The performance of the copulas in estimating the joint probability of extreme waves and surges along east coasts of the mainland China

Jiangxia Li a,b, Shunqi Pan c, Yongping Chen d,e,* Min Gan e

a School of Hydraulic Engineering, Changsha University of Science and Technology, Changsha, 410114, China
b Key Laboratory of Water-Sediment Sciences and Water Disaster Prevention of Hunan Province, Changsha, 410114, China
c Hydro-environmental Research Centre, School of Engineering, Cardiff University, Cardiff, CF24 3AA, United Kingdom
d State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, 210098, China
e College of Harbour, Coastal and Offshore Engineering, Hohai University, Nanjing, 210098, China

Abstract

In designing coastal and nearshore structures, the joint probability of the wave heights and storm surges is essential in determining the possible highest total water level. The key elements to accurately estimate the joint probability are the appropriate sampling of the extreme values and selection of probability functions for the analysis. This study is to provide a full assessment of the performance of the different methods employed in the joint probability analysis. The bivariate extreme wave height and surge samples are analysed using 2 different probability distributions and the performance of 4 copulas, namely: Gumbel–Hougaard copula, Clayton copula, Frank copula and Galambos copula, is assessed. The possible highest total water levels for 100-year return period along the coastline of the mainland China are estimated by the joint probability method with the Gumbel–Hougaard copula. The results show that the wave heights and surges are highly correlated in the areas of dense typhoon paths. The distributions of the possible highest total water levels show a higher value in the southeast coast and lower value in the north. The results also indicate that at the locations where the sea states are energetic, the joint probability approach can improve the accuracy of design.

Key words: Coast of the mainland China; Joint probability; Copula; Extreme wave height; Extreme surge level
1. Introduction

Designing the coastal and offshore structures requires the consideration of a broad range of ocean factors due to complexity of the environment surrounding them. Amongst them, extreme waves and storm surges are two main factors. Under severe meteorological conditions, such as those during typhoons or cold storms, the extreme waves and storm surges can be closely correlated due to their driving forces. Joint probability analysis commonly becomes essential in estimating the extreme water levels to ensure the effective and sustainable designs of coastal engineering structures as demonstrated in the studies of Serafin and Ruggiero (2014) and Wahl et al (2015).

In joint probability analysis, a wide range of probability distributions of simultaneous environmental variables are obtained with the bivariate methods, as used by Ferreira and Guedes Soares (2002), Galiatsatou and Prinos (2007). Furthermore, Bruun and Tawn (1998) compared the properties of two extreme value methods: the univariate structure variable method and multivariate joint probability method, and found that the latter provided more useful and accurate design information when applied to several sites along Dutch coastlines. Based on a marginal distribution function fitted to the water level and wave height and their dependence, Hawkes et al. (2002) conducted a joint probability analysis, which was seen to perform better than the commonly used structure variable approach and joint exceedance approach.

However, during the last few decades, the copula theory, which has been initially used in finance, insurance and other economic sectors, has been widely adopted for joint probability analysis in the fields of hydrology (Mikosch, 2006) and coastal engineering (Salvadori et al., 2015). A copula function can connect different environmental variables without any hypothesis about their marginal distributions, and provides a powerful tool for the joint analysis of multivariate data. Recent examples of adopting the copula theory in hydrology fields include the study of extreme rainfalls (Salvadori and De Michele, 2004; Zhang and Singh, 2007), flood frequency for rivers (Chen et al., 2012; Sraj et al., 2015) and droughts (De Michele et al., 2013).

In coastal engineering applications, the copula theory has been found to be useful in providing
increased flexibility in modelling the joint probabilities of ocean hydrodynamic variables. As stated in Coles et al (1999), quantifying dependence plays an importance role in the joint probability analysis. In dealing with the dependence between two variables, Wist and Myrhaug (2004) modelled two successive wave heights exceeding a certain threshold by a Gaussian copula and compared the results with field measurements and laboratory data. De Waal and van Gelder (2005) analysed the joint probability of extreme wave height and wave period using the Burr–Paterno–Logistic copula. Similar studies were also conducted by Montes–Iturrizaga and Heredia–Zavoni (2015), as well as Vanem and Erik (2016). Wahl et al. (2010) carried out a study between two storm surge parameters using the Gumbel–Hougaard (GH) copula. Chini and Stansby (2010) used an integrated modelling system to investigate the joint probability of the extreme wave heights and water levels at Walcott, on the eastern coast of UK for determining the changes in the overtopping rates. Gruhn et al. (2012) used the Frank copula to estimate the joint probability of the water level residuals and significant wave heights along the coast of the Baltic Sea. Wahl et al. (2012) applied Archimedean copula functions in the German Bight to determine the exceedance probabilities of storm surges and wind waves. Masina et al. (2015) used a copula-based approach to estimate the joint probability of the water levels and waves at the Ravenna coast in Italy. The probability of failure/inundation was estimated by the direct integration method, and the coastal flooding risks were calculated. Galiatsatou and Prinos (2016) applied the copula method to investigate the changes in the joint probabilities of extreme wave heights and corresponding storm surges with time in the Aegean Sea. Ward et al. (2018) used the copula models to analyse the dependence between sea level and river discharge as well as the probability of flooding events in global deltas and estuaries. Bevacqua et al. (2019) discovered a higher probability of compound flooding from precipitation and storm surge in Europe under climate change using a copula-based multivariate probability model.

