 doi:10.20944/preprints202012.0538.v1

Article

LiDAR Observations of Multi-modal Swash Probability Distributions on a Dissipative Beach

Caio Eadi Stringari [1,†,‡,*] and Hannah E. Power [1,‡]

[1] School of Environmental and Life Sciences, University of Newcastle, Australia
* Caio.EadiStringari@uon.edu.au
† Current address: France Energies Marines, Plouzané, France
‡ These authors contributed equally to this work.

Abstract: Understanding swash zone dynamics is of crucial importance for coastal management as the swash motion, consisting of the uprush of the wave on the beach face and the subsequent downrush, is responsible for driving changes the beach morphology trough sediment exchanges between the sub-aerial and sub-aqueous beach. Improved understanding of the probabilistic characteristics of these motions has the potential to allow coastal engineers to develop improved sediment transport models which, in turn, can be further developed into coastal management tools. In this paper, novel descriptors of swash motions are obtained by combining field data and statistical modelling. Our results indicate that the probability distribution function (PDF) of shoreline height ($p(\zeta)$) and trough-to-peak swash heights ($p(\rho)$) measured at a high energy, sandy beach were both inherently multimodal. Based on the observed multimodality of these PDFs, Gaussian Mixtures are shown to be the best method to statistically model them. Further, our results show that both offshore and surf zone dynamics are responsible for driving swash zone dynamics, which indicates unsaturated swash. The novel methods and results developed in this paper, both data collection and analysis, could aid coastal managers to develop improved swash zone models in the future.

Keywords: LiDAR, swash zone, nearshore waves, probability distributions, sandy beaches.

1. Introduction

The swash zone encompasses the transition region between the sub-aqueous and the sub-aerial beach [1]. It is a highly dynamic environment with alternating wet and dry conditions. Over the past five decades, this region has attracted increased research interest due to the significant role it plays in sediment dynamics and beach erosion [2]. The cross-shore shoreline oscillation, globally referred as to swash, can be divided into two main components: the uprush of the wave on the beach face and the subsequent downrush. Each of these two divisions can be described by their horizontal and vertical (height) components (Figure 1). A large proportion of swash zone research has focused on obtaining

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

empirical parametric formulae to describe extreme runup heights (see Atkinson *et al.* [3] and Power *et al.* [4] for recent reviews), however, as highlighted by Hughes *et al.* [5], this approach may not be fully satisfactory to provide information to coastal managers to develop operational tools other than inundation models.

The natural variability of the probability distribution functions (PDFs) of swash motions has received little research attention to date. To the authors' knowledge, only two studies have attempted to describe the variability of these PDFs in detail [5,6] both of which focused on comparing measured shoreline elevation height PDFs ($p(\zeta)$) to Cartwright and Longuet-Higgins' [7] theoretical PDF of a random variable. This theoretical PDF is a direct function of the spectral width (φ) of the analysed timeseries and reduces to the Rayleigh PDF for narrow bandwidth processes (φ=0) or to the Gaussian PDF for wide bandwidth processes (φ=1). Holland and Holman [6] found that their measured shoreline height PDFs matched Cartwright and Longuet-Higgins'[7] PDF for some values of φ, but they could not correlate the variability in their PDFs to environmental parameters. More recently, Hughes *et al.* [5] showed that both their shoreline height PDFs ($p(\zeta)$) and trough-to-peak swash height PDFs ($p(\rho)$) resembled, but were not statistically similar, to Cartwright and Longuet-Higgins' [7] theoretical PDFs. Hughes *et al.* [5] observed that, on aveage, PDFs were consistently right skewed when compared to the theory possibly due to the broad-band wave spectrum observed on natural beaches. It has been observed [8], however, that the spectral width parameter φ does not correlate with wave height PDFs in the surf zone and, by induction, should not correlate with swash heights either.

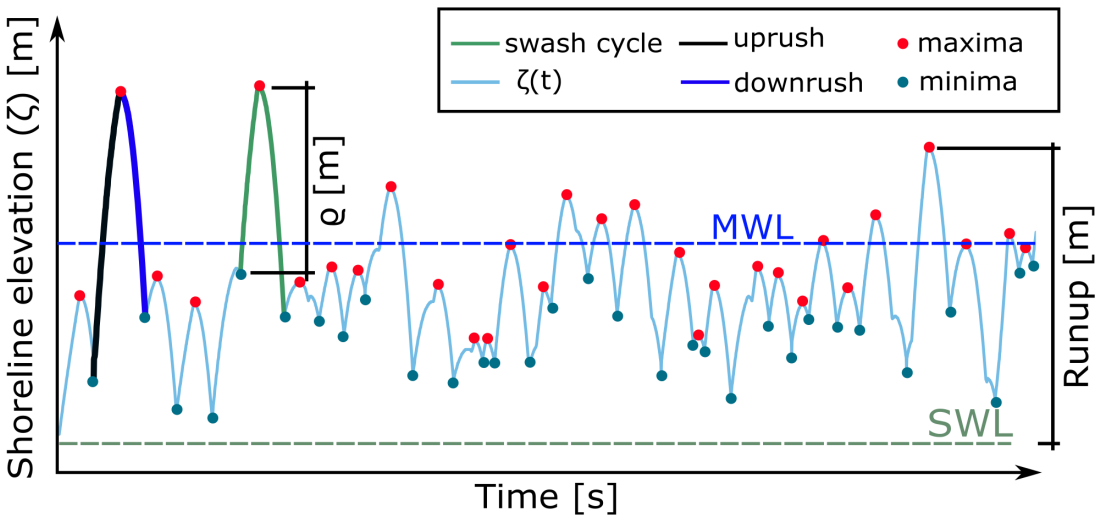

Figure 1. Swash zone definitions. Note that runup is defined relative to the still water level (SWL), the shoreline height (or elevation, ζ) is defined centered on the mean water level (MWL), and the trough-to-peak swash height (ρ) is defined for each swash cycle. Here, each swash cycle was defined by a local minima analysis.

In this paper, we provide novel field observations of swash motion PDFs obtained from a high resolution LiDAR system deployed at a high energy sandy beach. Specifically, the statistical properties

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

of shoreline height (ζ) and trough-to-peak swash heights (ρ) are investigated in detail. In addition, detailed surf zone and offshore data are used to the assess the patterns observed in the swash zone. The results obtained here deviate significantly from the theoretical predictions from Cartwright and Longuet-Higgins [7] and do not support the concept of swash saturation [9]. The present data indicate that a combination of surf zone and offshore forcing control swash zone dynamics. Finally, an approach that does not require prior knowledge of φ is investigated to predict the characteristics of swash motion PDFs. This paper is organised as follows. Section 2 describes the data collection methods and pre-processing, Section 3 presents the results of the field data collection with a focus on probabilistic descriptors for swash, Section 4 discusses the results, and Section 5 concludes.

2. Materials and Methods

2.1. Data Collection

Data were collected at at Seven Mile Beach, Gerroa, New South Wales, Australia, hereafter SMB. This beach is classed as modally dissipative in the Australian morphodynamic beach model [10,11] and for the duration of the present experiment, was characterised by a gently sloping profile (with slope $\beta = 0.03$) with no significant barred morphology, beach cusps, or alongshore variability. Video imagery, Pressure Transducer (PT), offshore spectral wave, Light imaging Detection And Ranging (LiDAR), Acoustic Doppler Velocimeter (ADV), and topographic data were collected during a field data collection experiment over six days in June 2018. In this paper, the focus will be on the LiDAR and offshore data. The experimental design is shown in Figure 2 and is summarised below. See Stringari *et al.* [12] and Stringari and Power [13] for further details.

