KEMENTERIAN SUMBER ASLI, ALAM SEKITAR
DAN PERUBAHAN IKLIM
Ministry of Natural Resources, Environment
And Climate Change

MALAYSIAN METEOROLOGICAL DEPARTMENT
MINISTRY OF NATURAL RESOURCES, ENVIRONMENT AND CLIMATE CHANGE

Technical Note No. 3/2022

Estimation of Maximum Wave Height at In-situ Wave Observation Sites in Malaysia

Yip Weng Sang and Nursalleh K Chang

TECHNICAL NOTE NO. 3/2022

Estimation of Maximum Wave Height at In-situ Wave Observation Sites in Malaysia

By
Yip Weng Sang and Nursalleh K Chang

Perpustakaan Negara Malaysia Data Pengkatalogan-dalam-Penerbitan

Published and printed by:
Jabatan Meteorologi Malaysia
Jalan Sultan
46667 Petaling Jaya
Selangor Darul Ehsan
Malaysia

Contents

Estimation of Maximum Wave Height at In-situ Wave Observation Sites in Malaysia

Yip Weng Sang and Nursalleh K. Chang

Abstract

The relationship between wave parameters is studied to find the most suitable method of estimating maximum wave height. Long-term in-situ wave observations in all seasons, and short-term in-situ wave observations during periods that coincides with strong winds and rough seas warnings issued by MET Malaysia, are analyzed. Correlation analysis of observations revealed that significant wave height observations are highly correlated with observed maximum wave height. On the other hand, the number of waves and wind speed is generally poorly correlated with observed maximum wave height. Kolmogorov-Smirnov significance test indicates that observed maximum wave height distribution does not follow either Rayleigh or Weibull distribution, at the 0.05 significance level. Estimation of maximum wave height using Rayleigh Distribution, regression, and neural networks are studied. The Rayleigh Distribution severely overestimates maximum wave height. On the contrary, regression and neural networks showed equal skill, and can estimate maximum wave height better. On average, the ratio of maximum wave height to significant wave height measured in shallow waters, where the depth is less than 20m, is 1.27 based on long-term, all-season observations. In deeper waters of depth 450m, the ratio is 1.50. At periods coinciding with rough seas and strong winds corresponding to thunderstorm warnings and marine warnings issued by MET Malaysia, the ratio increased to between 1.50 – 1.60. The accuracy of maximum wave height estimation by simply multiplying significant wave height by fixed 1.27 (depth less than 20m) or 1.50 otherwise, showed comparable mean absolute error, with regression or neural network.

1. Introduction

Maximum wave height (H_{max}) estimation is important in the safety design of oil platforms, ships, and other maritime structures. H_{max} is not a deterministic quantity, but a statistical quantity. Statistical distributions have been used to derive H_{max}.

Longuet-Higgins (1952) defined wave amplitude (a) as ½ the distance between the highest point of the wave to the lowest point of the previous wave. Wave trains with narrow frequency bandwidth that can be decomposed into numerous smaller, independent components are modelled. Based on this model and assumptions, the expected H_{max} is:

$$E(H_{max}) = 0.70621 \times H_s \times \left[\sqrt{(logN)} + \frac{\gamma}{2\sqrt{logN}}\right] - (1)$$

where H_{max} is the maximum wave amplitude, Hs is the significant wave height, N is number of waves equals to period of observation (Td) divided by mean wave period (Tmean). Meanwhile, the most probable H_{max} is:

$$\mu(H_{\max}) = 0.70621 \times \sqrt{logN} - (2)$$

This distribution is colloquially known as Rayleigh distribution, after Lord Rayleigh who modelled acoustic amplitudes from many random sound sources. However, Forristal (1978) reported that the Rayleigh distribution tends to overestimate height of the highest wave. Based on one hundred sixteen (116) hours of data in the Gulf of Mexico during hurricanes, Forristal (1978) fitted an empirical Weibull distribution:

$$E(x) = \exp\left(-\frac{x^{\alpha}}{\beta}\right) - (3)$$

$$x = \frac{H_o}{(m_o)^{1/2}} - (3a)$$

where H_o is wave height, m_o is mean-square of the wave profile, E(x) is the probability of wave height more than or equals to x. Coefficients α, β are calculated by linear-straight-line fitting. Based on Equation 3, the H_{max} is:

$$E(H_{\max}) = (\beta\theta_o)^{\frac{1}{\alpha}}\left(1 + \frac{\gamma}{\alpha\theta_0}\right) - (4)$$

where $E(x_{max})$is the expected H_{max}, $\theta_o = lnN$ where N is number of waves, $\gamma = 0.5772$ (Euler's constant). Forristal (1978) fitted the Empirical Weibull distribution (Equation 4) on 116 hours of wave data in the Guld of Mexico during hurricanes. A close fit was reported for the proposed Empirical Weibull distribution, and it performed better than the Rayleigh distribution. Forristal (1978) defined wave height as distance between trough and next crest.

Krogstad (1985) analyzed 3-hourly in-situ, wave data observations in the Norwegian Sea with records between 4-6 years. It was reported that the Rayleigh distribution failed to represent upper-tail wave heights, although it gave a reasonably well estimation of overall wave height. Wave parameters are estimated using zero up-crossing analysis. Krogstad (1985) attempted to fit the two-parameter Weibull distribution and observed that is suitable for wave height estimation in the upper tail of the distribution. The two-parameter Weibull distribution suggested is:

$$F(X) = \left[1 - \exp\left(-\frac{x^a}{\beta}\right)\right]^{N_o} \text{- (5)}$$

Where X is H_{max} normalized by $1/4$ significant wave height, F(X) is probability of X not exceeding x; α, β, N_o are constants solved by linear fitting, and N is number of waves that is duration of record divided by peak wave period. The expected ratio of H_{max} to H_s is:

$$E\left(\frac{H_{\max}}{H_s}\right) = 0.25(\beta logN)^{\frac{1}{\alpha}}\left[1 + \frac{0.5722}{\alpha logN}\right] - (6)$$

Muraleedharan (2007) studied wave data in the eastern Arabian seas during rough monsoon conditions for both deep and shallow depths. Zero-crossing wave analysis was used to compute H_s. 33,000 data over a period of 10 years were used. A modified Weibull distribution was proposed to estimate H_{max} and H_s. The modified distribution was reportedly more effective in simulating H_{max} distribution and other wave height parameters, compared to the standard Weibull and Rayleigh distributions. The equations proposed by Muraleedharan (2007) are detailed below:

$$P(h) = \exp\left(\alpha^b - \left(\frac{h}{a} + \alpha\right)^b\right) - (7)$$

where P(h) is probability of wave height exceeding h, α, a, β are coefficients to be fitted via Maximum Likelihood Estimation. The mean H_{max} is:

$$H_{\max} = a \cdot b^{-1} \cdot \Gamma\left(\frac{1}{b}\right) \cdot \left[n(1-\alpha^b) - \frac{n\cdot(n-1)}{2!\cdot 2^{\frac{1}{b}}} \times (1-2\alpha^b) + \cdots + \frac{(-1)^{r+1}}{n^{\frac{1}{b}}} \cdot (1 - n\cdot\alpha^b)\right] -$$
(8)

where Γ represents Gamma Function while ! is factorial. Most frequent maximum wave height is:

$$H_{mfm} = a\left[\left\{\frac{A \pm \sqrt{A^2 - 4\left(A - b\alpha^b + \alpha^b + 1\right)\alpha^b(b-1)}}{2(A - b\alpha^b + \alpha^b + 1)}\right\}^{\frac{1}{b}} - \alpha\right] - (9)$$

where $A = nb\alpha^b + b\alpha^b + nb - \alpha^b - 1$, n is the number of waves. Meanwhile, the return period (R_p) for H_{max} (h_L) is in **Equation 10**, where N is the number of daily maximum wave heights.

$$h_L = a \cdot \left[\left[\alpha^b - \ln\left\{ 1 - \left(1 - \frac{1}{R_p} \right)^{\frac{1}{N}} \right\} \right]^{\frac{1}{b}} - \alpha \right] \text{ - (10)}$$

On the other hand, Vandever et al. (2008) proposed the usage of spectral width parameter v to correct H_s, which is then used to calculate H_{max} using Rayleigh distribution. H_{max} is defined as maximum individual successive crest to trough distance. Spectral width parameter is defined by **Equation 11**:

$$\nu = \sqrt{\frac{m_o m_2}{m_1^2} - 1} \text{ – (11)}$$

where ν is the spectral width parameter known as the normalized radius of gyration. Ratio of $H_{1/3}$ to $m_o^{0.5}$ is plotted against v and linearly fitted to **Equation 12**:

$$H'_{1/3} = [\alpha - \beta\nu]\sqrt{m_o} \text{ – (12)}$$

where $H'_{1/3}$ is the bandwidth corrected significant wave height. Calculations of H_{max} using $H'_{1/3}$ showed less error with respect to observations taken from in-situ stations.

Kurtosis is a measure of how much data is in the tail or extreme end of the distribution. Janssen and Bidlot (2009) used kurtosis to estimate the H_{max} based on **Equation 13**:

$$< H_{\max} >= \sqrt{< z >} = \widehat{z_o} + \frac{\gamma}{2} + \frac{1}{2}\log\left[1 + C_4 \left\{ 2\widehat{z_o}(z_o - 1) - \gamma(1 - 2\widehat{z_o}) - \frac{1}{2}\left(\gamma^2 + \frac{\pi^2}{6} \right) \right\} \right]$$
$$\text{– (13)}$$

$$\widehat{z_o} = \frac{1}{2} log N \text{ – (13a)}$$

$$N = \frac{T_D}{T_p} \text{ – (13b)}$$

where $< H_{max} >$ is the expected maximum wave height, N is the number of waves, T_D is duration of record and T_P is the peak period, γ is the Euler's constant, and C_4 is kurtosis. Good agreement was reported between modelled and observed H_{max} in Canadian offshore buoys from 2006-2008 for 18-minute interval observations.

Feng (2014) also reported that H_{max} / H_s is linearly related to kurtosis as follows:

$$\frac{H_{max}}{H_s} = (0.49 \pm 0.01) * kurtosis \text{ – (14)}$$

Based on the relationship between H_{max} / H_s to number of waves N, Feng (2014) proposed **Equation 15** which is a modified Rayleigh distribution, to estimate H_{max}:

$$\frac{H_{max}}{H_s} = \frac{\sqrt{lnN}}{1.555} + \frac{1.7(H_s - 1.5)}{100} - (15)$$

It was reported that estimation of the largest waves has improved compared to conventional Rayleigh and Weibull distributions.

Zhuo and Sato (2015) analyzed two large typhoons that caused significant damage to concrete breakwater in Japan. Typhoon Wipha had lower wave height than Typhoon Man-yi but caused more damage. It was argued that H_s on its own could not fully account for extreme wave heights. Further analysis revealed that higher kurtosis tends to increase H_{max} to H_s ratio which leads to more extreme waves. Kurtosis is defined as

$$\mu_4 = \frac{1}{\eta_{rms}^4} \cdot \frac{1}{N} \Sigma_{i=1}^{N} (\eta_i - \bar{\eta})^4 - (16)$$

where η is surface elevation, and η_{rms} is root mean square elevation, N is the number of waves, and $\bar{\eta}$ is the mean surface elevation.

Chun and Suh (2019) proposed calculating H_{max} using the peaked-ness parameter Q_p , spectral wave period $T_{m-1,0}$, and significant wave height H_{m0} measured by wave spectral parameters. The least-squares method was used to fit H_{max} and derive **Equation 17**. H_{max} estimation agrees well with in-situ observation data from the East Sea, Japan, and southeast coast of Korea.

$$H_{\max} = 1.65 H_{m0} T_{m-1,0}^{-0.04} \left(Q_p - 1\right)^{0.03} - (17)$$

Barbariol et al. (2019) reported that H_{max} is influenced by waves steepness, kurtosis, and minimum of the autocovariance function which is one measure of spectral bandwidth. The following **equation 18** was proposed:

$$\bar{H}_{max,Na} = 2\sigma\sqrt{1 - \psi^*}(lnN)^{\frac{1}{2}} \left(1 + \frac{\gamma}{2lnN}\right) - (18)$$

where $\bar{H}_{max,Na}$is the expected maximum (crest-to-trough) wave height at single point with N waves, σ is sea surface elevation standard deviation, ψ^* is the minimum autocovariance function of the sea surface elevation $\psi(t)$, γ is the Euler-Mascheroni constant (0.5772). Values of estimated $\bar{H}_{max,Na}$have high correlation with OWS-P wave buoy observations in the North Pacific Ocean from June 2010 – December 2014.

Agrawal and Deo (2004) reported that regression and Neural Networks are useful tools for estimating H_{max} using H_s at an offshore observation site in India with data over a period of 1 year at 20 minutes interval.

This study aims to determine the most accurate method of determining H_{max}. The **Data Section** describes the in-situ observation data used in this study. It is followed by statistical analysis of H_s and H_{max} followed by correlation analysis of wave periods, wind speed, and H_s with respect to H_{max}. The regression methods and neural network used in estimating H_{max} are described next in the **Methodology Section**. Evaluation of the skill for each method in estimating H_{max} is contained in the R**esults and Analysis Section**. The **Conclusion Section** summarizes this study, the **Acknowledgement, and References Section** is the final section of this work.

2.0 Data

Four (4) Acoustic Doppler Current Profiler (ADCP) sensors were used in this study for in-situ wave observation. ADCP transmits pulses of acoustic waves, and uses Doppler shift to calculate sea current velocity, and measures wave height based on time taken to receive reflected acoustic waves. The Teledyne RDI Workhorse Sentinel ADCP 1,200kHZ system, with NEMO wave processing unit, was used in this study. Rorbaek and Anderson (2000) compared wave spectrum and height measurements made by a 1,200kHZ ADCP with a well-maintained S4 electromagnetic current meter in the Danish west coast. The period of measurements was from November 1999 to January 2000 when adverse weather conditions were observed. It was reported that overall statistical parameters observed by ADCP are reasonably consistent with the S4 current meter, and the ADCP can measure higher frequency waves than S4.

Meanwhile, a Wavescan buoy manufactured by Fugro OCEANOR was used to measure wave height, period, and direction. Wave parameters are calculated based on Heave, Roll, and Pitch observed by accelerometers, rate gyros, and magnetometers installed in the Wavescan buoy. The Wavescan buoy is located at the sea surface but anchored to the seabed. Schematics show that sea depth at the Wavescan buoy is 450m. Additionally, wind speed and direction are measured using propeller and vertical shafts with stainless steel, precision ball bearings. Current velocity is measured by an acoustic doppler sensor located in the Wavescan buoy but always in contact with the sea. In-situ observation sites are mapped in **Figure 1**. Station name, latitude, longitude, depth, location, interval of wave observation, number of non-missing datapoint, and period of observation are tabulated in **Table 1**.

Each station is given generic names. LT are long-term site observations with at least-1-year of data. ST are shorter term site observations of at most 5-days. Site observations LT1 to LT4 use ADCP sensors while site observation LT5 uses Wavescan buoy. Site observations LT1 to LT5 are long-term (4-9 years of hourly wave records) while site observations ST1 to ST5 are short-term (1-5 days of minute wave records) wave observations made during strong winds and rough seas. Sites LT1 to LT4 are shallow at less than 20m depth. On the other hand, ST4 is deepest (depth 1,550m), followed by ST5 and LT5 near 500m depth. Remaining sites ST1, ST2, and ST3 have sea depth between 60m to 80m. The LT1 observation site is the only station in the Straits of Malacca. Meanwhile, there are 4 observation sites, LT2, ST1, ST2 and

ST3 off the East Coast of Peninsula Malaysia in the South China Sea. Another 4 observation sites, LT3, LT5, ST4, and ST5 are located off the shores of Kota Kinabalu and Miri. LT4 is the only site in the Celebes Sea.

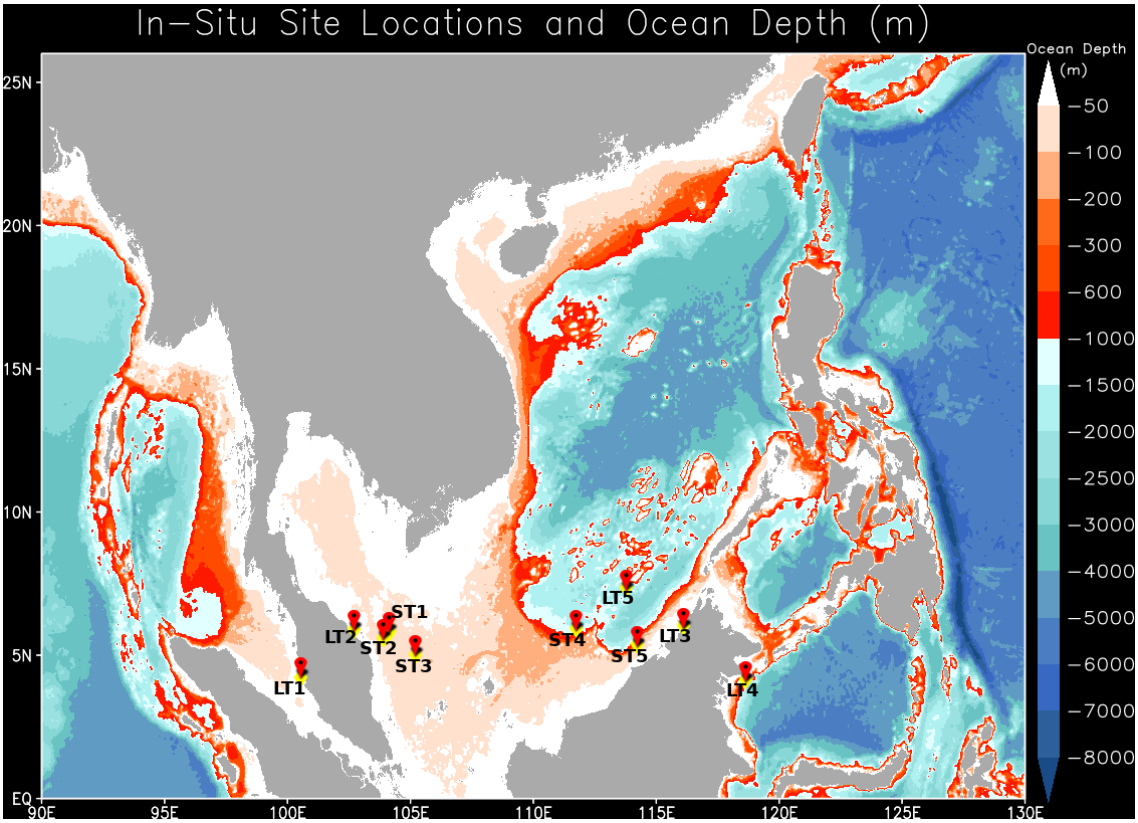


Figure 1. In-situ observation sites or station's locations

The depth of LT5 (**No. 5, Table 1**) has been given in a schematic diagram documented by MET Malaysia. The depth of the remaining observation sites is calculated based on their latitude and longitude coordinates from GEBCO at resolution 0.004º dataset, using minimum depth, at +/- 0.004º box centered at site latitude, longitude. The data points represent the number of non-missing wave observations.

This study considers the following wave data of significant wave height thereafter referred to as H_s, maximum wave height thereafter referred to as H_{max}, zero-crossing wave period (T_z), mean wave period (T_m), and peak wave period (T_P). The H_s is defined as the mean of the third highest waves measured, or four times the square root of the zeroth-order moment of the wave spectrum. It was first created by oceanographer Walter Munk as a statistical quantity that nearly matches visual wave observations. The World Meteorological Organization describes H_s as a common statistical description of the sea state. On the other hand, the H_{max} is the highest wave height observed within the period. It is an important parameter in designing marine structures, for example vertical breakwaters. The T_z is the record length divided by the number of wave-up-crosses (or down-crosses), the T_m is is the period associated with the mean frequency of the wave spectrum, while the T_P is the period associated with the most energetic wave at that specific point.

Table 1. Information about In-situ Observation Sites. Shallow Sites (depth<20m) in bold

No.	Site	Latitude (ºN)	Longitude (ºE)	Depth (m)	Location	Interval	Data Points	Duration
1	**LT1**	4.22475	100.53682	**18**	'Dolphin', Pulau Mentagor, Pangkor (Straits of Malacca)	Hourly	20,519	2009 / 02 / 15 – 2019 / 06 / 21
2	**LT2**	5.91025	102.70975	**19**	Bikon Jabatan Laut Pulau Perhentian (South China Sea)	Hourly	24,081	2009 / 03 / 06 – 2018 / 11 / 30
3	**LT3**	6.04060	116.10930	**13**	Jeti Institut Penyelidikan Marin Borneo, UMS Kota Kinabalu, Sabah	Hourly	16,510	2009 / 03 / 27 – 2019 / 04 /22
4	**LT4**	4.25393	118.63297	**14**	Seaventures Sipadan Resort, Mabul Semporna, Sabah (Celebes Sea)	Hourly	25,380	2009 / 03 / 31 – 2019 / 06 / 21
5	**LT5**	7.37500	113.79000	450	Pulau Layang-Layang (South China Sea)	Hourly	23,634	2013 / 01 / 01 – 2017 / 09 / 28
6	**ST1**	5.82903	104.15720	79	Petronas Site (South China Sea) off Terengganu	Minute	1,438	2022 / 06 / 29
7	**ST2**	5.60917	103.90610	66	Petronas Site (South China Sea) off Terengganu	Minute	1,440	2020 / 06 / 29
8	**ST3**	5.03019	105.20247	75	Petronas Site (South China Sea) off Terengganu	Minute	4,591	2022 / 02 / 23, 2022 / 02 / 28, 2022 / 04 / 03, 2022 / 04 / 08, 2022 / 04 / 27
9	**ST4**	5.93612	111.74350	1550	Petronas Site (South China Sea) off Miri, Sarawak	Minute	1,183	2022 / 04 / 08, 2022 / 04 / 27
10	**ST5**	5.38509	114.22094	441	Petronas Site (South China Sea) off Labuan	Minute	2,060	2022 / 06 / 28

2.1 Statistics of H_s and H_{max}

Out of 10 stations in this study, in-situ observations by the LT5 site showed the highest H_{max}. The most common (mode) $H_{max\ is}$ between 0.60 – 0.80m with median of 0.94m. The highest reported H_{max} measured by the LT5 site is 5.86m. This may happen because LT5 is in Layang-Layang Island, in the open seas with longer period of observation compared to other sites also in the open seas, such as ST4. In addition, site observations in the South China Sea may be exposed to intense northeasterly winds during the winter monsoon cold surges. They also risk exposure to tropical storms in the western North Pacific and South China Sea.

Meanwhile, in-situ observations at shallower waters (depth less than 20m), and located closer to land, are observed to have lower H_{max} and H_s. For example, Observations at LT1, LT2, LT3, and LT4 have the most common (mode) H_{max} of 0.20 – 0.40m, with median at most 0.30m. **Figures 2.** depicts the frequency histogram and probability density function (PDFs) of H_s and H_{max}. The PDFs show that the H_{max} and H_s are skewed the right (median>mode), especially for sites with long-term observations (5-10 years, they are LT1, LT2, LT3, LT4, and LT5). Right-skewed curves indicate potential for extreme wave heights.

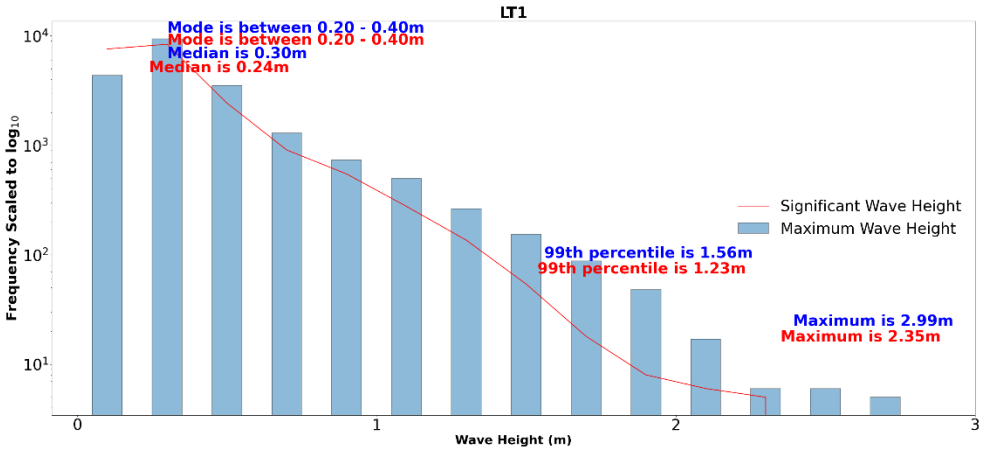


Figure 2a. Histogram of H_{max} (blue) and H_{sig} (red) for LT1 (20,519 data points)

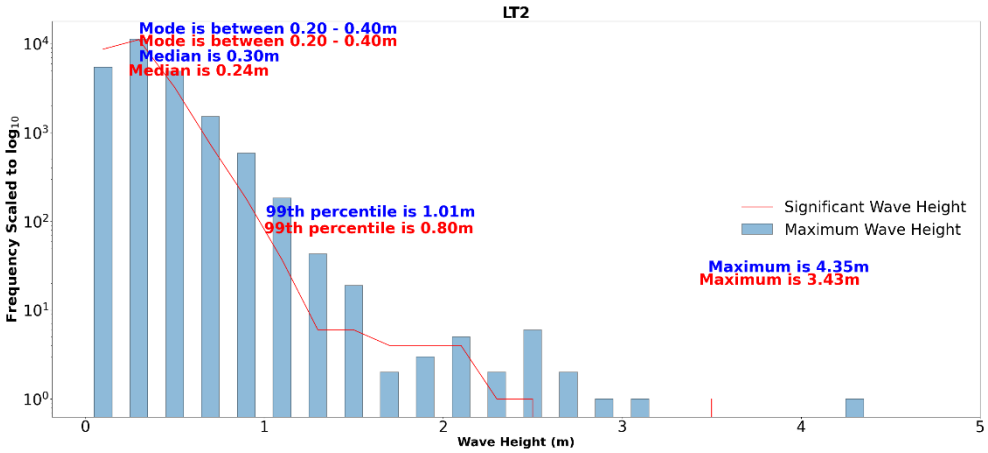


Figure 2b. Histogram of H_{max} (blue) and H_{sig} (red) for LT2 (24,081 data points).