For extreme events, Gudendorf and Segers (2010) proposed the extreme value copulas for extreme multivariate analysis due to their capability of describing the upper tail dependence well. Mazas and Hamm (2017) used an event-based approach for determining extreme joint probabilities of waves and sea levels by focusing on the sampling of extreme events. In their study, three extreme value copulas (GH copula, Galambos copula, Husler–Reiss copula) were compared, and their results showed that different extreme value copulas would yield similar results, but the sampling methods
could cause a large difference in the joint probability results. The samples could be selected by
different ways. For example, in the sampling of extreme wave heights and surges, some researchers
sample the extreme wave heights (or surges) and the simultaneous surges (or wave heights) by the
block maxima method (Li and Song, 2006). Others consider the “impact” of the events and select
the samples according to a defined response function, i.e. total water levels, overtopping and run-up
(Gouldby et al., 2014; Serafin et al., 2014; Rueda et al., 2016). Also, the extreme pairs of samples
by defining the storm events using certain thresholds of variables are used (Li et al., 2014; Wahl et
al., 2016).

For multivariate cases, the dependence among a large range of extreme ocean elements like wave
height, water level, wave period, storm duration, etc. was assessed. Corbella and Stretch (2012,
2013) investigated the dependence between storm parameters: significant wave height, peak wave
period, duration, inter-arrival time, and water level, by applying a copula-based statistical model
under varying climatic conditions. Li et al. (2014) analysed the variates of extreme storm events
(wave height, wave period, sea level, wave direction, and storm duration) under deep-water wave
conditions, where the Monte Carlo method and four other methods to construct the dependency
structures based on the copula functions, physical relationship, and extreme value theory were
adopted. It was found that the Gaussian copula model was the most suitable wave climate
simulation method for the Dutch coast. Rueda et al. (2016) used the generalized extreme value
(GEV) distributions and Gaussian copula to model the dependence between multivariate extremes
related to coastal floods for different weather patterns. Lin-Ye et al. (2016) applied a hierarchical
Archimedean copula to characterize storm intensity based on the storm energy, unitary energy, peak
wave period, and duration on the Catalan coast. Montes–Iturrizaga and Heredia–Zavoni (2016)
developed a multivariate model for the joint distributions of environmental variables using vine
copulas, which was applied to build trivariate environmental contours of the wave height, period,
and wind velocity at the Gulf of Mexico. Zhang et al. (2018) modelled multivariate ocean data
using asymmetric copulas and compared the results with those obtained by traditional copulas.

The applications to the coastal waters of China are also seen rapidly emerging in recent years. Tao
et al. (2013) developed a criterion to classify the intensity grade of a storm surge by the joint return
Yang and Zhang (2013) applied the GH copula to analyse the joint probability of extreme winds and wave heights at the Bohai Bay. Dong et al. (2015) used the Clayton copula to clarify the relations between the group height and length of ocean waves based on laboratory data and field wave data near the coast of Zhejiang province. Dong et al. (2017) studied the joint return probability of the wind speed and rainfall intensity in a typhoon-affected sea area close to Shanghai using the Weibull distribution and GH copula. More recently, Yin et al. (2018) estimated the extreme sea levels in the Yangtze estuary using the quadrature joint probability optimal sampling method (JPM-OS) with consideration of the typhoon field, wave height, and sea level in the studied region. Yang and Qian (2019) analysed the joint probability of typhoon-induced surges and rainstorms at Shenzhen and derived trivariate joint distributions and conditional distributions of these variables based on the copula method.

To estimate the desired design combination of wave height and surge accurately under extreme conditions can be rather challenging. Many studies have outlined that a univariate frequency analysis may not be capable of assessing the occurrence probability of extremes if the events are characterized by interrelated random variables (Chebana and Ouarda, 2011; Masina et al., 2015). According to Marcos et al. (2019), the return periods of extreme sea levels are underestimated in 30% of the coasts around the world if dependence is neglected. In particular, along the coasts of China, Li and Song (2006) analysed the correlations between the extreme wave heights and extreme water levels in the coastal waters of Hong Kong using the Gumbel–logistic model. The result proved that applying the commonly used empirical method to estimate the total water level (by directly adding the univariate extreme values) may not be sufficiently accurate to derive the coastal design criteria.

On the other hand, because of the lack of long-term matched oceanic data, most of the previous studies only focused on a limited area or specific observation station. Therefore, it is necessary to carry out further research to clarify the relationships between extreme wave heights and storm surges and devise a realistic and safe design in coastal and offshore engineering.

Built on the model results from the previous work of Chen et al. (2019), which used the GH copula in analysing the joint probability of the wave height and surge along the coast of the mainland China, this study is to fully examine the performance of four different types of copulas using the
existing model results from Chen et al. (2019) in estimating the joint probability. This study uses the annual N-largest sampling method with a detailed analysis of the predominance of joint extreme samples, in an attempt to effectively increase the available sample size compared to previous studies. Then a comprehensive analysis of the dependence between wave height and surge is conducted on the extreme samples obtained. The established joint probability model is subsequently applied to 87 selected locations representing the entire mainland China coast, to estimate the extreme combined water levels (CWLs) for flood risk assessment.

