Journal of Oceanology and Limnology
Vol. 44, No. 2, pp. 545-559, 2026
https://doi.org/10.1007/s00343-025-4329-4

Research Paper

An uncertainty evaluation for storm surge risk analysis based on information utilization efficiency*

Guilin LIU[1], Siyu DING[1], Shichun SONG[2, **], Bokai YANG[1], Pengyu ZHU[1], Liping WANG[3]

[1] *College of Engineering, Ocean University of China, Qingdao 266100, China*
[2] *Qingdao Military-Civilian Integration Development Group Co., Ltd., Qingdao 266500, China*
[3] *School of Mathematical Sciences, Ocean University of China, Qingdao 266100, China*

Received Dec. 5, 2024; accepted in principle Apr. 29, 2025; accepted for publication Jun. 6, 2025

Abstract With the development of methods for predicting extreme hydrological elements using probabilistic approaches, several commonly used methods have emerged for analyzing the risk of storm surge disasters, including the Annual Maxima method, the Peak-Over-Threshold method, the Gumbel distribution, and the Weibull distribution. Meanwhile, and emphases have been placed on assessing and comparing the applicability and stability of these various methods. To evaluate the rationality of different methods, we an entropy uncertainty analysis method was introduced based on information utilization efficiency, in which the sample Stochastic uncertainty is measured by the ratio of information entropy before and after sampling, i.e., the information extraction efficiency of the sampling method. Additionally, the cognitive uncertainty of the research method is assessed by the ratio of mutual information between the model and the sample to the information entropy of the sample, i.e., the information extraction efficiency of the mathematical model. Furthermore, we incorporated the group probability calculation method, information entropy and mutual information theory to analyze and calculate the entropy uncertainty more accurately. By applying this analysis to the design wave height and the recurrence period projected in the sea area west Guangdong of China, we believed that the most reasonable hazard assessment method shall be based on the over-threshold method combined with the Pareto distribution. Conversely, the assessment method based on the process extreme value method is deemed insufficiently reasonable and requires further research.

Keyword: uncertainty; storm surge; information entropy; mutual information; group probability calculation method

1 INTRODUCTION

The method for assessing storm surge hazard through probabilistic analysis involves three important parts: selection of samples, parameter estimation, and establishment of distribution models. Commonly used sampling methods include annual maxima (AM), peaks over threshold (POT), and storm process extreme value (SPE) method, among others. Parameter estimation methods encompass moment estimation, least squares estimation, maximum likelihood estimation, Bayesian estimation, and so forth. Distribution models used to reflect the statistical characteristics of the samples include Gumbel, Weibull, Pearson-III, Pareto, generalized Pareto distribution (GPD), and generalized extreme value distribution (GEV), among others (Carreau et al., 2009; Hussain et al., 2019; Gamet and Jalbert, 2022; Pan et al., 2022; Liu et al., 2023a; Marra et al., 2023). Faced with a wide variety of analytical methods, the question arises of how to compare the rationality of different methods in order to select the best one (Singh, 2019; Ebel and Moody, 2020;

* Supported by the National Natural Science Foundation of China (No. 52071306)
** Corresponding author: dsy18841262628@163.com

Martins et al., 2020; Anghel and Ilinca, 2022; Long et al., 2024).

Scientific research methods have performance limitations, and different methods generate varying levels of uncertainty, leading to different computational results (Huang et al., 2024; Li et al., 2024; Deng et al., 2025). Methods with lower uncertainty yield results that are closer to the real situation, making uncertainty a key metric for assessing the validity of scientific research methods. For example, based on particle swarm optimization, Acitas (2018) developed a new method for the estimating the Weibull distribution parameters and found that even with the same sample data and model distribution form, differences in model parameter estimation methods can significantly affect the final projection results. In recent years, in the field of storm surge hazard assessments, the degree of uncertainty that used to evaluate the discrepancy between scientific research results and actual phenomena has become an indicator for assessing the merits of assessment methods. The theory of uncertainty assessment for storm surge hazard assessments has been further refined. Some studies (Guan and Peng, 2015; Wang et al., 2016; Liu et al., 2021) have expanded the scope of the model, improved the accuracy of model parameter estimation and the precision of model prediction, etc. Lei et al. (2012). introduced the coefficient of variation (COV) value (standard deviation/mean) to measure the overall uncertainty, including both stochastic and cognitive uncertainties, of the results of marine environmental design parameter projections. Despite the diversity of uncertainty assessment methods, the aforementioned methods primarily evaluate the overall uncertainty of hazard assessment methods without independently considering stochastic and cognitive uncertainties.

Stochastic uncertainty refers to the challenge in scientific research methods to obtain true parameter values due to the randomness of data. Cognitive uncertainty denotes the discrepancy between research results and true values arising from performance limitations. The entropy uncertainty offers a flexible approach to consider either the overall uncertainty or both types of local uncertainties, which is a crucial aspect when offering improvement suggestions for specific steps in research methodologies. Consequently, the entropy uncertainty theory has progressively gained prominence in recent years. Alexander and Sarabia (2012) differentiated the uncertainty in risk assessment models into improper assumptions about statistical models and parameter uncertainty, and utilized the Shannon entropy (Shannon, 1948) to quantitatively assess the uncertainty of a one-dimensional model, thereby representing it as cognitive uncertainty. Chen et al. (2021) analyzed the various factors contributing to the uncertainty in the projection results of design wave heights in the Yellow Sea using the Monte Carlo method. Based on information entropy theory, Liu et al. (2023b) introduced a novel method for calculating the overall uncertainty of a multivariate marine environment design parameter extrapolation model; and meanwhile, they employed the Monte Carlo method to address the challenge of analyzing the impact of various factors on model uncertainty in small sample sizes. Currently, entropy uncertainty is often analyzed using a Monte Carlo-based distributed entropy uncertainty analysis method. However, in this algorithm, stochastic and cognitive uncertainties are intertwined, making it difficult to represent either uncertainty independently or flexibly. Furthermore, the calculation logic is grounded in the instability of data sources, resulting in complicated calculation steps that clash with commonly used probabilistic theories, such as maximum entropy theory.

In fact, in the past, when calculating entropy-based Stochastic uncertainty, the method used is to compute the information entropy variance. Such methods required establishing a distribution model for the data samples, followed by generating random samples to calculate both the information entropy and its variance. However, this process inadvertently introduced additional stochastic and cognitive uncertainties, posing limitations in terms of both the rationality and simplicity of the calculations. For this reason, this paper opts to use the group probability algorithm to calculate the information entropy. By omitting the modeling step, it avoids the introduction of cognitive uncertainty. Nevertheless, due to the inherent limitation of a typically small number of recorded storm surge events, the group probability method incurs a significant error in calculating information entropy. This paper opts to use the group probability algorithm to calculate the information entropy (Chao and Shen, 2003). By omitting the modeling step, it avoids the introduction of cognitive uncertainty. Nevertheless, due to the inherent limitation of a typically small number of recorded storm surge events, the group probability method incurs a significant error in calculating information entropy. Fortunately, Hausser

and Strimmer (2009) proposed an algorithm to reduce the error with less data, enables the independent estimation of stochastic entropy uncertainty within an acceptable error range.

The computational logic of entropy-aware uncertainty is often based on the degree of disorder in the data, i.e., measurement uncertainty. However, this measurement uncertainty does not fulfill the goal of assessing the discrepancy between the model and actual data. Instead, in hydrological numerical simulation methods, the theory of mutual information (Henderson and Vedral, 2000) is frequently employed to describe the error between the outcomes derived from the mathematical model and the real values. The outputs of numerical simulations and machine learning methods consist of datasets that can be used for mutual information calculations. Conversely, the outputs of probabilistic methods for hazard analysis are distribution models, which cannot be directly used to calculate mutual information with sample data. In this paper, we incorporate the Monte Carlo method to generate multiple sets of simulated data and solve for the average mutual information, thereby introducing the theory of mutual information into the cognitive uncertainty analysis of probabilistic methods.

