Lu, H.; Tang, L. Bivariate distributions of group height and length for ocean waves using copula methods. *Coast. Eng.* **2015**, *96*, 49–61. [CrossRef]
27. Iturrizaga, R.M.; Zavoni, E.H. Reliability analysis of mooring lines using copulas to model statistical dependence of environmental variables. *Appl. Ocean Res.* **2016**, *59*, 564–576. [CrossRef]

28. Kim, Y.T.; Park, J.H.; Choi, B.H.; Kim, D.H.; Kwon, H.H. A Bivariate Frequency Analysis of Extreme Wave Heights and Periods Using a Copula Function in South Korea. *J. Coast. Res.* **2018**, *85*, 566–570. [CrossRef]
29. Michele, C.D.; Salvadori, G.; Passoni, G.; Vezzoli, R. A multivariate model of sea storms using copulas. *Coast. Eng.* **2007**, *54*, 734–751. [CrossRef]
30. Corbella, S.; Stretch, D.D. Simulating a multivariate sea storm using Archimedean copulas. *Coast. Eng.* **2013**, *76*, 68–78. [CrossRef]
31. Zhang, Y.; Kim, C.W.; Beer, M.; Dai, H.; Soares, C.G. Modeling multivariate ocean data using asymmetric copulas. *Coast. Eng.* **2018**, *135*, 91–111. [CrossRef]
32. Mazas, F.; Hamm, L. An event-based approach for extreme joint probabilities of waves and sea levels. *Coast. Eng.* **2017**, *122*, 44–59. [CrossRef]
33. Li, J.; Pan, S.; Chen, Y.; Gan, M. The performance of the copulas in estimating the joint probability of extreme waves and surges along east coasts of the mainland China. *Ocean Eng.* **2021**, *237*, 109581. [CrossRef]
34. Kvingedal, B.; Bruserud, K.; Nygaard, E. Individual wave height and wave crest distributions based on field measurements from the northern North Sea. *Ocean Dyn.* **2018**, *68*, 1727–1738. [CrossRef]
35. Wu, Y.; Randell, D.; Christou, M.; Ewans, K.; Jonathan, P. On the distribution of wave height in shallow water. *Coast. Eng.* **2016**, *111*, 39–49. [CrossRef]
36. Forristall, G.Z. On the statistical distribution of wave heights in a storm. *J. Geophys. Res.* **1978**, *83*, 2353–2358. [CrossRef]