Quantitative metrics for evaluation of wave fields in basins

Keri M. Collins, Stuart Stripling, David J. Simmonds, Deborah M. Greaves

Abstract—Scale model testing in wave basins is a necessary part of the development of marine structures and marine renewable energy devices. Whilst many guidelines exist for the quality of experimentation and data acquisition, there are no standards for the basins themselves. We propose methodologies for assessing the quality of a wave field generated in a basin: a clustering parameter based on the variance of surface elevation at multiple gauges is used to score homogeneity and extended to a skill score for relative quality benchmarking. We use historic and recent data from the University of Plymouth's Ocean Basin as a case study for the methods. The quality metrics indicate that physical aspects of the basin itself, such as water depth, contribute the most to the accuracy and homogeneity with wave period performing better than height. Recommendations for experimentalists, such as using basins when operating in steady state, are presented and discussed.

***Index Terms*— Wave basins, benchmarking, skill score, wave generation.**

I. Introduction

The aim of laboratory testing is to replicate real-world conditions in a controlled manner. Wave basins are an important step in the development of ocean-deployed structures and devices: a variety of scale model tests can be conducted at a comparatively reduced cost from open water trials and under more controlled conditions.

The typical waves that a laboratory is trying to replicate are scaled representations of the wind-created gravity waves of the ocean that have typical periods of 1 s to 30 s [1]. Real wind waves cannot be predicted as the tides can, and so must be modelled according to a probability of occurrence. Much has been written on the subject of wave models and statistical representations of wind waves; a detailed state of the art review of wave modelling is given in [2].

Generating waves in a laboratory basin requires appropriate paddle-sets and software control systems that, together, impart energy to the water column. It is the ability to precisely control how such energy is imparted, and the theory behind wave making, that makes for a successful test facility.

A. Wavemaker theory

Wavemaker theory has its origins as far back as the 1920s with Havelock's work on forced surface-waves [3]. Current wavemaker technology is typically a computer-controlled wavemaker (either a piston-type or a flap/paddle-type) that is moved in such a way to create waves of a specified period and height in the basin according to linear theory. If a paddle's stroke can be determined mathematically for a particular type of wave, then its transfer function, ie a description of the amplitude of a wave for a unit input as a function of frequency, can be determined [4]. More recently, the transfer functions for 3D wave basins with force-controlled paddles have been developed [5].

B. Theory vs practice in wave generation

The discrepancies between a theoretical water wave and a physical measured quantity derive from numerous sources. For example, linear theory, which most basins still rely on for generation, assumes that water can be modelled as an ideal fluid, which itself imposes a long list of simplifying assumptions, and that waves have small amplitudes and a shape that is invariant in time and space [6].

The kinematics or dynamics of the wavemakers, basin geometry, reflections and measurement techniques all affect the wave and its quality. Even if the generation mode is of the first-order, energy imparted by the excitation of the paddles is not transmitted to the water column in a linear fashion [4, 8]. When energy is added, and a wave is generated, it can be accompanied by extra components: bound modes, which travel at the same celerity as the fundamental wave and free modes, which travel independently. Such higher-order waves are unwanted and can be a source of error.

Anderson *et al.* [9] conducted a review of available reflection analysis methods. They point out that one of the fundamental assumptions in the analysis procedure is that only the incident and reflected wave are present in the recorded signal and that any other wave components will introduce uncertainties. To address higher order effects, modification of the paddle motion was suggested, notably the introduction of a non-sinusoidal motion [7], and has since been developed for position-controlled paddles [10] and then extended for and implemented with force-controlled paddles [11, 12].

Finally, the waves themselves may evolve through spatial variations and instabilities due to reflections, diffraction and other physical mechanisms [13].

C. Laboratory testing

Ocean energy device developers need to have confidence that laboratory conditions are suitable for their tests. An understanding of the aspects and limitations of basin testing is necessary in order to reasonably support claims of performance. For a wave energy conversion device, an obvious place to start is with the wave height: the energy density of a sea is proportional to the square of the wave height, so deviations in this parameter may have a significant effect on its performance.