In this study, the James-Stein Shrinkage (JSS) algorithm is introduced to tackle the complexities of the previous entropy uncertainty calculation process and the issue of intertwined local uncertainties. Furthermore, the theory of mutual information is incorporated, and the group probability algorithm is innovatively applied to the computation of mutual information. By using multiple Monte Carlo simulations, the cognitive uncertainty of the distribution model is calculated based on the mutual information between the model and sample data. This addresses the inadequacy of previous cognitive uncertainty assessments. Combining these algorithms, we construct an uncertainty analysis method for storm surge hazard assessment, using information utilization efficiency as the evaluation criterion. This method can be employed to assess the rationale of probabilistic methods for various storm surge hazards.

In this paper, we employ the ratio of information entropy before and after sampling as a metric for the information utilization efficiency of the sampling method, which is used to evaluate its stochastic uncertainty. Similarly, the information entropy of the sample combined with the mutual information between the sample and the distribution model serves as a measure of the information utilization efficiency for the distribution model, thereby assessing its cognitive uncertainty. These two measures are then integrated to determine the overall uncertainty of the research methodology. Taking the storm surge elevation in the western Guangdong waters as a case study, we analyze the uncertainty of various storm surge hazard assessment methods using various commonly employed sampling methods and distribution models as our research subjects. The results demonstrate that the entropy uncertainty analysis method designed in this study can analyze Stochastic uncertainty and mathematical cognitive uncertainty in a simpler and more reasonable way.

The main innovation of this study is the development of an entropy uncertainty evaluation algorithm for storm surge hazard analysis methods, using information utilization efficiency as the criterion. This method enables more reasonable and efficient evaluation of the rationality of hazard analysis methods. It provides a reference for formulating storm surge disaster prevention strategies.

2 INFORMATION ENTROPY STOCHASTIC UNCERTAINTY

This chapter introduces the stochastic uncertainty calculation method of the sampling method, which takes the ratio of the information entropy of the sampling result to the information entropy of the original data as the information utilization efficiency of the sampling method. And combined with the non-parametric information entropy calculation method, the discrete information entropy of the sampling result is directly calculated using the group probability method. In this way, the stochastic uncertainty of the sampling method is measured.

2.1 Information entropy and uncertainty

In 1948, American mathematician C. E. Shannon published a seminal paper, integrating the concept of thermodynamic entropy into information theory and introducing the concept of “information entropy”, which serves as a quantitative measure of the amount of information present. Concurrently, information is interpreted as an agent that reduces uncertainty, suggesting that the degree of uncertainty can be assessed through information entropy.

Since information entropy signifies the informational richness of sample data, and the quantity of information indicates the extent of uncertainty reduced, it stands as a fitting method to

quantify uncertainty. Furthermore, literature underscores that a straightforward positive relationship exists between information entropy and uncertainty. Consequently, a thorough analysis of information entropy suffices to mirror the level of uncertainty.

It is worth noting, however, that the results of that study depict the degree of disorder of the information source, which typically signifies the measurement uncertainty arising during the observation process. In contrast, the uncertainty in scientific research diverges from measurement uncertainty, as it tends to characterize the discrepancy between research findings and the true value. Scientific research involves extracting information from the natural world and abstracting and summarizing it to derive laws of nature. The uncertainty in scientific research stems from the waste and misinterpretation of information resources during the abstraction process. Consequently, the entropy-related uncertainty in scientific research methods does not align with the functional relationship between measurement uncertainty and information entropy.

According to the principle of uncertainty generation, the uncertainty (U) of a scientific study can be described using the Eq.1:

$$U = E_d = H_d/H_0 = 1 - E_u = 1 - H_u/H_0, \quad (1)$$

where H_0 refers to the total information entropy of a research object in nature, H_d refers to the information entropy lost in the scientific research process, and H_u refers to the information entropy utilized in the scientific research process. E_d and E_u refer to the loss rate and utilization efficiency of information, respectively, which are calculated by Eqs.2–3:

$$E_d = H_d/H_0 = (H_0 - H_u)/H_0, \quad (2)$$

$$E_u = H_u/H_0 = (H_0 - H_d)/H_0. \quad (3)$$

As can be seen from the Eqs.2–3, the uncertainty of scientific research has a negative correlation with the efficiency of information utilization, so the uncertainty of research methods can be judged directly by the efficiency of information utilization.

In calculating the Stochastic uncertainty due to the sampling method, Eq.1 can be reduced to:

$$U_1 = E_{d1} = 1 - E_{u1} = 1 - (H_{u1}/H_0), \quad (4)$$

where U_1, E_{d1}, and E_{u1} refer to the calculation of Stochastic uncertainty, information loss efficiency and information utilization efficiency caused by the sampling method, and H_{u1} refers to the information entropy of the obtained sample.

2.2 Group probability estimation algorithm for information entropy

2.2.1 Group probability estimation algorithm

The calculation of information entropy can be primarily summarized into two types of approaches: "distribution-based" or "cell probability-based". Most previous studies on hydrological information entropy have utilized the "distribution-based" algorithm. This calculation method exhibits significant disadvantages when assessing the Stochastic uncertainty of samples: firstly, establishing the distribution model is cumbersome, and the equation for calculating information entropy using this model is complex and challenging to compute; secondly, the validity of the calculation results hinges on the flexibility of the distribution model and the well-distributed of the data, which can result in a blending of stochastic uncertainty and cognitive uncertainties, rendering the calculations insufficiently reasonable.

Different from the "distribution-based" approach, the "group probability-based" approach first categorizes the data into groups, calculates the probability within each group, and then computes the information entropy using these group probabilities. Liu et al. (2016) provided a detailed elaboration on the characteristics and advantages of this method: it involves fewer steps in the calculation process and the results are not constrained by the limitations of the distribution model. Therefore, it can be argued that group probability methods offer advantages when calculating stochastic uncertainty. This approach directly calculates the information entropy for the samples, which not only simplifies the steps involved in uncertainty calculation but also separates stochastic and cognitive uncertainties, forming an independent method for assessing stochastic uncertainty.

The random data is divided into p groups, and the probability of each group is θ_1, θ_2, ⋯, θ_p, and satisfies $\sum_{i=1}^{p}\theta_i = 1$. Then the information entropy of this random variable can be calculated by Eq.5:

$$H = -\sum_{i=1}^{N}\theta_i \ln \theta_i. \quad (5)$$

Obviously, the selection of the number of groups, p, is crucial for the accuracy of information entropy estimation. An inappropriate choice of p can result in reduced calculation accuracy. Lall et al. (1993) and Chiu (1996) gives several methods of determining group widths so that they can be used to derive the

number of subgroups p. But these methods are computationally complex and are only suitable for large sample sizes. Zhuang (2005) provided an equation for determining the number of groups, p, specifically under conditions of small sample sizes:

$$p \approx 1.87(n-1)^{2/5}, \tag{6}$$

where n is the number of random samples.

2.2.2 Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation is the most commonly used information entropy estimation method due to its simple calculation theory and easy calculation steps. The calculation is done by Eq.7:

$$\hat{H}^{\mathrm{ML}} = -\sum_{i=1}^{p} \hat{\theta}_i^{\mathrm{ML}} \ln \frac{y_i}{n}. \tag{7}$$

Denote the likelihood function L as:

$$L(y_1, y_2, \cdots, y_p; \theta_1, \theta_2, \cdots, \theta_p) = \frac{n!}{\prod_{i=1}^{p} y_i} \prod_{i=1}^{p} \theta_i^{y_i}, \tag{8}$$

where y_i is the variable group count, which is the number of data having in the i^{th} group.

The maximum likelihood estimate of the group probability can be obtained:

$$\hat{\theta}_i^{\mathrm{ML}} = \frac{y_i}{n}. \tag{9}$$

The information entropy of the great likelihood estimation can be obtained by substituting $\hat{\theta}_i^{\mathrm{ML}}$ into the information entropy calculation by Eq.10:

$$\hat{H}^{\mathrm{ML}} = -\sum_{i=1}^{p} \hat{\theta}_i^{\mathrm{ML}} \ln \hat{\theta}_i^{\mathrm{ML}}. \tag{10}$$

The maximum likelihood estimation is a simple and effective method for calculating information entropy when the amount of data is large. However, its calculation accuracy cannot be guaranteed when the data is limited. Studies have confirmed that, in cases of insufficient data, the estimated value of information entropy will be significantly lower than the true value.