2. Study area and data

In this study, the model results of the significant wave heights (Hs) and surge levels (S) over a 35-year (1979–2013) period as detailed in Li et al. (2018) are used. For the sake of completeness, the model setup and applications are briefly presented here. The computational domain covered an area from 105 °E to 140 °E and from 15° N to 41 °N, as shown in Fig. 1. A coupled wave (FVCOM-SWAVE) and hydrodynamic (FVCOM) model (Qi et al., 2009), which was well calibrated and validated in the nearshore and offshore area by Li et al. (2018), was used. The model used an unstructured mesh with a spatial resolution of 1 degree at the open boundaries and finer than 0.1 degree in the coastal areas. Along the open boundaries, the model was driven by the tide conditions obtained from TPXO database. The modified ECMWF re-analysis wind data with a parametric typhoon model in order to account for the effects of 862 typhoons during the simulation period was used as the sea surface forcing. Hourly wave height and surge data from the model at nine nearshore locations which are identical to those in Chen et al. (2019), as shown in Fig. 1, are extracted from the model results and used for the joint probability analysis in this study. The selection of those nearshore locations is mainly due to the availability of field measurements for validating the hydrodynamic model.
3. Methodology

3.1 Sampling method

For extreme analysis, sampling of the extreme values from the time series is a key step. When the data length is sufficiently long, the annual maximum (AM) method is commonly used to select the joint extreme samples to ensure the independence of extreme samples (Sraj et al., 2015; Yang and Zhang, 2013). However, according to the studies of Bernardara et al. (2014) and Mazas and Hamm (2017), for effective bivariate analysis, the sample size should be normally more than 300. Therefore, in most cases, the AM method may only generate a small sample size of extreme events, insufficient to effectively capture the information of the dependence between the variables. To overcome this, the peak over threshold (POT) method can be the effective one for selecting multivariate samples (Li et al., 2014; Mazas et al, 2014). Compared to the block maxima approach, POT method is advantageous when selected peaks result from different storm events. However, the POT-based joint sampling methods can present with the major difficulty in determining the values of the thresholds, particularly in the cases of highly variable hydrodynamic conditions temporally and spatially over a large study area such as this study.
Based on the block maximum sampling method for univariate analysis (Galiatsatou, 2011), in this study, an annual N-largest (ANL) joint extreme sampling method is proposed. This method selects the top N samples in each year such that it can capture more information than the AM method. Unlike the POT methods, the number of samples selected per year can be pre-determined in the ANL method, so that the extreme conditions can be fairly represented over the study area. In addition, to ensure the independence of the extreme events selected, a standard storm length covering both sides of each peak is considered. The standard storm length generally ranges from 24 to 72 hours in coastal storm analysis, following several previous studies (Basco and Walker, 2010; Martzikos at al., 2021; Marclos at al., 2019). It is set to be 48 hours in this study after conducting a sensitivity test suggested by Tawn (1988): provided the storm length is approximately correct the estimates of quantiles should not change too much by making small changes to this length. The simultaneous S is selected within the standard storm length along with the N-largest Hs to account for the possible time lag between extreme Hs and S. The number of samples per year (N) can be set accordingly to meet the required sampling size. Thus, in this study, by considering data length available over the 35-year period and the required sample size for joint probability analysis suggested by Mazas et al. (2014), N = 10 is used.

3.2 Univariate probabilistic distributions

Before establishing the dependence between wave height (Hs) and surge level (S), a frequency analysis would be required for each variable to define its marginal distribution. The two probabilistic distributions as shown in Table 1 are tested in this study for searching the best fit of the samples:

<table>
<thead>
<tr>
<th>Distribution</th>
<th>CDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson-III (P3)</td>
<td>$F_p(x) = \frac{\frac{2}{\bar{x} C_s} + \frac{4}{C_s^2}}{\Gamma(4)} \int_{-\infty}^{x} \frac{\exp\left(-\frac{2}{\bar{x} C_s} (x - \bar{x} + \frac{2C_s}{\bar{x}})\right)}{C_s^{4-r}} dx$</td>
</tr>
</tbody>
</table>

where, $\Gamma$ is the gamma function; $\bar{x}$ is the mean value of the samples;
$C_v$ and $C_s$ are the coefficients of variation and skewness.

**Generalized Extreme Value (GEV)**

$$F_{\text{gev}}(x) = \exp\left(-\left(1 + k \frac{x - \mu}{\sigma}\right)^{-\frac{1}{k}}\right)$$

where, $\mu$, $\sigma$ and $k$ are the location, scale and shape parameters respectively.

3.3 Copulas

According to the theory of Sklar (1959), there exists a copula, $C$, that can connect the marginal distributions, $u_1 = F_X(x)$ and $u_2 = F_Y(y)$, to form the CDF (Genest and Favre, 2003) expressed as:

$$F(x, y) = C(F_X(x), F_Y(y)) \quad (1)$$

The commonly used copula families include Gaussian copula, $t$-copula, extreme value copula (EV-copula) and Archimedean copula. Among them, the Archimedean copula family has been frequently applied to the hydrologic fields. Meanwhile, Gudendorf and Segers (2010) suggested that EV-copula could also well describe the upper tail dependence for an extreme multivariate analysis.

Thus, in this study, three commonly used copulas under the Archimedean family: Gumbel–Hougaard (GH) copula, Frank copula, and Clayton copula, together with an EV-copula, Galambos copula, are examined. The EV-copula is a type of copula which not only satisfies all the definitions and properties of copulas, but also meets the max-stable property for fixed integer $n$, i.e.

$$\lim_{n \to \infty} C_{\text{ev}}(u_1^{1/n}, \ldots, u_d^{1/n})^n = C(u_1, \ldots, u_d), \quad (u_1, \ldots, u_d) \in [0,1]^d \quad (2)$$

In fact, GH copula fits the properties of both Archimedean copula and EV-copula groups.