The PT data collection consisted of 30 PTs (RBR Solo and INW P2X) deployed on the seabed in a cross-shore orientation. The LiDAR (SiCK LMS511) was mounted on a scaffold frame and recorded in the same cross-shore orientation as the PT transect (see Figure 2-a and b). A Datawell waverider buoy was deployed offshore of the transect line at the 10m isobath. The beach was surveyed several times each day using a Trimble S5 total station and a Trimble R4 RTK GPS, and representative beach profile is shown in Figure 2-c. Figure 3 shows the offshore conditions for the duration of the field campaign. This dataset was ultimately chosen for the analyses presented in this paper because it overcomes the limitations of classical remote-sensing datasets (e.g., pixel miss-registration; see Vousdoukaset al.'s [14] Figure 7), it has precise and unique offshore conditions, and it has a high degree of offshore wave variability for comparable tidal water levels and beach slopes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

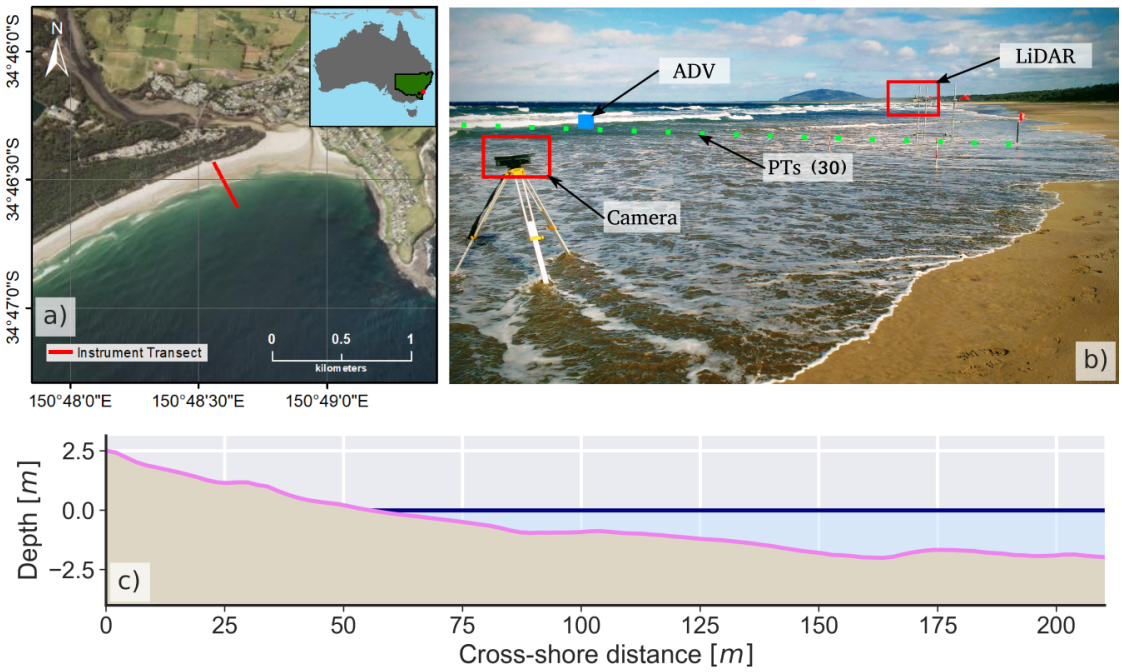

Figure 2. a) Experiment location. The red line shows the instrumentation transect. b) Photo of the experimental setup (19/06/2018). c) Representative beach profile (16/08/2018).

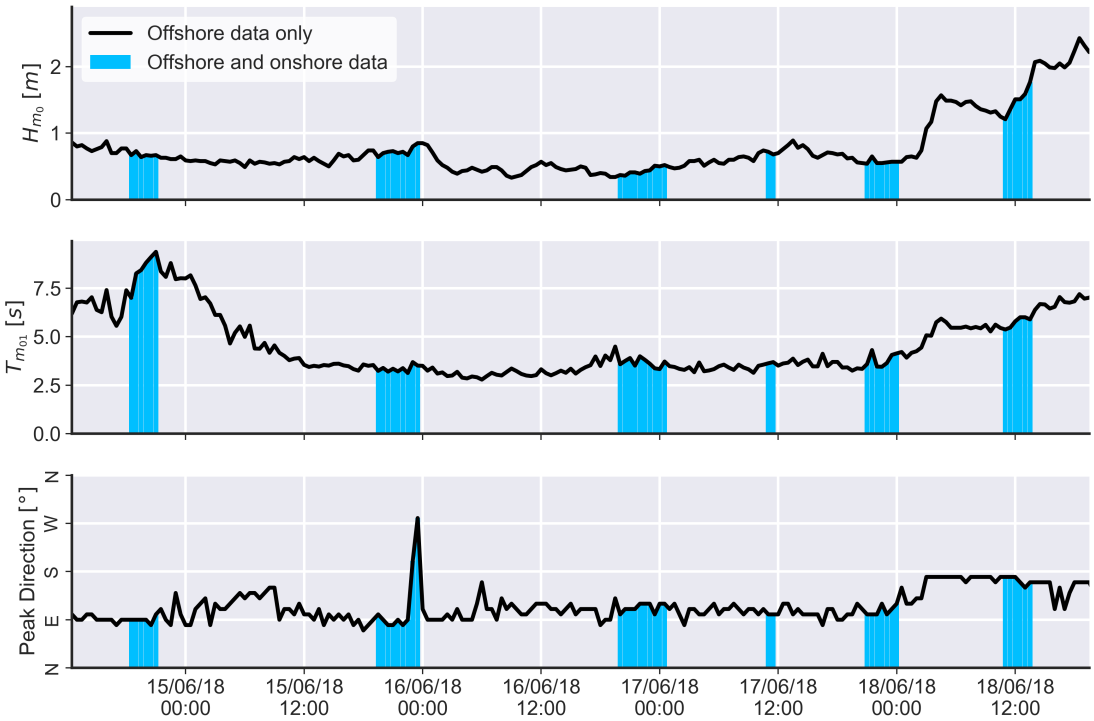

Figure 3. Offshore data (spectral significant wave heights, H_{m0_∞}, and periods, T_{m01_∞}), for the duration of the field campaign. The filled blue regions indicate periods of simultaneous offshore, LiDAR, and PT data collection.

 doi:10.20944/preprints202012.0538.v1

2.2. Data Processing

From the raw LiDAR data (Figure 4-a), timeseries of the cross-shore evolution of the water surface elevation were extracted at 10Hz and stacked in time, resulting a dataset similar to a video-derived timestack [15] (Figure 4-b, color scale). Incoming waves were tracked using a modified version of the method from Stringari, Harris and Power [12]. For each LIDAR timestack, the vertical Sobel [16] edge detector was applied to the timestack, pixel intensity peaks in the resulting image were extracted and then clustered using the DBSCAN algorithm [17]. Unlike the original method, no color-based machine learning was applied to the dataset. The absence of the color-learning step resulted in a significant increase in the number of false-positive cases of wave crests being detected. These erroneous wave crests were manually corrected in QGIS to ensure that no errors were propagated into subsequent analyses. Optimal wave paths were then obtained as per the original tracking algorithm [12]. The tracking algorithm was set to stop tracking waves if the water elevation above the bed was less than 0.015m, which is significantly lower than other recent works [18,19].

The temporal evolution of the shoreline position was obtained in three steps: 1) the uprush was obtained from the tracked wave paths as described above, 2) the downrush was obtained via edge detection using the horizontal Sobel [16] operator and, 3) the continuous shoreline timeseries was obtained by combining the results of steps 1) and 2) and interpolating the data a to regular time vector with sample frequency of 5Hz using a Gaussian Radial Basis Function (RBF) interpolation. Finally, horizontal shoreline excursion timeseries were converted to shoreline height (ζ) timeseries and to trough-to-peak swash heights (ρ) using the measured beach profiles and a local minima analysis (see Figure 1 for definitions). The Australian Height Datum (AHD, [20]) was used as the vertical reference.