2.2.3 Shrinkage estimation

Shrinkage estimation is a novel estimation method that applies JSS to the maximum likelihood group probability $\hat{\theta}_i^{\mathrm{ML}}$. Hausser and Strimmer (2009) conducted simulation experiments to validate the algorithm's advantages in both accuracy and speed, concluding that it is a highly desirable method for estimating information under small sample conditions. James-Stein-type shrinkage involves a weighted average between a high-dimensional model characterized by low bias but high variance, and a low-dimensional model with high bias but low variance. The group probability obtained through the JSS method is a weighted average of the maximum likelihood-estimated group probability and the shrinkage target t_k. The equation for calculating this weighted average of group probabilities is:

$$\hat{\theta}_i^{\mathrm{JSS}} = \lambda t_k + (1-\lambda)\hat{\theta}_i^{\mathrm{ML}}, \tag{11}$$

where t_k=1/p; t_k is the deflation strength, $t_k \in [0.1]$.

Consider Bias$\left(\hat{\theta}_i^{\mathrm{ML}}\right) = 0$, by unbiased estimator:

$$\widehat{\mathrm{Var}}(\hat{\theta}_i^{\mathrm{ML}}) = \frac{\hat{\theta}_i^{\mathrm{ML}}(1-\hat{\theta}_i^{\mathrm{ML}})}{n-1}, \tag{12}$$

which in turn yields an estimate of λ:

$$\hat{\lambda} = \frac{\sum_{i=1}^{p} \widehat{\mathrm{Var}}(\hat{\theta}_i^{\mathrm{ML}})}{\sum_{i=1}^{p} (t_k - \hat{\theta}_i^{\mathrm{ML}})^2} = \frac{1 - \sum_{i=1}^{p} (\hat{\theta}_i^{\mathrm{ML}})^2}{(n-1)\sum_{i=1}^{p} (t_k - \hat{\theta}_i^{\mathrm{ML}})^2}. \tag{13}$$

Substituting the resulting $\hat{\theta}_i^{\mathrm{JSS}}$ into the information entropy equation, the information entropy shrinkage estimation can be obtained:

$$\hat{H}^{\mathrm{JSS}} = -\sum_{i=1}^{p} \hat{\theta}_i^{\mathrm{JSS}} \ln \hat{\theta}_i^{\mathrm{JSS}}. \tag{14}$$

Obviously, the JSS method is an improvement upon the MLE method. Liu et al. (2016). compared several methods, proved that the JSS method excels in estimation and calculation. Furthermore, it has been demonstrated that when estimating the information entropy of hydrological data, the JSS method yields comparable results to the MLE method for larger datasets, but significantly outperforms the MLE method for smaller datasets. Therefore, this paper introduces the JSS method for calculating the information entropy of hydrological data.

3 MUTUAL INFORMATION COGNITIVE UNCERTAINTY

In this chapter, we established a criterion for calculating the cognitive uncertainty of the research method by focusing on the efficiency of the mathematical model in extracting information from sample data. Specifically, we defined this efficiency as the ratio of the mutual information between the mathematical model and the sample data to the information entropy of the sample data. By adopting this ratio, we obtained a measure for assessing cognitive uncertainty.

3.1 Mutual information and uncertainty

Previously, mutual information has predominantly been employed to quantify the uncertainty in numerical simulation models, with limited application in the uncertainty analysis of probabilistic methods. We combined the Monte Carlo simulation with the computation of cognitive uncertainty to determine specifically the mutual information between simulated data from a distribution model and real-world samples. To account for the inherent randomness in the Monte Carlo methods, we utilized multiple sets of simulation samples to mitigate the Stochastic uncertainties and adopt the average mutual information as our metric. The ratio of this mutual information to the sample information entropy was then calculated to assess the information utilization efficiency of the research method, to which we interpreted as a measure of its cognitive uncertainty.

Mutual information quantifies the shared information content between two or more variables. The higher the mutual information, the stronger the correlation among the variables. Conceptually, mutual information serves as an extension of the correlation coefficient to scenarios involving high-dimensional and nonlinear relationships. The interplay between information entropy, mutual information, and conditional information entropy is depicted in the following:

The mutual information of X and Y in Fig.1 is the amount of information shared by X and Y. It also refers to the known value of Y, of which how much information about X is known.

In the case of discrete random variables, the mutual information of two discrete random variables X and Y can be defined as:

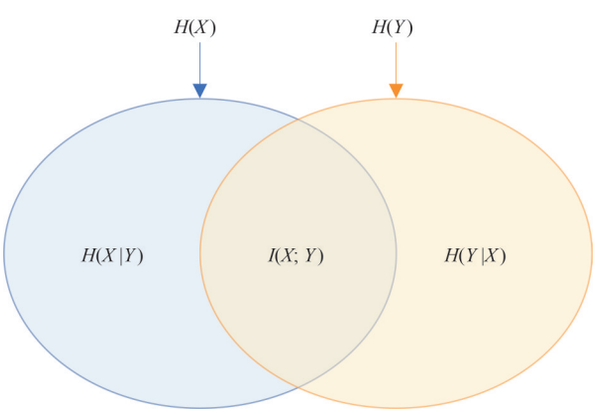

Fig.1 Relationship between information entropy, mutual information, and conditional information entropy

$$I(X;Y)\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \quad (15)$$

Knowing the input X_{in} and the output X_{out}, there is a mutual information inequality:

$$H(X_{in})\geq I(X_{in};X_{out}). \quad (16)$$

The equation equals only when all the information in X_{in} is fully used by the mathematical model. However, it is clear that in scientific research, there are rarely mathematical models that perfectly harness all the information. Even if such models exist, they tend to lack practicality owing to computational inefficiency and strict data requirements. Therefore, in most cases, $H(X_{in})>I(X_{in};\ X_{out})$. We define the information wasted as a result of model imperfections as cognitive uncertainty, with the information loss being calculated using Eq.17:

$$H_d=H(X_{in})-I(X_{in};X_{out}). \quad (17)$$

Cognitive uncertainty quantifies the discrepancy between the information utilized by the mathematical model, denoted as $I(X_{in};\ X_{out})$, and the information inherent in the data, denoted as $H(X_{in})$. This metric offers insight into the model's capacity to leverage sample data effectively. In contrast to stochastic uncertainty, cognitive uncertainty pertains to the model's inefficiency in harnessing information from the sample data, and it is characterized by Eq.18:

$$U_2=E_{d2}=1-E_{u2}=1-\ I(X_{in};X_{out})/H_{u1}, \quad (18)$$

where U_2, E_{d2}, and E_{u2} refer to the cognitive uncertainty, information loss efficiency and information utilization efficiency caused by the mathematical model. X_{in} refers here to the sample data, X_{out} refers here to the Monte Carlo simulation data of the mathematical model, and $I(X_{in};\ X_{out})$ refers to the mutual information of the sample and the mathematical model.

3.2 Group probability algorithm for mutual information

According to the one-dimensional information entropy group probability algorithm, it is generalized to two dimensions. The two-dimensional mutual information group probability algorithm can be obtained. The grouping of data is calculated by the Eq.19:

$$p_i\approx 1.87(n_i-1)^{2/5};i=1,2, \quad (19)$$

where p_1 and n_1 are the number of groups and

number of samples, and p_2 and n_2 are the number of groups and number of simulated data.

The sample and the random data are corresponded one-to-one as two-dimensional data, and the data are grouped two-bit into $p_1 \times p_2$ groups, each with probability θ_{11}, θ_{21}, ⋯, $\theta_{p_1 p_2}$ and satisfying $\sum_{j=1}^{p_2}\sum_{i=1}^{p_2}\theta_{ij}=1$. The mutual information of this random variable can be calculated by Eq.20:

$$I(X;Y)=\sum_{j=1}^{p_2}\sum_{i=1}^{p_1}\theta_{ij}(x,y)\ln\frac{\theta_{ij}(x,y)}{\theta_i(x)\theta_j(y)}, \tag{20}$$

where X and Y denote the sample and the random data, respectively, $\theta_{ij}(x, y)$ represents the joint group probability of the sample and the random data. $\theta_i(x)$ and $\theta_j(y)$ represent the marginal group probabilities of the sample and the random data, respectively, and are computed using Eqs.21–22:

$$\theta_i(x)=\sum_{j=1}^{p_2}\theta_{ij}, \tag{21}$$

$$\theta_j(x)=\sum_{i=1}^{p_1}\theta_{ij}. \tag{22}$$

Its maximum likelihood estimation method is similar to Section 2.3.3 of this paper and will not be repeated here.