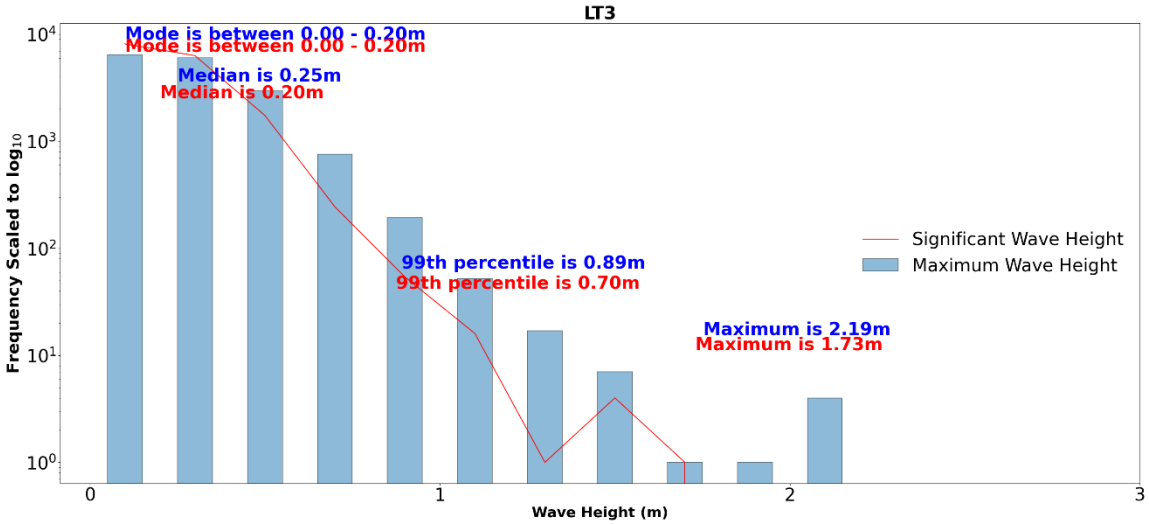


Figure 2c. Histogram of H_{max} (blue) and H_{sig} (red) for LT3 (16,510 data points).

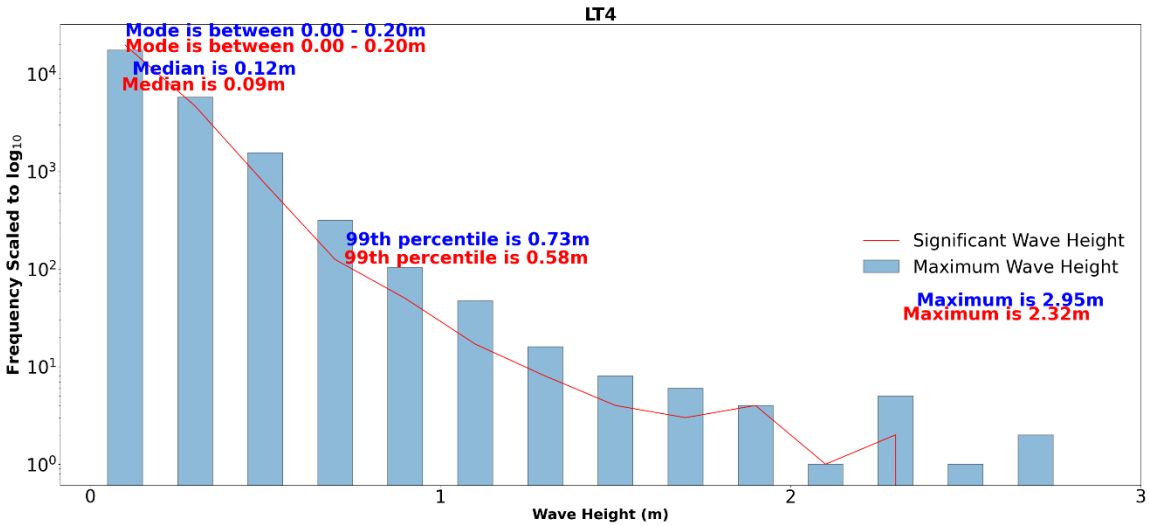


Figure 2d. Histogram of H_{max} (blue) and H_{sig} (red) for LT4 (25,380 data points).

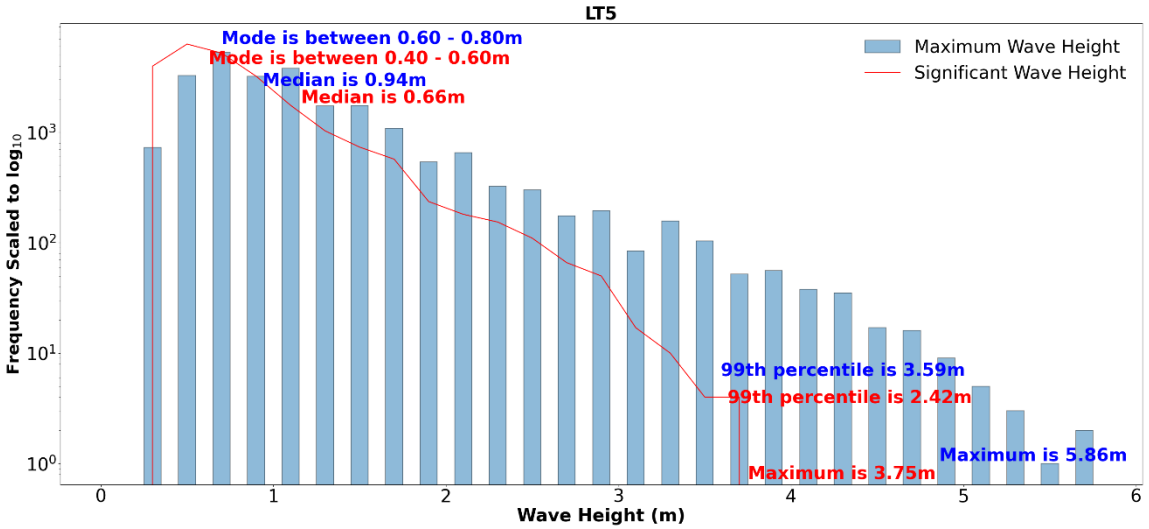


Figure 2e. Histogram of H_{max} (blue) and H_{sig} (red) for LT5 (23,634 data points).

Extreme value analysis is crucial for risk assessment. It is important to know the highest maximum wave height that will occur after N years. Goda (2000) suggested that the standard method of extreme value analysis is by fitting the best probability distribution function to the dataset. In this study, we attempted to fit H_{max} to Rayleigh and Weibull Distributions. Their cumulative distribution functions are:

$$Rayleight\ CPDF, F(x) = 1 - \exp\left[-\frac{x^2}{2\sigma^2}\right] - (19a)$$

$$Weibull\ CPDF, F(x) = 1 - \exp\left[-\left(\frac{x-B}{A}\right)^k\right] - (19b)$$

where F(x) is the probability of H_{max} being less than or equals to x. The coefficients are determined by the Maximum Likelihood Method, and the goodness-of-fit are evaluated using the Kolmogorov-Smirnov (KS) test. The D-statistic, which is the maximum absolute difference between cumulative probability density function (CPDF) between observed H_{max} and H_{max} generated by fitting the distribution is computed.

$$D_j = maximum\left|CPDF(H_{\max}) - F_j(H_{\max})\right| - (20)$$

where D_j is the D-statistic, CPDF(H_{max}) is the observed CPDF of H_{max}, and F_j(H_{max}) is the CPDF of H_{max} estimated by fitting distribution of type *j* to observed H_{max}.

The null hypothesis is observed H_{max} belongs to distribution *j*. D-statistic is compared with critical value taken from Masey (1952) for large sample sizes at 0.05 significance level:

$$D_{critical} = 1.36/\sqrt{N} - (21)$$

where N is the number of observations. The null hypothesis is rejected if the D-statistic exceeds $D_{critical}$. **Table 2** reveals the results of KS-test goodness of fit applied to H_{max}.

Table 2: KS-test for Rayleigh, Weibull, compared to critical D value

		Distribution		Critical Value for KS-test at 0.05 significance level
In-situ Site	**No. of Samples**	**Rayleigh**	**Weibull**	$D_{critical} = 1.36/\sqrt{N}$
LT1	20519	0.25099	0.12136	0.0095
LT2	24081	0.09234	0.05749	0.0088
LT3	16510	0.13344	0.05350	0.0106
LT4	25380	0.29055	0.11387	0.0085
LT5	23634	0.13492	0.11708	0.0088
ST1	1438	0.44565	0.12612	0.0359
ST2	1440	0.43220	0.14057	0.0358
ST3	4591	0.22664	0.18595	0.0201
ST4	1183	0.36388	0.33702	0.0395
ST5	2060	0.28443	0.13904	0.0300

Both Rayleigh and Weibull distribution D-statistics exceed critical D value. Therefore, distribution of H_{max} for each in-situ site cannot be described by either Weibull or Rayleigh distribution. As the underlying distribution of H_{max} is unknown, the return period is determined based on frequency analysis of observed H_{max} itself (**Equations 22**). **Table 3 (4)** summarizes the statistics of H_{max} (H_s) for each in-situ observation site. Rows with shaded columns indicate sites with too few observations to calculate return period.

$$Frequency\ of\ Exceeding\ Wave\ Height, H\ is\ F(H) = No.of\ Times\ Wave\ Heights\ h, exceeds\ H\ in\ the\ Entire\ Record \quad -(22a)$$

$$Probability\ of\ Exceeding\ H, P(H) = \frac{F(H)}{No.of\ Observations\ in\ the\ Entire\ Record} \quad -(22b)$$

$$Return\ Period\ of\ Exceeding\ H, T(H) = \frac{1}{P(H)} \quad -(22c)$$

Table 3: Significant Wave Height Statistics of All In-Situ Observation Sites

		Significant Wave Height (m)				
No.	**Station Name**	**Mode**	**Median**	**99th Percentile**	**Maximum**	**14-17 Day Return Period Value**
1	**LT1**	0.20 – 0.40	0.24	1.44	2.36	1.97
2	**LT2**	0.20 – 0.40	0.24	0.80	3.43	0.99
3	**LT3**	0.00 – 0.20	0.20	0.70	1.73	0.88
4	**LT4**	0.00 – 0.20	0.09	0.58	2.32	0.85
5	**LT5**	0.40 – 0.60	0.66	2.42	3.75	2.85
6	**ST1**	0.80 – 1.00	0.96	1.34	1.46	
7	**ST2**	0.80 – 1.00	0.85	1.20	1.26	
8	**ST3**	1.60 – 1.80	1.62	2.50	2.63	
9	**ST4**	1.40 – 1.60	1.46	1.71	1.84	
10	**ST5**	0.40 – 0.60	0.50	0.94	1.05	

Table 4: Maximum Wave Height Statistics of All In-Situ Observation Sites

No.	Station Name	Maximum Wave Height (m)				
		Mode	Median	99th Percentile	Maximum	14-17 Day Return Period Value
1	LT1	0.20 – 0.40	0.30	1.83	2.99	2.50
2	LT2	0.20 – 0.40	0.30	1.01	4.35	1.26
3	LT3	0.00 – 0.20	0.25	0.89	2.19	1.11
4	LT4	0.00 – 0.20	0.12	0.73	2.95	1.08
5	LT5	0.60 – 0.80	0.94	3.59	5.86	4.30
6	ST1	1.40 – 1.60	1.51	2.27	2.38	
7	ST2	1.20 – 1.40	1.31	2.23	2.23	
8	ST3	2.40 – 2.60	2.49	3.90	4.32	
9	ST4	2.20 – 2.40	2.27	3.06	3.16	
10	ST5	0.60 – 0.80	0.75	1.54	1.66	

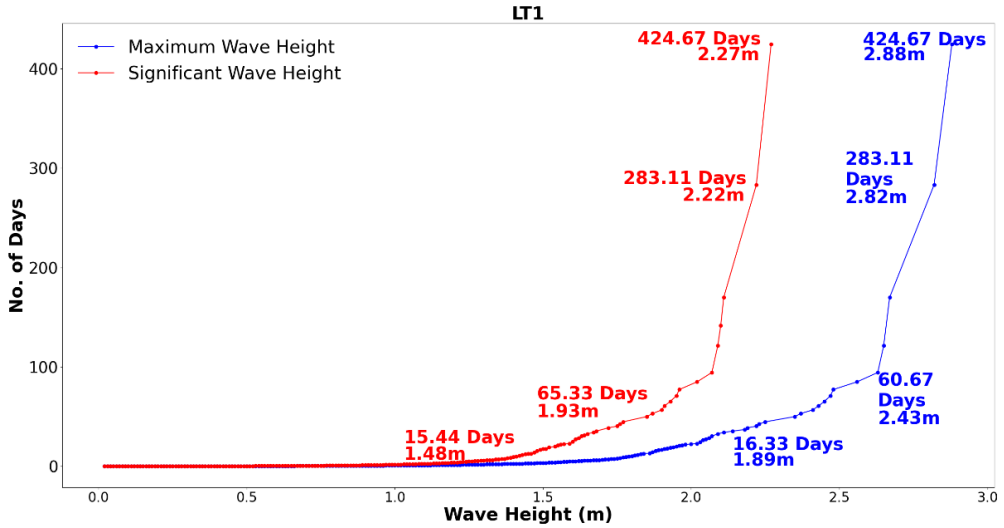


Figure 3a. Observed return period curve for LT1

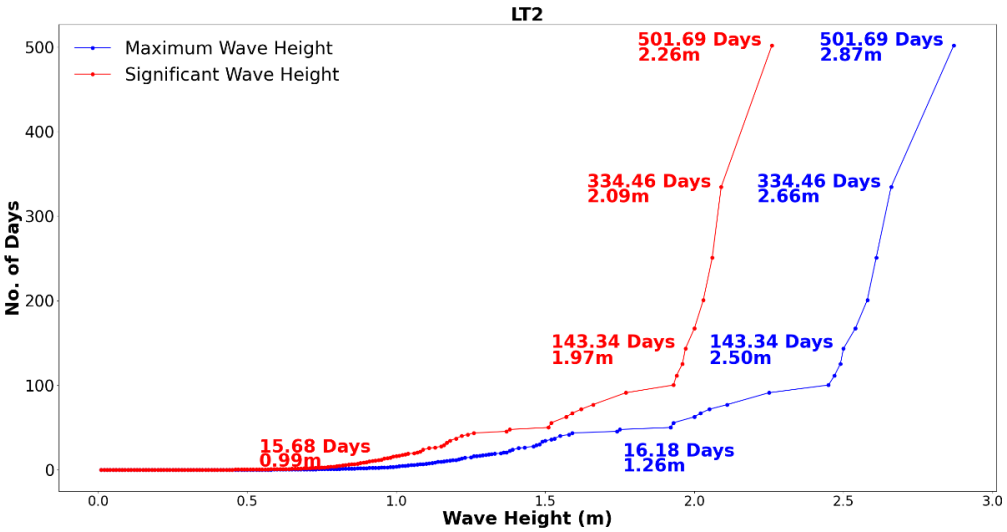


Figure 3b. Observed return period curve for LT2

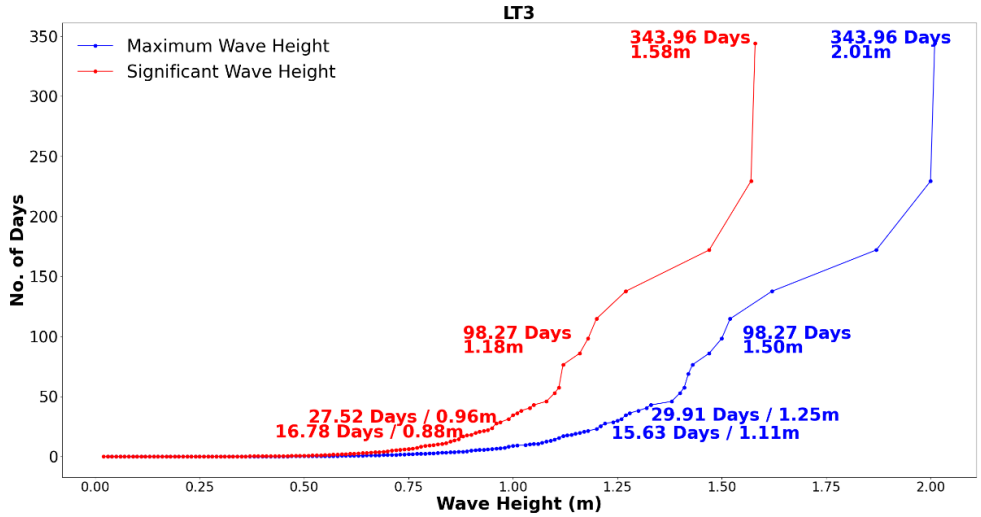


Figure 3c. Observed return period curve for LT3

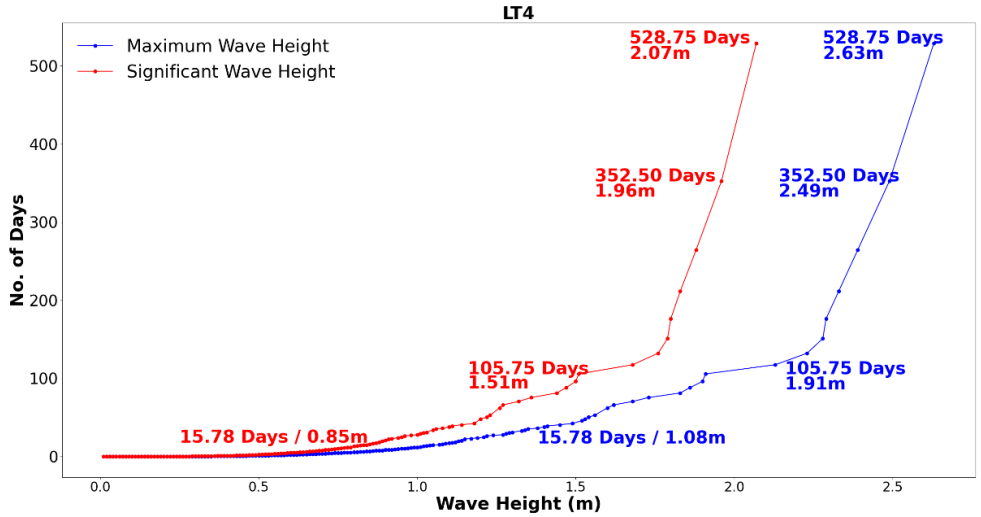


Figure 3d. Observed return period curve for LT4

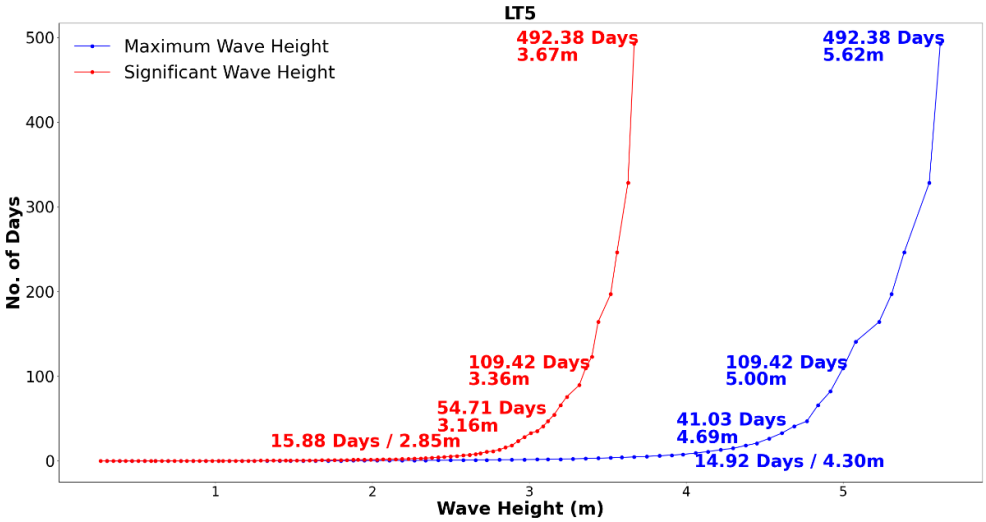


Figure 3e. Observed return period curve for LT5

2.2 Correlation Analysis with Respect to H_{max}

The estimation of H_{max} is the primary goal of this study. Correlation analysis is performed to determine the most accurate predictor of H_{max}. The coefficient of determination or R^2 is used to evaluate the goodness of each predictor. In this study, the predictors considered are as follows: number of waves (N), wind speed (m/s), and H_s.

The number of waves is defined as duration of wave of observation divided by the mean wave period, T_m given in Equation 22:

$$Number\ of\ waves, N = \frac{Duration\ of\ Wave\ Observation\ (s)}{Mean\ Wave\ Period, T_m} - (22)$$

Previous studies reported that H_{max} is a function of H_s and N (Feng et al., 2014, Goda, 2000, and Forristall, 1978). The Rayleigh distribution has been reported to give a good approximation of the distribution of individual wave heights (Goda, 2000). Most probable value, arithmetic mean value, and probability of exceeding μ, of H_{max} are given in Equations 23, 24, and 25. They are based on the Rayleigh distribution.

$$\left(\frac{H_{max}}{H_s}\right)_{mode} = 0.706\sqrt{lnN} \text{ - (23)}$$

$$\left(\frac{H_{max}}{H_s}\right)_{mean} = 0.706\left[\sqrt{\ln N} + \frac{0.5772}{2\sqrt{\ln N}}\right] \text{ - (24)}$$

$$\left(\frac{H_{max}}{H_s}\right)_{\mu} = 0.706\sqrt{\ln\left[\frac{N}{\ln\frac{1}{(1-\mu)}}\right]} - (25)$$

Fetch is an area over the ocean where wind blows in a constant direction. Stronger winds mean larger fetch which generates higher waves. In this study, we attempt to study the correlation between wind speed and H_{max}. It is assumed that local winds measured in-situ approximates wind speeds over a larger area of the ocean.

The coefficient of determination, R^2 measures the strength of linear relationship between two variables. R^2 of 1 shows perfect linear fit between the variables while R^2 of zero shows that they are not linearly correlated. R^2 score of less than 0.5 shows that the mean value of y is a better predictor than predictor x. R^2 score of 0.5 shows that predictor x is just as good as the mean value of y in predicting y. R^2 is depicted in **Equation 26**.

$$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} - (26)$$

$$y - Observed\ value\ (predictand)$$

$$\hat{y} - Linear\ Regression\ (x)\ where\ x\ is\ the\ predictor$$

$$\bar{y} - mean\ value\ of\ y$$

Scatterplot analysis between H_{max} against wind speed (W_s) is in **Figures 4**. The R^2 score between H_{max} and W_s is summarized in **Table 5**. Based on **Figures 4** and **Table 5**, W_s is a poor predictor of H_{max} with R^2 score less than 0.50. Meanwhile, scatterplot analysis between H_{max} and N are shown in **Figure 5.** R^2 score between H_{max} and N is depicted in **Table 6**. Analysis revealed in-situ observations in ST1, ST2, ST3, ST4, and ST5 during strong wind and rough seas warnings issued by MET Malaysia, have high R^2 score exceeding 0.5. Refer to **Table 7 / Figure 6** for warning category and dates. However, LT1, LT2, LT3, LT4, and LT5 site observations over long-term for all seasons revealed low R^2 score. This may indicate N is a parameter useful for estimating H_{max} only during strong wind and rough seas. Analysis between H_{max} and H_s (**Figure 7**), and R^2 score (**Table 8**) indicated H_s has strong linear relationship to H_{max}. Nearly all in-situ observations exceed 0.95 R^2 except ST3 (0.81) and ST5 (0.72). As a result, this study will mainly focus on using H_s to estimate H_{max}.

Table 5. R^2 Score between H_{max} and Wind Speed in decreasing R^2 order

In-Situ Wave Observation Site	R^2 Score
ST1	0.45
ST4	0.28
ST2	0.27
LT5	0.19
ST5	0.11
ST3	0.05

Table 6. R^2 Score between H_{max} and Number of Waves (N)

In-Situ Wave Observation Site	R^2 Score
LT1	0.07
LT2	0.01
LT3	0.17
LT4	0.26
LT5	0.15
ST1	0.53
ST2	0.54
ST3	0.81
ST4	0.91
ST5	0.72

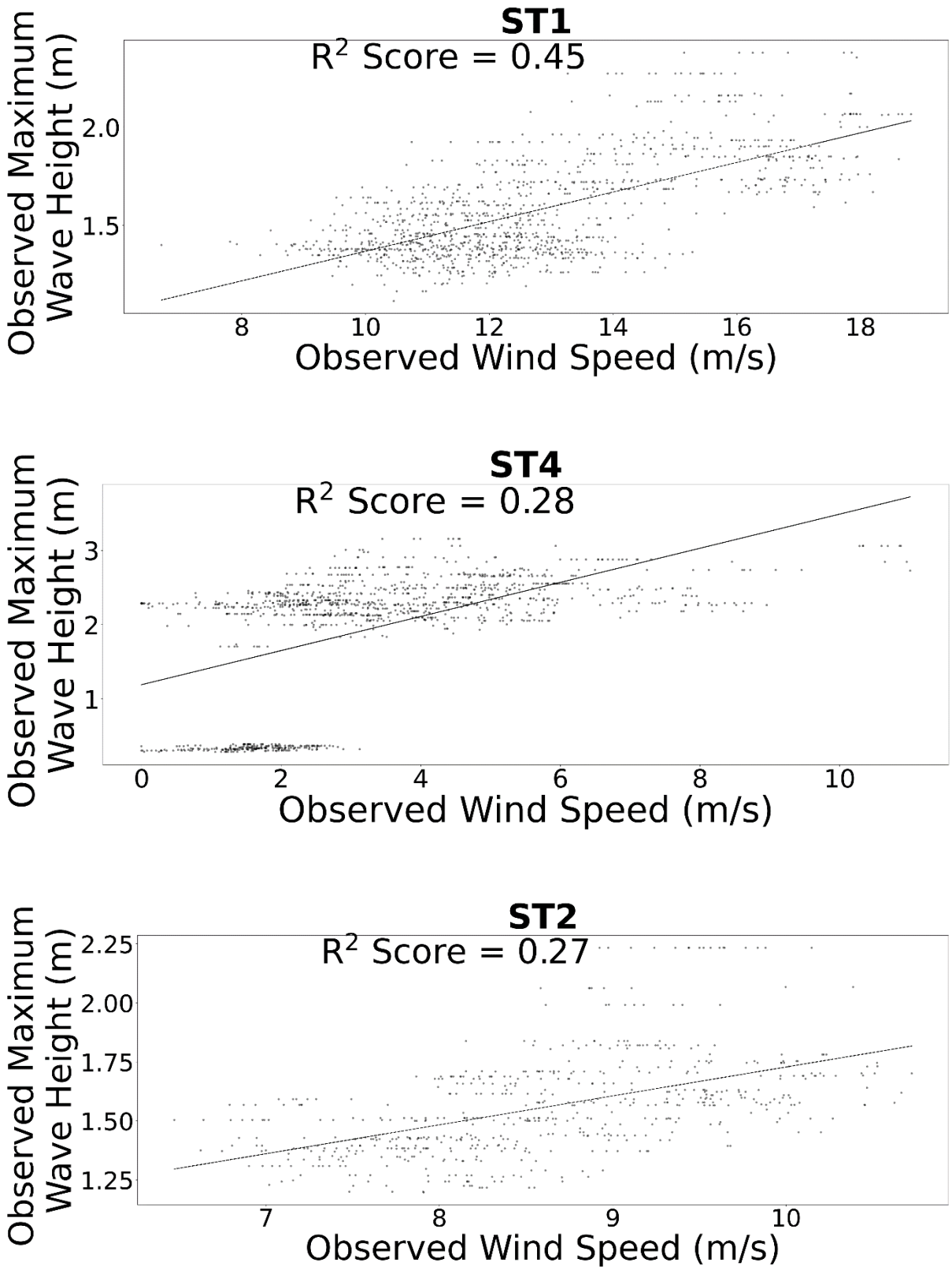


Figure 4a. Scatterplots of H_{max} vs. W_s for ST1 (top), ST4 (middle), and ST2 (bottom) in decreasing order of R^2.