3. Results

3.1. Surf Zone Dynamics

The cross-shore variation of surf zone significant wave heights (H_{m_0}) and the fraction of broken waves (Q_b) were used to assess the surf zone dynamics (Figure 5). In this paper, Q_b was calculated as follows: for each data run in which there were unique offshore, surf zone, and swash zone data available, 10 minutes of PT data were extracted from the raw records and, from these records, individual waves were extracted using a local minima analysis as per Power *et al.* [21] and classified as broken or unbroken using the neural network from Stringari and Power [13]. The neural network was updated with field data from Seven Mile Beach to increase the classification performance for the present dataset. The updated neural network accuracy score reached 95% when classifying waves in a test dataset

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

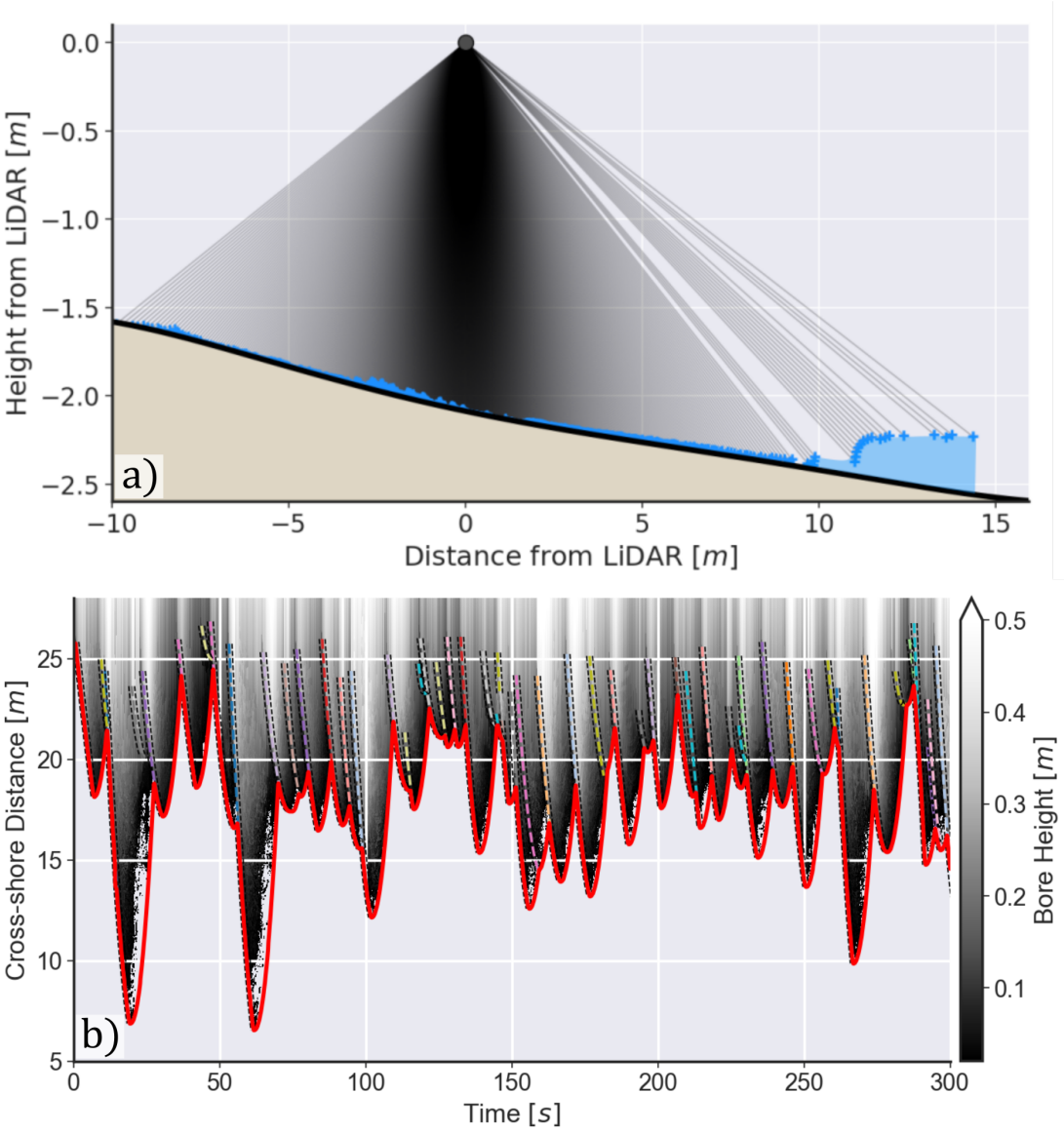

Figure 4. a) Raw LiDAR data showing a bore running up the beach profile. b) Example of LiDAR timestack showing the tracked wave paths (coloured dashed lines) and the resulting time-varying shoreline position (thick red line). The grey scale indicates the bore height (that is, water depth) in relation to the measured profile in a).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

(that is, data that the neural network had never seen) and correlation scores (r^2) were >0.95 for Q_b predictions (not shown).

Data from each Q_b curve were segmented into three clusters using the *k-means* algorithm (Figure 5-a): one cluster that was representative of the outer surf zone, one cluster representative of the mid surf zone, and one cluster representative of the inner surf zone. The probability distribution of Q_b ($p(Q_b)$) was then calculated for each class (Figure 5-b) which showed that in the outer surf zone, most of the waves were unbroken ($p(Q_b) < 0.2$), in the mid surf zone, about half of the waves were broken ($p(Q_b) \approx 0.5$), and in the inner surf zone, most waves were broken ($p(Q_b) > 0.8$). This result is consistent with the conceptual hydro-kinematic model for gently sloping beaches [22]. Interestingly, Q_b values close to the surf-swash boundary were never $Q_b = 1$, which indicates that small unbroken waves reach the swash zone, even on a dissipative beach such as SMB (average Iribarren Number [23] $\xi_\infty = \frac{\tan\beta}{\sqrt{H_{m0_\infty} L_\infty}} = 1.21$, where L_∞ is the wave length calculated as $L_\infty = \frac{g}{2\pi} T^2_{m_{01_\infty}}$, and averaged $\Omega_\infty = \frac{H_{m0_\infty}}{T_{m_{01_\infty}} W_s} = 3.77$, where W_s is the sediment fall velocity). Based on the observed distributions of Q_b, three locations in the surf zone were chosen to assess wave heights: $Q_b = 0.95$ (inner surf zone), $Q_b = 0.50$ (mid surf zone), and $Q_b = 0.05$ (outer surf zone).

The analysis of the correlation between offshore (H_{m0_∞}) and surf zone (H_{m_0}) wave heights showed that there was a direct correlation between H_{m_0} and H_{m0_∞} across the full width of the surf zone (Figure 5-c to f). Following the definition of surf zone saturation from Power *et al.* [24], the observed correlations strongly suggest that the surf zone was unsaturated during experiment, despite the dissipative nature of the beach. Finally, the wave height to water depth ratio (γ_{sig}) was compared to the offshore wave height normalised by averaged water depth for each 10-min data run ($H_{m0_\infty}/\overline{h}$) (Figure 5-g). The results from this analysis are analogous to Figure 11 in Power *et al.* [21] and indicate that: 1) the surf zone was unsaturated, and 2) there was a terminal bore height reaching the surf-swash boundary. Following from the analysis in Figure 5-a that $Q_b \neq 1$, this terminal bore height could represent either broken or unbroken waves. These results are significant because if the surf zone is unsaturated, it is probable that the swash zone is also unsaturated. This is discussed further in Section 4.

3.2. Shoreline Height PDFs

For each data run in which there were unique offshore and LiDAR data, 10-minute timeseries were extracted from the raw LiDAR record and the time-varying shoreline was obtained using the method described in Section 2.2. The PDFs of normalised ($p((\zeta - \mu)/\sigma)$) and non-normalised ($p(\zeta)$) shoreline height were then obtained via histograms and kernel density estimations (KDEs). The use of KDEs to obtain PDFs is advantageous over more traditional histogram methods because: 1) they are a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