4 PRACTICAL EXAMPLE

In this study, we selected several sampling techniques, including the commonly used fixed-window methods (the annual, quarterly, and monthly extreme value methods), POT, and the SPE methods. Subsequently, we extracted the fitted optimal distributions from each sampling result and apply them as distinct hazard assessment methods, respectively. For each sampling result, we derived the fitted optimal distribution, which we then adopted as a distinct hazard assessment methodology. To illustrate our approach, we used storm surge data from the Naozhou observation station. We evaluated the Stochastic uncertainty and cognitive uncertainty associated with each hazard assessment method separately. Ultimately, we integrated these two uncertainties into a comprehensive evaluation criterion. Uncertainty evaluation criterion was used to assess the rationality and precision of the hazard assessment methods.

4.1 Sources of data

Naozhou Island in Zhanjiang, Guangdong, South China, is renowned for its island eco-tourism. However, Guangdong is one of the regions in China most affected by tropical cyclones and typhoon disasters. In this study, we used storm surge data from the Naozhou observation station as a case study to estimate the uncertainty of a hazard analysis model. The raw data for this study come from tropical cyclone surge records spanning from 1980 to 2016. These records were independently observed for each tropical cyclone that occurred in the Western Guangdong Sea (WGS). It is worth noting that the storm surge data are available during tropical cyclones only I as extreme wave heights occur during these periods only, and most sampling methods exclude non-extreme values during non-cyclone periods. The extreme value theory was generalizable and valid for data from various regions were used to construct hazard analysis methods. The selected storm surge data can be applied to build hazard analysis models. The purpose of using data from this region was first to establish various hazard analysis methods and then to calculate uncertainties for different methods, thereby demonstrating that the uncertainty calculation method designed in this study can evaluate the rationality of different hazard analysis methods, thus the data is suitable for the uncertainty analysis cases in this study.

4.2 Stochastic uncertainty

Tropical cyclone surge data from the Naozhou Island station were sampled using the annual extreme method for the years 1980–2016, excluding 2004 and 2007. The results of this sampling are shown below as Fig.2. The blue line in the figure depicts all the tropical cyclone surge data collected at the Naozhou Island station between 1980 and 2016. Meanwhile, the data points enclosed in red circles represent the annual extreme surge values obtained through the application of the annual extreme method.

It can be clearly seen that the annual extreme method takes less samples. It is difficult to accurately estimate the stochastic uncertainty of the annual extreme method sample using the traditional group probability method.

Here the JSS method is chosen to estimate the stochastic uncertainty of the annual extreme sample. Using Eq.6 to calculate p, the annual extreme data are grouped according to the grouping number p. The grouping results are shown below as Fig.3.

The maximum likelihood estimate of the group probability, derived from Eq.9 of the maximum likelihood estimation method, is simply the group

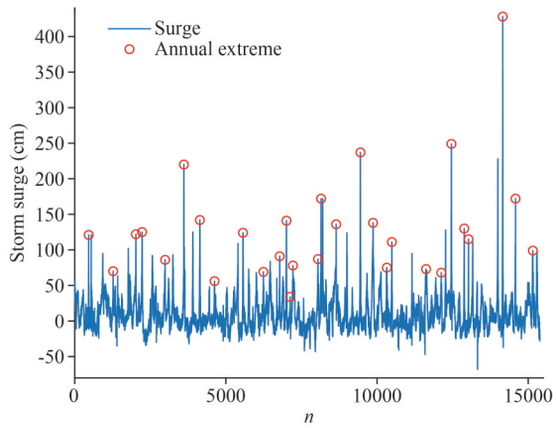

Fig.2 Annual extreme method sampling map

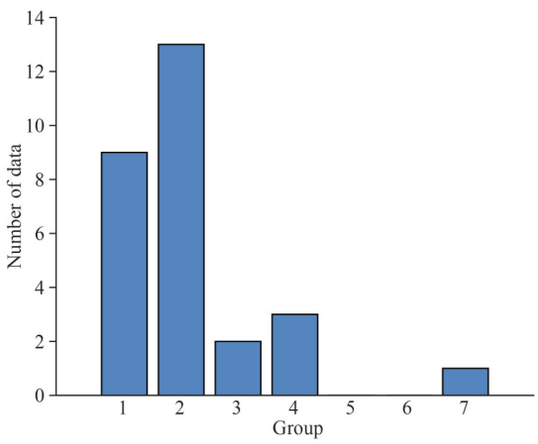

Fig.3 Annual polar method grouping diagram

probability itself. Next, based on the group probabilities of each subgroup, we used Eq.11 of the JSS method to correct these values. The result of this correction process is the information entropy of the annual extreme data.

Similarly, the tropical cyclone water increase data from the salammoniac test station for the years 1980–2016 were sampled. This sampling was done using both the quarterly extreme value method and the monthly extreme value method. The results of these sampling methods are shown in Fig.4.

In Fig.4, the data were obtained through sampling using two methods: the quarterly extreme value method and the monthly extreme value method. The sampling results are highlighted in red. It is evident that both the quarterly and monthly extreme value sampling methods yield significantly more results compared to the annual extreme method. This theoretically allows for the retention of more information. However, at the lower data range, the samples selected by these two sampling methods may lack representativeness. This is due to

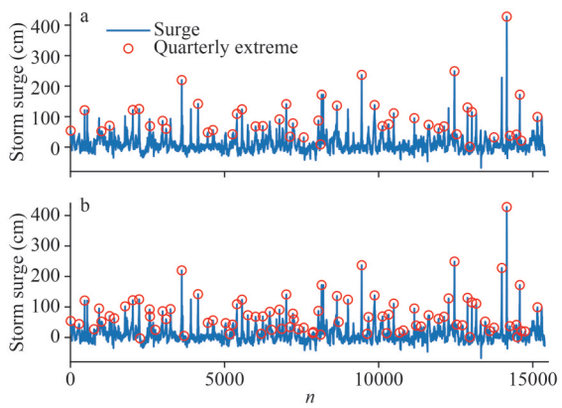

Fig.4 Sampling plots of quarterly extreme value method (a) and monthly extreme value method (b)

the absence of storms in certain quarters or months. Given their respective strengths and weaknesses, it is challenging to conclusively determine which method is superior solely through visualization. Therefore, calculating the uncertainty is necessary. This will help ascertain which time window is more sensible for sampling.

The data is then sampled based on the over-threshold and process extreme value methods.

The threshold selection process is as follows:

(1) Calculate the mean value $\overline{X}$ and kurtosis K_n of the samples:

$$\overline{X} = \frac{(X_1 + X_2 + X_3 + \cdots + X_n)}{n}, \tag{23}$$

$$K_n = \frac{(X_i - \overline{X}_n)^4}{(E - (X_i - \overline{X}_n)^2)^2}, i = 1, 2, 3, \cdots, n. \tag{24}$$

(2) Evaluate the kurtosis K_n, if $K_n \geq 3$, remove the X_i with the maximum $|X_i - \overline{X}_n|$.

(3) Repeat Steps 1 and 2 until $K_n < 3$.

Select the maximum $|X_k|$ from the remaining data as the threshold.

The kurtosis method is an automatic threshold selection method for the POT method, eliminating interference from manual threshold selection. This ensures that the calculated uncertainty in the method remains unaffected by researcher involvement, which helps in more effectively analyzing the uncertainty of hazard analysis methods.

The sampling results are shown in Fig.5.