The generator function, CDF and probability density function (PDF) of these copulas are listed in Table 2, where $u_1$ and $u_2$ are the marginal distributions and $\theta$ is the parameter of copula which describes the dependencies. The Galambos copula which belongs to EV-copulas does not have a generator function.

**Table 2 The generator function, CDF and PDF of four copulas**

<table>
<thead>
<tr>
<th>Copula</th>
<th>Function</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Dependence

Several methods are available to determine the dependence structure between two random variables X and Y. They are commonly used to calculate the correlation coefficients, for example, Pearson’s r correlation coefficient, Spearman’s $\rho$ coefficient, or Kendall’s $\tau$ coefficient. In this study, Kendall’s $\tau$ coefficient is chosen to quantify the dependence between the Hs and S samples. It describes the dependence between the samples by ranking the variables with the following expression:
\[ \tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2} \] (3)

where \( n \) is the total number of pairs. Any pair of observations, \((x_i, y_i)\) and \((x_j, y_j)\), where \( i \neq j \), is reckoned to be concordant if the ranks for both the elements agree, i.e., both \( x_i < x_j \) and \( y_i < y_j \) holds or both \( x_i > x_j \) and \( y_i > y_j \) holds, and otherwise is regarded as the discordant pair.

Therefore, \( \tau = 0 \) indicates the perfectly independent cases and \( \tau = 1 \) indicates perfectly dependent cases.

Generally, in the extreme analysis, the dependency is determined for the extreme values. However, the correlation coefficients for the extreme values can be less capable of fully capturing the asymptotic dependency (Mazas et al., 2014). Thus, in this study, the chi-plots are used as graphical tools to assess the dependence between the extreme Hs and S. It supplements an ordinary scatterplot of the data by providing a graph that has characteristic patterns depending on whether the variates are independent, with some degree of monotone relationship or more complex dependence structure.

Two variables \((\Lambda_i, X_i)\) as suggested by Fisher and Switzer (1985, 2001) are used in the scatterplots as:

\[ \Lambda_i = 4S_i \max \left\{ \left( F_i - \frac{1}{2} \right)^2, \left( G_i - \frac{1}{2} \right)^2 \right\} \] (4)

\[ X_i = \frac{(H_i - F_i)G_i}{\left\{ F_i(1-F_i)G_i(1-G_i) \right\}^{1/2}} \] (5)

where,

\[ S_i = \text{sign}\left\{ \left( F_i - \frac{1}{2} \right) \left( G_i - \frac{1}{2} \right) \right\} \] (6)

\[ F_i = \sum_{j \neq i} I(x_j \leq x_i) / (n-1) \] (7)

\[ G_i = \sum_{j \neq i} I(y_j \leq y_i) / (n-1) \] (8)

\[ H_i = \sum_{j \neq i} I(x_j \leq x_i, y_j \leq y_i) / (n-1) \] (9)

and \( I \) is the indicator function.
The relationships between Kendall’s coefficient $\tau$ and the correlation index, $\theta$, for copulas introduced in Section 3.3 are listed in Table 3.

### Table 3 Relationships between Kendall's coefficient $\tau$ and parameter $\theta$ for different copulas

<table>
<thead>
<tr>
<th>Copula</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gumbel–Hougaard copula</td>
<td>$\tau = 1-1/\theta$</td>
</tr>
<tr>
<td>Clayton copula</td>
<td>$\tau = \theta / (\theta + 2)$</td>
</tr>
<tr>
<td>Frank copula</td>
<td>$\tau = 1 + \frac{4}{\theta} \left[ \int_0^\theta \frac{1}{t} \left( e^{-t} - 1 \right) dt \right]$</td>
</tr>
<tr>
<td>Galambos copula</td>
<td>$\tau = \frac{\theta + 1}{\theta} \int_0^1 \left( \frac{1}{s^{\theta}} + \frac{1}{(1-s)^{\theta}} \right)^{-1} ds$</td>
</tr>
</tbody>
</table>

3.5 Return period

In joint probability analysis, the bivariate return period can be defined. The OR return period ($T_o$) indicates that at least one of the variable exceeds a certain value, and the AND return period ($T_a$) indicates that both the variables exceed a certain value. They can be calculated using the following expressions:

\[
T_o(x, y) = \frac{1}{1 - F(x, y)} \quad (10)
\]

and

\[
T_a(x, y) = \frac{1}{1 + F(x, y) - F_x(x) - F_y(y)} \quad (11)
\]

where $F_x(x)$ and $F_y(y)$ are the marginal distributions and $F(x, y)$ is calculated by Eq. (1) by combining the CDF of the copula and corresponding marginal distributions.

### 4. Results

4.1 Dependence of extreme samples

For the joint probability analysis, it is necessary first to examine the dependency between the extreme $H_s$ and $S$. As an example, the extreme wave height ($H_s$) and surge ($S$) sampled at nine
nearshore stations are shown in Fig. 2. It indicates that those two variables are partly related as the
data points present a clear linear relation at all stations, but with a high degree of scattering.
Relatively stronger dependencies between the extreme Hs and S are found at Haikou, Zhapo, Hong
Kong, Xiamen and Kanmen stations because the scatters show a more obvious linear trend, but at
other stations, such dependency appears relatively weaker. It is also noticed that the stations with
stronger dependencies are located in the coastal areas facing the open sea and are easily affected by
typhoon events. Stations Beihai and Dongfang are to some extent sheltered by the land. Stations
Lvsi and Shijiusuo are located in the mid-north coast where fewer typhoon events occur. This result
indicates that the dependencies between the extreme Hs and S at certain locations can be influenced
by typhoon events.