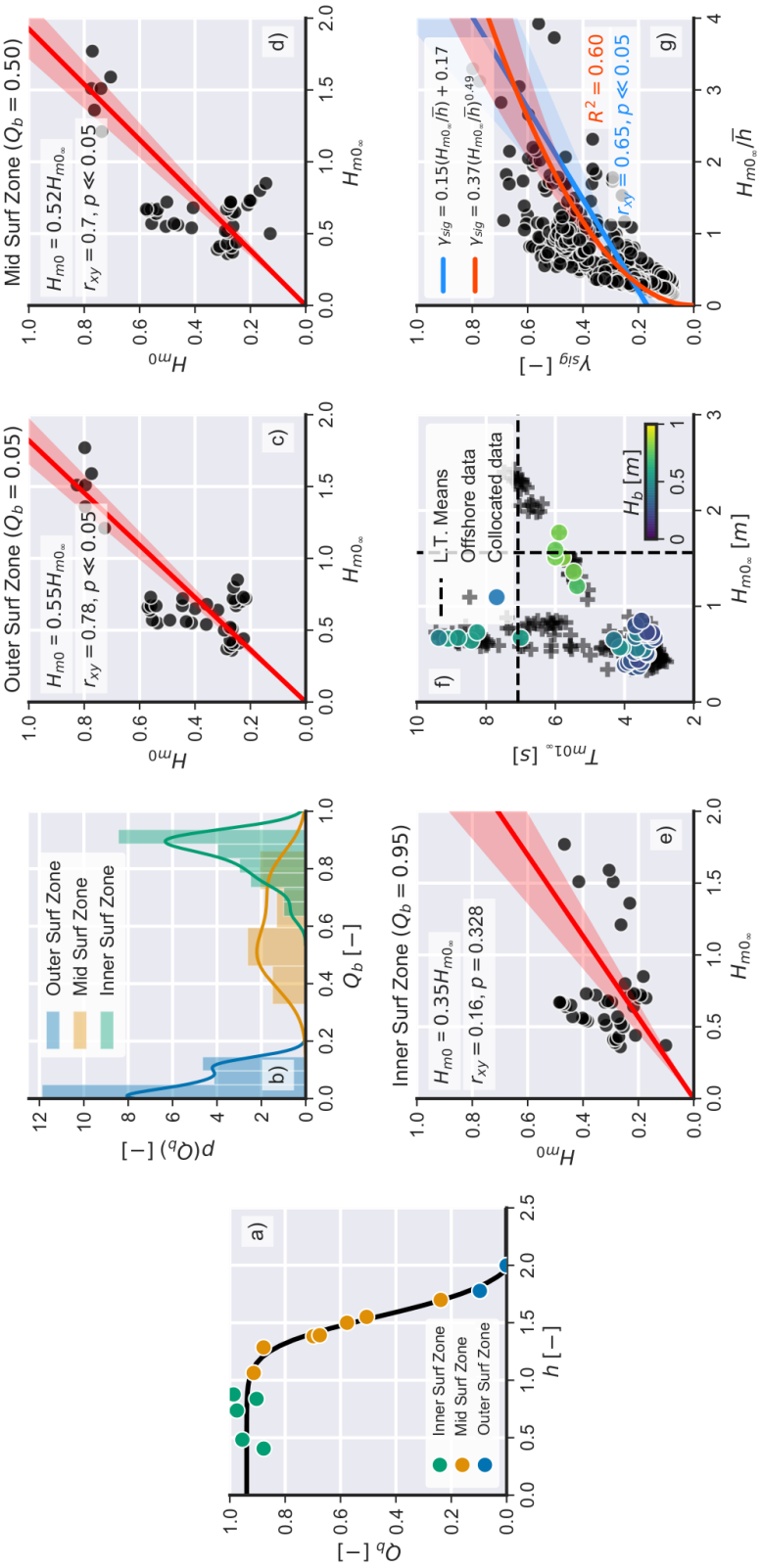

Figure 5. a) Example of Q_b curve segmentation using *k-means* for one 10-minute data run. b) Algorithm clustered Q_b PDFs. The number of bins for these histograms were calculated using the Freedman and Diaconis [25] rule. Correlations between offshore (H_{m0_∞}) and surf zone (H_{m_0}) wave heights in c) the outer surf zone, d) the mid surf zone, and e) the inner surf zone. The red swath shows the 95% confidence interval for the linear regression. f) Comparison between offshore conditions (H_{m0_∞} and T_{m01_∞}) and break point wave height (H_b). The crosses show all the measured offshore data, and the dashed lines show the long term H_{m0_∞} and T_{m01_∞} averages for the nearest offshore wave buoy (Port Kembla) [26]. g) Analysis of γ_{sig} against $H_{m0_\infty}/\bar{h}$ (analogous to Figure 11 in Power *et al.* [21]). The coloured swaths show the for 95% confidence interval for the regressions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

fully non-parametric approach, and 2) they are able to identify fluctuations in the data's distribution that are usually not seen when using histograms with potentially non-ideal bin sizes.

Analysis of individual normalised shoreline height PDFs indicated a high degree of variability between runs and that the majority of PDFs were multimodal (97.5%). The Shapiro and Wilk [27] test at the 95% confidence interval confirmed that none of the analysed timeseries were normally distributed. The observed PDFs were then grouped into three clusters based on the observed offshore conditions (Figure 6-a). Cluster A represents average wave heights conditions with short periods, Cluster B represents calm conditions (low wave heights and short wave periods), and cluster C represents calm conditions with long wave periods. Figure 6-b shows the PDFs for cluster A, Figure 6-c shows the PDFs for cluster B and Figure 6-d shows the PDFs for cluster C. By averaging the PDFs in each cluster, a right-skewed PDF similar to Hughes et al.'s [5] ensemble PDF was observed (see their Figure 2). To assess the effect of the normalisation strategy and offshore conditions on the shape of the shoreline height PDFs, each PDF was compared to every other PDF in the same cluster, and then to every PDF in each of the other two clusters using the Kullback and Leibler [28] divergence as similarity measurement. The results from this analysis indicated that: 1) non-normalised PDFs ($p(\zeta)$) are dissimilar within and between clusters (Figure 6-h), except PDFs in cluster C which were strongly similar to each other, and 2) normalised PDFs ($p((\zeta-\mu)/\sigma)$) are strongly similar within and between clusters (Figure 6-i). For further discussion see Section 4.

Two other methods were assessed for obtaining a function (or combination of functions) to describe the observed PDFs. This was done because KDE is a non-parametric method that requires prior knowledge of the input timeseries, thus preventing an assessment of correlations between descriptors of the analysed PDFs and environmental parameters. Note that the results presented below were invariant regardless of which PDF ($p((\zeta-\mu)/\sigma)$ or $p(\zeta)$) was being modelled. The first method consisted of fitting all PDFs available in the SciPy library [29] to the observed data (96 PDFs were available as of December 2020) and using three metrics to evaluate the fitted PDFs: the sum of squared errors, the Akaike information criterion [30], and the Kullback-Leibler divergence [28]. The results from these analyses indicated that none of the best-fit PDFs were able to statistically satisfactorily describe the majority ($>50\%$) of the observed PDFs, regardless of the metric adopted to rank them. The analytical PDF that best fitted the greatest number of observed PDFs ($\approx 35\%$) was the non-central Student's T (NCT) PDF, which is a complicated four-parameter function [31] and thus is impractical. Examples of the NCT fit to the data are shown in Figure 6-e to g (blue lines). Given the poor overall performance of the NCT and given that this PDF cannot describe the multimodal characteristics of the data, this strategy was not pursued further.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