It can be seen that both methods obtain more samples and enables better retention of data information. However, during a storm event, the POT method records data throughout the entire process, resulting in sample clustering. This phenomenon can lead to an overestimation of

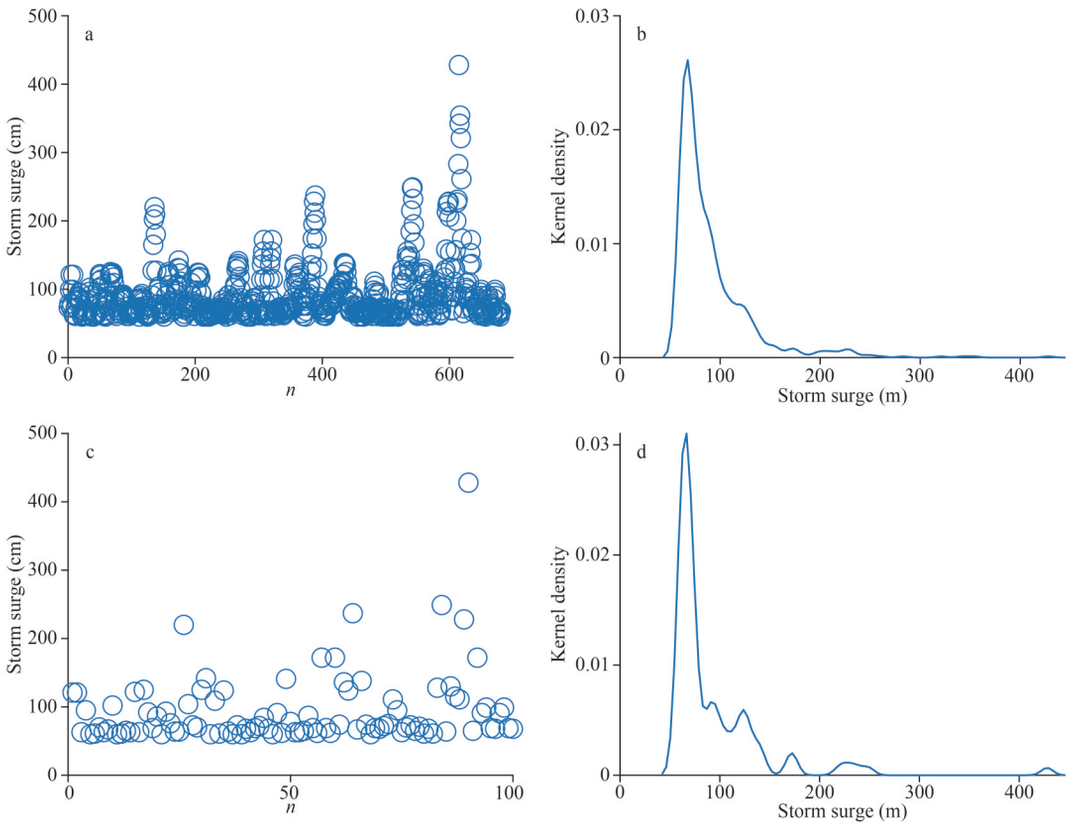

Fig.5 Sampling charts for POT and SPE methods

a. sampling charts for POT methods; b. sampling charts for SPE methods; c. Kernel density estimation of the POT sample; d. Kernel density estimation of the SPE sample.

extreme events. In contrast, clustering does not occur in extreme event samples selected by the method of SPE. However, further research is needed on how to adjust the determination of the storm process. Its reasonableness likewise requires computational uncertainty to be visualized.

The sampling results of the four sampling methods were grouped using the group probability method and the grouping results were shown in Fig.6.

It is clear that the group probability values of the sample data are not predictable like those of an extreme value distribution function. This highlights the necessity of using the group probability method to calculate the information entropy of the samples. By similarly grouping all data in the database to calculate the group probability, the sample information entropy and stochastic uncertainty for each sampling method can then be obtained according to Eq.14. The results are shown in Table 1.

Theoretically, the fixed window method, especially the annual maxima method, has lower data utilization efficiency. As a result, the sample information entropy is smaller, and the stochastic uncertainty is higher. By contrast, the POT method and the SPE method make more efficient use of data. They produce higher sample information entropy and lower stochastic uncertainty. Interestingly, even though the POT method retains the most information, it does not achieve the lowest stochastic uncertainty. Instead, the SPE method, which theoretically balances sample representativeness and quantity best, achieves the lowest stochastic uncertainty. These results confirm that the stochastic uncertainty method developed in this study is effective. It assesses both the quantity and representativeness of sampling results, offering a sound basis for evaluating the rationality of sampling methods.

4.3 Cognitive uncertainty

In marine engineering, commonly used one-dimensional models for fitting extreme value data include the Gumbel, Weibull, Pearson- III , Pareto, GPD, and GEV models. We selected these six models to fit the sample data, using the Kolmogorov-Smirnov test to determine the optimal fit distribution.

The KS test results were used to evaluate the distribution model that best fits the sampling

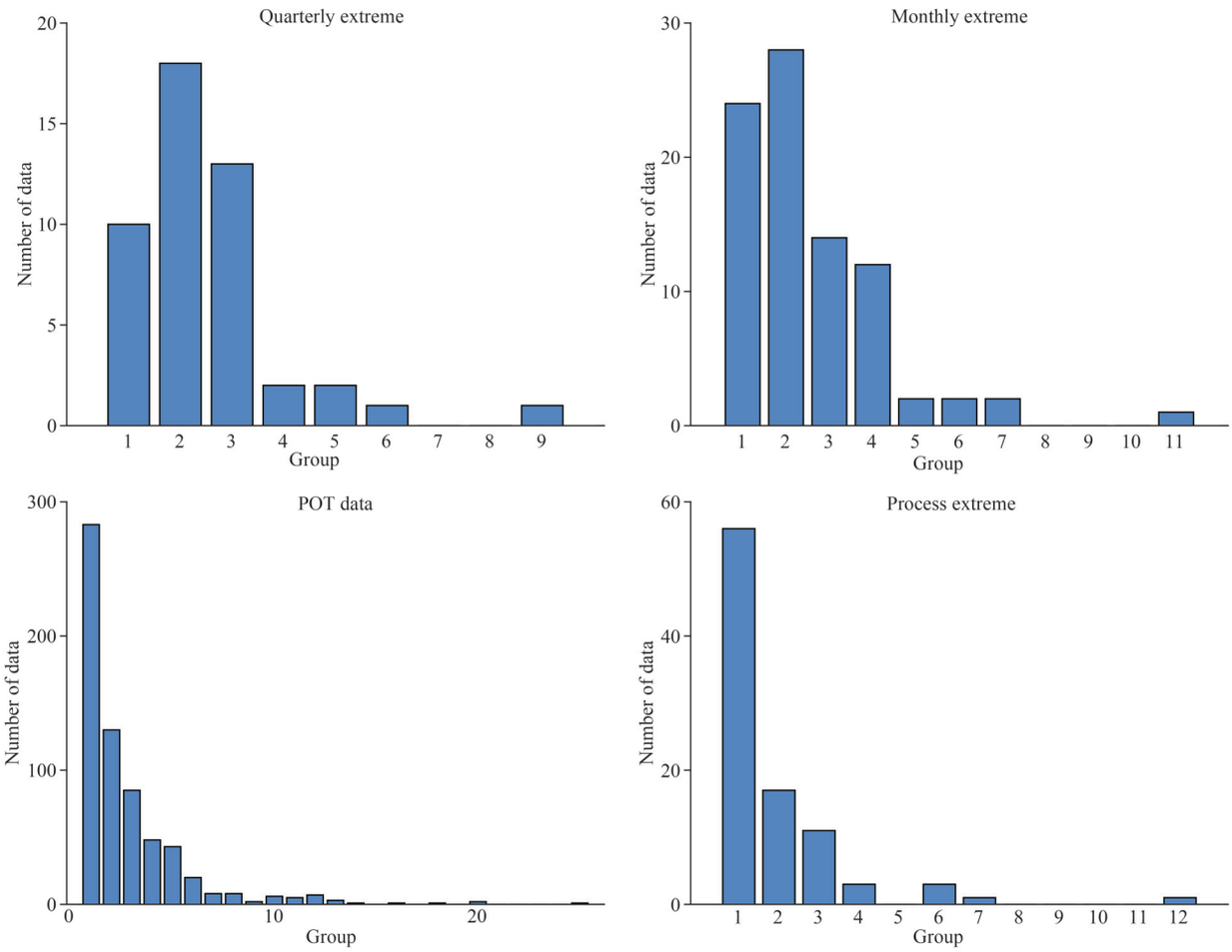

Fig.6 Quarterly extreme value method grouping diagram (a); monthly extreme value method grouping diagram (b); threshold method grouping diagram (c); storm process extreme value method grouping diagram (d)

Table 1 Information entropy and stochastic uncertainty in sampling methods

Variable	Information entropy	Uncertainty (U_1)
Annual extreme	1.456 5	0.463 4
Quarterly extreme	1.647 2	0.393 1
Monthly extreme	1.732 4	0.361 8
Overthreshold data	1.744 1	0.357 4
Process extreme	2.034 6	0.250 4

Table 2 *P* for the KS test

Distribution	Annual extreme	Quarterly extreme	Monthly extreme	POT	Process extreme
Gumbel	0.880 5	0.857 9	0.852 7	1.356 8	2.570 2
Weibull	0.915 4	0.646 6	0.851 6	3.903 3	1.261 9
GEV	0.979 5	0.974 3	0.914 9	0.003 3	**0.266 5**
Pearson3	**0.545 8**	0.759 1	0.879 4	1.432 0	4.399 2
Pareto	0.000 4	7.151 9	5.999 2	**0.130 3**	0.029 2
GPD	0.831 2	**0.078 0**	**0.571 9**	3.363 5	2.075 3

Bold values indicate the smallest *P* for each method-sample combination.

method (Table 2).