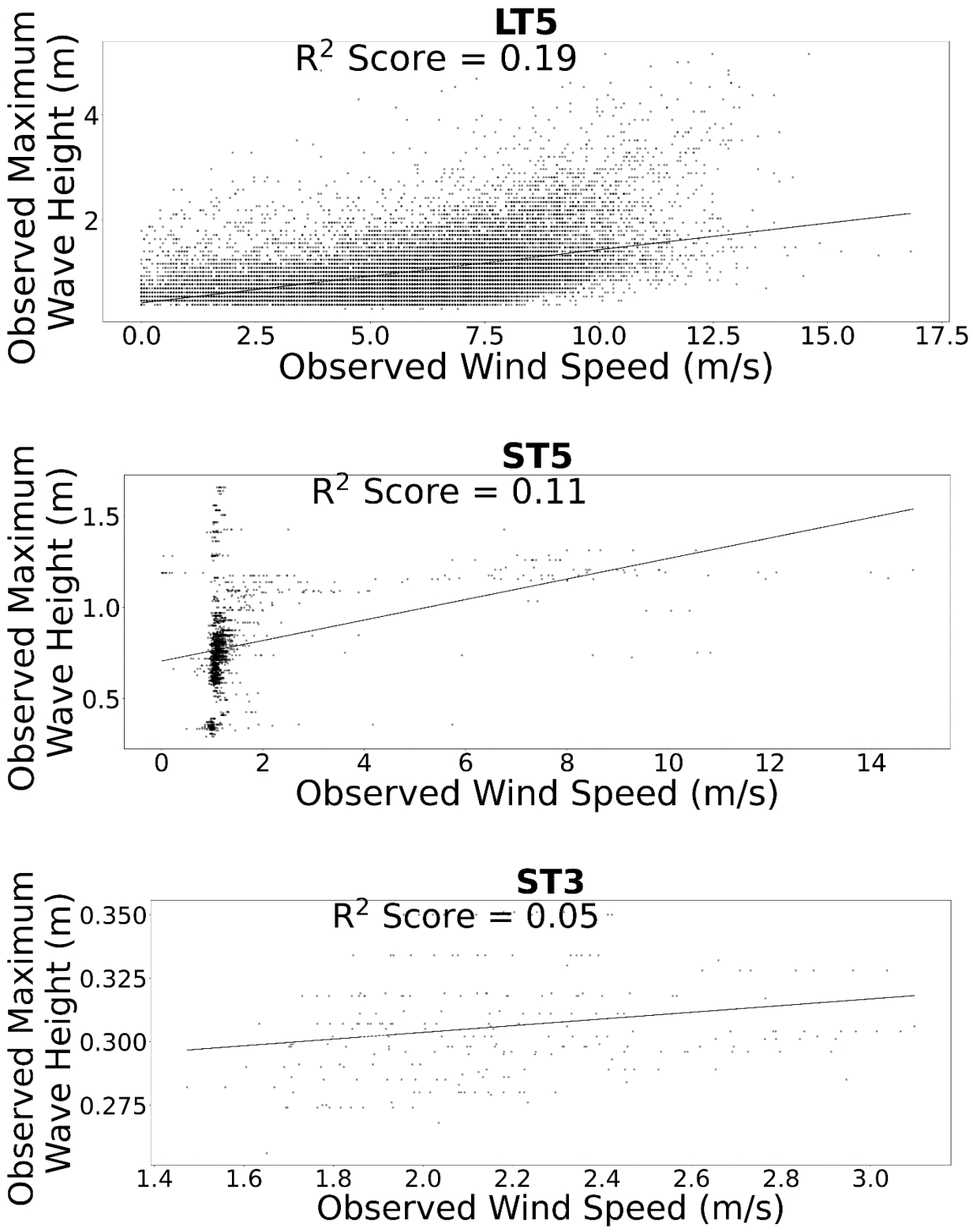
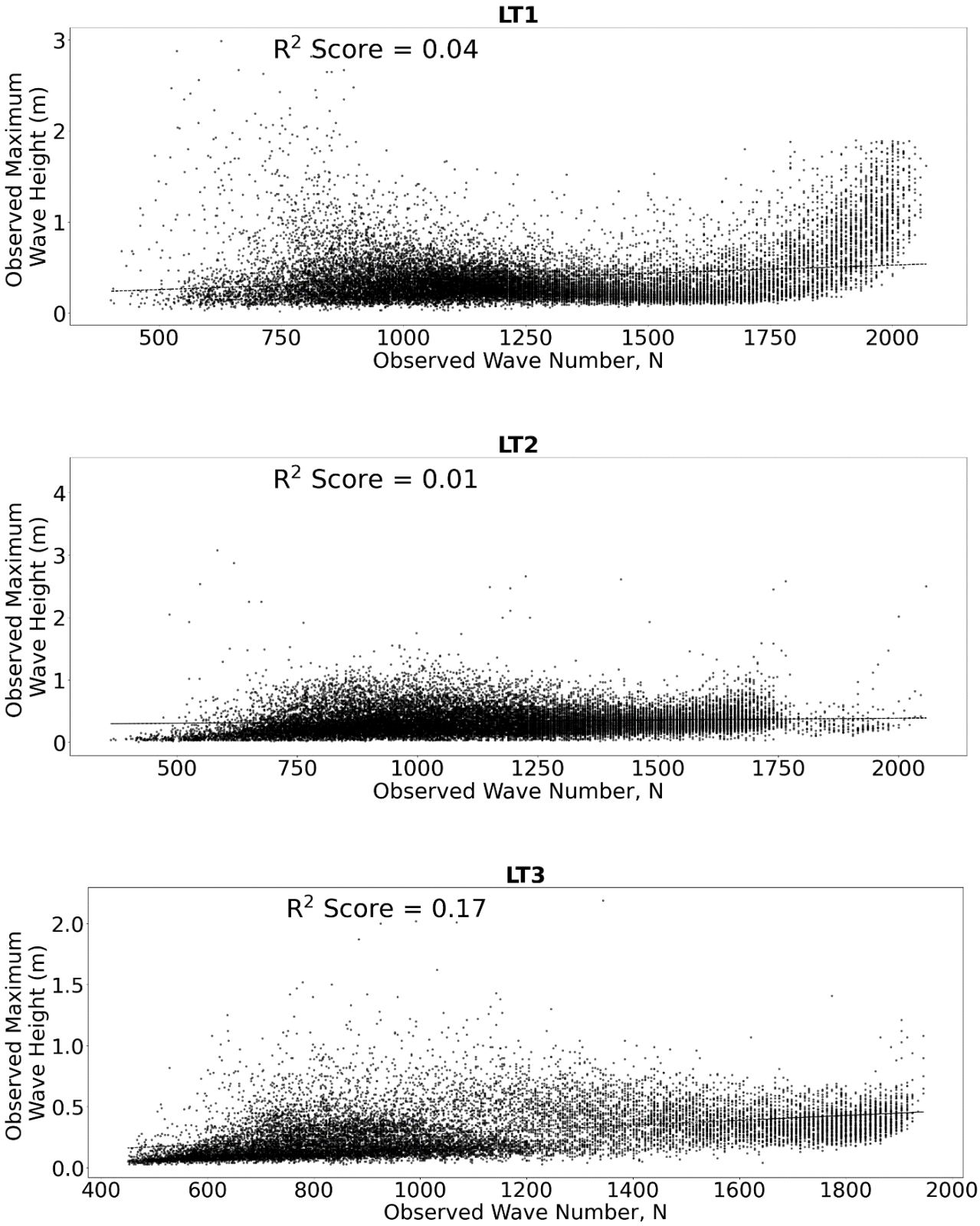
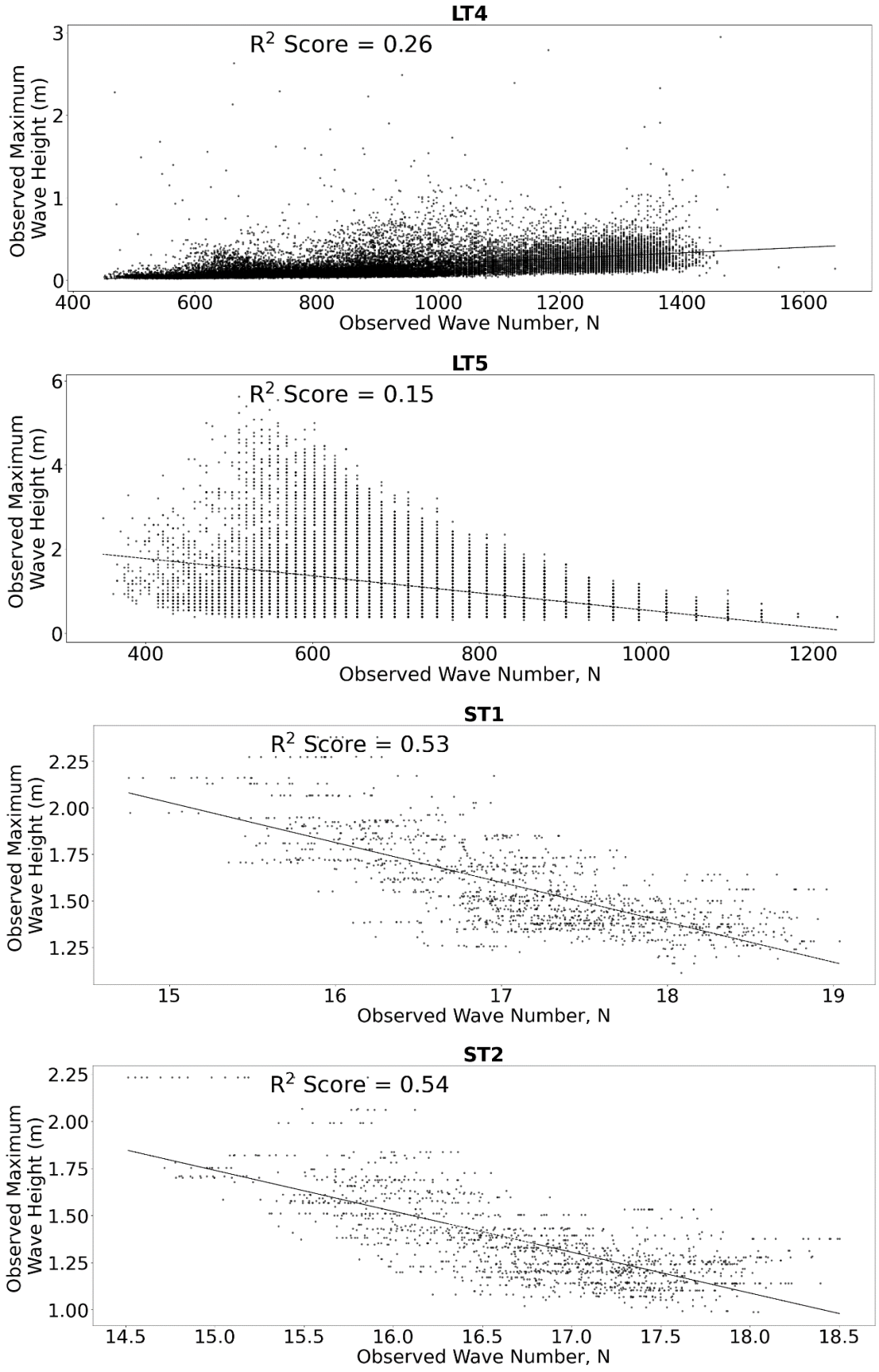


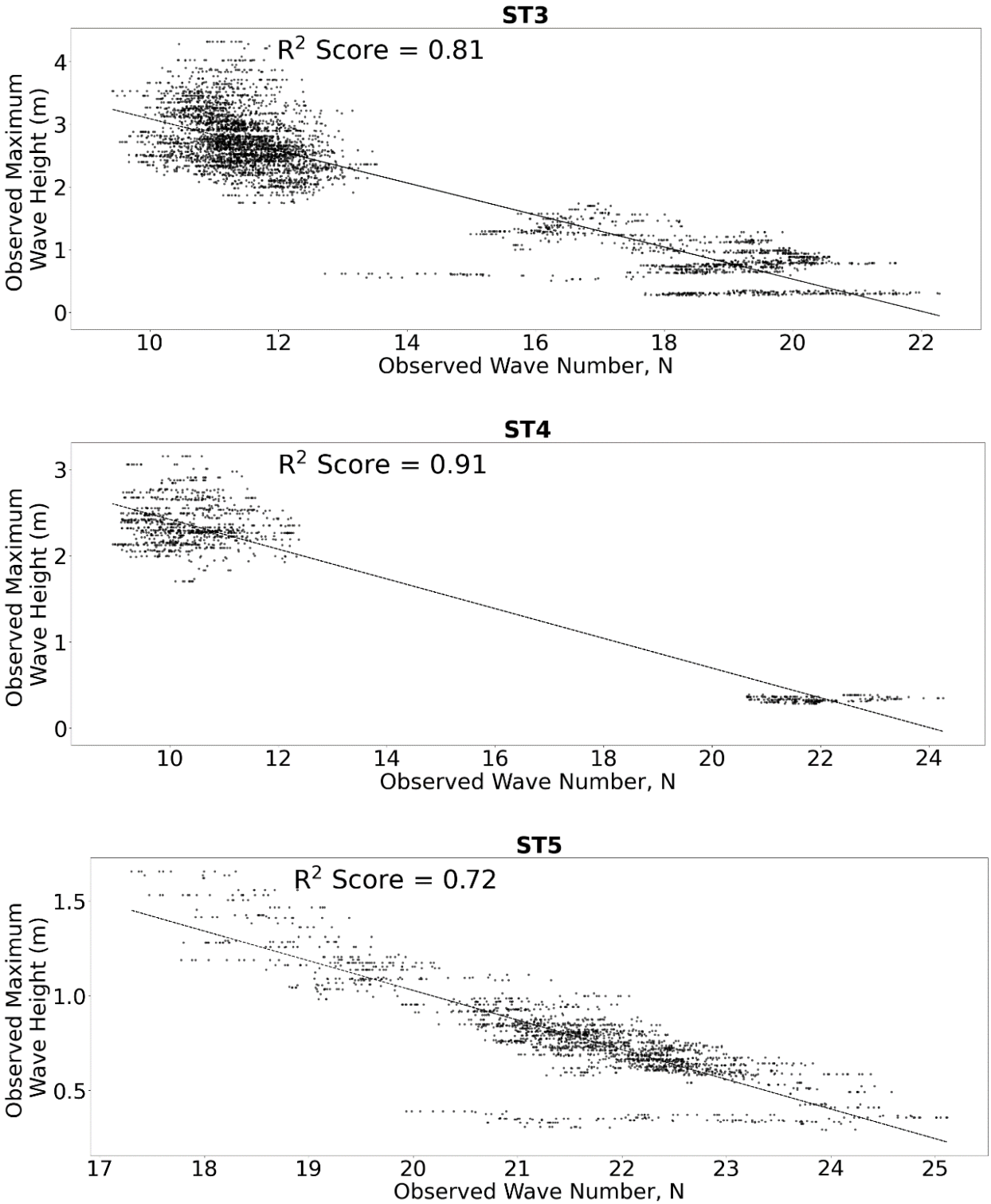
Figure 4b. Scatterplots of H_{max} vs. W_s for LT5 (top), ST5 (middle), and ST3 (bottom) in decreasing order of R^2.



Figures 5a. H_{max} vs. N for LT1 (top), LT2 (middle), and LT3 (bottom).



Figures 5b. H_{max} vs. N for LT4, LT5, ST1, and ST2 (top to bottom).



Figures 5c. H_{max} vs. N for ST3, ST4 and ST5 (top to bottom).

Table 7a. Site, location, date, and selected warnings (MET Malaysia)

No.	Site	Location	Date of issue corresponding to the data	Category of Warning	Location of Warning (Applicable to or close to Site)
1	**ST1**	South China Sea (off coast of Terengganu)	2022-06-29	2	Northeastern part of Condore, northern part of Reef North, Layang-Layang, Palawan
2	**ST2**		2020-06-29	Thunderstorms	Waters off Perlis, Kedah, Eastern Johore, Pahang, Eastern Sabah, and Lahad Datu
3	**ST3**		2022-02-23	3	Southeastern part of Samui, Tioman, Condore, northern part of Bunguran, Reef North, Layang-Layang
			2022-02-28	2	Condore, Reef North, Layang-Layang
			2022-04-03	1	Condore, northeastern and southwestern parts of Reef North, Layang-Layang
			2022-04-08	2	Western part of Condore, Bunguran, eastern part of Reef North, Reef South, Layang-Layang, Labuan, Palawan
			2022-04-27	Thunderstorm	Waters off Selangor, Johor, Pahang, Sarawak, Sabah

Table 7a. Site, location, date, and selected warnings (MET Malaysia)

No.	Site	Location	Date of issue corresponding to the data	Category of Warning	Location of Warning (Applicable to or close to Site)
4	**ST4**	South China Sea (off Miri, Sarawak)	2022-04-08	2	Waters off Sarawak, western Sabah, and Labuan
			2022-04-27	Thunderstorms	Waters off Selangor, Johor, Pahang, Sarawak, Sabah
5	**ST5**	South China Sea (off Labuan)	2022-06-28	Thunderstorms	Eastern part of Phuket, Straits of Malacca, northwestern part of Samui, southeastern part of Bunguran, Reef North, Labuan, Palawan, Sulu

Table 7b. Category of Warning, Wind speeds, and Wave heights (MET Malaysia)

Category of Warning	Wind speeds (kmph)	Wave heights (m)
3	Exceeding 60	Exceeding 4.5
2	50 – 60	3.5 – 4.5
1	40 – 50	2.5 – 3.5
Thunderstorms for Shipping	Up to 50	Up to 3.5m

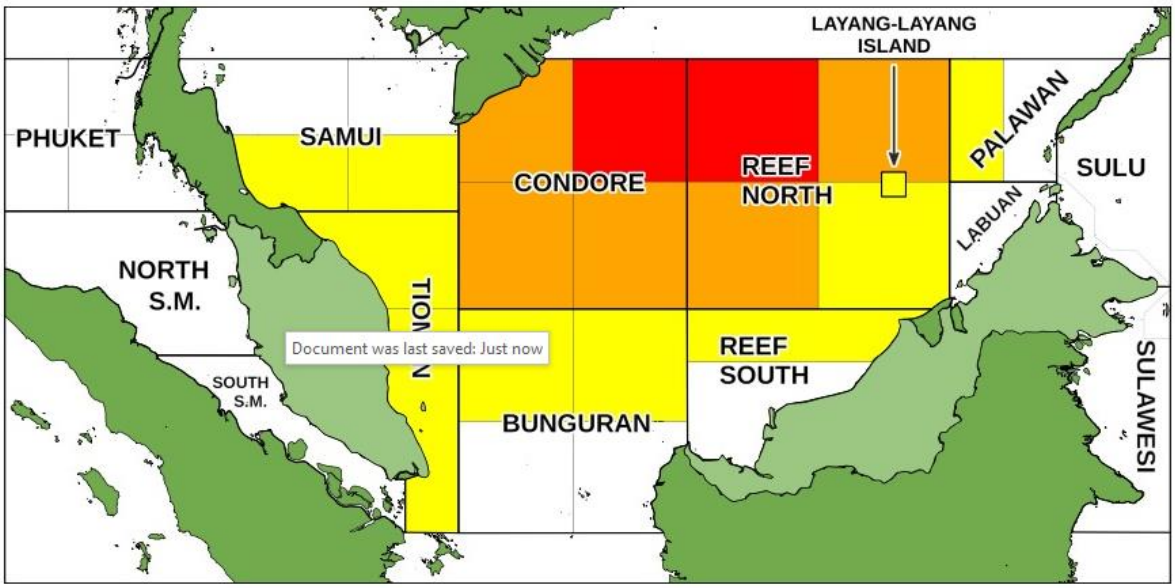


Figure 6. Location of warnings under the responsibility of MET Malaysia.

Table 8. R^2 Score between H_{max} and H_s

In-Situ Wave Observation Site	R^2 Score
LT1	1.00
LT2	1.00
LT3	1.00
LT4	1.00
LT5	0.96
ST1	0.81
ST2	0.78
ST3	0.96
ST4	0.95
ST5	0.95

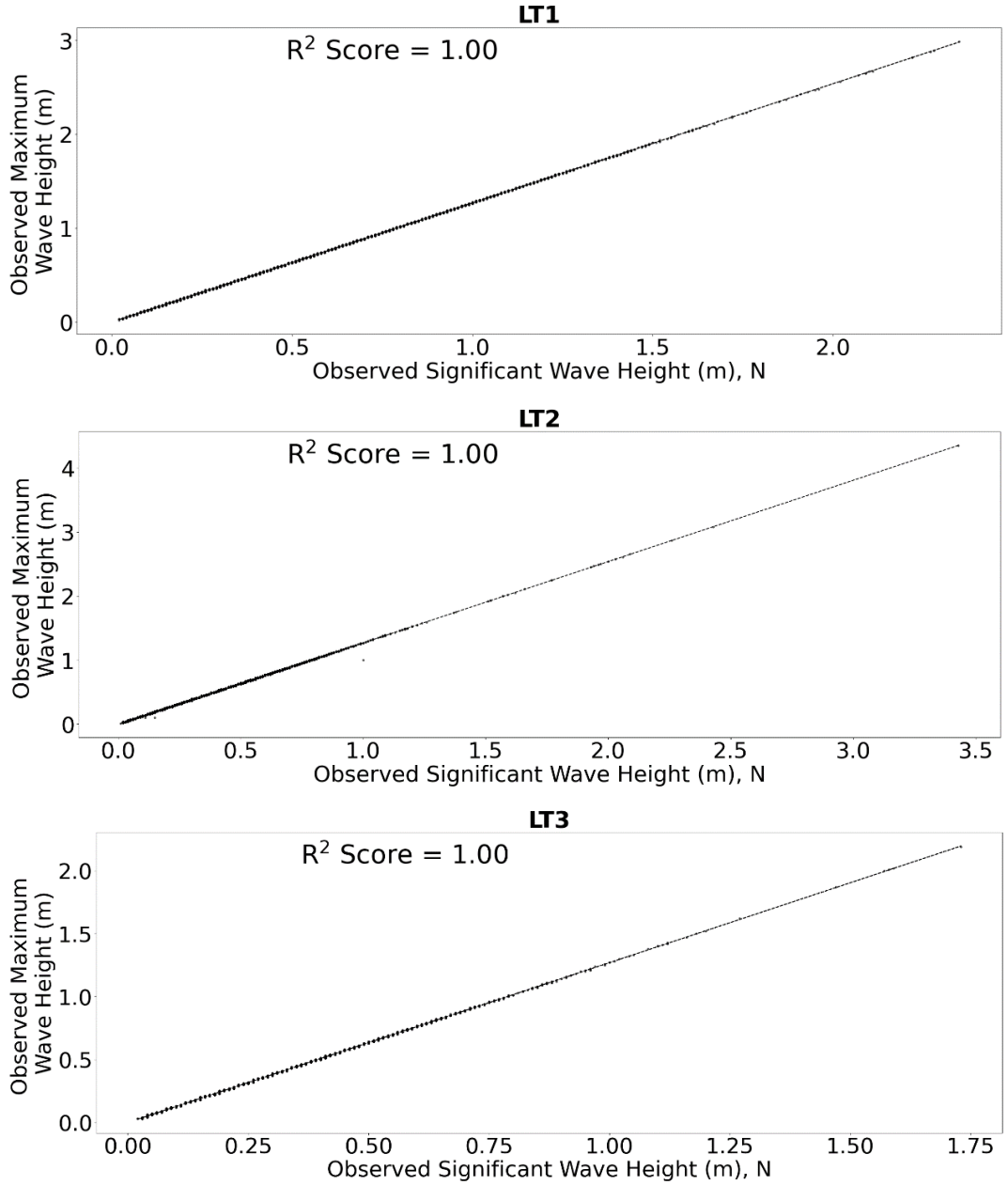


Figure 7a. Scatterplots, H_{max} vs. H_s, site observations LT1, LT2, and LT3 (top to bottom).