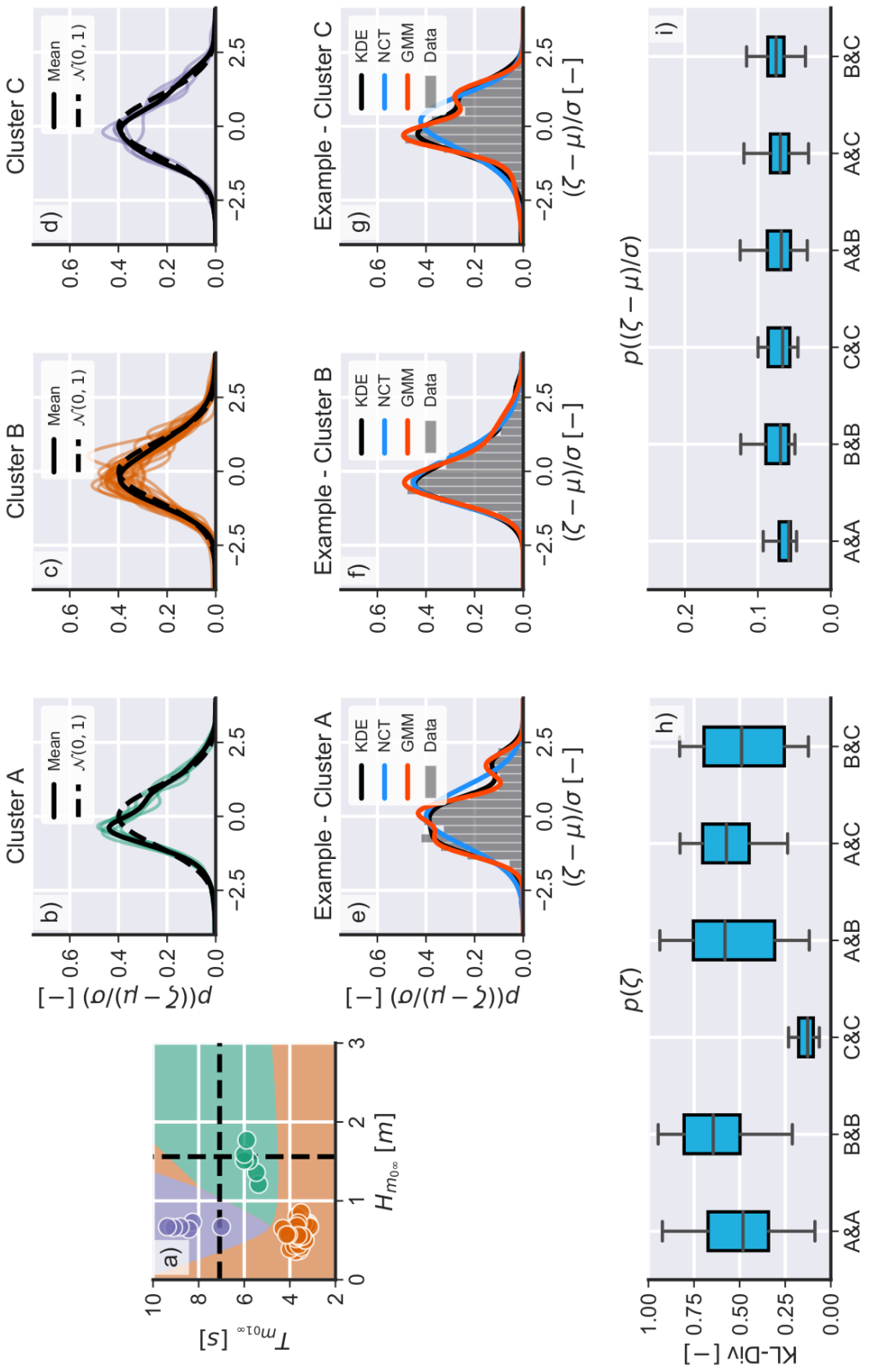

Figure 6. a) Clustering analysis of offshore wave conditions (H_{m0_∞}, T_{m01_∞}). The dashed lines show the long term H_{m0_∞} and T_{m01_∞} averages for Port Kembla wave buoy [26]. KDE approximations, mean KDEs, and standard Gaussian PDF ($\mathcal{N}(0,1)$) for b) cluster A , c) cluster B and d) cluster C. Representative examples of PDFs for e) cluster A, f) cluster B and g) cluster C showing the KDE approximations (black), NCT fits (blue), and GMM approximations (red). The number of bins for these histograms were calculated using the Freedman-Diaconis rule [25]. Analysis using the Kullback and Leibler [28] divergence (KL-Div) for h) non-normalised PDFs ($(p(\zeta))$, and i) normalised ($p((\zeta-\mu)/\sigma)$) to assess PDF similarity within each cluster and between pair of clusters. In h) and i) lower values indicate more similar PDFs. Note that the KL-Div has no upper-bound value.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

To account for the multimodality observed in the data, a second approach to obtain analytical descriptions of $p(\zeta)$ and $p((\zeta-\mu)/\sigma)$ was used. In this method, the analysed PDFs were approximated by the sum of a number of Gaussian PDFs, each described individually by their mean (μ), standard deviation (σ) and mixing weight (α), that is, a Gaussian Mixture Model (GMM) [32]. This approach was able to precisely reproduce the observed multimodality in all the shoreline height PDFs (see Figure 5-e to g, for example) and the Kolmogorov–Smirnov test confirmed that the PDFs predicted by the GMMs were statistically similar to the observed PDFs at the 95% confidence level (which is expected, given the characteristics of the method). A Gaussian mixture model is, however, a parametric method that requires prior knowledge of the number of mixtures to be used. By using the Akaike Information Criterion [30] as an evaluation metric, a mixture with three components was found to be the optimal value to statistically satisfactorily represent the majority of the observed data ($\geq$90%) whilst maintaining model simplicity. As GMMs provide the parameters μ, σ, α and the optimal number of mixtures (N_{mix}), it becomes possible to correlate these parameters to known variables in a predictive way, thus overcoming the major limitation of KDEs and methods such as Cartwright and Longuet-Higgins' [7]. A model using surf zone and offshore parameters to assess the variability observed in $p(\zeta)$, assuming that such variability is directly correlated to the optimal number of Gaussian mixtures (N_{mix}) for each PDF, is discussed in Section 4.

3.3. Trough-to-peak Swash Height PDFs

In the previous section, $p(\zeta)$ was observed to be multimodally distributed and, consequently, to deviated from the expected theoretical PDFs. Based on this, it is therefore reasonable to assume that PDFs derived from a swash-by-swash analysis would follow a similar pattern. In this section, the trough-to-peak swash height (ρ) was used as proxy variable for such analysis (see Figure 1 for definitions). By applying the wavelet decomposition method detailed in Stringari and Power [33] (see their Appendix A) it was possible to classify each swash event as occurring under infragravity or sea-swell wave dominant forcing. For each timestack, ρ was calculated for each individual swash cycle and compared to the time-varying infragravity and sea-swell energy levels obtained using data from the PT in the surf zone that was closest to the surf-swash boundary in each data run. If energy in the infragravity band was greater than energy in the sea-swell band (that is, $E_{ig}(t) > E_{sw}(t)$) during the time of swash excursion, the swash event was considered to be dominated by an infragravity wave, otherwise, the swash event was considered to be dominated by a sea-swell wave. Due to the characteristic long-period of infragravity motions, there was no need to account for time offsets between the shoreline and nearest surf zone PT timeseries. Finally, it is worth noting that the approach

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

used here is equivalent to Guza and Thornton's [9] classical approach, only more robust, as it considers both time and frequency domains whereas the classical approach only works in the frequency domain.

As with the analyses shown in Section 3.1, both $p(\rho)$ and $(p((\rho - \mu)/\sigma))$ in both frequency bands presented great variability, were mostly multimodal (>95%), and were significantly statistically different ($p < 0.05$ using the Kolmogorov–Smirnov test) from the theoretical PDFs predicted by Cartwright and Longuet-Higgins [7]. To be in accordance with the theory, the observed PDFs should have collapsed to a Rayleigh PDF for non-normalised PDFs ($p(\rho)$, Figure 7-a to c), or to a Gaussian PDF for normalised PDFs ($p((\rho - \mu)/\sigma)$, Figure 7-d to f), but this was not observed. The mean PDF for each frequency band in each cluster was also obtained (black lines in Figure 7-a to f) and these mean PDFs also deviated from the PDFs predicted by the theory. It is worth noting, however, that non-normalised PDFs ($p(\rho)$) in cluster C closely approached but were not statistically similar to a Rayleigh PDF as assessed by the Kolmogorov–Smirnov test ($p \approx 0.05$). Further, when all the data in each frequency band were aggregated (Figure 7-g and f), the observed PDF did not collapse into the expected Gaussian PDF either, with the observed aggregated PDFs in both the sea-swell and infragravity bands being right-skewed and statistically similar to the Beta PDF, consistent with Hughes et al.'s [5] results (their Figure 7). Similar results to Figure 7-g and f were observed when aggregating the data in each frequency band based on the offshore clusters (not shown). See Section 4 for further discussion on the correlation between offshore parameters and the observed swash height PDFs.

Finally, Figure 7-i shows an analysis similar to that of Guza and Thornton's [9] (their Figure 7) which has been widely used in the literature to support the concept of swash saturation in the sea-swell frequency band. For each data run, the trough-to-peak significant swash height (ρ_{sig}) in each frequency band was calculated and compared to the observed offshore wave height. In contrast to Guza and Thornton's [9] data, the data analysed here showed a correlation between increases in the offshore wave height and increases in the significant trough-to-peak swash height in both the sea-swell and infragravity frequency bands. These results do not support, therefore, the assumption of swash saturation in the sea-swell band. For further discussion on swash saturation, see Section 4.