Table 2 shows the *P* of the KS test for combinations of various samples and distribution models. Combinations in bold are rejected by the KS test. The method-sample combination with the smallest *P* is the best-fitting distribution, serving as the cognitive capability limit for each hazard assessment method.

First, take the annual extreme sample as an example. The optimal fitted distribution is selected. Using the Monte Carlo method, simulated samples are generated to match the number of original samples. Each simulated data point is then paired one-to-one with a real sample data point. This forms a two-dimensional corresponding dataset, as shown in Fig.7.

The two-dimensional dataset is concentrated in a smaller region, with fewer extreme events. This distribution aligns with the actual behavior of

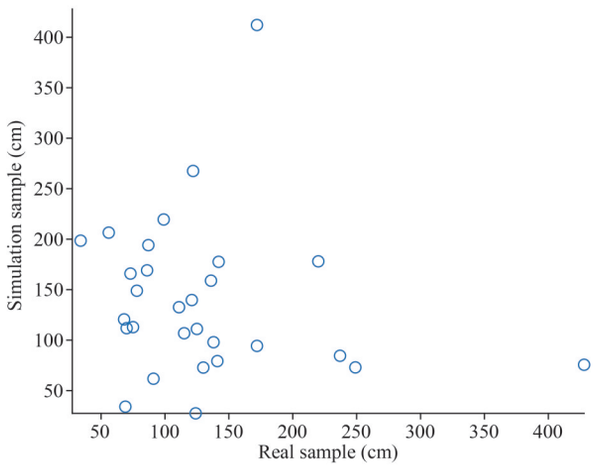

Fig.7 Annual extremes method simulation corresponds to real sample data

hydrological extremes. Next, the data is grouped in two dimensions, and the group frequencies are calculated. The results of these frequency calculations are shown in Fig.8.

The grouping structure aligns with the data distribution. This reflects the statistical characteristics of hydrological data. Group frequency was used as the maximum likelihood estimate for group probability. Mutual information was then calculated using the group probability method.

Similarly, the corresponding datasets of simulated and real data for the quarterly maxima method, monthly maxima method, POT method, and SPE method are shown in Fig.9.

It can be seen that the distribution of all the data are consistent with the distribution of hydrological extreme events. By grouping them in two dimensions and calculating the group frequency, the results can be obtained as shown in Fig.10.

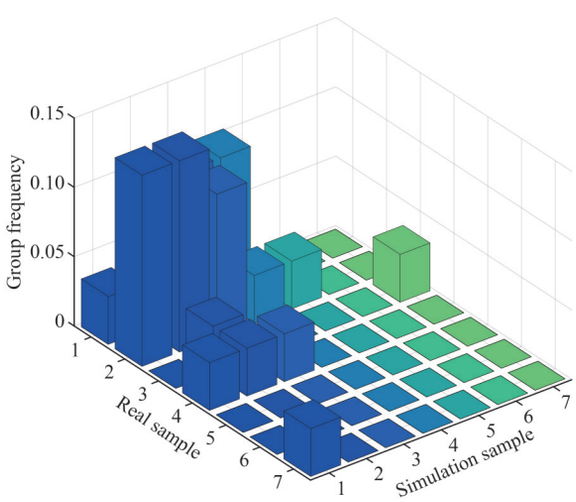

Fig.8 Two-dimensional grouping diagram of the Annual Extreme method

As shown in Fig.10, this method can determine the number of groups for sample sets of different sizes. During grouping calculations, the POT method with more samples has the most groups, which ensures its calculation accuracy. In contrast, the annual extreme value method and quarterly extreme value method with fewer samples have coarser grouping, failing to guarantee the calculation accuracy of information entropy. Therefore, it is necessary to introduce the JSS algorithm to correct the group probability in this study.

The results of the cognitive uncertainty of the research methodology through group frequency calculations are shown in Table 3.

Cognitive uncertainty (U_2) reflects the ability to match sampling results with the tested distribution model. If the distribution model lacks flexibility, it may result in less efficient utilization of information from the sample data. Additionally, if certain parameters of the sampling method (such as the window width in the fixed window method or the threshold in the peak-over-threshold method) are unreasonable, this can further reduce the degree of matching. Moreover, if the distribution model does not align well with the sampling method from a probabilistic theory perspective, cognitive uncertainty may increase, which is consistent with the research implications of cognitive uncertainty.

As shown in Table 3, the samples from the annual extreme value method, quarterly extreme value method, and monthly extreme value method demonstrate an increasing trend in mutual information and a corresponding decrease in cognitive uncertainty as they are fitted to the optimal distribution. In contrast, the peak-over-threshold method exhibits the highest mutual information in the results, resulting in the lowest cognitive uncertainty. Meanwhile, the storm peak-over-threshold method produces the highest cognitive uncertainty.

This study selected six classic and commonly used distribution models, representing the cognitive limits of conventional hazard assessment methods. The calculated results show significant differences in cognitive uncertainty. This may be due to the low compatibility between traditional distribution models and certain sampling methods. If better-performing composite distribution models or extreme value extension models are used, they could