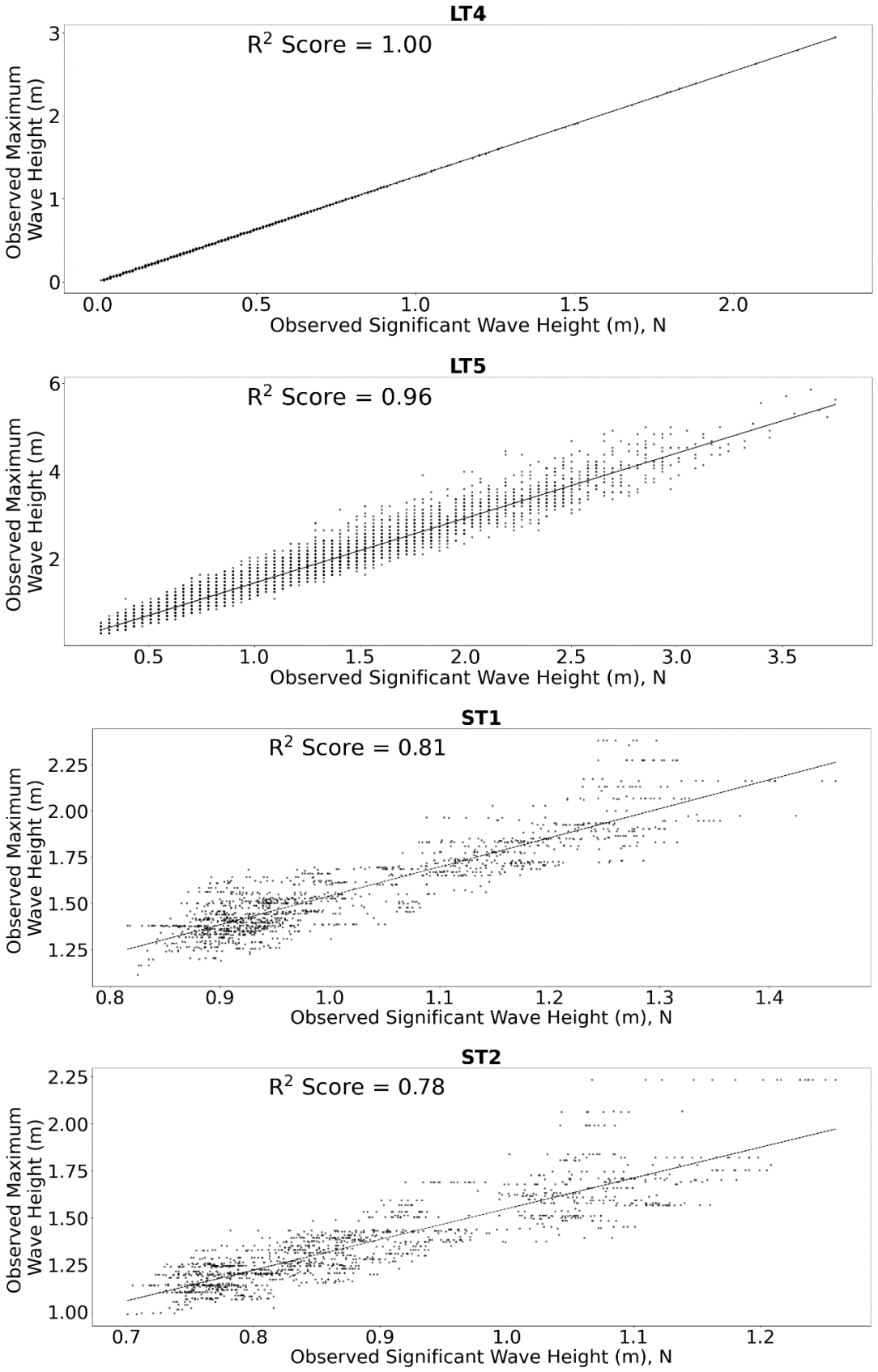
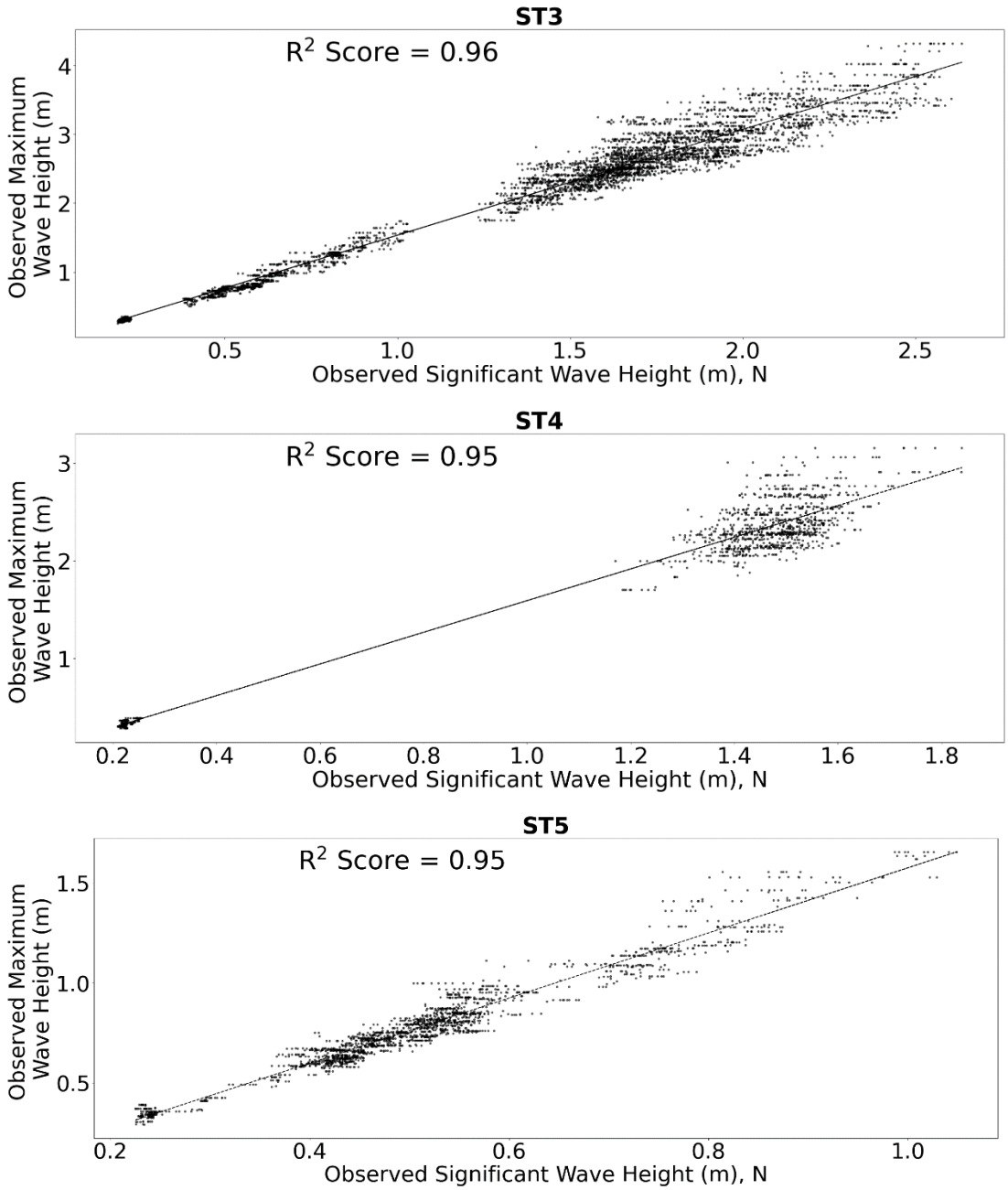


Figure 7b. Scatterplots, H_{max} vs. H_s, site observations LT4, LT5, ST1 and ST2 (top to bottom).



Figures 7c. Scatterplots, H_{max} vs. H_s, site observations ST3, ST4, and ST5 (top to bottom).

3.0 Methodology

3.1 Ratio of H_{max} to H_s

H_{max} is non-deterministic. Estimations of H_{max} are normally within 1.6 to 2.0 of H_s.

$$H_{max} = (1.6 \sim 2.0) H_s - (27)$$

Measurements of H_{max} to H_s ratio are performed for each site and the results are depicted in **Table 9** and **Figures 8**. Using long-term wave observations, sites in

shallow waters (depths < 20m) and close to the shore (LT1, LT2, LT3, and LT4) have ratios between 1.26 – 1.27. On the other hand, long-term wave observations for deeper depths (450m) in the open seas, LT5 have ratio of 1.44. However, for episodes of strong winds and rough seas (**Tables 7**) the ratios tend to 1.52 – 1.58 (ST1, ST2, ST3, ST4, and ST5). Histogram analysis (**Figures 9**) reveals that sites at deeper waters (LT5, ST1, ST2, ST3, and ST5) have ratios between 1.40-1.60, while sites in shallower waters (depth <20m) such as LT1, LT2, LT3, and LT4 have ratios 1.20 – 1.40.

Figure 10 shows the CPDF of H_{max}/H_s ratio. The safest ratio is suggested to be 1.8, with zero (0) probability of underestimating the H_{max}. However, ratio of 1.50 showed least overall error in estimating H_{max}, but it may underestimate H_{max} in rough seas and strong wind conditions. This is demonstrated by the average ratio for ST1, ST2, ST3, ST4 and ST5 exceeding 1.50 during strong winds and rough seas condition.

Table 9. Average Ratio of Hmax to Hs

In-Situ Site	Ratio of Maximum Wave Height (H_{max}) to Significant Wave Height (H_s)
LT1	1.27
LT2	1.27
LT3	1.27
LT4	1.26
LT5	1.44
ST1	1.54
ST2	1.54
ST3	1.53
ST4	1.58
ST5	1.52

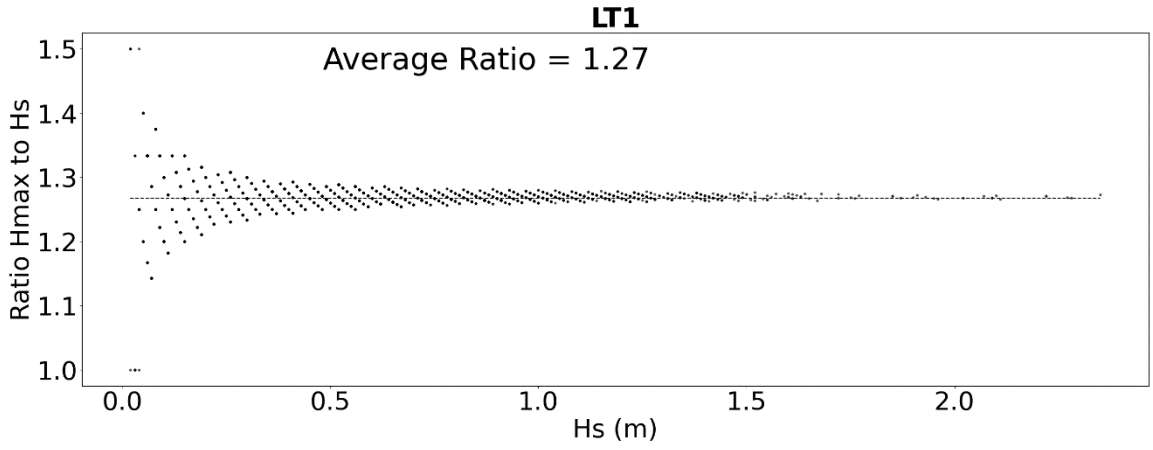


Figure 8a. Ratio H_{max} to H_s vs. H_s in LT1.

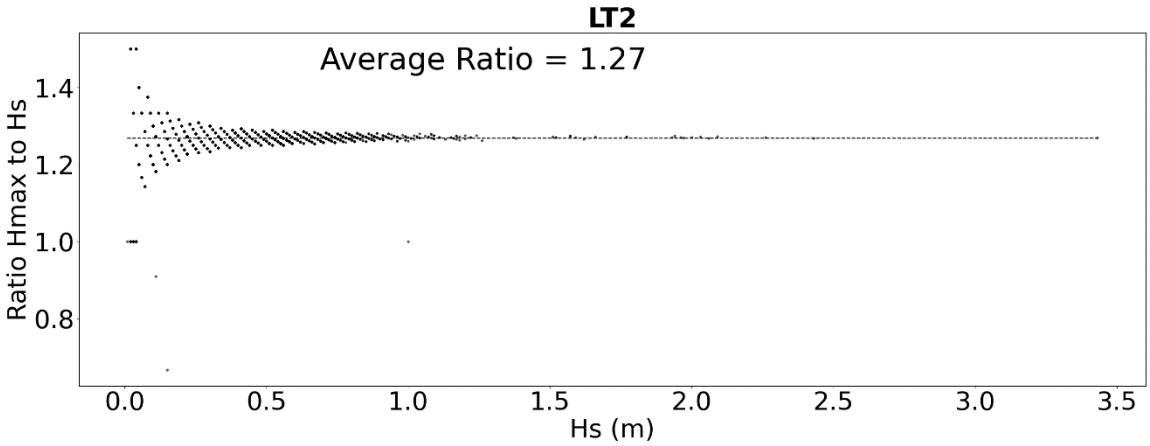


Figure 8b. Ratio H_{max} to H_s vs. H_s in LT2.

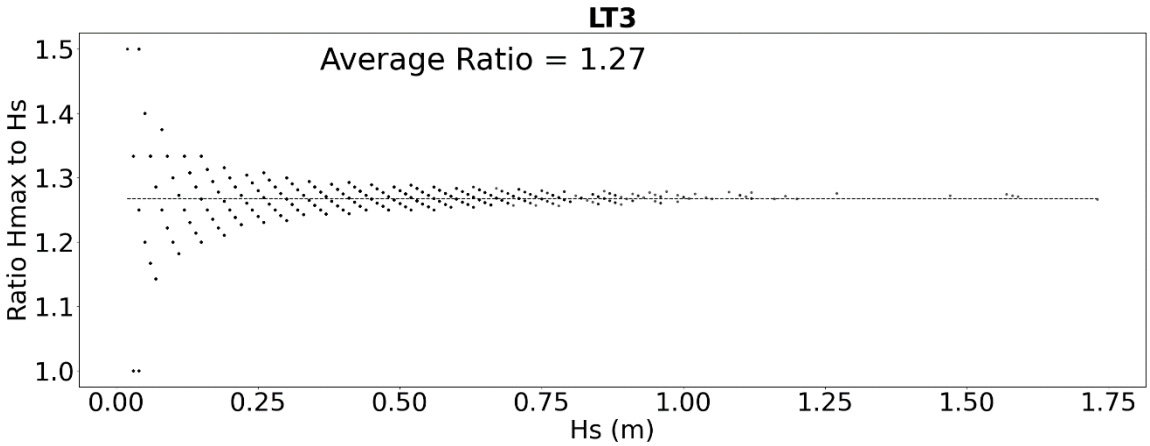


Figure 8c. Ratio H_{max} to H_s vs. H_s in LT3.

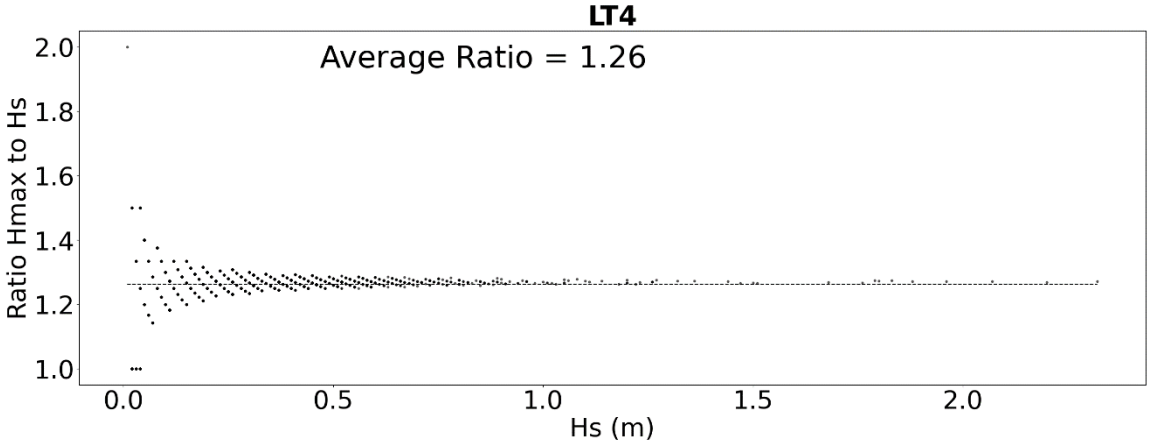


Figure 8d. Ratio H_{max} to H_s vs. H_s in LT4.

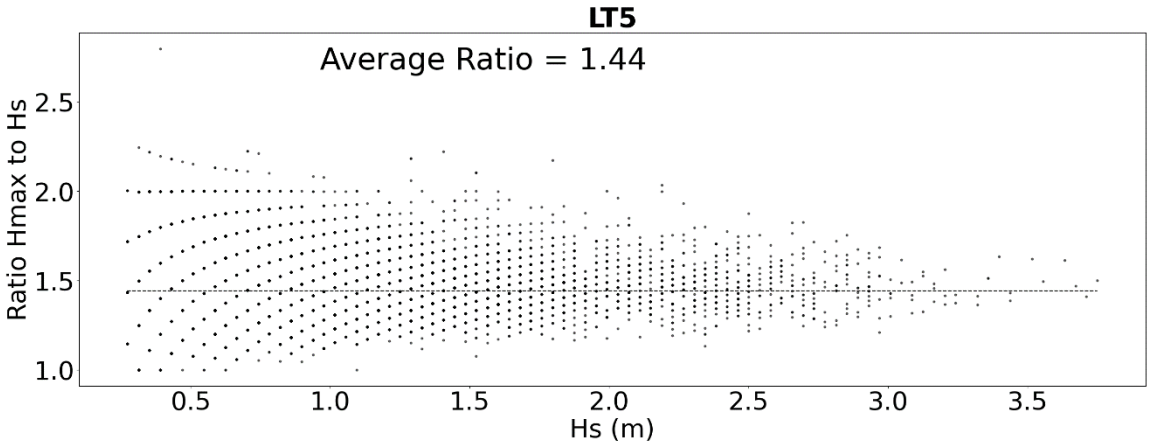


Figure 8e. Ratio H_{max} to H_s vs. H_s in LT5.

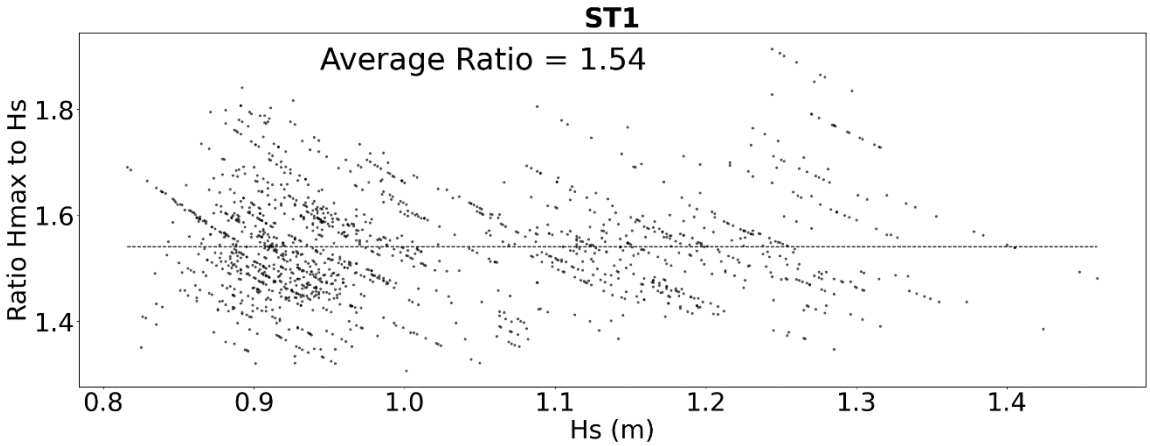


Figure 8f. Ratio H_{max} to H_s vs. H_s in ST1.

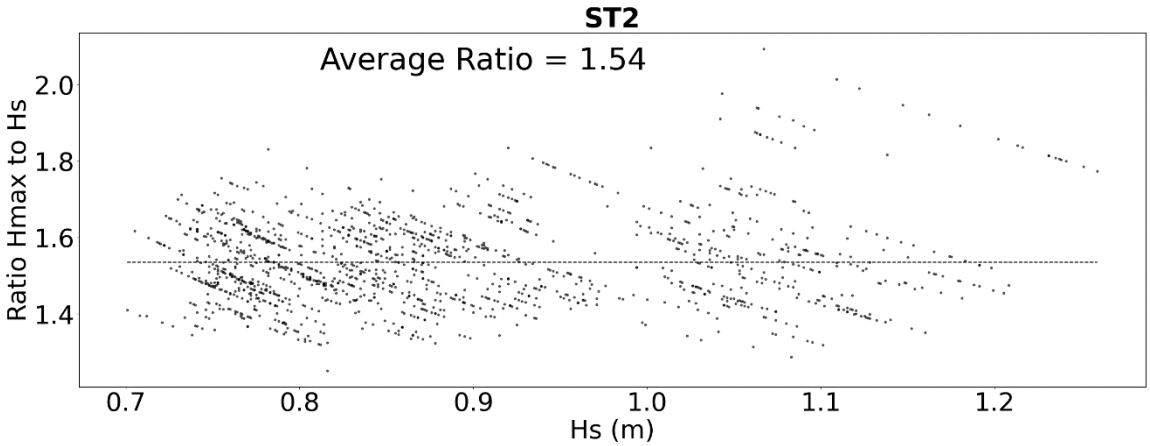


Figure 8g. Ratio H_{max} to H_s vs. H_s in ST2.

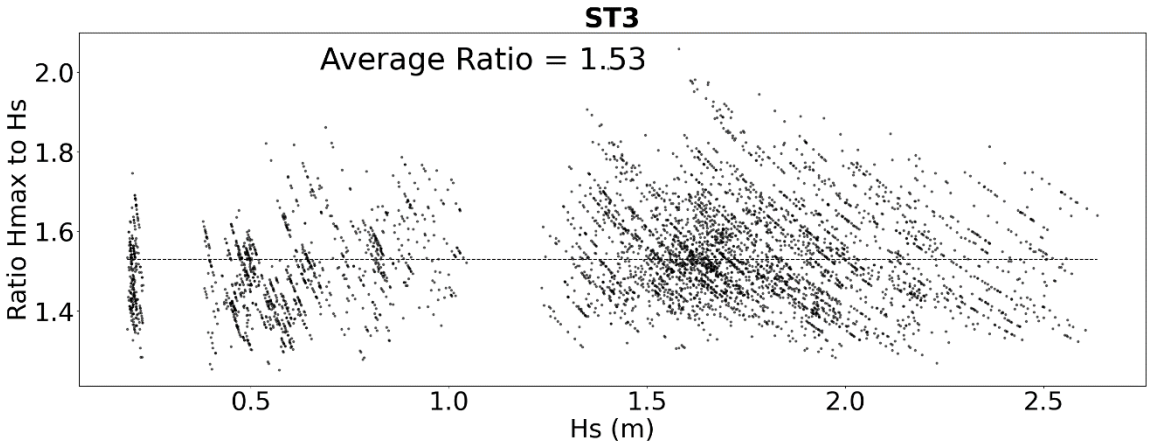


Figure 8h. Ratio H_{max} to H_s vs. H_s in ST3.

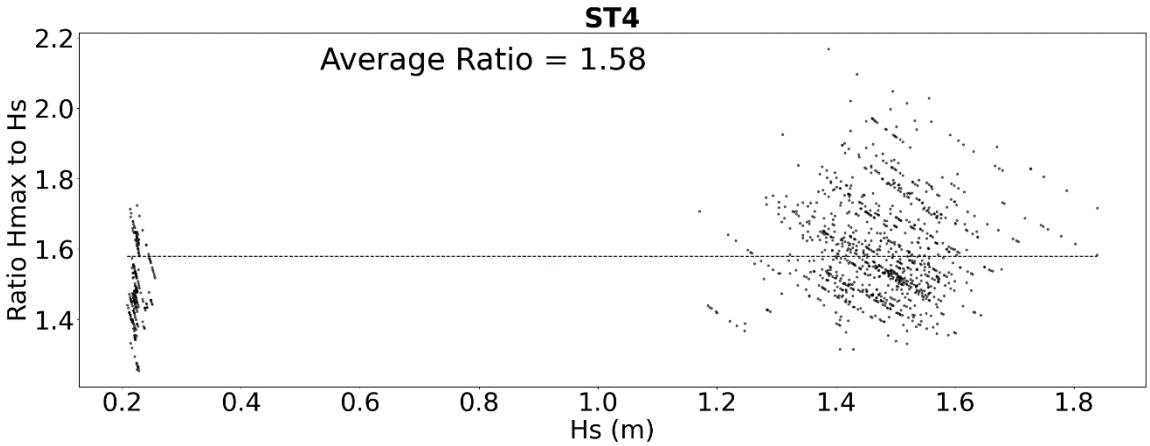


Figure 8i. Ratio H_{max} to H_s vs. H_s in ST4.

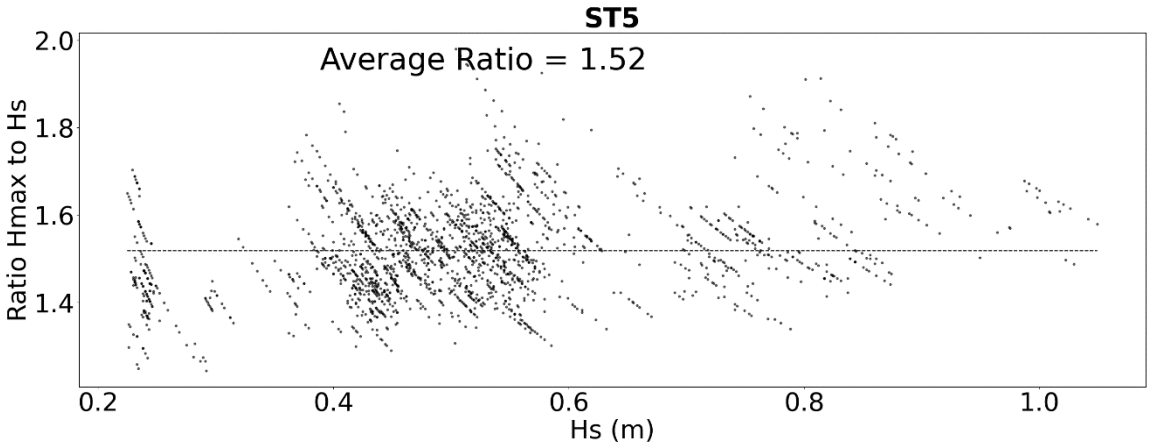


Figure 8j. Ratio H_{max} to H_s vs. H_s in ST5.

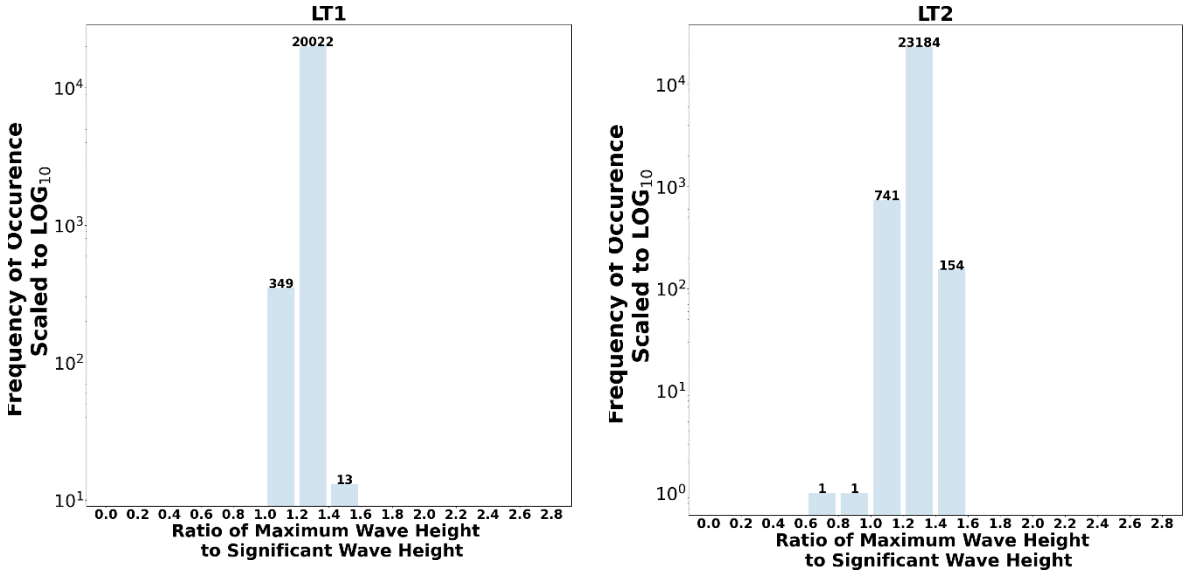


Figure 9a. Histogram of Frequency, Ratio of H_{max} to H_s for LT1 (left), LT2 (right).

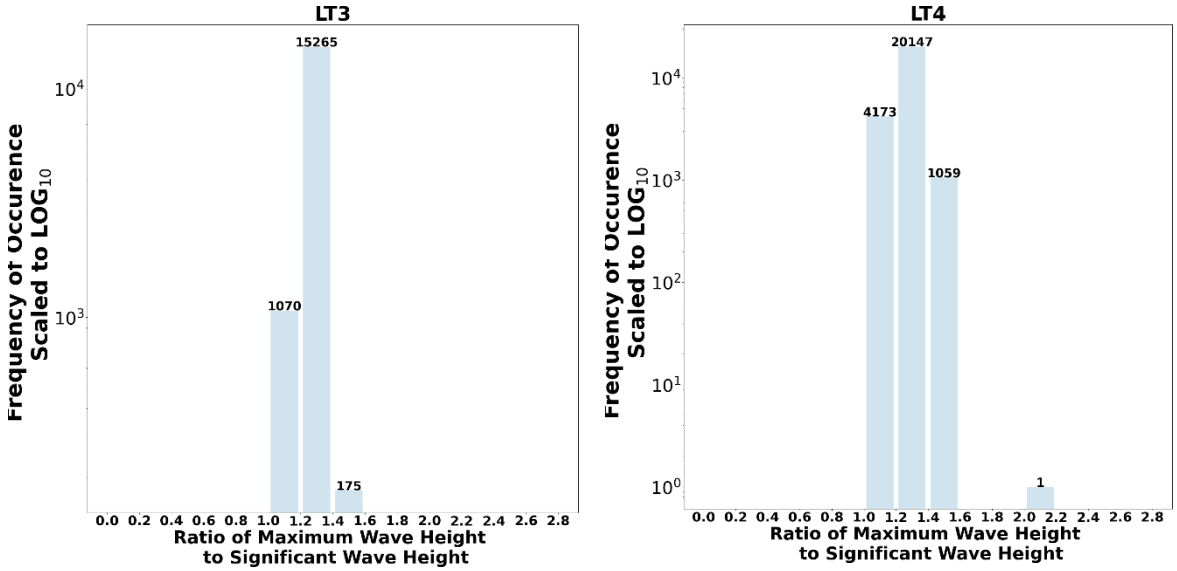


Figure 9b. Histogram of Frequency, Ratio of H_{max} to H_s for LT3 (left), LT4 (right).