4. Discussion

This paper has presented a novel, data-driven approach for analysing the probability distribution functions of swash motions. Both shoreline height (ζ) and trough-to-peak swash height (ρ) PDFs were observed to be strongly multimodal, highly variable, and systematically statistically different from expected theoretical PDFs. Previous previous research [5,6] has shown that PDFs of different swash motions can deviate from Cartwright and Longuet-Higgins' [7] theory but, to the authors' knowledge, multimodal $p(\zeta)$ and $p(\rho)$ have not been reported before. Given the observed multimodality of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

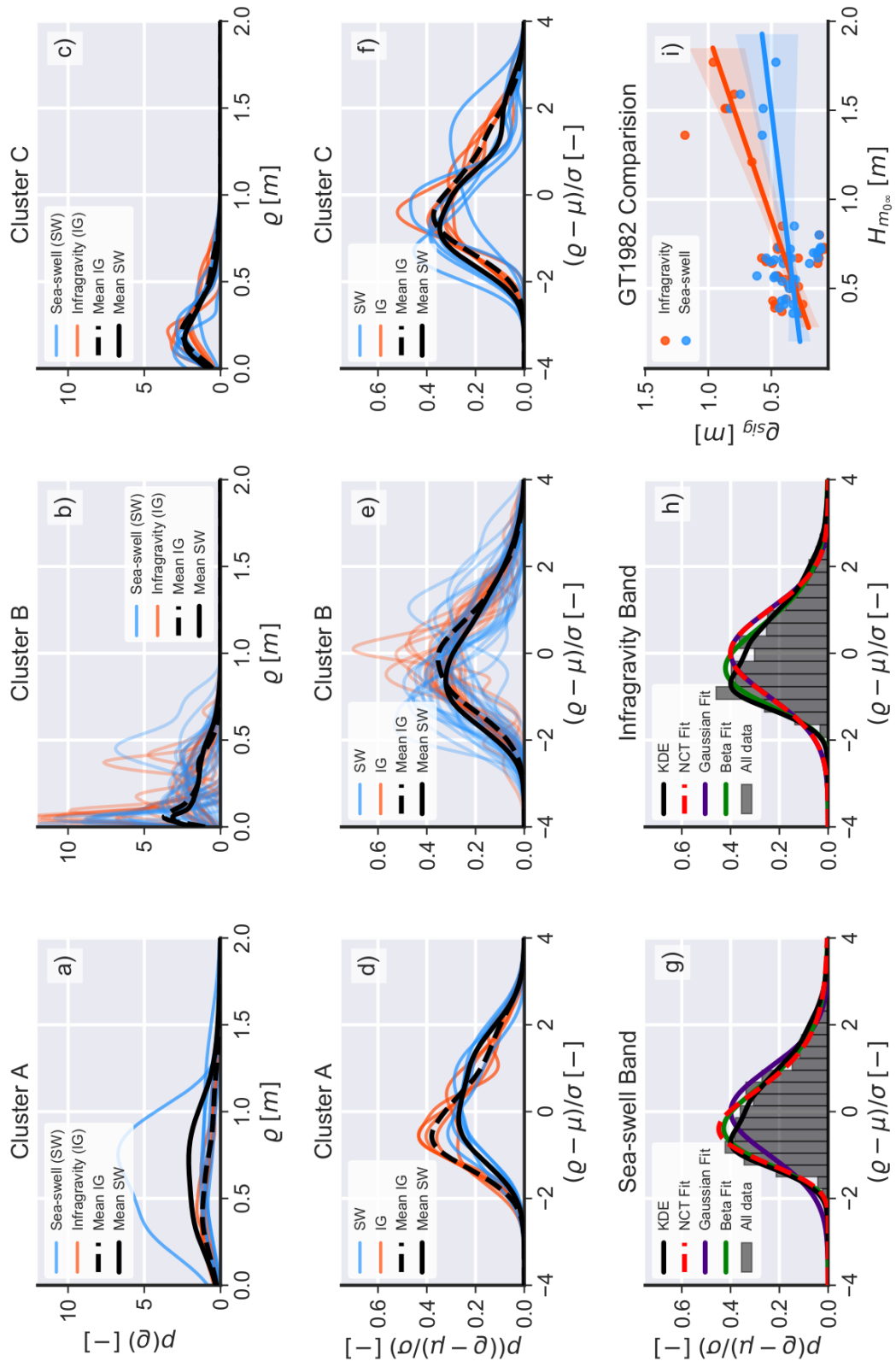

Figure 7. Non-normalised trough-to-peak swash height PDFs ($p(\rho)$) at infragravity (red) and sea-swell (blue) frequency bands for offshore a) cluster A, b) cluster B and c) cluster C. Normalised trough-to-peak swash height PDFs ($p((\rho-\mu)/\sigma)$) at infragravity (red) and sea-swell (blue) frequency bands for offshore d) cluster A, e) cluster B and f) cluster C. The black lines in a) to f) show the mean PDF for each frequency band. g) Normalised trough-to-peak swash height PDF for all data from panels a) to c) in the sea-swell frequency band. e) Normalised trough-to-peak swash height PDF for all data from panels a) to c) in the infragravity frequency band. In g) and h), the black lines show the KDE approximation to the data, the red dashed lines show the NCT PDF fit to the data, the blue lines show the Gaussian PDF fit to the data, and the green lines show the Beta PDF fit to the data. The number of bins in the histograms were calculated according to the Freedman-Diaconis rule [25]. i) Correlation between offshore wave height and significant trough-to-peak swash height (ρ_{sig}) for infragravity and sea-swell frequency bands. The coloured swaths in e) show the 95% confidence intervals. The regression lines are $\rho_{sig} = 0.48H_{m0_\infty} + 0.07$ ($r_{xy} = 0.64$, $p \ll 0.05$) in the infragravity band and $\rho_{sig} = 0.17H_{m0_\infty} + 0.24$ ($r_{xy} = 0.35$, $p = 0.02$) in the sea-swell band. Note that $\rho_{sig} \neq 0$ at $H_{m0_\infty} = 0$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

shoreline height PDFs, Gaussian Mixture Models (GMMs) were shown to be the best method to approximate $p((\zeta - \mu)/\sigma)$ (e.g., Figure 6), which are easily transferable to model $p(\zeta)$, $p(\rho)$, and $p((\rho - \mu)/\sigma)$. Interestingly, when the data were normalised, the shoreline height PDFs ($p((\zeta - \mu)/\sigma)$) collapsed into very similar PDFs, indicating that environmental forcing directly correlates with the shape of the non-normalised PDFs, further supporting the clustering approach based on offshore conditions. The influence of offshore wave conditions on swash motion PDFs is further supported by three other observations: 1) that shoreline height PDFs in cluster C which had a narrow offshore wave height band were very similar to each other regardless of data normalisation (see Figure 6-h); 2) that the width of $p(\rho)$ directly increased with increasing offshore height in both frequency bands; and 3) that the mean $p(\rho)$ PDFs in cluster C were only marginally statistically different to the predicted Rayleigh PDF (see Figure 7-c) which is consistent with the narrow offshore wave height band of this cluster. Ultimately, these results suggest that the swash zone was unsaturated in both infragravity and sea-swell frequency bands for the data analysed here.

The multimodality observed in both shoreline height and trough-to-peak swash height PDFs can theoretically be linked to the observation by Guza and Thornton [9] that energy in different frequency bands will result in distinct density peaks at different swash height elevations. This assumption is consistent with the analysis presented in Section 3.3, in which clear density peaks in $p(\rho)$ are observed at different frequency bands (e.g., note the separation between the mean PDFs in Figure 7-d to f). Therefore, the fact that GMMs were the only method that satisfactorily reproduced the observed PDFs may be a direct consequence of this (physical) phenomenon and not necessarily a result of pure statistical inference. In contrast to the observations of Guza and Thornton [9], however, the data analysed here does not support swash saturation in the sea-swell frequency band (see Figure 7-f). It is worth nothing, however, that Guza and Thornton's (1982) data were from a beach more dissipative than SMB and, therefore, the present results may not be directly comparable to theirs. The results in this paper showed, nonetheless, that as a consequence of the surf zone being unsaturated, the swash zone was also unsaturated, which is supported by the correlations between the offshore clusters and swash motion PDFs. This result is consistent with recent results from Hughes *et al.* [34] who also showed that swash saturation is not always the case on natural beaches.