![Fig. 2 Scatterplot of the N-Largest joint samples](image)

The chi-plots for all nine stations are shown in Fig. 3. In the chi-plot, \( \lambda \) measures the distance of a
pair of variables from their medians: a positive (negative) value implies that both variables are on
the same (opposite) side of their respective medians and a value close to 1 (0) implies they are
larger or smaller relative to (close to) their respective medians, and $X_i$ measures the dependence: a positive (negative) value describes a positive (negative) dependence, while a value close to zero suggests independence (Mazas et al., 2014). From Fig. 3, it can be seen that there is a clear positive dependence between the extreme Hs and S at all the stations. However, for the events where $\Lambda_i$ is negative, there is only one population at all stations, whilst for positive $\Lambda_i$, two different populations are observed, namely, the upper and lower “lobes” as suggested by Fisher and Switzer (2001). The upper “lobe” corresponds to pairs where both the Hs and S are larger than their median, exhibiting a relatively strong dependence. This is because higher Hs and S are generally caused by the same extreme atmospheric event. In contrast, the lower “lobe” corresponds to a pair where both the Hs and S are smaller than their medians, exhibiting weak dependence. At most stations, there are two distinct upper and lower lobes, which indicates the bimodal dependence of wave height and surge due to relatively large events (such as typhoons) or weaker events. In other words, this bimodal dependence could be caused by two extreme situations: typhoon related extremes and non-typhoon related extremes. At the Shijiusuo station, however, the boundaries of the two “lobes” are obscure, which may be attributed to the low frequency of typhoon events at this location.
Furthermore, the distribution of Kendall’s coefficient $\tau$ over the computational domain is also calculated, as shown in Fig. 4a. The results clearly show that the coefficients in the southeast area of the computational domain are remarkably larger than those at other locations, which coincides well with the areas along the paths of frequent typhoons during the 35-year (1979–2013) period (Fig. 4b). Specifically, the dependence between the extreme $H_s$ and $S$ increases in the areas where the sea states are more energetic, which was also reported in Hawkes et al. (2002). It is found that this character could not be fully revealed with the AM sampling method as used in Chen et al. (2019), which also serves as an indication of the improvement when the ANL sampling method is used.
More specifically, the Kendall’s coefficient $\tau$ at the nine nearshore locations is shown in Fig. 5. The results indicate that at Beihai, Dongfang, Lvsì, and Shijiusuo stations, values of $\tau$ are generally smaller compared to other stations, just below 0.35, suggesting relatively weak dependence between the extreme $H_s$ and $S$ at these stations. At other stations, particularly Zhapo and Hong Kong, the dependence between the extreme $H_s$ and $S$ is strong. According to Fig. 4 and Fig. 5, it seems that the stations in the areas frequently affected by typhoons tend to have large $\tau$ coefficients, as expected since typhoon events can be a significant cause for the extreme $H_s$ and $S$.

An advantage of applying the copula theory to bivariate or multivariate probability analysis is that copulas allow different types of marginal distributions to be used for different variables. To examine
the performance and fitness of copulas, in this study both Hs and S in the joint extreme samples at
the nine nearshore locations are fitted with two univariate distributions as introduced previously.
Since the study area is observed to be wave-predominated (discussed further in Section 6), extreme
Hs data is sampled by the ANL method, and extreme S data is sampled based on the sampled Hs.
By subsampling the N-largest data to annual maxima, two probability distributions introduced in
Section 3.2 are used to fit the samples. The parameters in GEV are estimated by the maximum
likelihood estimation method. In Fig. 6, the fittings of the Hs and S in the joint extreme samples
with different distributions at the Kanmen station are plotted. Fig. 6 (a) shows that the P3
distribution fits better the extreme Hs samples at Kanmen than GEV distribution, whereas in Fig. 6
(b) the GEV distribution can comparatively better fit the extreme S samples.

To quantify the fitting results, Pearson’s coefficient \( r \) (Pearson, 1895) between the samples (dot in
Fig. 6) and theoretical values (line in Fig. 6) are calculated at Kanmen station and listed in Table 4.
The Pearson’s coefficient \( r \) could be calculated by,

\[
r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}.
\]

where, \( X_i \) and \( Y_i \) are the sample values and the theoretical values; \( \bar{X} \) and \( \bar{Y} \) are the averaged
values of \( X_i \) and \( Y_i \). The largest correlation coefficients coincide with the best fit distribution
chosen by Fig. 6.
Fig. 6 Fitting of the samples with different distributions at Kanmen station: (a) wave height; (b) surge level

Table 4 Correlation coefficients between the samples and different distributions at Kanmen station (the best fit distributions are indicated in bold)

<table>
<thead>
<tr>
<th></th>
<th>GEV</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wave height</td>
<td>0.9876</td>
<td><strong>0.9877</strong></td>
</tr>
<tr>
<td>Surge level</td>
<td><strong>0.9921</strong></td>
<td>0.9833</td>
</tr>
</tbody>
</table>

By combining the results in Fig. 6 and Table 4, the best fit distributions for the nine nearshore stations are summarized in Table 5. It can be seen that GEV distributions fit the extreme Hs samples better than P3 at 6 out of 9 stations, and the GEV distribution fits the extreme S samples better at all stations in the study area. Although not shown here, the 95% confidence intervals of the selected marginal are also examined to ensure a proper fit. It is reasonable to see that the confidence intervals increase from the lower tail to the upper tail. Therefore, the distributions of the Hs and S are determined by the selected probability distributions in this study.