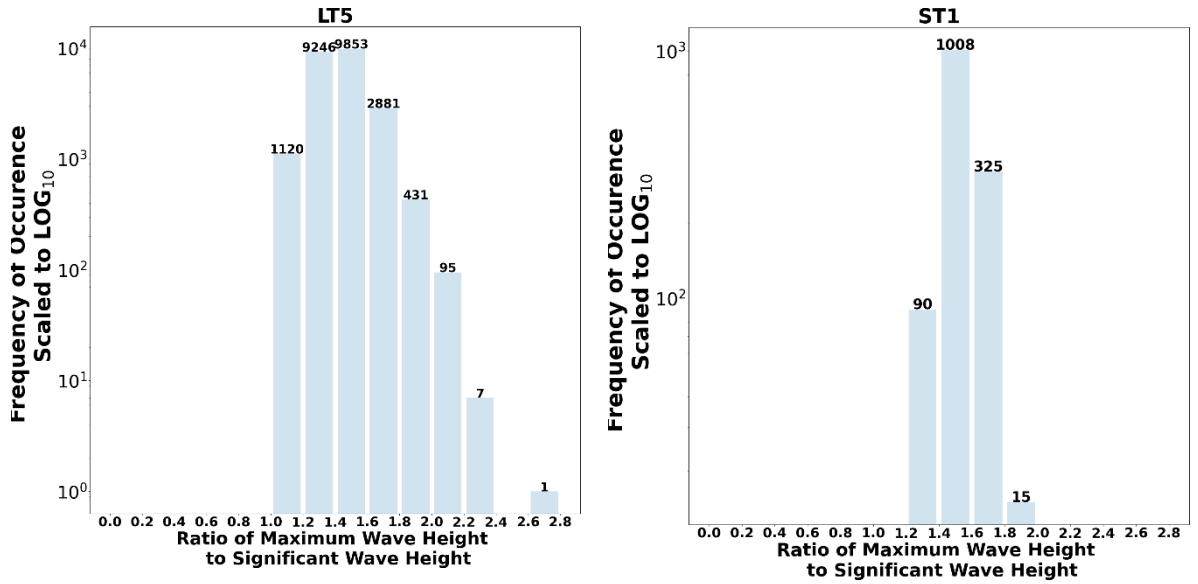


Figure 9c. Histogram of Frequency, Ratio of H_{max} to H_s for LT5 (left), ST1 (right).

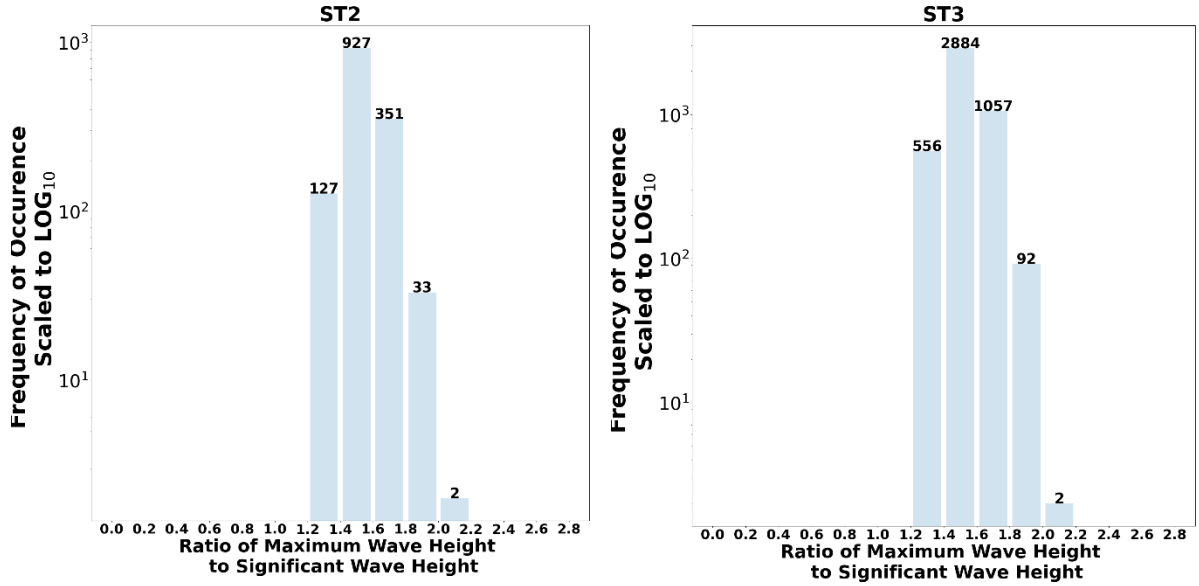


Figure 9d. Histogram of Frequency, Ratio of H_{max} to H_s for ST2 (left), ST3 (right).

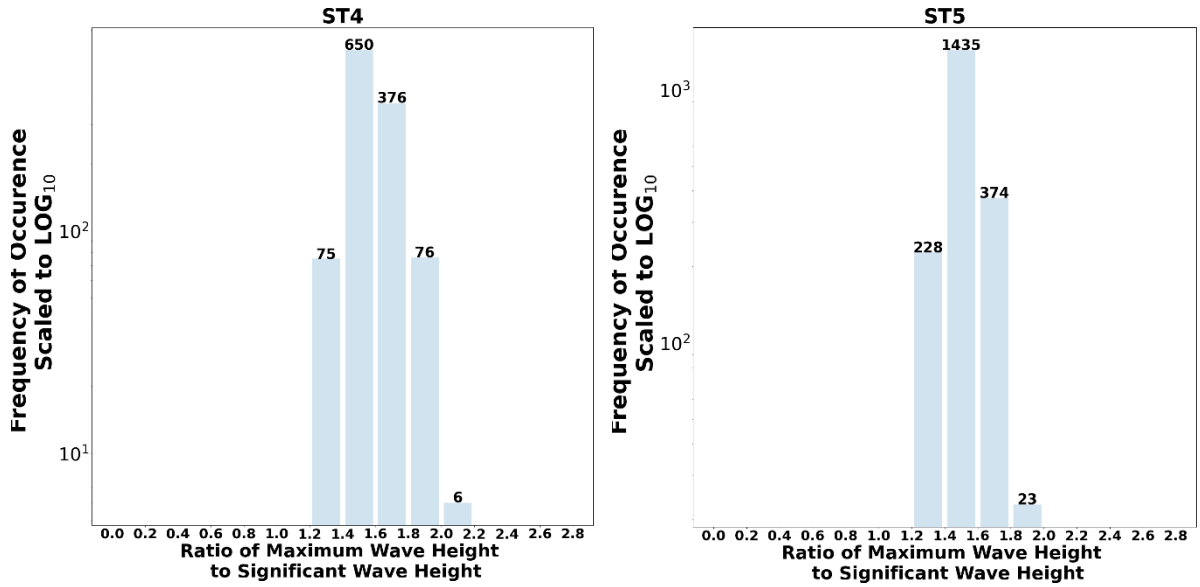


Figure 9e. Histogram of Frequency, Ratio of H_{max} to H_s for ST4 (left), ST5 (right).

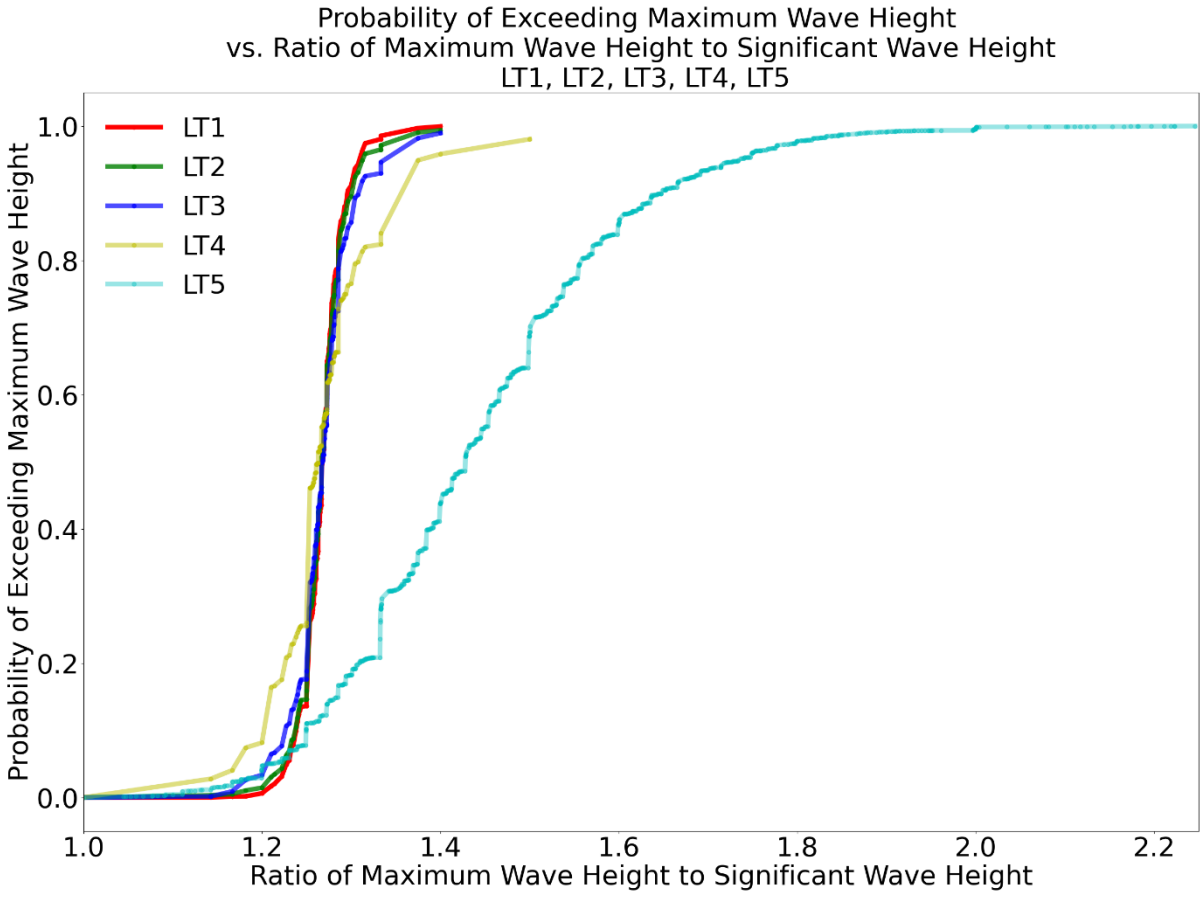


Figure 10a. Exceedance Probability of H_{max} plotted against H_{max}/H_s for LT1, LT2, LT3, LT4 and LT5.

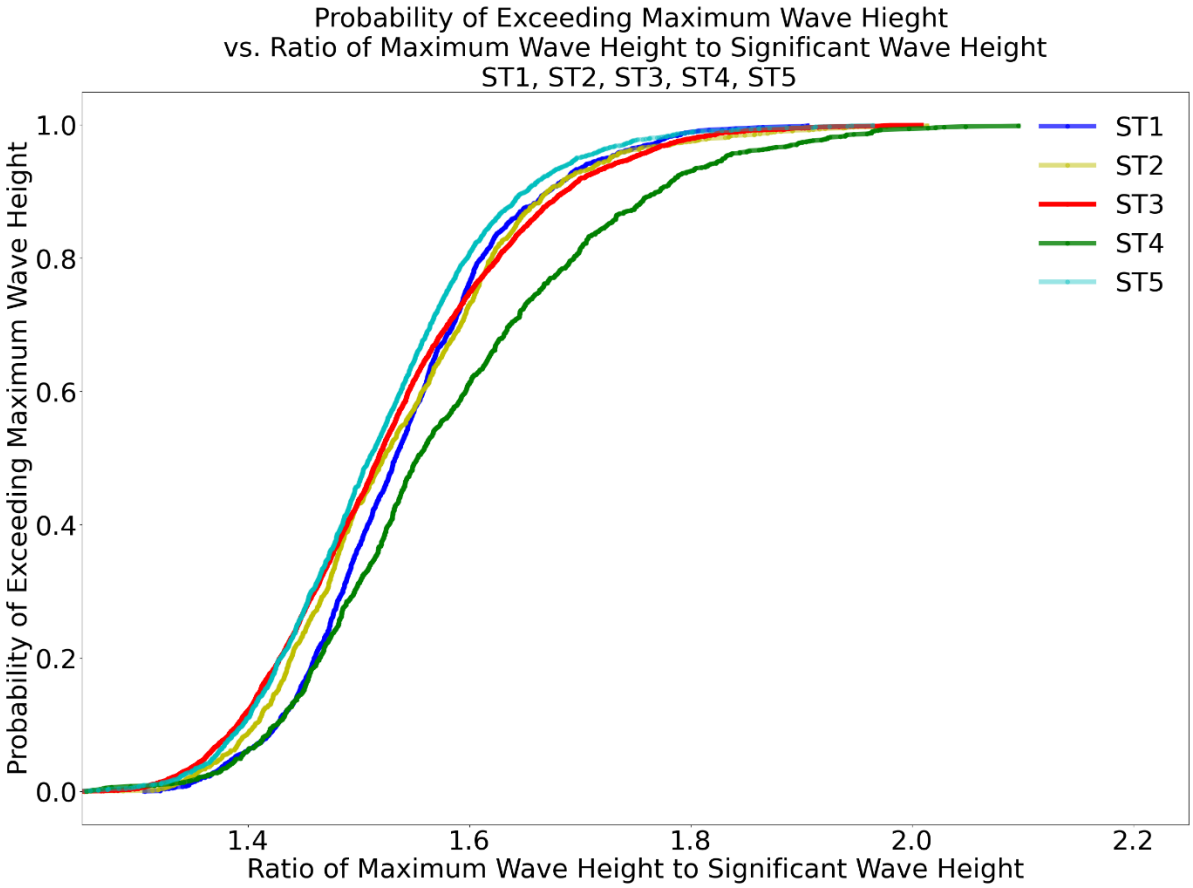


Figure 10b. Exceedance Probability of H_{max} plotted against H_{max}/H_s for ST1, ST2, ST3, ST4, and ST5.

3.2 Methods of Estimating H_{max}

3.2.1 Rayleigh Distribution

Individual wave heights with sufficiently narrow spectrum could be approximated by the Rayleigh Distribution. Although real-life individual waves have spectral spread which deviates from this assumption, it has been reported (Goda, 2000) that the Rayleigh Distribution still provides a good estimation of the distribution of individual wave heights. This study aims to compare the observed H_{max} with the H_{max} calculated using the Rayleigh Distribution. The equations for calculating H_{max} have already been described in **Section 2.2**, **Equations 23, 24,** and **25**. The maximum wave height in this study is calculated at probability μ = 0.01 (refer to **Equation 25** in **Section 2.2**).

3.2.2 Linear Regression

Section 2.2, **Table 8** reported that the H_s has strong linear correlation with respect to the H_{max}. Therefore, it is reasonable to consider linear regression as a potential method for accurate estimation of H_{max}. The equation for linear regression is given in **Equation 28**:

$$y = mx + c - (28)$$

where y is the predictand which is the H_{max}, x is the predictor which is H_s, m is a coefficient of gradient, and c is the coefficient of intersect. The optimal parameters are found by minimizing the least squared error between true values of y with predicted values by regression ($\hat{y}$). In our study, all forms of regression coefficients are fitted by the method of iteratively reweighted least squares.

3.2.3 Polynomial Regression

A weakness of linear regression is that it only models' linear relationships between predictor (x) and predictand (y). In other words, linear regression is not suitable for non-linear relationships between x and y. This limitation is depicted in **Figure 11**. In this example, we note that linear regression is incapable of correctly plotting the best-fit curve for simple quadratic relationship, $y = x^2$ even though the R^2 score is high.

To model non-linear relationships, the method of polynomial regression can be used. In this method, predictand y is modelled as a function to the nth degree polynomial of predictor x. There is a risk of overfitting if the degree (n) is too high. In this study, we evaluate polynomial regression to the power of 2 or n = 2. The polynomial regression equation is given in **Equation 29**.

$$y = \beta_2 x^2 + \beta_1 x + \beta_0 - (29)$$

where y is the predictand (in this study, H_{max}), x is the predictor (H_s in our study).

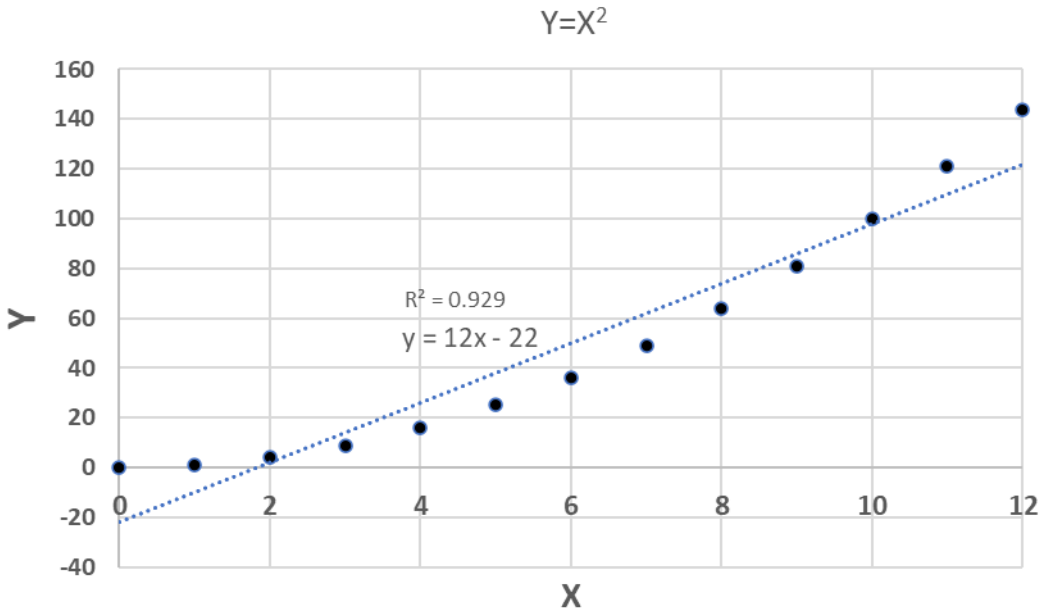


Figure 11. Linear Regression best-fit line against non-linear (quadratic) relationship.

3.2.4 Power Regression

Although polynomial regression (**Section 3.2.3**) can fit non-linear relationships between predictand y and predictor x, it uses more coefficients (3 coefficients) to explain the relationship between 2 variables (y and x). The relationship modelled by

polynomial regression (**Equation 29**) is also difficult to explain because it is not one-on-one, but one to two, that is y with respect to x^2, x^1, and x^0. An alternative to polynomial regression is power regression.

Power regression can fit non-linear relationships as well. In addition, power regression uses only 2 coefficients (a and b) to explain the relationship between 2 variables (x and y). Power regression is also easier to explain because it models y to x in a one-on-one relationship. In a nutshell, power regression models y as a function of x raised to the power of b (**Equation 30**):

$$y = ax^b - (30)$$

3.2.5 Multiple Linear Regression (MLR)

MLR attempts to model the predictand y by calculating linear relationships between set y and set $\{x_1, x_2, x_3, \ldots, x_N\}$ as described in **Equation 31**:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots b_nx_n - (31)$$

MLR is an extension of linear regression to multiple predictands. This method assumes that predictand y is linearly dependent on predictor set $X = \{x_1, x_2, \ldots, x_N\}$. Additionally, the residuals are assumed to follow a normal distribution. The advantage of MLR is that the relationship between predictand and predictors are easy to explain.

In this study, for in-situ observations labelled LT1, LT2, LT3, LT4, ST1, ST2, ST3, ST4, and ST5, the predictors are (H_s), zero-crossing wave period (T_z), and peak wave period (T_p). Meanwhile, for in-situ observation labelled LT5 the predictors are (H_s), period of the highest wave (Thmax), estimated mean wave period in respect to fundamental zeroth moment and first moment, m1 (Tm01), estimated mean wave period in respect to fundamental zeroth moment and first moment, m2 (Tm02), estimated mean wave period in respect to fundamental m0 and first moment m1 in lower frequency band (Tm02a), and estimated mean wave period in respect to fundamental m0 and first moment m1 in mid-frequency band (Tm02b).

3.2.6 Deep Learning

Regression-based techniques assume that the relationship between H_s to H_{max} obeys a pre-defined relationship. However, this may not be the case. Artificial Neural Networks (ANN) can estimate predictands without any underlying assumption of the relationship. ANNs have the advantage of being able to model non-linear relationship between predictors and predictands. It was reported that the (H_s) relationship with average zero-cross wave period (T_z) and peak-spectral period (T_p) was more accurately modelled using ANNs while H_s relationship to maximum spectral energy density (E_{max}) and (H_{max}) is equally well modelled between regression and ANNs (Agrawal and Deo, 2004). This study compares the performance of ANNs with regression methods in estimating (H_{max}) using (H_s). ANNs have the disadvantage of

being difficult to explain the mathematical relationship in a concise and precise manner as they consist of several layers of equations. In addition, ANNs also have the disadvantage of consuming more computational power than regression techniques during the training phase.

The predictor, (H_s) and predictand, (H_{max}) are scaled to values of 0 to 1 before ingested into the ANN. This study uses multilayer feedforward Neural Network consisting of 3 layers that contain 300, 400, and 1 node respectively, with learning rate of 0.00001. The Rectified Linear Unit or ReLU activation function was used in our study. The ReLU activation function avoids the vanishing gradient problem associated with sigmoid and tanh activation functions. Meanwhile, the loss function of this study is the *mean absolute error* (MAE) with Nadam optimizer used to minimize the MAE. The Nadam optimizer incorporates a momentum component to improve ANN convergence speed and quality. Verification of Nadam optimizer reports smaller Mean Square Error (MSE) than other common optimizers (Dozat, 2016).

4.0 Results and Analysis

4.1 Scatterplot and Boxplot Analysis

The algorithms mentioned in **Section 3.2** namely Rayleigh distribution, linear regression, polynomial regression, power regression, multiple linear regression and neural networks are used to estimate the H_{max} based on H_s and wave period. The estimated H_{max} is then plotted against the observed H_{max}. **Figures 12a-j** depict scatterplots for LT1, LT2, LT3, LT4, LT5, ST1, ST2, ST3, ST4, and ST5 respectively. R^2 score and mean absolute error (MAE) were used to quantify the goodness of fit between estimated to observed maximum wave height.

Long term wave observations gathered for all months of the year at LT1, LT2, LT3, LT4 and LT5 indicate that the Rayleigh mode and mean tends to overestimate H_{max}. The overestimation is especially acute for Rayleigh H_{max} at 0.01% probability. Linear regression, polynomial regression, power regression, multiple linear regression, and neural network (H_s) produce the most accurate estimation of H_{max} compared to observation. Surprisingly neural networks with (H_s), zero-crossing wave period (T_z), and peak wave period (T_p) have lower skill compared to neural networks using only (H_s). Estimating (H_{max}) by multiplying (H_s) with constant 1.27 for LT1, LT2, LT3, and LT4 and constant 1.50 for LT5 worked almost as well as regression and neural network method.

Short-term wave observations in strong winds and rough seas recorded in ST1, ST2, ST3, ST4, and ST5 showed that Rayleigh mode and mean underestimates H_{max}. This contrasts with long term observations recorded in LT1, LT2, LT3, LT4, and LT5. Meanwhile, Rayleigh H_{max} at 0.01% probability slightly overestimates H_{max}. Regression showed negligible difference in skill compared to neural networks (H_s). Three (3) parameter neural network (H_s, T_z, T_p) showed worse skill compared to one (1) parameter neural network (H_s). Regression techniques and single parameter neural network (H_s) performed equally well with a high R^2 score. Estimating (H_{max}) by multiplying (H_s) with 1.50 showed nearly as good R^2 and MAE skill as regression and neural networks.

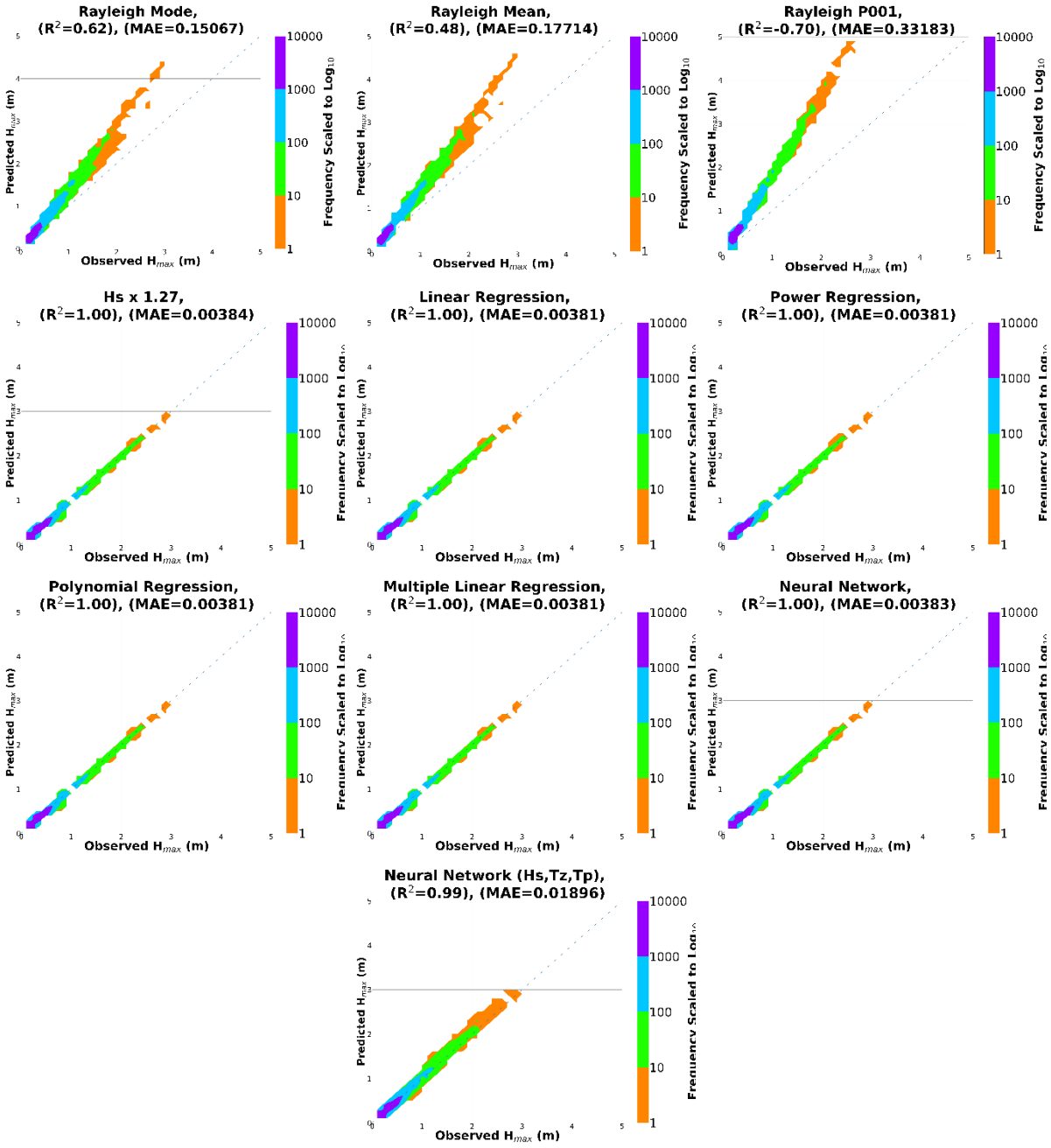


Figure 12a. Scatterplot of Estimated H_{max} (y-axis) vs.Observed H_{max} (x-axis) for LT1.