Finally, an investigation of which environmental parameters best explained the variability seen in $p(\zeta)$ was conducted. Assuming that optimal number of mixtures (N_{mix}) is a direct proxy for the degree of variability and, consequently, the complexity of $p(\zeta)$, a model that ranks which environmental parameters best explained N_{mix} was constructed. This analysis provides an initial insight into which variables are most important for describing the trends seen in the data and aims to further support our observations that the observed surf zone dynamics were directly controlling the swash zone. A

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

random forest model was chosen to accomplish this task (see Appendix 1 for details). Note that, in contrast to Section 3.2, the maximum number of mixtures was not restricted to three and was, therefore, chosen based on the lowest AIC for each 10 minute data run (although the N_{mix} is unbounded here, the highest number of optimal mixtures observed was six because models with a too large number of mixtures gets heavily penalized by AIC). As inputs for the model, wave heights and periods at four cross-shore locations were used (offshore, Q_b=0.05, Q_b=0.50 and Q_b=0.95). The model was trained one hundred times to account for statistical variability and the feature importance for each variable was obtained. The same approach can be used to predict which parameters best explain μ, σ and α but this was not attempted here due to the small size of the dataset (see Section 6.4 in Stringari [35] for an attempt at this using model data). The results shown in Figure 8 indicate that a combination of several parameters were responsible for best explaining N_{mix}, with the wave height at seaward end of the surf zone ($H_{m0_{5\%}}$) consistently being the most important parameter for the model. In general, this result agress with the results from Section 3.1 as N_{mix} directly correlates with surf zone wave heights which implies that, as a consequence of the surf zone being unsaturated, the swash zone is unsaturated and, therefore, driven by incoming bores with non-negative terminal heights, as previously shown by two recent studies [19,21]. As more data become available in the future, models based on the present approach could provide a robust predictor for shoreline statistical properties based solely on known parameters, which will be valuable tools for coastal managers.

5. Conclusions

In this paper, analysis of swash motions from a gently sloping sandy beach under varying offshore forcing showed that the majority of observed PDFs (both $p(\zeta)$ and $p(\rho)$) were multimodally distributed and were statistically different from the PDFs predicted by the theory. Hence, Gaussian Mixtures were shown to be the best approach to model $p((\rho-\mu)/\sigma)$, which could be easily extended to other swash processes. The parameters of the Gaussian Mixtures that described these swash motions were closely correlated to wave conditions in the surf zone and further offshore, which is indicative of unsaturated swash. Analysis of the correlation between significant trough-to-peak swash heights (ρ_{sig}) and offshore wave heights further confirmed unsaturated swash in both short and long wave frequency bands. The field data collection and statistical methods used in this paper were shown to overcome the limitations of more traditional methods and allowed for a novel statistical descriptions of swash motions. These approaches, although preliminary and limited by a small dataset, should provide a robust basis for coastal managers when developing improved swash zone models in the future.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0538.v1

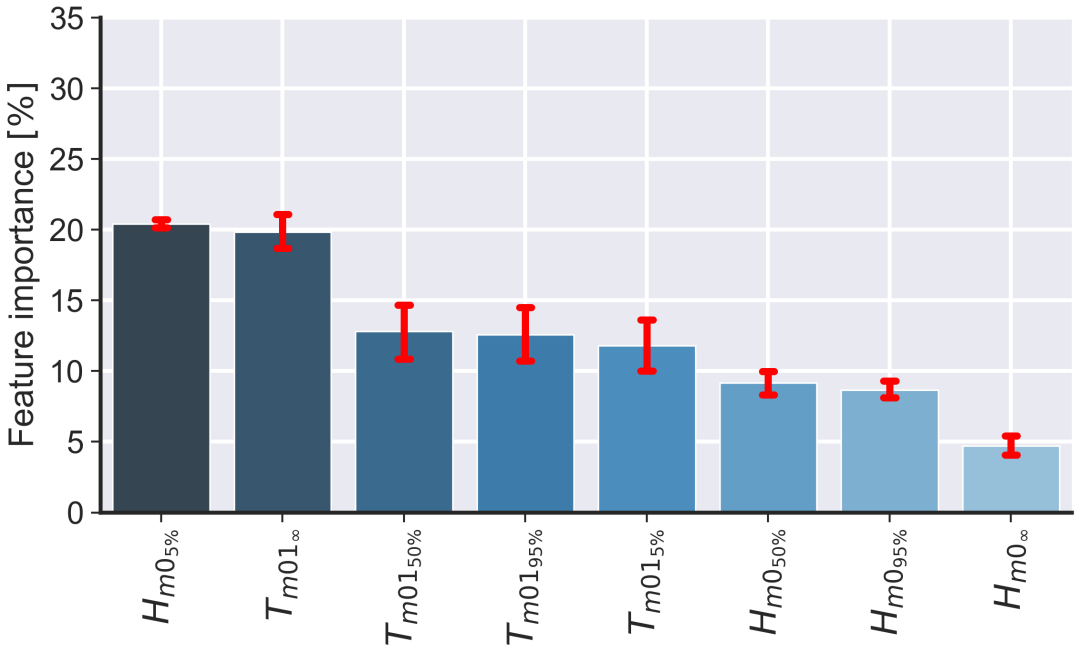

Figure 8. Feature importance of the random forest model. In this plot, $H_{m0_{\infty}}$ and $T_{m01_{\infty}}$ are the significant wave height and significant wave period offshore of the surf zone, $H_{m0_{5\%}}$, $T_{m01_{5\%}}$, $H_{m0_{50\%}}$, $T_{m01_{50\%}}$, $H_{m0_{95\%}}$, and $T_{m01_{95\%}}$ are the significant wave height and significant wave period at the Q_b value indicated by indexes where $Q_b = 5\%$ is representative of the outer surf zone, $Q_b = 50\%$ is representative of the mid surf zone and $Q_b = 95\%$ is representative of the inner surf zone.

Appendix A

This appendix describes the predictor for the optimal number of Gaussian Mixtures for a given sea-state. The eXtreme Gradient Boost (XGB) model [36] was chosen as the classifier. The goal was to obtain a non-linear function that maps input features into the predicted number of Gaussian Mixtures. Mathematically, this relationship can be written as:

$$\hat{N_{mix}} \simeq f\left(H_{m0_{\infty}}, T_{m01_{\infty}}, H_{m0_{5\%}}, T_{m01_{5\%}}, H_{m0_{50\%}}, T_{m01_{50\%}}, H_{m0_{95\%}}, T_{m01_{95\%}}\right) \tag{A1}$$

in which $H_{m0_{\infty}}$ and $T_{m01_{\infty}}$ are the significant wave height and significant wave period offshore of the surf zone respectively and, $H_{m0_{5\%}}$, $T_{m01_{5\%}}$, $H_{m0_{50\%}}$, $T_{m01_{50\%}}$, $H_{m0_{95\%}}$, and $T_{m01_{95\%}}$ are the significant wave heights and significant wave periods at the Q_b value indicated by the subscripts. These features were chosen based on the results results from Sections 3.1, 3.2, and 3.3.

The model is then defined as:

$$\hat{N_{mix}} = \sum_{k=1}^{K} f_k(X_i), f_K \in \mathcal{G} \tag{A2}$$

 doi:10.20944/preprints202012.0538.v1

where $\hat{N_{mix}}$ is the predicted number of mixtures, $f(X_i)$ is a function (in this case, a decision tree) that takes input training samples (X_i), and $\mathcal{G}$ is the space of functions containing all decision trees. The objective (obj) of the model is to learn the best function(s) that minimises a loss function (l) while, at the same time, keeping the model ensemble as simple as possible. This is done by considering a regularisation parameter (ω_r):

$$obj = \sum_{i}^{N} l(y, \hat{y}) + \sum_{k=1}^{K} \omega_r(f_k) \tag{A3}$$

The model is then trained using the greedy algorithm know as adaptive training [37]. The loss function for the model was the mean absolute error (MAE):

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{A4}$$

where y_i is the predicted number of mixtures and x_i is the observed number of mixtures. For the training step, the data were randomly split into training (70%) and testing (30%) datasets and the model was run 100 hundred times for each combination to account for statistical variability. The R^2 for all models always reached values greater than 95%.