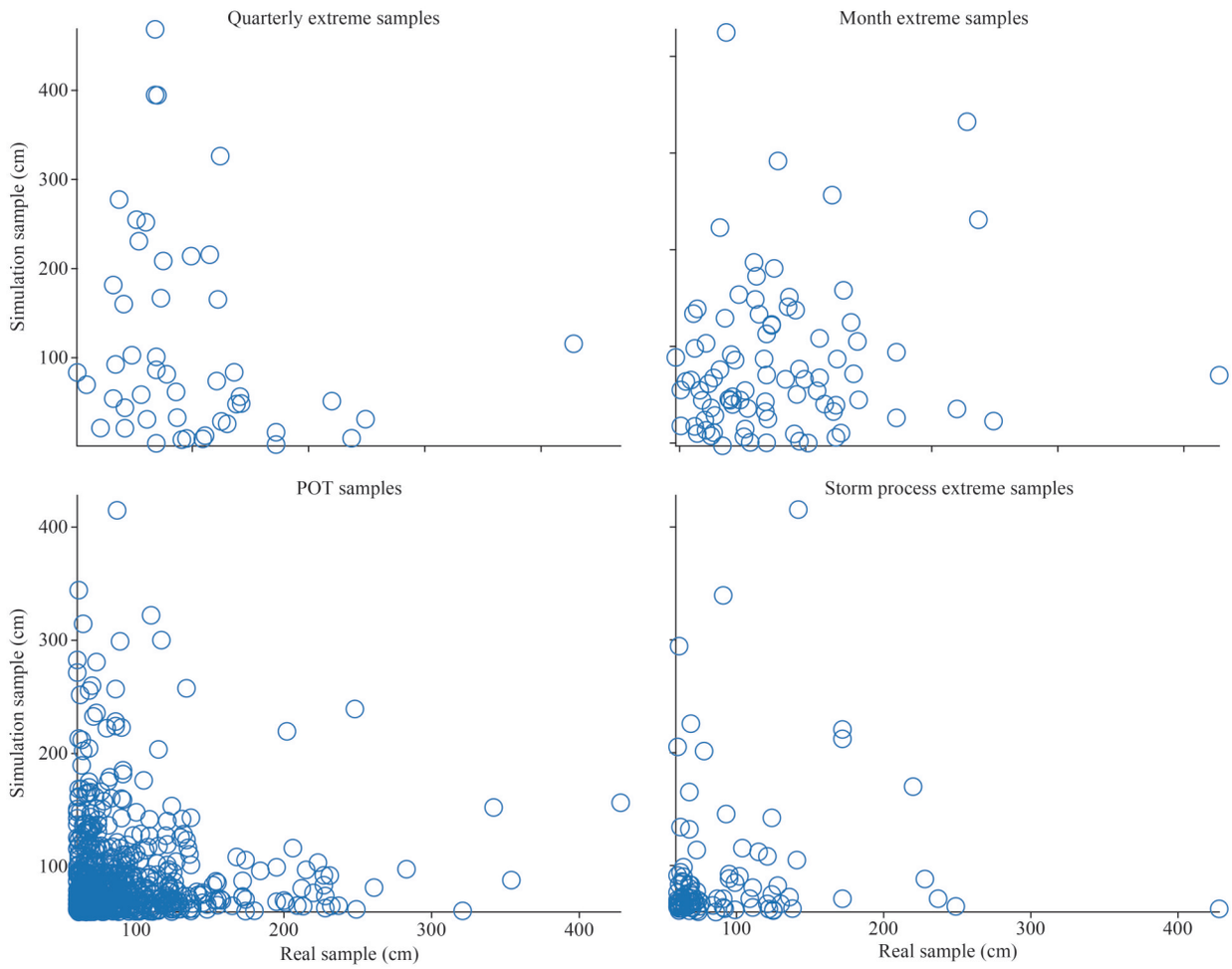

Fig.9 Quarterly extreme value method data correspondence graph (a); monthly extreme value method data correspondence graph (b); threshold crossing method data correspondence graph (c); storm process extreme value method data correspondence (d)

leverage richer sample information. As a result, this may lead to lower cognitive uncertainty.

4.4 Overall uncertainty

In Eq.1, the overall uncertainty U can be obtained by substituting the mutual information I (X_{in}; X_{out}) of the samples ad the model into H_d. Additionally, the information entropy of all the data in the database is included in H_0. This calculation provides the overall uncertainty of various hazard assessment methods, as shown in Table 4.

Obviously, the fixed-window method reduces the overall uncertainty in hazard analysis as the window width decreases, making the research method more rational. The lowest uncertainty is still observed in the over-threshold method combined with the Pareto distribution. In contrast, the SPE method exhibits the highest uncertainty. This may indicate that further research is needed on the current correction methods for assessing storm processes and selecting thresholds. Alternatively, the commonly used traditional extreme value distributions may lack the flexibility required to accurately represent the statistical characteristics of the process extreme value samples.

Therefore, this study provided better references for selecting hazard analysis methods. In practice, the scientific optimization of storm surge hazard analysis methods is a core link to enhance the reliability of risk assessment. As storm surge research deepens, detailed analyses of different regions will become a new research direction. With the refinement of study areas and the diversification of research methods, selecting superior analysis methods can effectively improve the reliability of risk prevention and control. The uncertainty analysis

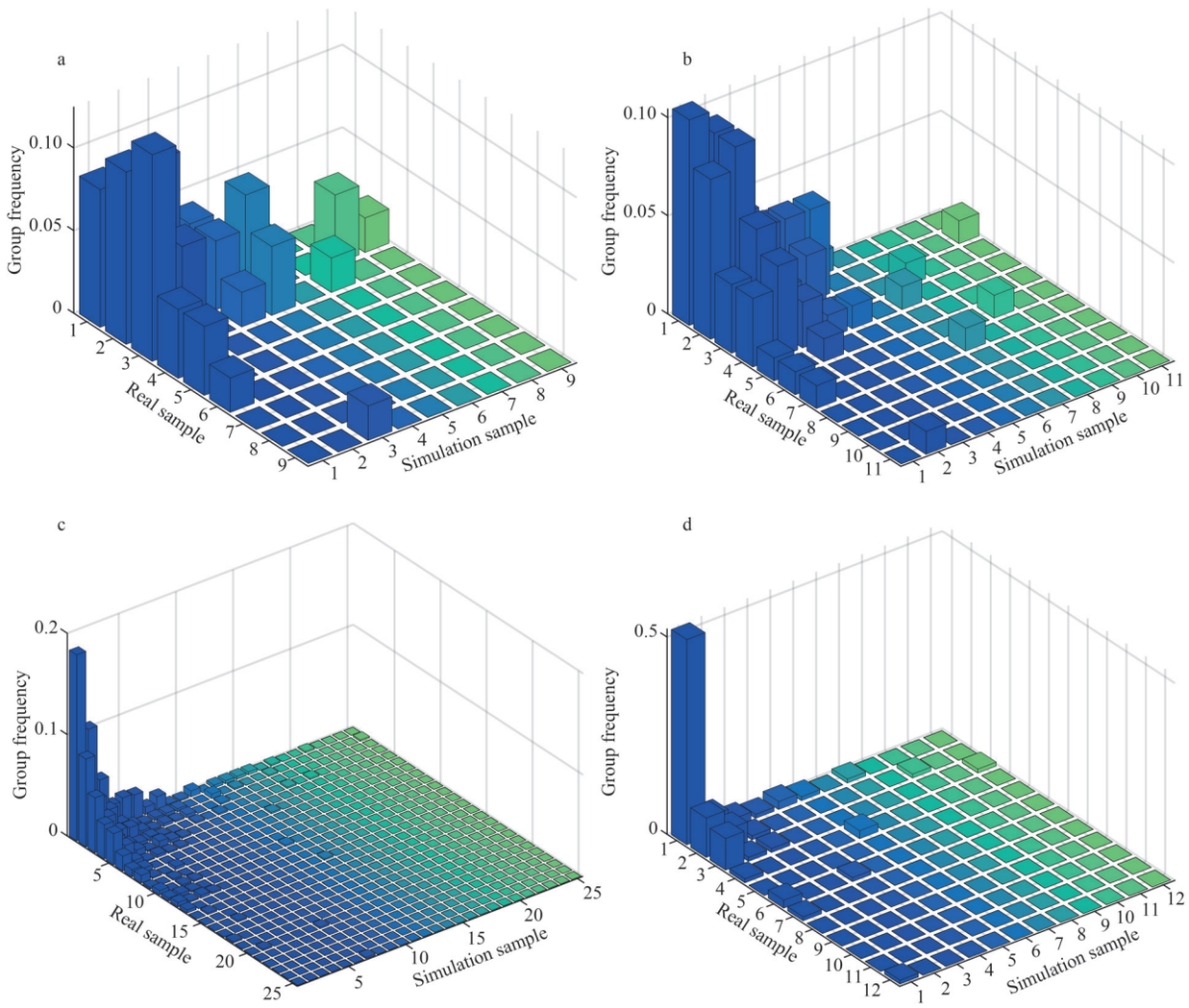

Fig.10 Two-dimensional grouping diagram by quarterly extreme value method (a); two-dimensional grouping diagram by monthly extreme value method (b); two-dimensional grouping diagram by threshold crossing method (c); two-dimensional grouping diagram by extreme value method for storm process (d)

Table 3 Model-sample mutual information and cognitive uncertainty

Variable	Distribution	Mutual information	Uncertainty (U_2)
Annual extreme	Weibull	0.943 8	0.350 8
Quarterly extreme	GPD	1.300 2	0.185 1
Monthly extreme	GPD	1.584 9	0.084 9
Overthreshold data	Pareto	1.740 0	0.002 0
Process extreme	GEV	0.924 6	0.546 1

Table 4 Stochastic uncertainty, cognitive uncertainty, and overall uncertainty for each method

	Distribution	U_1	U_2	U
Annual extreme	Weibull	0.463 4	0.350 8	0.651 6
Quarterly extreme	GPD	0.393 1	0.185 1	0.505 5
Monthly extreme	GPD	0.361 8	0.084 9	0.415 9
Overthreshold data	Pareto	0.357 4	0.002 0	0.358 7
Process extreme	GEV	0.250 4	0.546 1	0.659 8

method propsed in this study can accommodate data from different regions or incorporate more hazard analysis methods. It provides objective and reasonable references for optimizing hazard analysis methods, thereby enhancing the effectiveness of decision-making on disaster prevention.