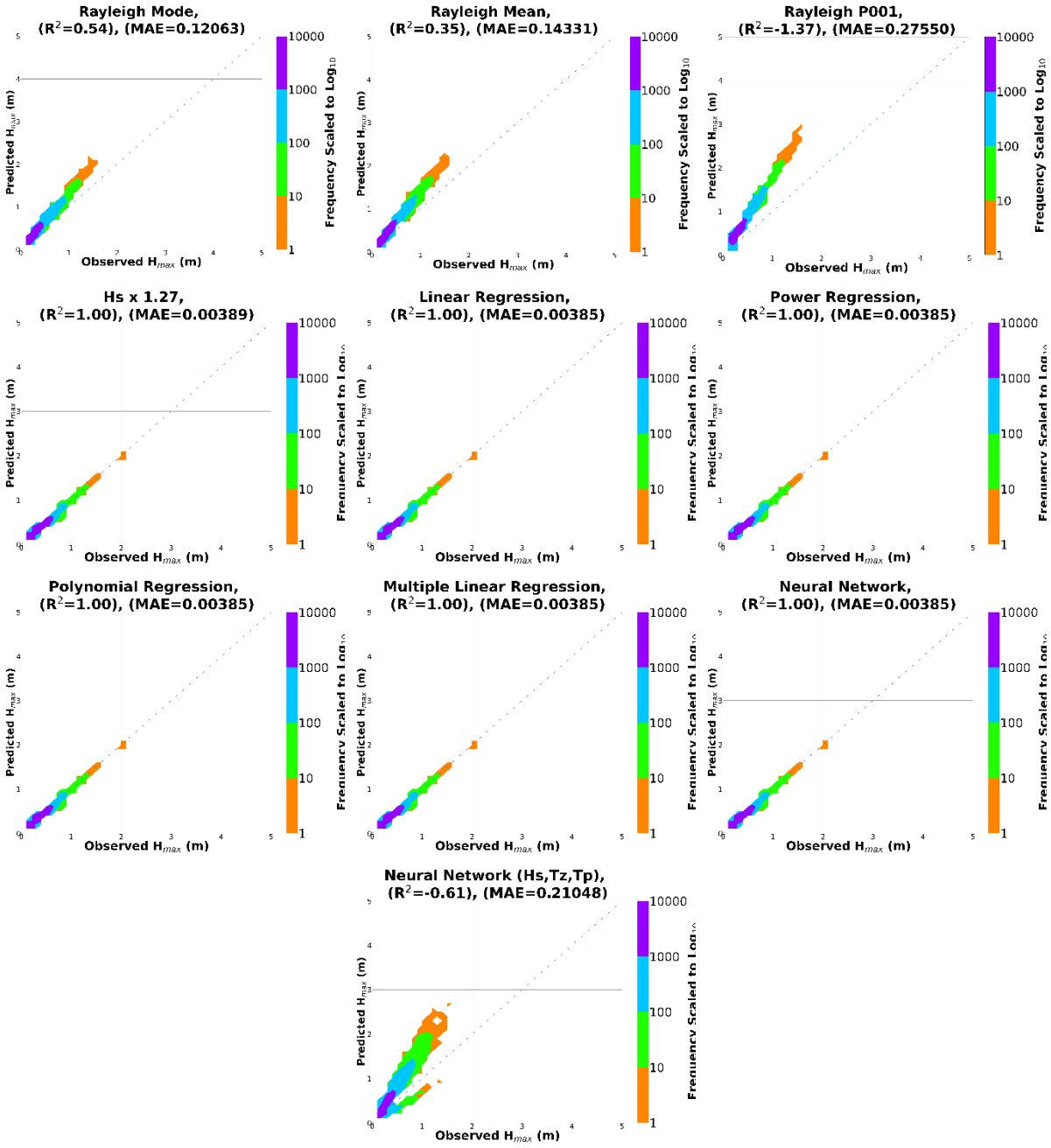


Figure 12b. Scatterplot of Estimated H_{max} (y-axis) vs. Observed H_{max} (x-axis) for LT2.

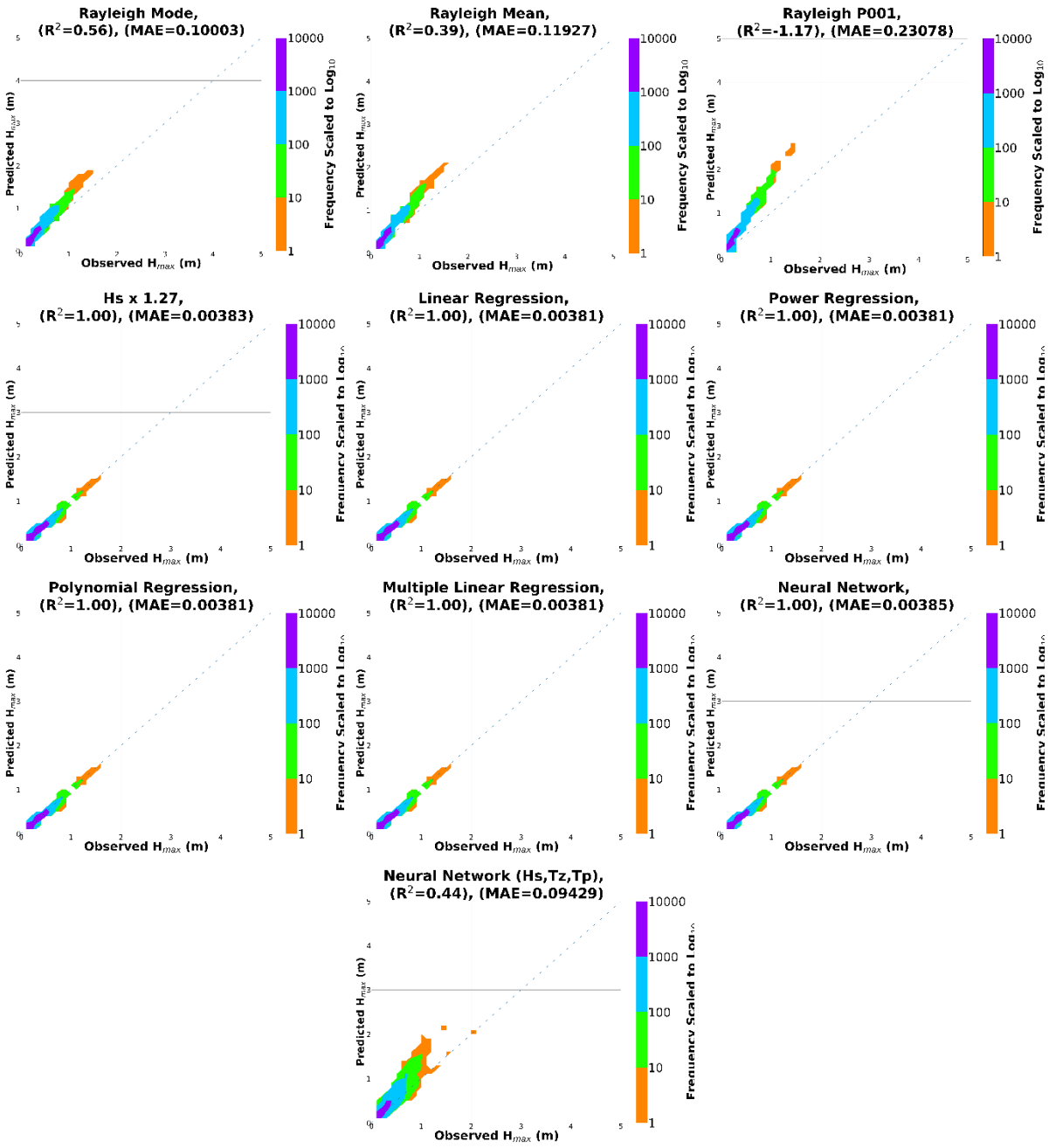


Figure 12c. Scatterplot of Estimated H_{max} (y-axis) vs. Observed H_{max} (x-axis) for LT3.

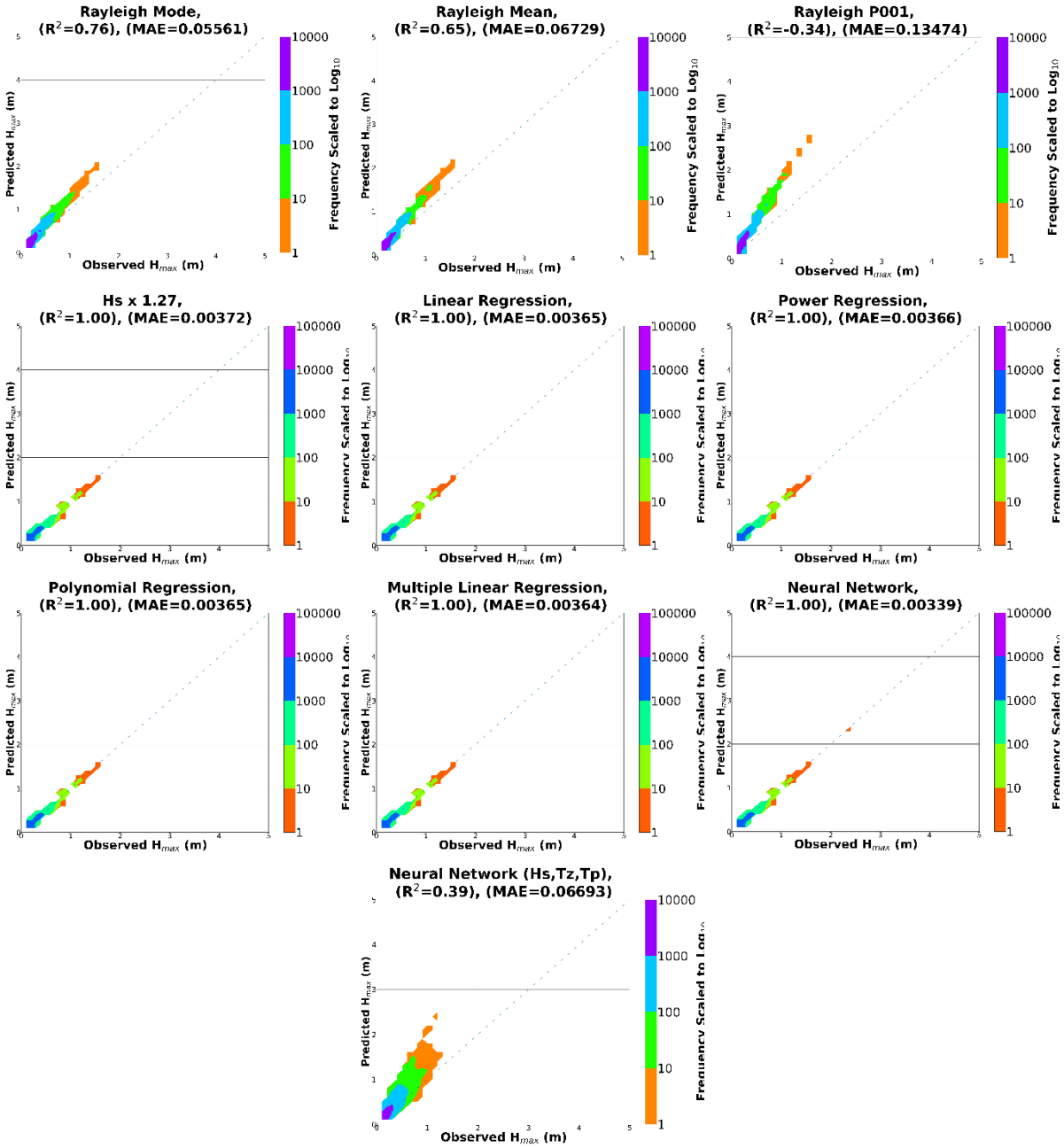


Figure 12d. Scatterplot of Estimated H_{max} (y-axis) vs. Observed H_{max} (x-axis) for LT4.

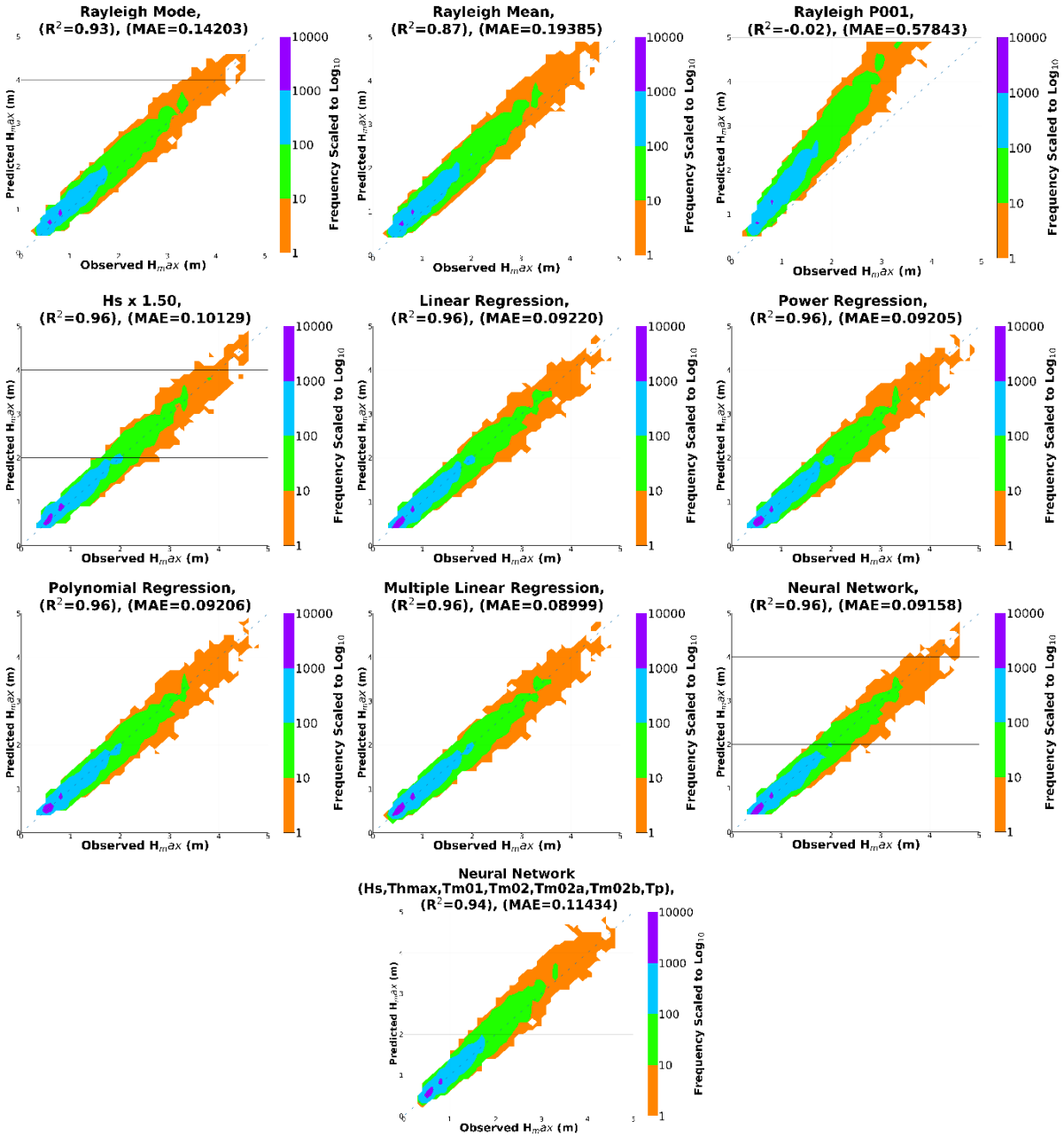


Figure 12e. Scatterplot of Estimated H_{max} (y-axis) vs. Observed H_{max} (x-axis) for LT5.

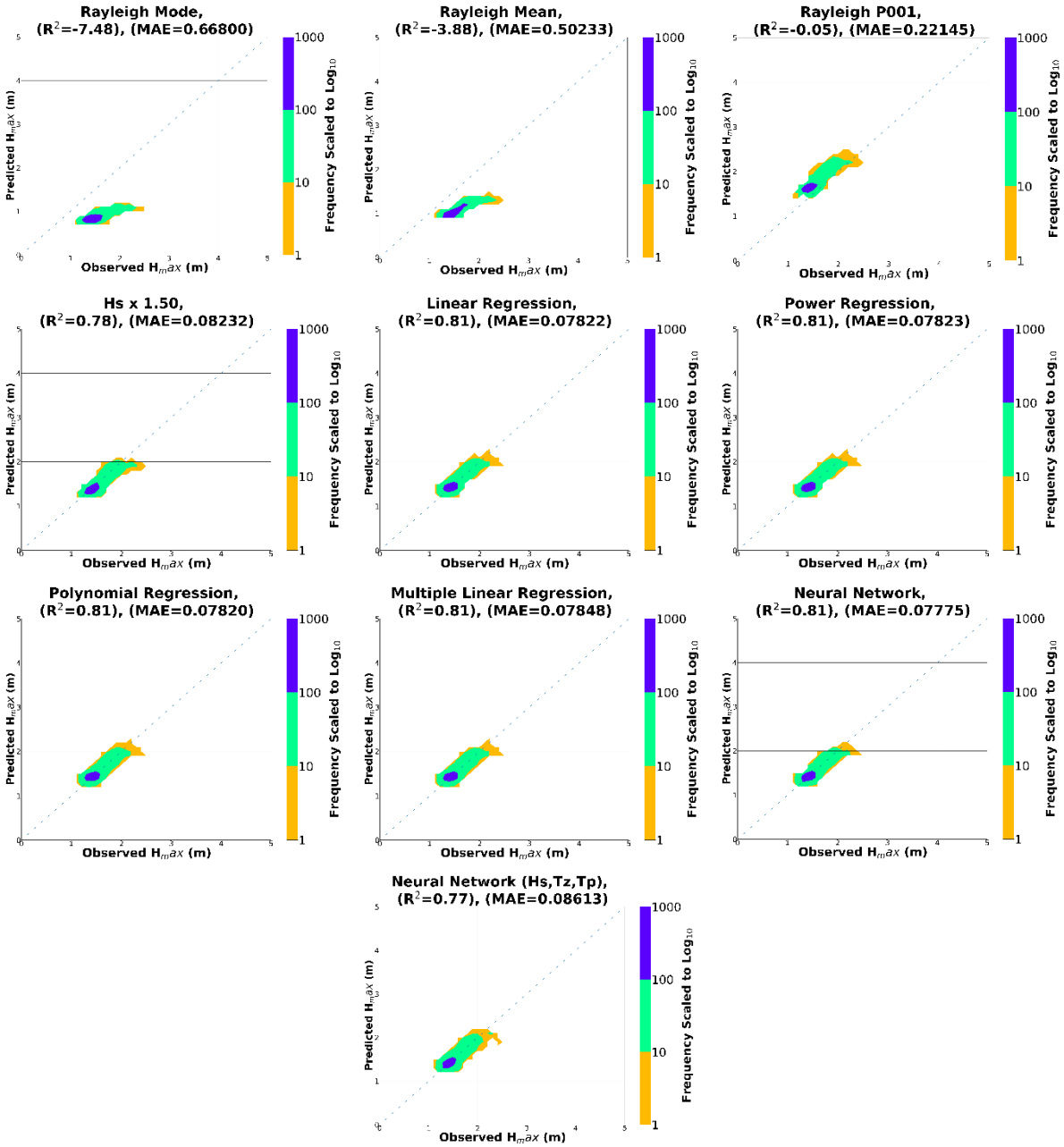


Figure 12f. Scatterplot of Estimated H_{max} (y-axis) vs. Observed H_{max} (x-axis) for ST1.

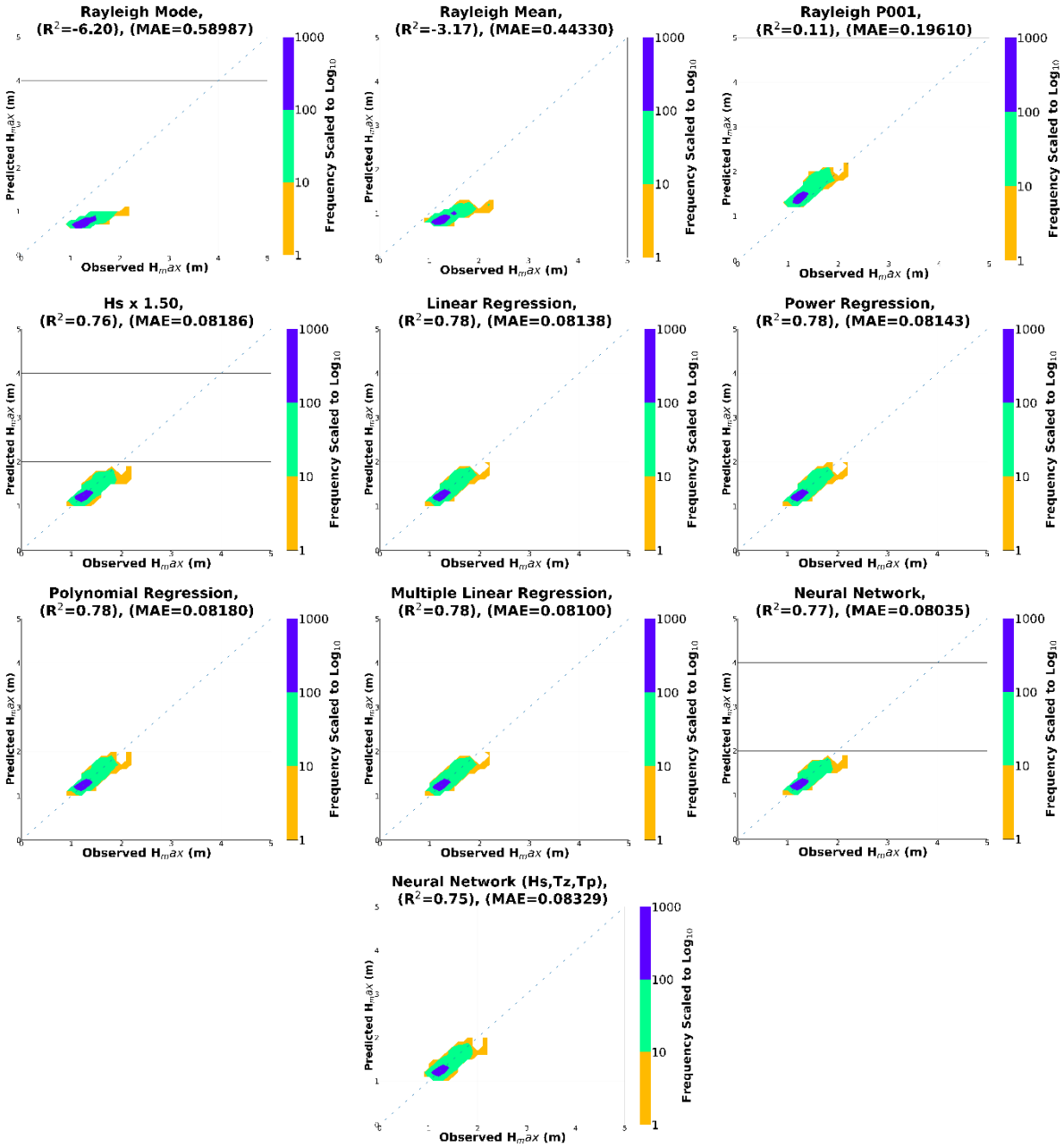


Figure 12g. Scatterplot of Estimated H_{max} (y-axis) vs. Observed H_{max} (x-axis) for ST2.

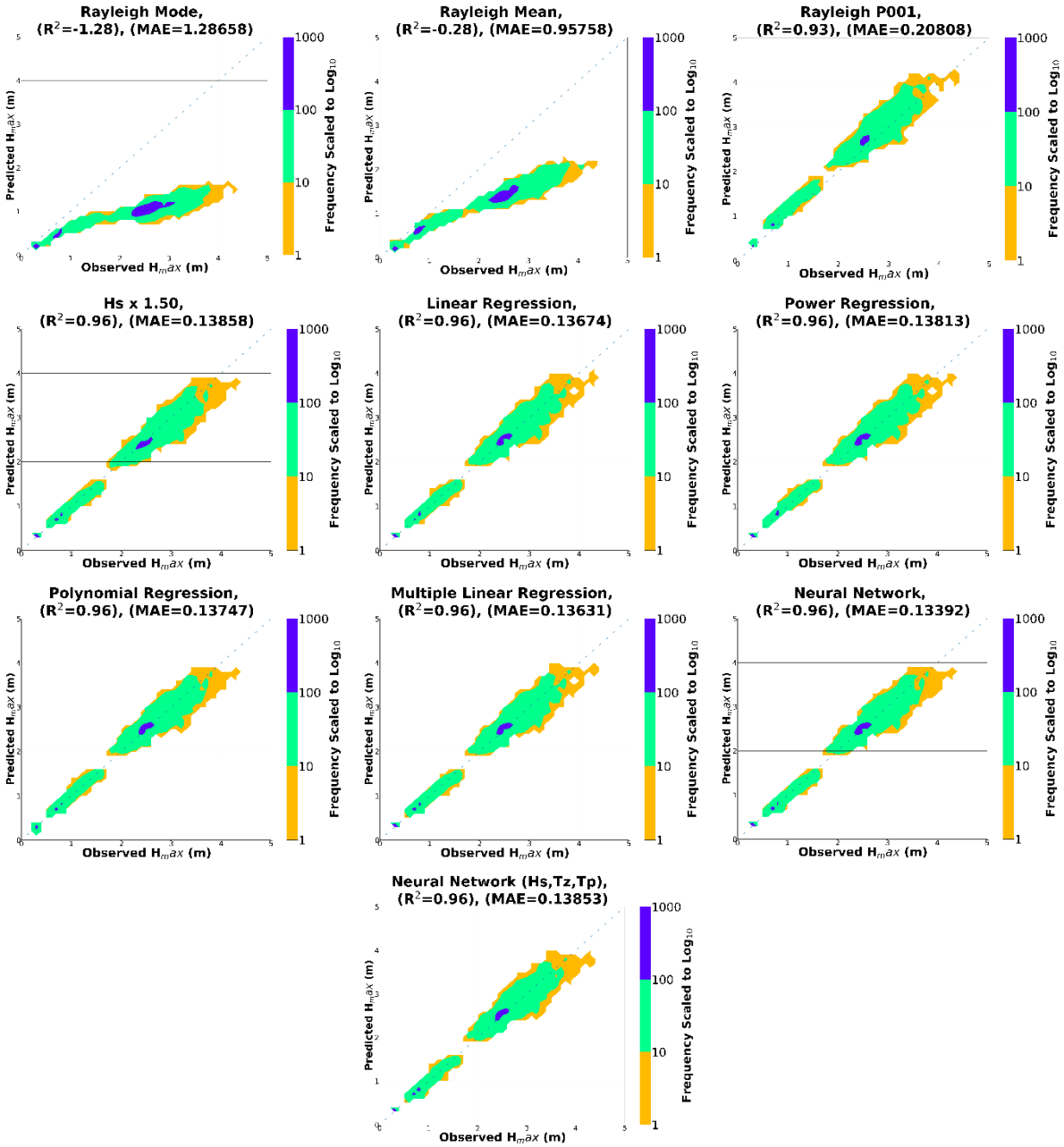


Figure 12h. Scatterplot of Estimated H_{max} (y-axis) vs. Observed H_{max} (x-axis) for ST3.