Funding: C.E.S. was funded by a University of Newcastle Research Degree Scholarship and a Central and Faculty Scholarship (5050UNRS).

Acknowledgments: The authors are grateful to Tom Doyle, Kaya Wilson, Madeleine Broadfoot, Murray Kendall, Kendall Mollison and David Schmidt who assisted with the field data collection. We kindly thank Tom Baldock from University of Queensland who lent some of the pressure transducers used for data collection. We are particularly thankful to David Hanslow and Mike Kinsela from New South Wales Department of Planning Industry, and Environment (DPIE) who conducted the offshore data collection. The authors are also thankful to the Academic Research Computing Support Team, particularly Aaron Scott, at the University of Newcastle for support with the I.T. infrastructure and to Robert Holman and Evan Goldstein whose helpful comments shaped the final version of this paper.

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

References

1. Komar, P.D. *Beach Processes and Sedimentation*; Prentice-Hall, 1976.
2. Masselink, G.; Puleo, J.A. Swash-zone morphodynamics. *Continental Shelf Research* **2006**, *26*, 661–680. doi:10.1016/j.csr.2006.01.015.
3. Atkinson, A.L.; Power, H.E.; Moura, T.; Hammond, T.; Callaghan, D.P.; Baldock, T.E. Assessment of runup predictions by empirical models on non-truncated beaches on the south-east Australian coast. *Coastal Engineering* **2017**, *119*, 15–31. doi:10.1016/j.coastaleng.2016.10.001.
4. Power, H.E.; Gharabaghi, B.; Bonakdari, H.; Robertson, B.; Atkinson, A.L.; Baldock, T.E. Prediction of wave runup on beaches using Gene-Expression Programming and empirical relationships. *Coastal Engineering* **2019**, *144*, 47–61. doi:10.1016/j.coastaleng.2018.10.006.
5. Hughes, M.G.; Moseley, A.S.; Baldock, T.E. Probability distributions for wave runup on beaches. *Coastal Engineering* **2010**, *57*, 575–584. doi:10.1016/j.coastaleng.2010.01.001.

6. Holland, K.T.; Holman, R.A. The Statistical Distribution of Swash Maxima on Natural Beache. *Journal of Geophysical Research* **1993**, *98*, 271–278.
7. Cartwright, D.E.; Longuet-Higgins, M.S. The Statistical Distribution of the Maxima of a Random Function. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **1956**, *283*, 212–232.
8. Goda, Y. Reanalysis of Regular and Random Breaking Wave Statistics. *Coastal Engineering Journal* **2010**, *52*, 71–106. doi:10.1142/S0578563410002129.
9. Guza, R.T.; Thornton, E.B. Swash Oscillations on a Natural Beach. *Journal of Geophysical Research* **1982**, *87*, 483–491.
10. Wright, L.D.; Guza, R.T.; Short, A.D. Dynamics of a high-energy dissipative surf zone. *Marine Geology* **1982**, *45*, 41–62. doi:10.1016/0025-3227(82)90179-7.
11. Wright, L.D.; Short, A.D. Morphodynamic variability of surf zones and beaches: A synthesis. *Marine Geology* **1984**, *56*, 93–118. doi:10.1016/0025-3227(84)90008-2.
12. Stringari, C.E.; Harris, D.L.; Power, H.E. A Novel Machine Learning Algorithm for Tracking Remotely Sensed Waves in the Surf Zone. *Coastal Engineering* **2019**, *147*, 149–158.
13. Stringari, C.E.; Power, H.E. The Fraction of Broken Waves in Natural Surf Zones. *Journal of Geophysical Research: Oceans* **2019**, *124*, 1–27. doi:10.1029/2019JC015213.
14. Vousdoukas, M.I.; Kirupakaramoorthy, T.; Oumeraci, H.; de la Torre, M.; Wübbold, F.; Wagner, B.; Schimmels, S. The role of combined laser scanning and video techniques in monitoring wave-by-wave swash zone processes. *Coastal Engineering* **2014**, *83*, 150–165. doi:10.1016/j.coastaleng.2013.10.013.
15. Aagaard, T.; Holm, J. Digitization of Wave Run-up Using Video Records. *Journal of Coastal Research* **1989**, *5*, 547–551.
16. Sobel, I. An Isotropic 3x3 Image Gradient Operator, 1968.
17. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. Density-Based Clustering Methods. *Comprehensive Chemometrics* **1996**, *2*, 635–654. doi:10.1016/B978-044452701-1.00067-3.
18. Fiedler, J.W.; Brodie, K.L.; McNinch, J.E.; Guza, R.T. Observations of runup and energy flux on a low-slope beach with high-energy, long-period ocean swell. *Geophysical Research Letters* **2015**, *42*, 9933–9941. doi:10.1002/2015GL066124.
19. Bergsma, E.W.J.; Blenkinsopp, C.E.; MArtins, K.; Almar, R.; de Almeida, L.P.M. Bore collapse and wave run-up on a sandy beach. *Continental Shelf Research* **2019**, *174*, 132–139. doi:10.1016/j.csr.2019.01.009.
20. Roelse, A.; Granger, H.W.; Graham, J.W. Technical Report 12: The Adjustment of the Australian Levelling Survey 1970-1971. Technical Report 2, 1975.
21. Power, H.E.; Hughes, M.G.; Aagaard, T.; Baldock, T.E. Nearshore wave height variation in unsaturated surf. *Journal of Geophysical Research: Oceans* **2010**, *115*, 1–15. doi:10.1029/2009JC005758.
22. Svendsen, I.A.; Madsen, P.A.; Hansen, J.B. Wave Characteristics in the Surf Zone. Proceedings of of the 16th international conference on coastal engineering, 1978, pp. 520–539.
23. Iribarren, C.R.; Nogales, C.M. Protection des ports, 1949.
24. Power, H.E.; Holman, R.A.; Baldock, T.E. Swash zone boundary conditions derived from optical remote sensing of swash zone flow patterns. *Journal of Geophysical Research: Oceans* **2011**, *116*, 1–13. doi:10.1029/2010JC006724.
25. Freedman, D.; Diaconis, P. On the histogram as a density estimator:L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **1981**, *57*, 453–476. doi:10.1007/BF01025868.
26. New South Wales Department of Planning Industry, and Environment. Ocean Wave Data Collection Program. https://www.mhl.nsw.gov.au/Station-PTKMOW. Accessed: 2020-11-11.
27. Shapiro, S.S.; Wilk, M.B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **1965**, *52*, 311–319.
28. Kullback, S.; Leibler, R.A. On infromation and sufficiency. *Ann. Math. Statist.* **1951**, *37*, 688–697. doi:10.1214/aoms/1177705148.
29. Jones, E.; Oliphant, T.; Peterson, P.; Others. SciPy: Open source scientific tools for Python, 2001.
30. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* **1974**, *19*, 716–723. doi:10.1109/TAC.1974.1100705.
31. Hogben, D.; Pinkham, R.S.; Wilk, M.B. The Moments of the Non-Central t-Distribution. *Biometrika* **1961**, *48*, 465–468.

32. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics, Springer New York Inc.: New York, NY, USA, 2001.
33. Stringari, C.E.; Power, H.E. Quantifying Bore-bore Capture on Natural Beaches. *Journal of Geophysical Research: Oceans* **2020**, *125*, 1–16. doi:10.1029/2019JC015689.
34. Hughes, M.G.; Baldock, T.E.; Aagaard, T. Swash saturation: an assessment of available models. *Ocean Dynamics* **2018**, *68*, 911–922. doi:10.1007/s10236-018-1170-8.
35. Stringari, C.E. Data-driven investigations of broken wave behaviour in the surf and swash zones. PhD thesis, University of Newcastle, 2020. doi:http://hdl.handle.net/1959.13/1411217.
36. Chen, T.; Guestrin, C. XGBoost: Reliable Large-scale Tree Boosting System. Conference on knowledge discovery and data mining, 2016, pp. 1–6.
37. Friedman, J. Greedy Function Approximation: A Gradient Boosting Machine, 2001.