Table 5 Chosen distributions for the Hs and S in the joint samples at the nine nearshore stations

<table>
<thead>
<tr>
<th>Station</th>
<th>Beihai</th>
<th>Dongfang</th>
<th>Haikou</th>
<th>Zhapo</th>
<th>Hong Kong</th>
<th>Xiamen</th>
<th>Kanmen</th>
<th>Lvsi</th>
<th>Shijiusuo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs</td>
<td>GEV</td>
<td>GEV</td>
<td>GEV</td>
<td>P3</td>
<td>P3</td>
<td>GEV</td>
<td>P3</td>
<td>GEV</td>
<td>GEV</td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.3 Selection of copulas

To determine the best fit copulas for the data sets in this study with the chosen marginal distributions of the extreme Hs and S as described previously, it is essential to examine the characteristics of each copula. Fig. 7 shows the probability density distributions of the GH copula, Clayton copula, Frank copula, and Galambos copula. It is clear that both GH and Galambos copulas have a pronounced upper tail density, suggesting that they are capable of describing the dependence in the upper tail of the distribution, i.e. upper tail dependence. However, the density distribution of the Clayton copula has a thick lower tail density, suggesting that it can better describe the dependence in the lower tail of the distribution, i.e. lower tail dependence. The Frank copula has a symmetric tail, i.e. no tail dependence, which can only be suitable for the symmetrical distributed samples.

![Fig. 7 Probability density distributions of (a) GH copula, (b) Clayton copula, (c) Frank copula, and (d) Galambos copula](image)

To achieve the best match of the characteristics of the copulas shown in Fig. 7 with the samples in
this study, the extreme samples at all nine stations are examined with the binary frequency
histograms of the Hs and S. As shown in Fig. 8, at all stations, a thick upper tail density can be
observed, although the frequency distributions are slightly different at different stations. In general,
there is a clear suggestion that the GH copula and Galambos copula can be chosen in the probability
analysis as they match well with all density distributions at those stations.

However, for the completeness of analysis, all four copulas are also used to fit the joint extreme
samples using Kendall’s coefficient as introduced in Section 3.5. Fig. 9 shows their joint cumulative
probabilities in comparison with those of the empirical copula at all nine stations. As the probability
of the empirical copula is directly calculated based on the samples, any copula in the test that has
the best fit with the empirical copula will be regarded as the optimal copula for the samples. It can
be seen from the comparisons that the contours of four copulas provide a very similar fit in the
mid-range of probabilities. However, Clayton and Frank copula perform poorly with tendency of
underestimating the probability in the upper tail region while overestimate the probability in the
lower tail region. This is related to the density distribution of those tested copulas. The results clearly show a general trend of good match of the GH and Galambos copulas with the empirical copula, better than the other two copulas, while Frank copula has the worst fit.

Fig. 9 Comparison of joint probability of four copulas with that of empirical copula

In addition, the Cramér-von Mises (CVM) test is carried out to compare the performance of the four copulas with that of the empirical copula quantitatively, using the following equation (Mazas and Hamm, 2017; Genest and Rivest, 1993):

\[ S_n = \sum_{i=1}^{N} \left( C_n(U_i, V_i) - C_\theta(U_i, V_i) \right)^2, \]  

(13)

where, N is the sample size, \((U_i, V_i)\) is the sample of the normalized ranks, \(C_n\) is the copula in test, and \(C_\theta\) is the empirical copula. The CVM statistics at all stations are shown in Fig. 10. It is clear that CVM values for the Galambos and GH copulas are the lowest amongst all 4 copulas, while GH copula preforms slightly better than Galambos copula. The results again confirm the
outcomes of the probability density analysis of these copulas as shown in Fig. 7 and Fig. 8.

From the results presented in Fig. 10, it can be concluded that both GH and Galambos copulas, which have the lowest CVM values amongst all, are deemed to be the optimal ones for studying the joint probability of the extreme Hs and S along the east coast of the mainland China. It also highlights the necessity of using an EV-copula to conduct the joint probability analysis of extreme values. Considering that the GH copula has a simpler function than the Galambos copula, therefore it is decided that the GH copula is adopted in this study.

4.4 Joint probability

For the joint probability, both AND and OR return periods are assessed at all station. As an example, the isolines of the joint events with both return periods at the Kanmen station are shown in Fig. 11.

In general, for the same joint event, the AND return period is found to be larger than the OR return period. Specifically, when calculating the joint probability of the variables, the selection of the different types of return period should be according to the aim of the study. In the following analysis in this study, the AND return period is applied. Concurrently, according to a previous study (Chen et al., 2019), the shapes of the isolines are diverse at different locations because the joint probability is location-specific, particularly in the nearshore areas. Because the distributions of the joint events at different locations are discussed in detail in a previous study (Chen et al., 2019), the isolines of the joint events at other stations are not provided here.
From Fig. 11, it can be seen that different combinations of the Hs and S can have the same return period along an isoline when calculating the joint events by using a cumulative probability. Thus, to search for the most probable joint event for a certain return period, the joint probability density is the best function to be used. The combined water level (CWL) which is the sum of the Hs and S is analysed in this study for engineering application. To determine the most probable CWL, the joint probability density is calculated to obtain the failure probability by integration over the failure region (Masina et al., 2015; Chen et al., 2019). Along the isoline of the failure probability, the point corresponding to the highest probability density is the most probable extreme event, which is the tangential point between the isoline of the failure probability (indicated by straight lines in Fig. 12) and a particular isoline of the probability density (indicated by curves in Fig. 12). Then the extreme CWL is calculated by adding the Hs and S of the most probable extreme event. Fig. 12 shows the isolines of the joint probability density and failure probability at the nine representative nearshore stations. The most probable joint events with a 50-year and 100-year return period at the nine nearshore stations are then determined according to Fig. 12, as shown in Table 6.
Table 6 Most probable 50-year and 100-year return level joint events