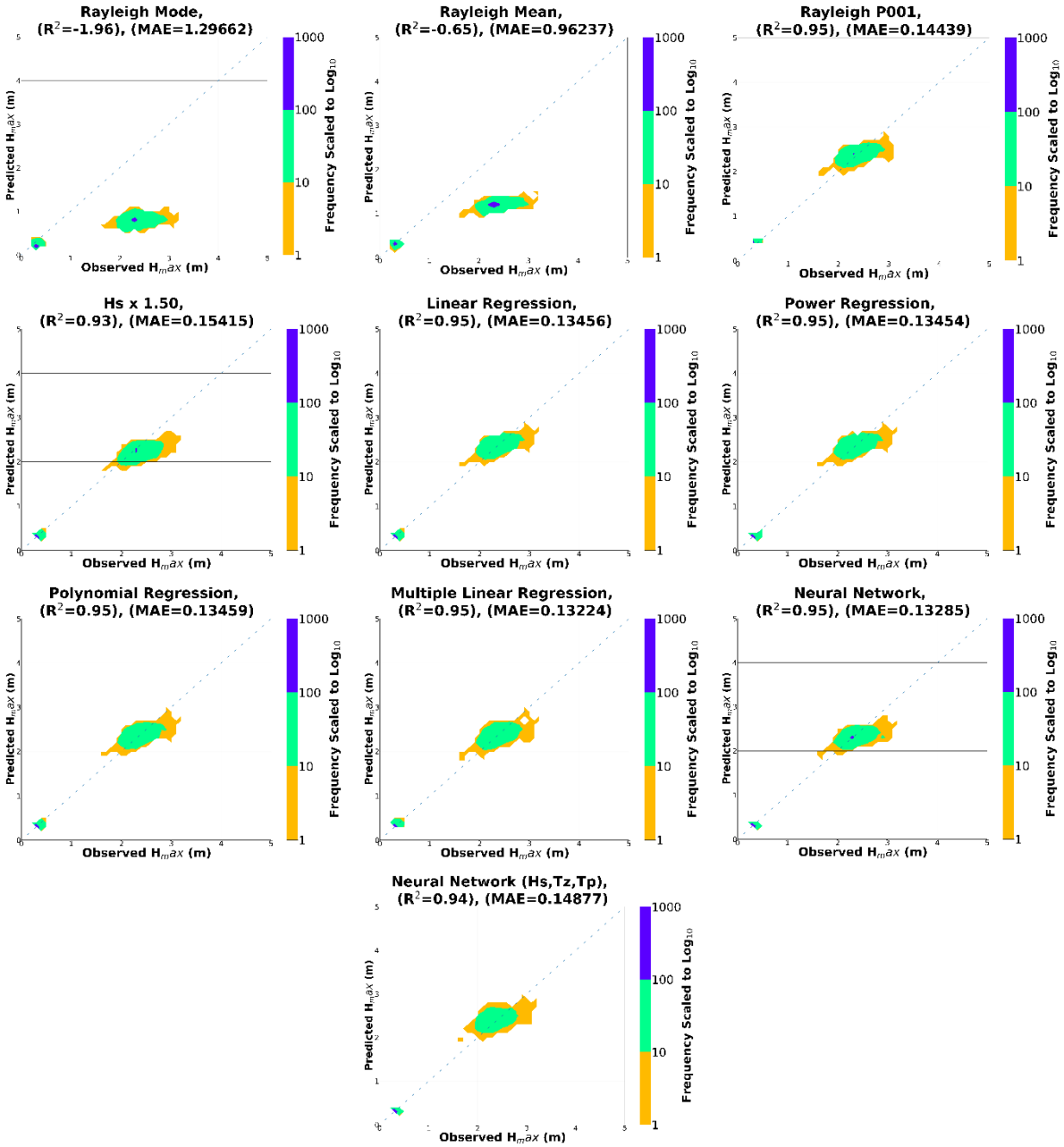


Figure 12i. Scatterplot of Estimated H_{max} (y-axis) vs. Observed H_{max} (x-axis) for ST4.

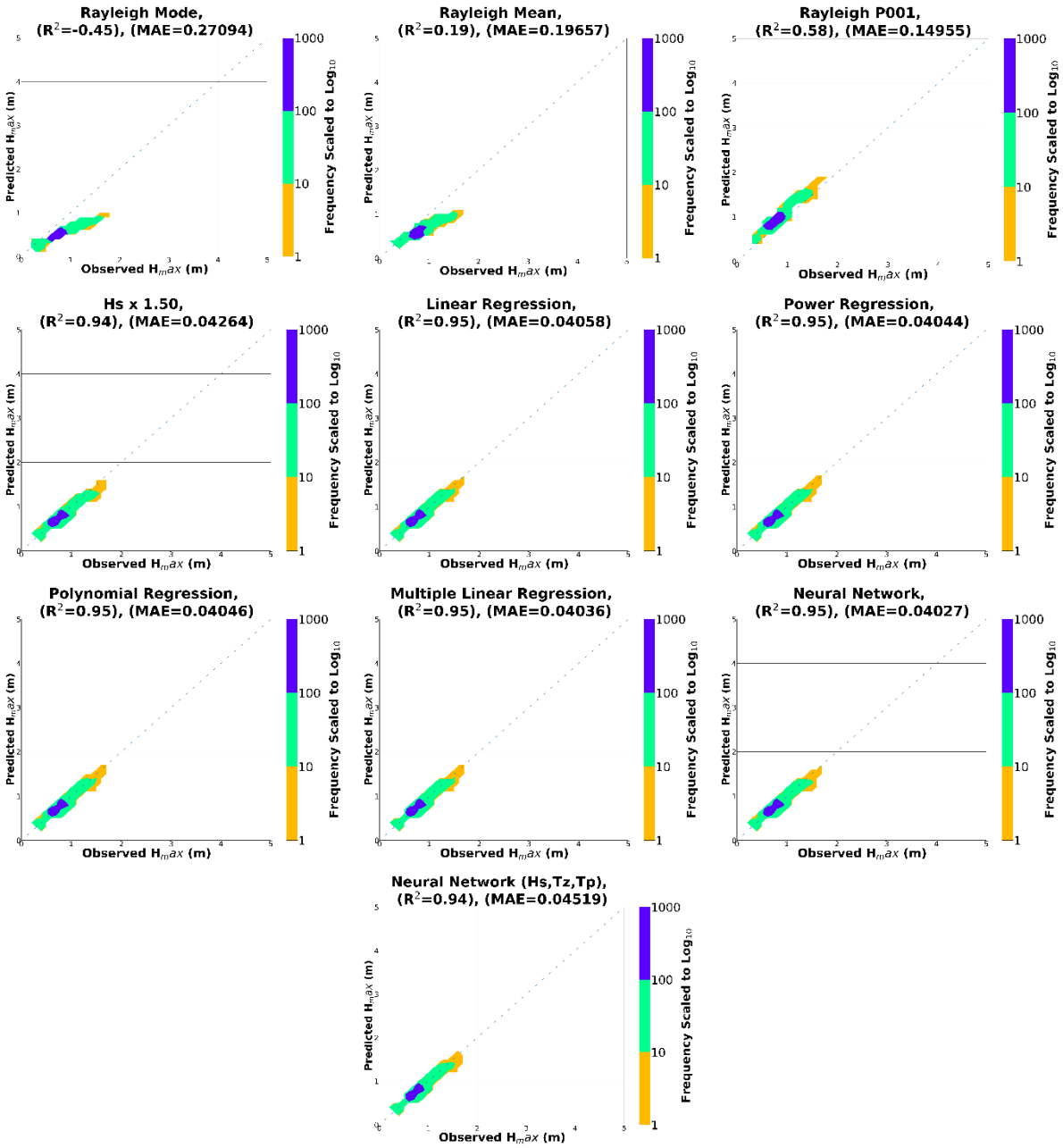


Figure 12j. Scatterplot of Estimated H_{max} (y-axis) vs. Observed H_{max} (x-axis) for ST5.

Analysis of the distribution for each in-situ observation site shows that multiplication of H_s by 1.27 for sites at shallow water (LT1, LT2, LT3, and LT4) and 1.50 for sites at deep water (LT5), or strong winds and rough seas (ST1, ST2, ST3, ST4, and ST5) is nearly as good as regression analysis and neural network prediction. **Figures 13a** to **13j** depict box plots for each site. MAE refers to mean absolute error while std MAE is the standard deviation of the MAE.