5 CONCLUSION

This study used information utilization efficiency as the standard for evaluating uncertainty, with information entropy and mutual information as the measures of information quantity, which established independent assessment criteria for stochastic

uncertainty and cognitive uncertainty, and provided a reference for assessing the rationality of hazard assessment methods. Through evaluating the uncertainties associated with several hazard assessment methods, the following conclusions are drawn:

In hazard assessment methods using fixed-window sampling, the stochastic uncertainty of the sampling method decreased with the window width decrease, so did the cognitive uncertainty of the model. Therefore, among all the methods, the monthly extreme value method provided the best reasonable results.

When using the six traditional distribution models in this study, the over-threshold method combined with the Pareto distribution model formed a more reasonable hazard assessment method. This approach had higher information utilization efficiency, and the results are closer to actual conditions.

Hazard assessment methods using the SPE approach are the least efficient in information utilization and have the highest uncertainty. However, the SPE method itself has the lowest stochastic uncertainty due probably to limited research on the method or to the performance of the distribution model.

The innovation of this study lies on constructing a more reasonable entropy uncertainty analysis method for storm surge hazard analysis. The introduction of the JSS algorithm, group probability, and mutual information concepts addressed the performance limitations of traditional entropy uncertainty algorithms, which provided better references for selecting hazard analysis methods and presented an uncertainty analysis method for hazard analysis. This method was used to evaluate the rationality of various hazard analysis methods for the Naozhou region, showing that the method could effectively determine the most reasonable hazard analysis method, thereby improving the reliability of disaster prevention strategies.

6 DATA AVAILABILITY STATEMENT

Due to the subsequent research needs of the research group, the datasets generated and/or analyzed during the current research period are not made public. However, at the reasonable request, they can be obtained from the corresponding author.

References

Acitas S. 2018. A new weighted distribution as an extension of the generalized half-normal distribution with applications. *Journal of Statistical Computation and Simulation*, **88**(12): 2325-2341, https://doi.org/10.1080/00949655.2018.1462812.

Alexander C, Sarabia J M. 2012. Quantile uncertainty and value-at-risk model risk. *Risk Analysis*: *An Official Publication of the Society for Risk Analysis*, **32**(8): 1293-1308, https://doi.org/10.1111/j.1539-6924.2012.01824.x.

Anghel C G, Ilinca C. 2022. Parameter estimation for some probability distributions used in hydrology. *Applied Sciences*, **12**(24): 12588, https://doi.org/10.3390/APP122412588.

Carreau J, Naveau P, Sauquet E. 2009. A statistical rainfall-runoff mixture model with heavy-tailed components. *Water Resources Research*, **45**(10): W10437, https://doi.org/10.1029/2009WR007880.

Chao A N, Shen T J. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, **10**(4): 429-443, https://doi.org/10.1023/A:1026096204727.

Chen B Y, Kou Y, Zhao D et al. 2021. Maximum entropy distribution function and uncertainty evaluation criteria. *China Ocean Engineering*, **35**(2): 238-249, https://doi.org/10.1007/S13344-021-0021-4.

Chiu S. 1996. A comparative review of bandwidth selection for kernel density estimation. *Statistica Sinica*, **6**(1): 129-145, https://www.jstor.org/stable/24306003.

Deng W, Li X Y, Xu J J et al. 2025. BFKD: blockchain-based federated knowledge distillation for aviation Internet of Things. *IEEE Transactions on Reliability*, **74**(2): 2626-2639, https://doi.org/10.1109/tr.2024.3474710.

Ebel B A, Moody J A. 2020. Parameter estimation for multiple post-wildfire hydrologic models. *Hydrological Processes*, **34**(21): 4049-4066, https://doi.org/10.1002/hyp.13865.

Gamet P, Jalbert J. 2022. A flexible extended generalized Pareto distribution for tail estimation. *Environmetrics*, **33**(6): e2744, https://doi.org/10.1002/ENV.2744.

Guan Q S, Peng W. 2015. Parameter estimation for geometric-gumbel compound extreme-value distribution based on the pi-th quantiles of samples. *In*: Proceedings of the 2015 Conference on Informatization in Education, Management and Business. Atlantis Press, Guangzhou, China. p.60-64, https://doi.org/10.2991/iemb-15.2015.12.

Hausser J, Strimmer K. 2009. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, **10**: 1469-1484.

Henderson L, Vedral V. 2000. Information, relative entropy of entanglement, and irreversibility. *Physical Review Letters*, **84**(10): 2263-2266, https://doi.org/10.1103/PhysRevLett.84.2263.

Huang C, Wu D Q, Zhou X B et al. 2024. Competitive swarm optimizer with dynamic multi-competitions and convergence accelerator for large-scale optimization problems. *Applied Soft Computing*, **167**: 112252, https://doi.org/10.1016/J.ASOC.2024.112252.

Hussain T, Bakouch H S, Chesneau C. 2019. A new probability model with application to heavy-tailed hydrological data. *Environmental and Ecological Statistics*, **26**(2): 127-151, https://doi.org/10.1007/s10651-019-00422-7.

Lall U, Moon Y I, Bosworth K. 1993. Kernel flood frequency estimators: bandwidth selection and kernel choice.

Water Resources Research, **29**(4): 1003-1015, https://doi.org/10.1029/92WR02466.

Lei F H, Xie B T, Wang J Q. 2012. Uncertainty analysis of marine environment elements calculation. *The Ocean Engineering*, **30**(4): 109-117, https://doi.org/10.1175/JCLI-D-14-00470.1. (in Chinese with English abstract)

Li X Y, Zhao H M, Xu J J et al. 2024. APDPFL: anti-poisoning attack decentralized privacy enhanced federated learning scheme for flight operation data sharing. *IEEE Transactions on Wireless Communications*, **23**(12): 19098-19109, https://doi.org/10.1109/TWC.2024.3479149.

Liu D F, Wang D, Wang Y K et al. 2016. Entropy of hydrological systems under small samples: uncertainty and variability. *Journal of Hydrology*, **532**: 163-176, https://doi.org/10.1016/j.jhydrol.2015.11.019.

Liu G L, Cui K, Jiang S et al. 2021. A new empirical distribution for the design wave heights under the impact of typhoons. *Applied Ocean Research*, **111**: 102679, https://doi.org/10.1016/J.APOR.2021.102679.

Liu G L, Fang D, Song S C et al. 2023a. Calculation of design wave height and its application base on independent sample. *Journal of Marine Sciences*, **41**(3): 73-82, https://doi.org/10.3969/j.issn.1001-909X.2023.03.007. (in Chinese with English abstract)

Liu G L, Zhou X S, Kou Y et al. 2023b. Uncertainty analysis for the calculation of marine environmental design parameters in the South China Sea. *Journal of Oceanology and Limnology*, **41**(2): 427-443, https://doi.org/10.1007/S00343-022-2052-Y.

Long H Y, Chen T, Chen H Y et al. 2024. Principal space approximation ensemble discriminative marginalized least-squares regression for hyperspectral image classification. *Engineering Applications of Artificial Intelligence*, **133**: 108031, https://doi.org/10.1016/J.ENGAPPAI.2024.108031.

Marra F, Amponsah W, Papalexiou S M. 2023. Non-asymptotic Weibull tails explain the statistics of extreme daily precipitation. *Advances in Water Resources*, **173**: 104388, https://doi.org/10.1016/J.ADVWATRES.2023.104388.

Martins A L A, Liska G R, Beijo L A et al. 2020. Generalized Pareto distribution applied to the analysis of maximum rainfall events in Uruguaiana, RS, Brazil. *SN Applied Sciences*, **2**(9): 1479, https://doi.org/10.1007/s42452-020-03199-8.

Pan X, Rahman A, Haddad K et al. 2022. Peaks-over-threshold model in flood frequency analysis: a scoping review. *Stochastic Environmental Research and Risk Assessment*, **36**(9): 2419-2435, https://doi.org/10.1007/S00477-022-02174-6.

Shannon C E. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, **27**(3): 379-423, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Singh S K. 2019. Dynamic parameter estimation for hydrological model. *International Journal of Hydrology Science and Technology*, **9**(2): 124-136, https://doi.org/10.1504/IJHST.2019.098159.

Wang L P, Chen B Y, Chen C et al. 2016. Application of linear mean-square estimation in ocean engineering. *China Ocean Engineering*, **30**(1): 149-160, https://doi.org/10.1007/s13344-016-0007-9.

Zhuang C Q. 2005. Basis of Applied Mathematical Statistics. South China University of Technology Press, Guangzhou. p.222-305.