<table>
<thead>
<tr>
<th>Station</th>
<th>50-year</th>
<th>100-year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hs(m)</td>
<td>S(m)</td>
</tr>
<tr>
<td>Beihai</td>
<td>1.55</td>
<td>0.95</td>
</tr>
<tr>
<td>Dongfang</td>
<td>3.30</td>
<td>0.60</td>
</tr>
<tr>
<td>Haikou</td>
<td>4.25</td>
<td>0.95</td>
</tr>
<tr>
<td>Zhapo</td>
<td>5.25</td>
<td>1.65</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>3.00</td>
<td>1.20</td>
</tr>
<tr>
<td>Xiamen</td>
<td>1.95</td>
<td>1.95</td>
</tr>
<tr>
<td>Kanmen</td>
<td>2.90</td>
<td>1.90</td>
</tr>
<tr>
<td>Lvsi</td>
<td>1.35</td>
<td>1.45</td>
</tr>
<tr>
<td>Shijiusuo</td>
<td>3.55</td>
<td>0.95</td>
</tr>
</tbody>
</table>

In engineering practice, when lacking the analysis of joint probability, the joint event for a certain period is typically estimated by an addition of the single event with specified return period. For example, a 100-year return level joint event is sometimes approximated by the sum of a 100-year
Hs (100Hs) and 10-year S (10S), the sum of a 10-year Hs (10Hs) and 100-year S (100S), or the sum of a 50-year Hs (50Hs) and 50-year S (50S) (Li and Song, 2006), or an addition of 100-year Hs and S (Code of Hydrology for Harbour and Waterway, JTS 145-2015, China). To compare the outcome of these combinations and the joint probability results, the Hs and S sampled by the univariate method without considering their dependence are used to calculate the Hs and S with 100-year, 50-year and 10-year return periods at all nine locations. The CWLs calculated by four empirical combinations described above are compared with those calculated by the joint probability method with the 50-year and 100-year return periods at the nine nearshore stations, as shown in Fig. 13, where the ranges of the CWLs from the 50-year to 100-year return levels calculated by joint probability method are presented for the sake of clarity. It can be seen from the figure that the 100-year CWLs calculated by joint probability method are larger than the “100Hs+10S,” “50Hs+50S,” and “10Hs+100S” combinations but are smaller than the “100Hs+100S” combination. In general, the “100Hs+10S” and “50Hs+50S” combinations are close to the 50-year return level CWLs calculated by joint probability which could be recommended to estimate the 50-year return level situation when the joint probability data is unavailable. Meanwhile, “50Hs+50S” combination is always within the range of the 50-year and the 100-year return levels calculated by joint probability method which could be a meaningful indicator for joint events between 50-year and 100-year return levels. The “10Hs+100S” combination is relatively smaller than other combinations, especially at Beihai, Dongfang, Haikou, Zhapo, and Shijiusuo stations, which indicates a strong wave predominant property.

The result suggests that three of the empirical combinations may lead to an unsafe design, with only the “100Hs+100S” combination being safe for an engineering design at all nine locations. However, the method of using the “100Hs+100S” combination to estimate a 100-year joint event is proposed based on the assumption that the Hs and S are independent random variables. This assumption may be unrealistic because it has been proved that the Hs and S are partly dependent in this study, as described in Section 4. Thus, it is necessary to conduct a joint probability analysis when designing engineering structures.
Fig. 13 The ranges of the CWL calculated by the joint probability method with GH copula for 50-year and 100-year return periods (shown as a box) in comparison with the CWLs calculated by the empirical combinations.

5. Discussion

From the detailed comparison of the CWLs calculated by the empirical combination and joint probability method along the east coast of the mainland China, it is clear that the “100Hs+100S” combination is the only method which can lead to a safe design among the four empirical combinations. If this approach is adopted for the entire coastline, the extreme CWLs can be estimated for wide engineering applications. For the purposes of inter-comparison, Fig. 14 shows the distributions of the 100-year return level Hs and S at the studied coastline with 87 uniformly distributed locations. It can be seen from Fig. 14 (a) that the 100-year Hs along the southeast coast of the mainland China are remarkably larger than those at the other sites. However, in Fig. 14 (b), the distributions of the extreme S are rather uniform along the entire coast, with S being generally larger than 2 m from the mid-east to the south coast.

Fig. 14 Distributions of the 100-year return level: (a) wave height and (b) surge level along the coasts of the mainland China without considering their dependence.
To investigate the hydrodynamic conditions for different areas in detail, a response coefficient (D) is defined as:

\[ D = \frac{(H_s + S)}{H_s} \]  

where \(H_s\) is the 100-year return level wave height and \(S\) is the 100-year return level surge.

Although partly related, wave and surge are characterized by different dynamic, and have different magnitudes and spatial scales. As a coastal environment is usually defined as wave-predominated or surge-predominated based on the relative contributions of the wave and surge on coastal processes studied, as well as on coastal morphodynamics, the coefficient \(D\) could give a first impression on the relative significance of these two variables.