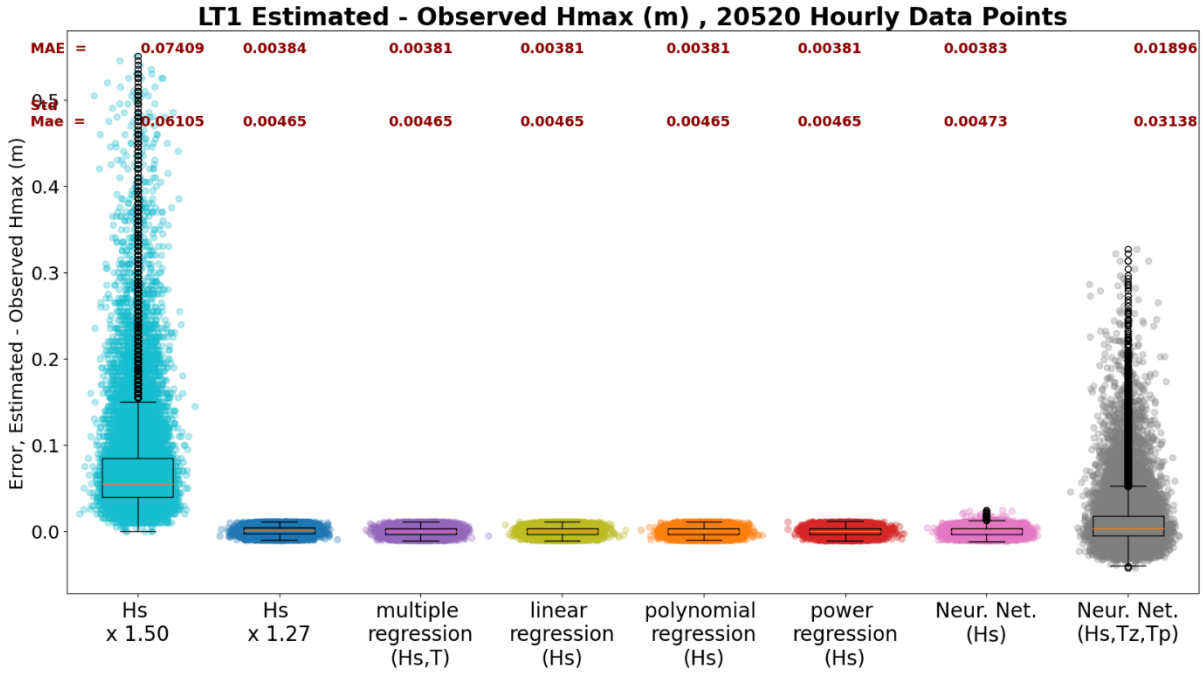


Figure 13a. Boxplot Analysis for LT1

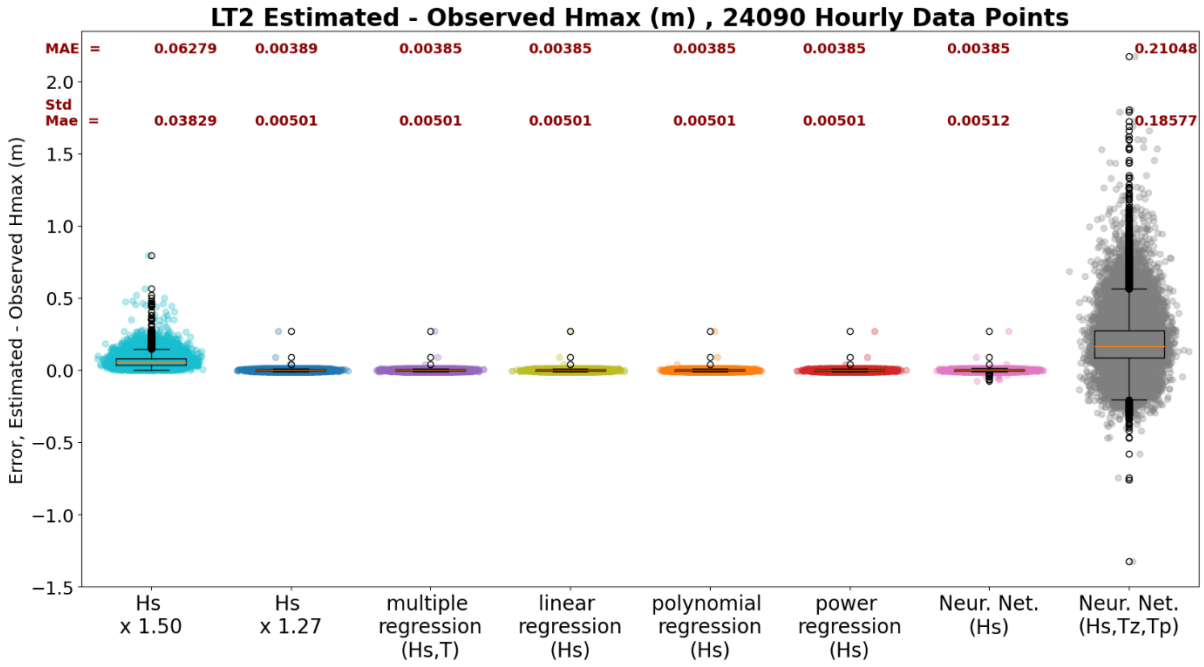


Figure 13b. Boxplot Analysis for LT2

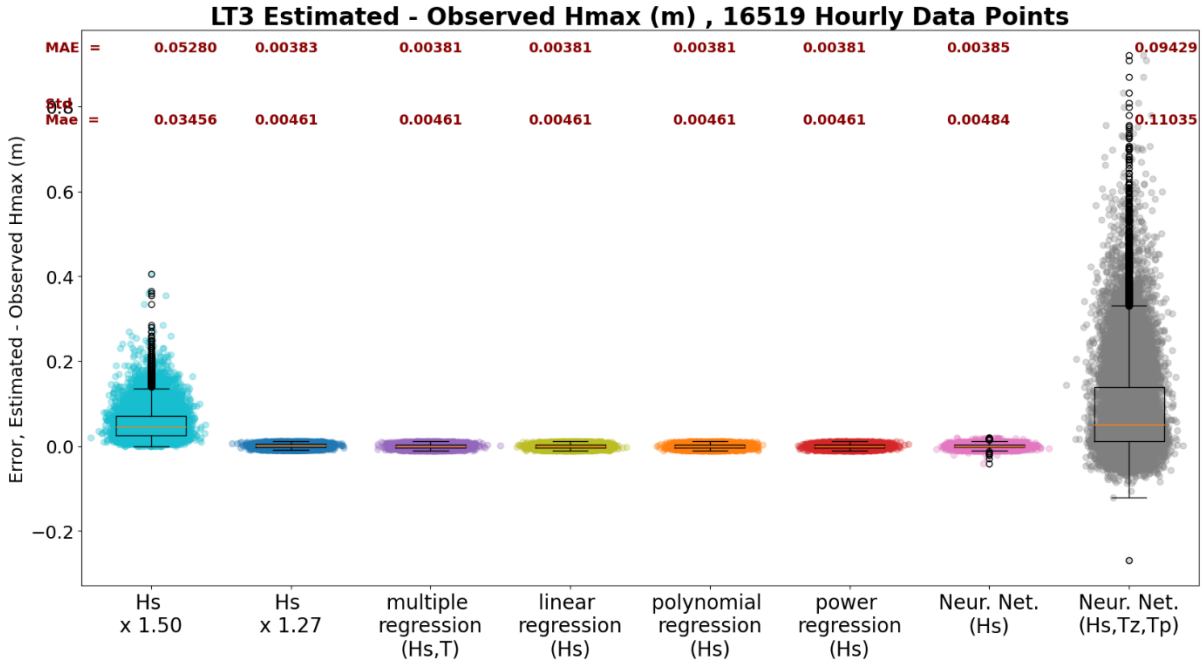


Figure 13c. Boxplot Analysis for LT3

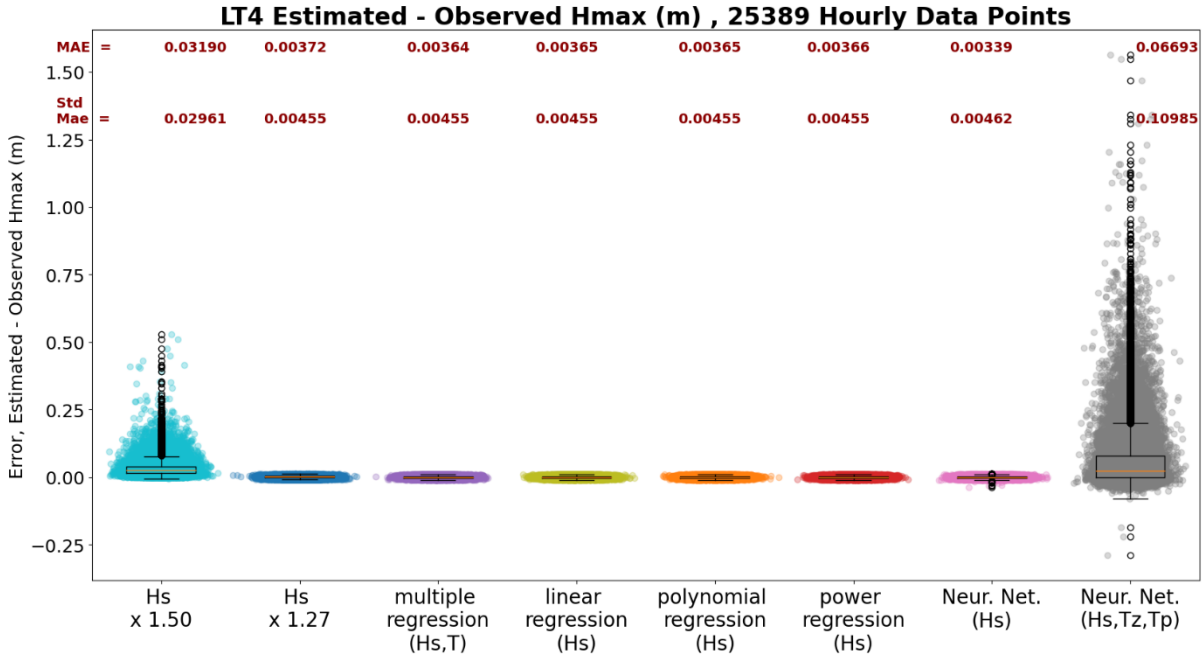


Figure 13d. Boxplot Analysis for LT4

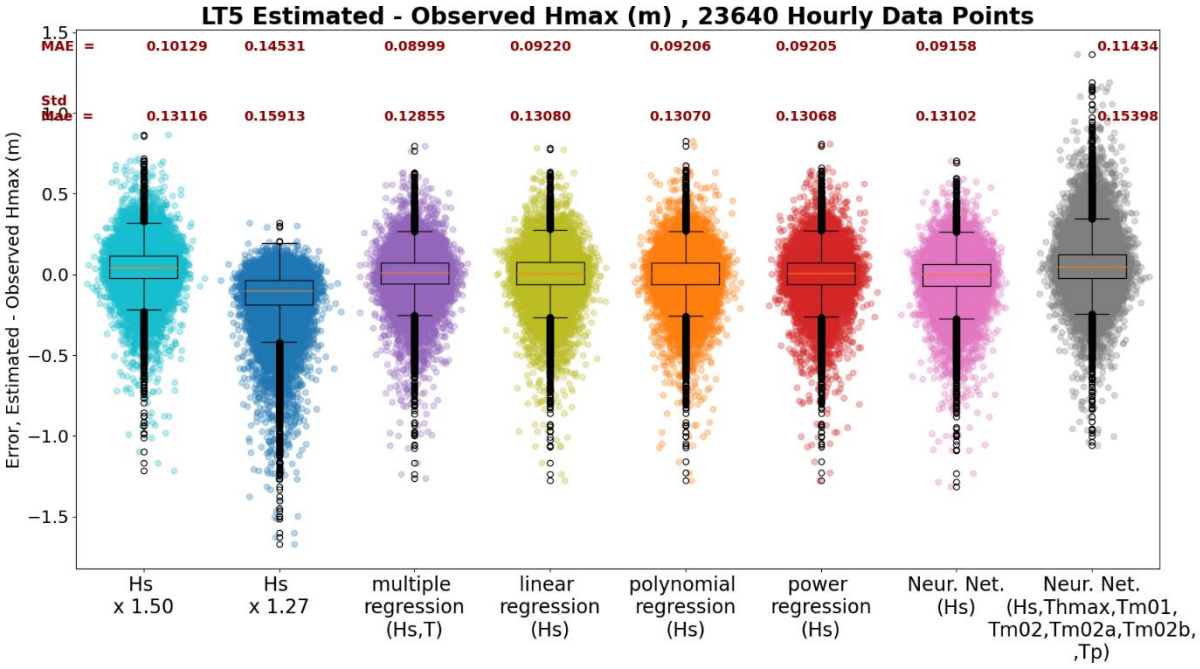


Figure 13e. Boxplot Analysis for LT5

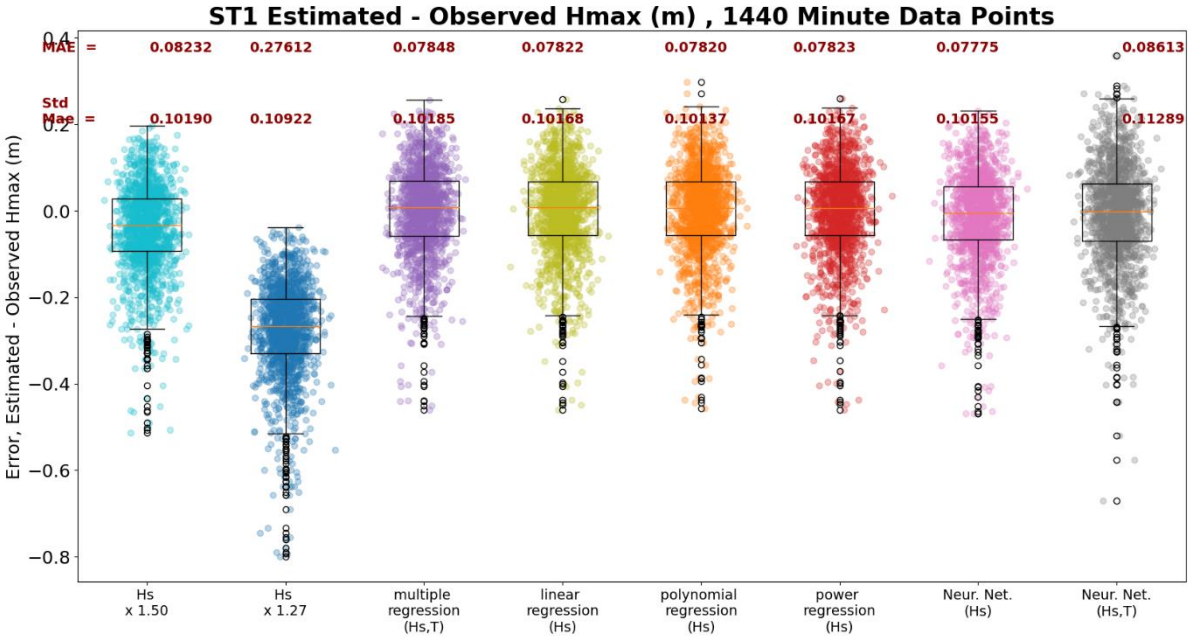


Figure 13f. Boxplot Analysis for ST1

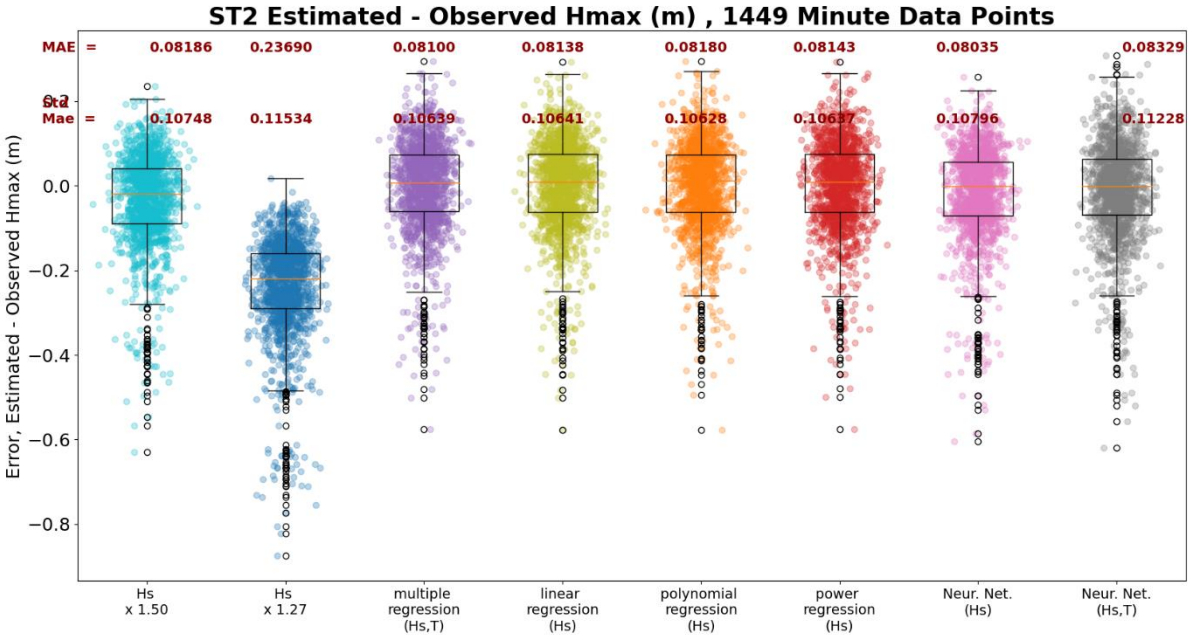


Figure 13g. Boxplot Analysis for ST2

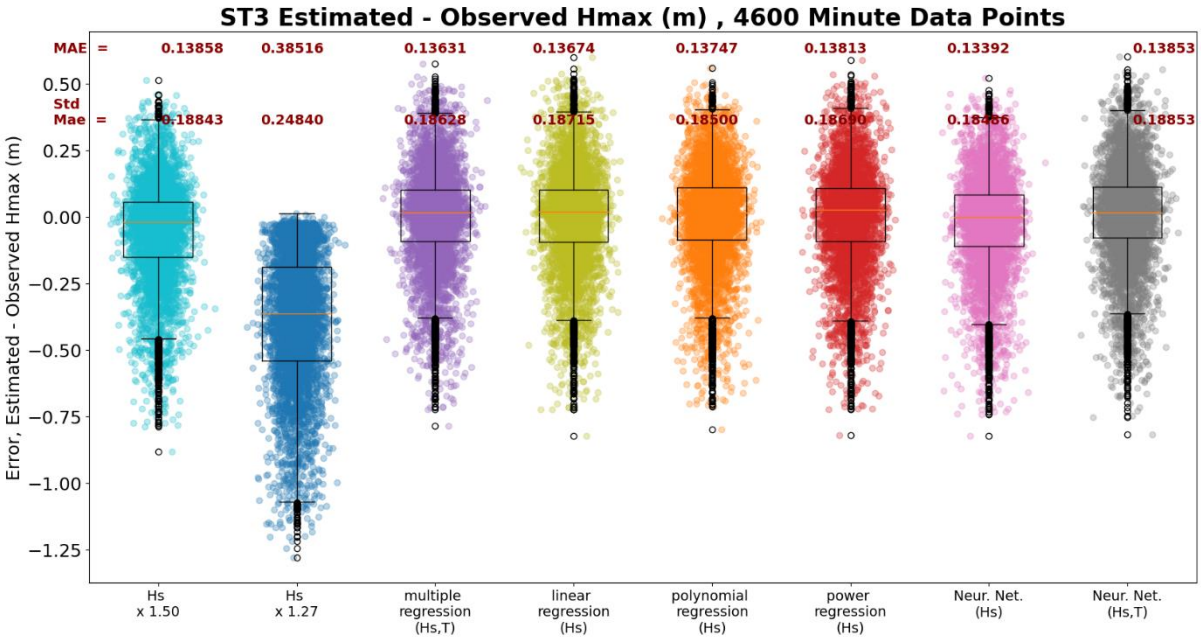


Figure 13h. Boxplot Analysis for ST3

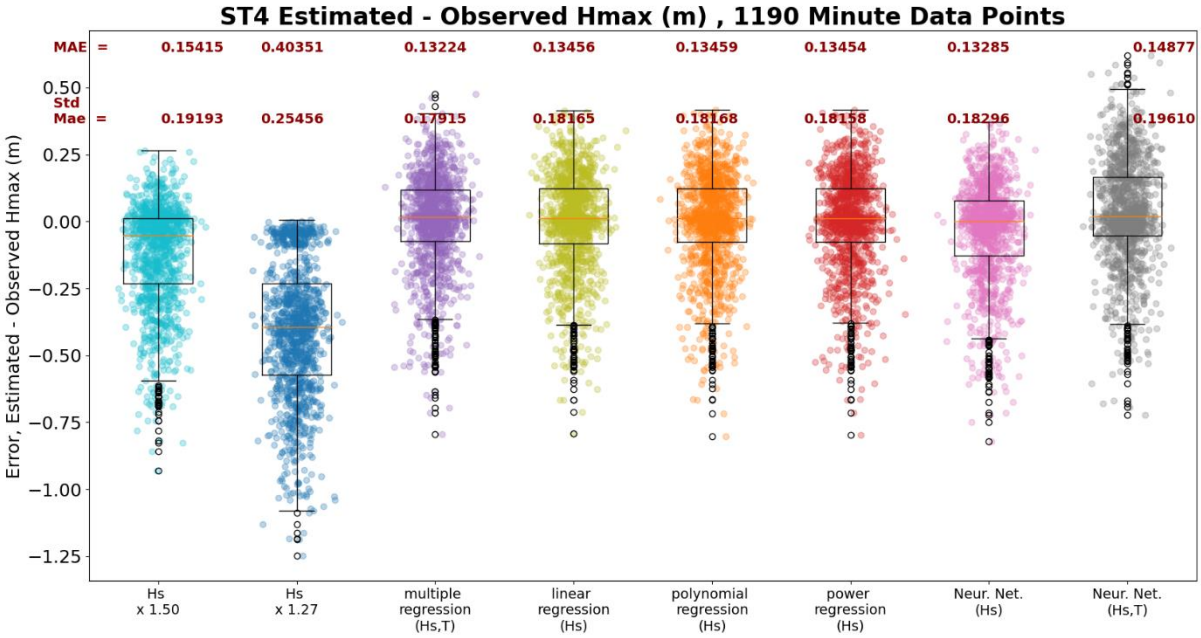


Figure 13i. Boxplot Analysis for ST4

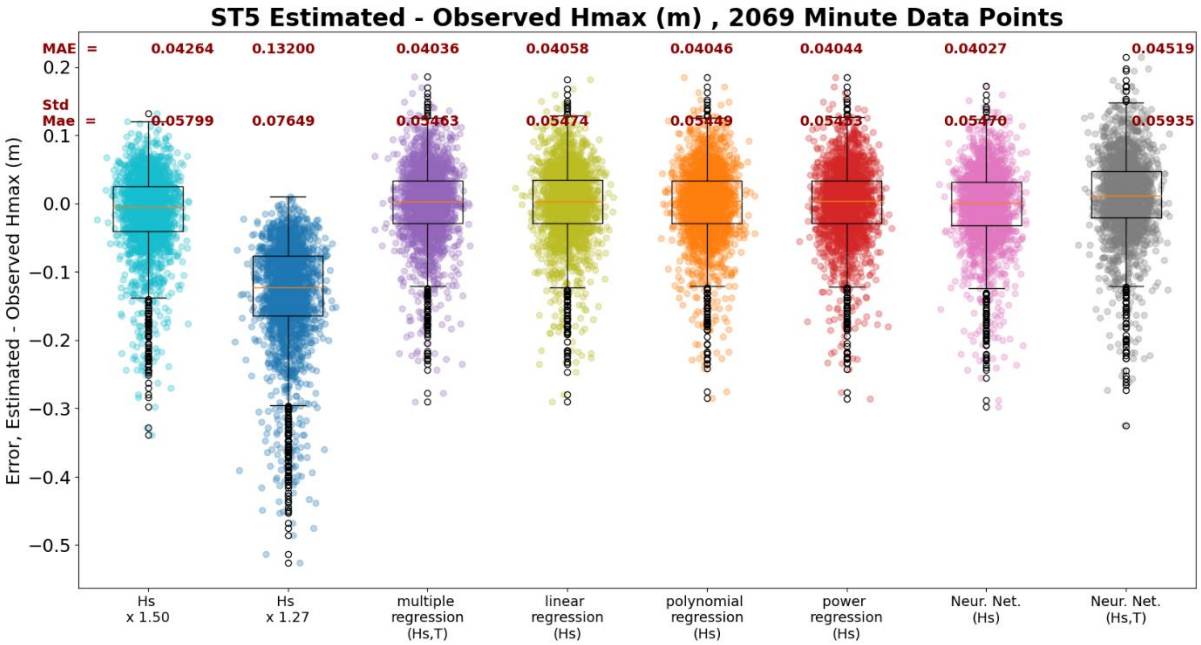


Figure 13j. Boxplot Analysis for ST5

4.2 Regression Analysis

The regression equation between H_{max} as predictand with respect to its predictors are analysed. Regression equations give insight into the relationship between predictand and each predictor. The coefficient of regression describes the rate of change of predictand with respect to each predictor.

Table 10. Regression Equation of $H_{\max}$ at Each In-Situ Observation Sites.

LT1, LT2, LT3, and LT4	
Linear Regression	$H_{\max} = 1.27 \times H_s$
Polynomial Regression	$H_{\max} = 0.00 \times H_s^2 + 1.27 \times H_s + 0.00$
Power Regression	$H_{\max} = 1.27 \times {H_s}^{1.00}$
Multiple Linear Regression	$H_{\max} = 1.27 \times H_s + 0.00 \times T_z - 0.00 \times T_p - 0.00$
LT5	
Linear Regression	$H_{\max} = 1.48 \times H_s - 0.02$
Polynomial Regression	$H_{\max} = 0.02 \times H_s^2 + 1.44 \times H_s - 0.01$
Power Regression	$H_{\max} = 1.45 \times {H_s}^{1.02}$
Multiple Linear Regression	$H_{\max} = 1.46 \times H_s - 0.00 \times Thmax - 0.06 \times Tm01 + 0.05 \times Tm02 - 0.00 \times Tm02a + 0.05 \times Tm02b - 0.00 \times T_p - 0.16$
ST1	
Linear Regression	$H_{\max} = 1.57 \times H_s - 0.03$
Polynomial Regression	$H_{\max} = 0.55 \times {H_s}^2 + 0.39 \times H_s + 0.59$
Power Regression	$H_{\max} = 1.54 \times {H_s}^{1.02}$
Multiple Linear Regression	$H_{\max} = 1.62 \times H_s - 0.05 \times T_z + 0.01 \times T_p + 0.03$
ST2	
Linear Regression	$H_{\max} = 1.63 \times H_s - 0.09$
Polynomial Regression	$H_{\max} = 0.60 \times {H_s}^2 + 0.50 \times H_s + 0.43$
Power Regression	$H_{\max} = 1.55 \times {H_s}^{1.06}$
Multiple Linear Regression	$H_{\max} = 1.64 \times H_s - 0.01 \times T_z + 0.02 \times T_p - 0.14$

Table 10. (Continued)

ST3	
Linear Regression	$H_{\max} = 1.54 \times H_s - 0.00$
Polynomial Regression	$H_{\max} = -0.08 \times {H_s}^2 + 1.74 \times H_s - 0.09$
Power Regression	$H_{\max} = 1.55 \times {H_s}^{0.99}$
Multiple Linear Regression	$H_{\max} = 1.47 \times H_s + 0.05 \times T_z - 0.00 \times T_p - 0.12$
ST4	
Linear Regression	$H_{\max} = 1.62 \times H_s - 0.03$
Polynomial Regression	$H_{\max} = 0.06 \times {H_s}^2 + 1.52 \times H_s - 0.01$
Power Regression	$H_{\max} = 1.58 \times {H_s}^{1.04}$
Multiple Linear Regression	$H_{\max} = 1.88 \times H_s - 0.08 \times T_z - 0.02 \times T_p + 0.22$
ST5	
Linear Regression	$H_{\max} = 1.63 \times H_s - 0.05$
Polynomial Regression	$H_{\max} = 0.14 \times H_s^2 + 1.47 \times H_s - 0.01$
Power Regression	$H_{\max} = 1.59 \times {H_s}^{1.07}$
Multiple Linear Regression	$H_{\max} = 1.56 \times H_s + 0.06 \times T_z - 0.00 \times T_p - 0.17$

Legend	
Variable Name in Regression Equation	**Description of Variable Name**
$H_{\max}$	Maximum Wave Height (m)
H_s	Significant Wave Height (m)
T_z	Zero-crossing wave period (s)
T_p	Peak wave period (s)
$Thmax$	Period of the highest wave (s)
$Tm01$	Estimated mean wave period in respect to fundamental zeroth moment and first moment, m1
$Tm02$	Estimated mean wave period in respect to fundamental zeroth moment and first moment, m2

Table 10. indicates that long-term wave observations near the shore, and in shallow waters such as LT1, LT2, LT3, and LT4 have the same regression equation. For these sites, the H_{max} is equal to 1.27 times the H_s. On the other hand, the LT5 site is in the open seas and in deeper waters. LT5 has long term records too. Compared to the LT1-LT4 sites in shallow waters, the LT5 site has a higher ratio of H_{max} to H_s of approximately 1.50. For both LT1-LT4 and LT5 long term records, the regression coefficient of wave periods is negligible in comparison with the H_s. LT1-LT4 sites showed zero regression coefficient (to 2 decimal places) for zero-crossing wave period, and peak wave period with respect to H_s. Meanwhile, LT5 mean period Tm01, Tm02, and Tm02b showed regression coefficients of -0.06, 0.05, and 0.05 respectively. They are 24.3, and 29.2 times smaller than the regression coefficient for H_s. This indicates that the wave periods do not influence the H_{max}.

Meanwhile, sites with short-term wave observations (at most 5 days) during strong winds and rough seas, such as ST1, ST2, ST3, ST4 and ST5 have a higher ratio of H_{max} to H_s closer to 1.60. Nonetheless, consistent with long-term wave observations, the zero-crossing wave period (T_z) and peak wave period (T_p) does not influence the (H_{max}). This is indicated by the very small regression coefficients of T_p and T_z.

4.3 K-Fold Stratified Cross Validation

The methods outlined in this study are used to estimate the H_{max}. Subsequently, the estimated H_{max} is compared against the observed H_{max}. A widely used method of comparing an estimate with the ground truth is by cross-validation.

It is not possible to use the same data to train the model and validate the model. This will cause overfitting whereby the model is specifically trained only for that dataset. Even noise and other errors will be modelled. Although the overfitted model may have very low error, it is only with respect to the dataset it has trained with. When that model is deployed on datasets it has not seen, such as in real-time operation, that model will fail.

To prevent this issue, the model needs to be validated or compared against dataset excluded from training. In this study, the entire dataset is first shuffled randomly. Random shuffling increases the likelihood that each subset of dataset population has approximately the same distribution as other subsets of the population. This is also known as stratification.

Then, the dataset is split into k = 10 subsets. The model is trained using k-1 subsets and validated against the remaining subset. Subsequently, the error, E is calculated. This procedure is repeated until each k subset has been part of the validation dataset. As a result of stratification by random shuffling, each k subset population density should represent the population dataset. The 10-fold stratified cross-validation is applied in this study. **Figure 14**. depicts the implementation of 10-

fold cross-validation used in this study. **Equation (32)** shows the average mean absolute error after 10-fold cross-validation, where i refers to the i-th subset of the data, which is divided into 10 subsets.

$$MAE_{cross\ validated} = \frac{1}{10}\sum_{i=1}^{10}|Forecast - Observed|_i \quad -(32)$$

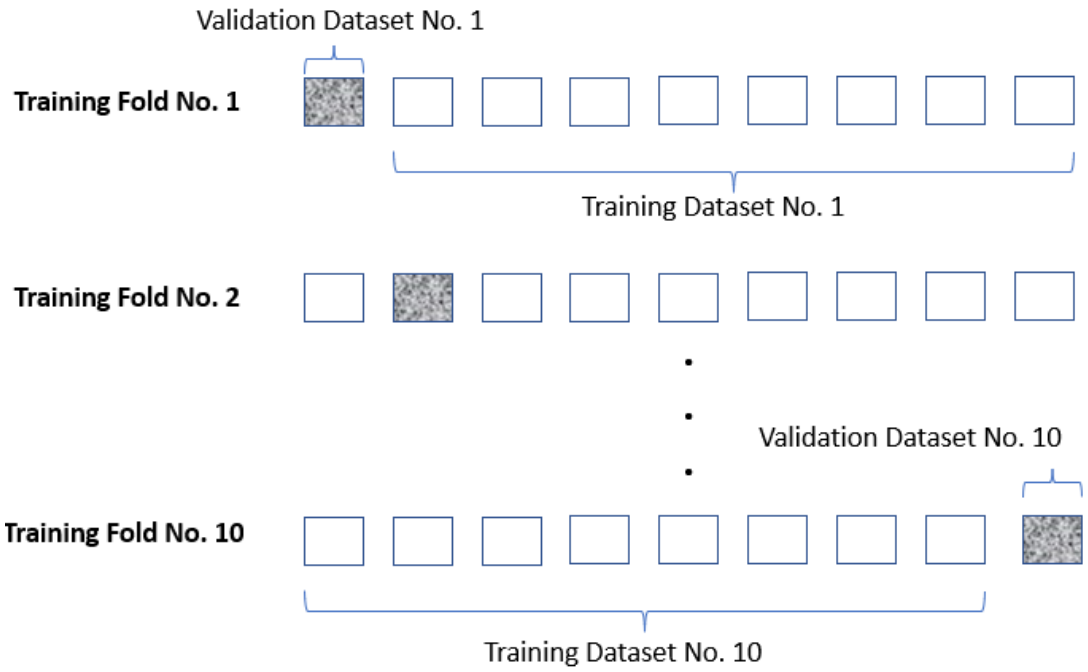


Figure 14. 10-fold cross validation. Dataset of each station are randomly shuffled and split into 10 subsets. The model is trained 10 times using 9 of the subsets and validated against the remaining 1 subset.

Based on 10-fold cross validation, the MAE is calculated for each method of estimating H_{max}, that is H_s times 1.50, H_s times 1.27, multiple regression, linear regression, polynomial regression, power regression, neural network with respect to H_s, and neural network with respect to both H_s and wave period. The MAE for each method at each in-situ observation site is tabulated in **Figure 15**.

For the LT1, LT2, LT3, and LT4 stations that have long-term wave observations and are in shallow waters, the method of regression and method of neural network with respect to just H_s has the least MAE. The simplest method of estimating H_{max} for these stations is by multiplying H_s by 1.27.

On the other hand, the worst method of estimating H_{max} for stations with short-term wave observations at conditions of strong winds and rough seas, is multiplication by 1.27. This is because the ratio between H_{max} and H_{sig} is observed to be closer to 1.50 – 1.60 for these stations (ST1, ST2, ST3, ST4, and ST5). Method of multiple regression, linear regression, polynomial regression, and neural network (with respect to H_s) are equally the best methods of H_{max} estimation.

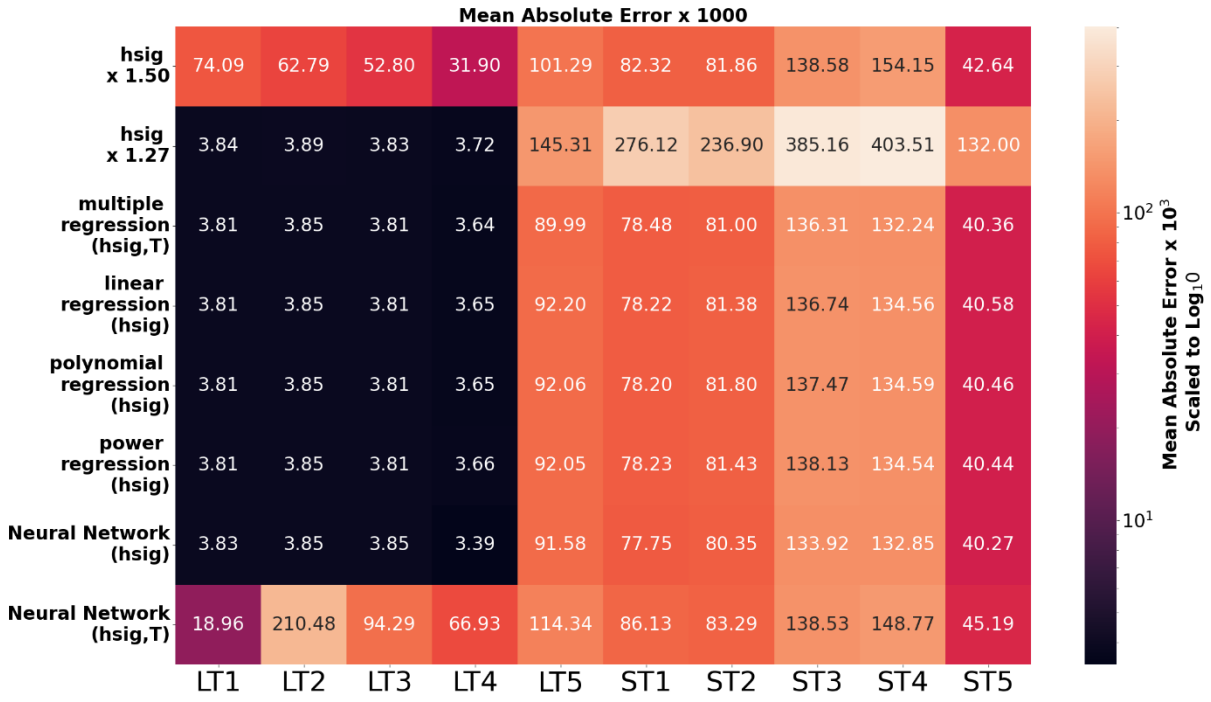


Figure 15. Heatmap indicating Mean Absolute Error x 1000 for each method of estimating H_{max} (vertical axis) with respect to in-situ observation site at the horizontal axis.

5. Conclusion

Ten (10) wave observation sites were used in this study. Four were in shallow waters with depth less than 20m, (LT1, LT2, LT3, and LT4) while the remaining were in deeper waters (depth at least 66m). Four observation sites had long-term data spanning at 10 years (LT1, LT2, LT3, and LT4), one observation site had long-term data spanning 5 years (LT5), and remaining observation site had short-term observations of at most 5 days (ST1, ST2, ST3, ST4, and ST5) during strong winds and rough seas conditions, based on warnings issued by MET Malaysia (refer to **Table 7, Section 2.2**).

The ratio of H_{max} to H_s in shallow waters was 1.27 while the ratio was 1.50 otherwise. On the other hand, the ratio increased to between 1.50-1.60 during strong winds and rough seas conditions. Calculations of H_{max} by Rayleigh mode, mean, and probability $\mu = 0.01$ tended to overestimate observed H_{max} for long-term wave observations. On the contrary, H_{max} calculated by Rayleigh mode, and mean tended to underestimate H_{max} observed during strong winds and rough seas conditions. Rayleigh H_{max} at probability $\mu = 0.01$ tends to slightly overestimate observed H_{max} during strong winds and rough seas condition.

Observed H_{max} is poorly correlated to wind speed, wave periods and number of waves. On the other hand, observed H_{max} is highly correlated with H_s observation. Estimations of H_{max} by linear, polynomial, power, and artificial neural networks with H_s as predictor showed less Mean Absolute Error (MAE) than multiplying H_{max} by a

constant. Nevertheless, the simplest rule of thumb for estimating H_{max} is multiplication of H_s by 1.27 in shallow waters near the shore (LT1, LT2, LT3, and LT4), multiplying H_s by 1.50 in deep waters under normal sea conditions (LT5), and multiplying H_s by between 1.50-1.60 under strong winds and rough sea conditions (warning issued by MET Malaysia) as shown in ST1, ST2, ST3, ST4 and ST5 observation sites. Ratio of H_{max} to H_s may increase when wave distribution is shifted towards the tails (increased kurtosis), during rough seas and strong wind conditions. Kurtosis may be explored to increase accuracy of H_{max} measurement in the future.

More observations are needed for future work. Longer term observations are needed to improve statistical significance of return period analysis. The underlying distribution of H_{max} must be determined empirically using good quality wave observations to accurately determine return period of H_{max}. Knowing the underlying distribution of wave observations is crucial in calculating more accurate H_{max}.

6. Acknowledgement

The authors would like to thank the Technical Weather and Geophysics Division (BTCG) of MET Malaysia, and the Meteorological Instrumentation Division (BIM) of MET Malaysia, for providing long term wave observation data at stations LT1, LT2, LT3, LT4, and LT5.

We also express our appreciation to METOCEAN, PETROLIAM NASIONAL BERHAD (PETRONAS), GROUP TECHNICAL SOLUTIONS (GTS), PETRONAS, and PROJECT DELIVERY AND TECHNOLOGY (PD&T), PETRONAS, for the short-term observation data at stations ST1, ST2, ST3, ST4, and ST5.

7. References

1. Agrawal, J.D., and Deo, M.C. (2004). Wave parameter estimation using neural networks. Marine Structures Volume 17, Issue 7, 536-550, ISSN 0951-8339. https://doi.org/10.1016/j.marstruc.2005.01.001
2. Barbariol, F., Bidlot, J. R., Cavaleri, L., Sclavo, M., Thomson, J., & Benetazzo, A. (2019). Maximum wave heights from global model reanalysis. Progress in oceanography, 175, 139-160. https://doi.org/10.1016/j.pocean.2019.03.009
3. Chun, H., & Suh, K. D. (2019). Empirical formulas for estimating maximum wave height and period in numerical wave hindcasting model. Ocean Engineering, 193, 106608. https://doi.org/10.1016/j.oceaneng.2019.106608
4. Dozat, T. (2016). *Incorporating Nesterov Momentum into ADAM.* Workshop track- ICLR 2016. Retrieved from https://openreview.net/pdf?id=OM0jvwB8jIp57ZJjtNEZ
5. Feng, X., Tsimplis, M. N., Quartly, G. D., & Yelland, M. J. (2014). Wave height analysis from 10 years of observations in the Norwegian Sea. Continental Shelf Research, 72, 47-56. http://dx.doi.org/10.1016/j.csr.2013.10.013
6. Forristall, G.Z. (1978). *On the Statistical Distribution of Wave Heights in a Storm.* Journal of Geophysical Research, 83 (C5). Paper number 7C1100. 0148-0227/78/057C-1100803.00. DOI: 10.1029/JC083iC05p02353
7. Goda, Y. (2000). Random Seas and Design of Maritime Structures. Advanced Series on Ocean Engineering (15). World Scientific Publishing Co. Pte. Ltd. ISBN 981-02-3256-X.
8. Janssen, P. A. E. M., & Bidlot, J. R. (2009). On the extension of the freak wave warning system and its verification (p. 42). Reading, UK: European Centre for Medium-Range Weather Forecasts. Retrieved on 16 December 2022 from https://www.ecmwf.int/sites/default/files/elibrary/2009/10243-extension-freak-wave-warning-system-and-its-verification.pdf
9. Krogstad, H. E. (1985). Height and period distributions of extreme waves. *Applied Ocean Research*, *7*(3), 158-165. doi: 10.1016/0141-1187(85)90008-2
10. M. S. Longuet-Higgins, “On the statistical distributions of the heights of sea waves,” J. Marine Res. IX (3) (1952), pp. 245-266.
11. Muraleedharan, G., Rao, A. D., Kurup, P. G., Nair, N. U., & Sinha, M. (2007). Modified Weibull distribution for maximum and significant wave height simulation and prediction. Coastal Engineering, 54(8), 630-638. doi: 10.1016/j.coastaleng.2007.05.001
12. Rorbaek, K., & Andersen, H. (2000, September). Evaluation of wave measurements with an acoustic Doppler current profiler. In OCEANS 2000 MTS/IEEE Conference and Exhibition. Conference Proceedings (Cat. No. 00CH37158) (Vol. 2, pp. 1181-1187). IEEE. DOI: 10.1109/OCEANS.2000.881761

13. Vandever, J. P., Siegel, E. M., Brubaker, J. M., & Friedrichs, C. T. (2008). Influence of spectral width on wave height parameter estimates in coastal environments. Journal of waterway, port, coastal, and ocean engineering, 134(3), 187-194. DOI: 10.1061/(ASCE)0733-950X(2008)134:3(187)
14. Zhuo, Z., & Sato, S. (2015). Characteristics of Wave Grouping and Freak Wave Observed by Two Typhoons. Procedia Engineering, 116, 277-284. https://doi.org/10.1016/j.proeng.2015.08.291

MALAYSIA METEOROLOGICAL DEPARTMENT
JALAN SULTAN
46667 PETALING JAYA
SELANGOR DARUL EHSAN
Tel : 603-79678000
Fax : 603-79550964
www.met.gov.my

ISBN 978-967-2327-10-3