With the response coefficient (D) presenting the relative contributions of the \(H_s\) and \(S\) for the same return period at different locations, the hydrodynamic conditions there can then be described as wave-predominated or surge-predominated. \(D\) is generally larger than 1. If \(D\) is between 1 and 2, the location could be described as wave-predominated since the \(H_s\) has larger impact than \(S\); otherwise, if over 2, it is surge-predominated. The value of \(D\) could reflect the relative value of \(H_s\) and \(S\). Higher values indicate a larger impact of \(S\). Fig. 15 shows the distribution of coefficient \(D\) along the mainland China coast. It can be seen that coefficient \(D\) at most of the sites along the mainland China coast is between 1 and 2, which suggests that most of the areas along the mainland China coast are wave-predominated. The extreme wave height is obviously larger than surge level, which indicates a larger wave impact at these locations. This justifies the way that the joint extreme samples were selected in a wave-predominated manner in Section 3.1. For the southeast coast, the coefficient \(D\) is a little bit larger than 1, as these areas are facing open seas and are found in deep waters, which enhance the wave energy and mitigate the surge. However, in a few sites, \(D\) coefficients are far larger than 2, for example, the points in the Yangtze River estuary and Hangzhou bay. The water depths are small at these locations, and the shape of the estuary coastline may have caused surge to concentrate, resulting in those sites becoming surge-predominated.
Furthermore, Fig. 16 shows the distributions of the 100-year return level CWLs, calculated by the empirical method (100Hs+100S) and joint probability method. The distributions of the extreme CWLs calculated by the joint probability method show a relatively higher value in the southeast coast and lower value in the north. Although with the similar distribution pattern, it is clear that using the empirical method by assuming Hs and S being independent random variables can yield a higher water level, but using the joint probability method can yield relatively more economical design conditions.

As an indication of the improvement made between the 100-year CWLs calculated by the empirical and joint probability methods, a parameter, Q, is introduced and defined as:

\[ Q = \frac{CWL_\text{e} - CWL_i}{CWL_\text{e}} \times 100\% , \]  

(15)
where, $CWL_e$ is the water level calculated by the empirical method, and $CWL_j$ is the water level calculated by the joint probability method.

The distribution of the Q parameter shown in Fig. 17 indicates that the improvements in the joint probability method compared with the empirical method are not notable in the north coast (Bohai Sea coast) and south coast with a Q of under 6%, which means that the use of empirical combinations at these locations is relatively reasonable. However, for the mid-east mainland coast and southeast mainland coast, Q is relative larger, over 25% at its maximum. Thus, using the design water level calculated by the empirical method at these locations may be inaccurate. In other words, the joint probability method can yield better results at the sites where the hydrodynamic conditions are generally complex or energetic. For example, two major estuaries (Yangtze and Hangzhou Bay) are located the mid-east coast. The southeast coast is frequently affected by typhoon events, particularly near the Taiwan Strait, which can incur stronger hydrodynamic processes and cause larger diversity between water levels calculated by empirical and joint probability methods.

6. Conclusions

This study uses long-term (35 years) model results to examine the suitability and performance of 4 copulas in the joint probability analysis of the extreme wave height ($H_s$) and surge ($S$) along the coasts of the mainland China. The extreme data is extracted with the annual N-largest sampling
method and the dependencies between the Hs and S in the joint extreme samples at the nine selected nearshore stations are fully analysed. The performance of the four commonly used copulas, i.e. Gumbel-Hougaard, Clayton, Frank and Galambos copulas, in estimating the joint probability of extreme samples are assessed. The optimal copula identified is used for predicting combined water levels (CWLs, sum of Hs and S) in the study area with 50- and 100-year return periods and the accuracy is quantified.

Two theoretical univariate probabilistic distributions, i.e. GEV and P3, are used to fit the marginal of Hs and S samples. The results show that either GEV or P3 distributions could appropriately fit the extreme wave samples which depends on their location, while the GEV distribution provides the best fit to the extreme surge samples for all the selected locations along the mainland China coast. After assessing the performance of the copulas, the extreme value copula group is found to be the optimal copula group to describe the joint probability of extreme Hs and S. The Gumbel–Hougaard copula that belongs to the extreme value copula group is finally chosen to conduct the joint probability analysis of the Hs and S along the mainland China coast owing to its precision and conciseness.

By adopting the GEV/P3 distribution and applying the copula theory, the joint exceedance probabilities and joint probability densities at the nine representative nearshore stations are calculated. The results at these locations show that there are no uniform distribution patterns of joint distributions at different locations. The failure probability analysis is applied to calculate the most probable CWLs. The analysis is also extended to the entire coastline of the study site at 87 uniformly distributed locations, where the coastline is clearly identified with the predominance of the waves and surges. The empirical value of “100Hs+10S” and “50Hs+50S” combinations is recommended to estimate the 50-year return level situation when the joint probability data is unavailable and the “50Hs+50S” combination could be a meaningful indicator for events between 50-year and 100-year return levels. In comparison with the commonly used empirical design approaches, the improvement coefficient (Q) is introduced and calculated, which suggests that applying the joint probability approaches to the mid-east coast and southeast coast can improve the accuracy in predicting extreme combined water levels with the given return period. The results from
this study provide reliable and realistic design guidelines for coastal engineering applications.

Acknowledgements

The work was partly supported by the National Key R&D Program of China (2017YFC0405401); the National Natural Science Foundation of China (51909013, 51979076) and the Fundamental Research Funds for the Central Universities, China (B200204017, 2019B04814)

References