The main driver of laboratory testing is the controllability of the parameters that allow hypotheses to be tested scientifically. As such, a legitimate, and often asked question, is: how accurate is the basin at producing waves? This is complicated to answer and is fundamentally the wrong question; a better question would be: how can the quality of the waves in the basin be objectively assessed? Tank testing guidelines are available to assist in the identification of sources of error for uncertainty analysis during experimentation [14, 15] but these focus on the experimental process rather than the characteristics of the basin involved.

The goal of this paper is to provide and describe new methods to assess the quality of waves as generated in a physical basin. Here quality is used as shorthand for the accuracy and precision of the wave height and period. We develop a new skill score that is better adapted to the questions concerning wave basin quality than other measures available in the literature, reviewed in Section II. These measures are applied to the University of Plymouth (UoP) Ocean Basin in Section III to demonstrate their application in terms of quality assessment. To achieve this, three data sets are used: the first to demonstrate the functioning of the quality assessment methods with the second to provide an "after" comparison for the skill scores. The third data set was taken under different operating conditions and is used to examine the influence of the physical basin geometry. The application of the quality measures is discussed with the results in Section III and a more general discussion of the methods themselves is found in Section IV.

II. Scores for quality evaluation

Evaluating the quality of a wave basin is not straightforward: there are many parameters that can be optimised although the importance of these parameters will depend on the experiment being performed [13]. That said, fundamentally, experiments are conducted in a laboratory setting so that the parameters are controllable. The ideal wave basin produces the waves that are programmed in by the user, which can be calibrated as little or as much as desired, classified in terms of accuracy: how close to the demanded value a certain parameter is; and in terms of precision: how narrow is the spread of results. Since these relate to the quality of the basin in both time and space, we can further add homogeneity, stationarity and ergodicity to the list of desirable parameters, along with repeatability. In reality, we must decide on the acceptable limits for these parameters and a measure that allows us to decide whether a system has improved across a large range of values.

Currently, there are no guidelines as to how accurate or precise wave parameters in a wave basin should be, nor is there a consensus of how that should be quantified. The ITTC recommends that the experimental outcomes be taken into consideration when assessing accuracy and this is the approach taken by some research facilities. Homogeneity of parameters in a wave basin was investigated in the context of wave energy converter (WEC) array testing at the Queen's University Marine Laboratory Portaferry basin in Northern Ireland [16, 17]. Both physical modelling and numerical simulation were used to compare wave amplitude around the basin. Irregular waves showed less variation of H_s, the significant wave height, than shown in the amplitude of the monochromatic waves [16], primarily because irregular spectra are combined measures where variation in individual wave components may average out. The later paper investigating the Portaferry basin [17] concluded with a set of protocols for array testing in order to ensure quality data acquisition and numerical model validation. More recently published research by the same group covers the process of testing different configurations of end and side beaches with a view to enhancing the quality of the wave climate in a coastal wave basin [18].

A. Measures of accuracy and precision

In the context of predictive models, bias describes the tendency of a model to over- or under-predict the observations and is based on the mean, although it is not strictly a measure of accuracy since it does not provide information as to the magnitude of the predictions [19]. Equation (1) shows how the bias, B, is given by the mean difference between the predictions, p, and observations, o, at N different values, with the nth point occurring at the same time and position in space, with $\langle\cdot\rangle$ denoting the mean value. The main drawback of the bias is that similar amounts of over- and under-prediction may cancel to give a low bias score.

$$\boldsymbol{B = \frac{1}{N}\sum_{n=1}^{N}(p_n - o_n) = \langle P\rangle - \langle O\rangle} \tag{1}$$

There are several measures of accuracy with the two most common being the mean absolute error, MAE, and the mean square error, MSE, [20] defined in (2) and (3). In addition, the root-mean-square error, RMSE, is given by the square root of (3).

$$\mathbf{MAE}(\boldsymbol{P},\boldsymbol{O}) = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{n}=\mathbf{1}}^{\boldsymbol{N}}|\boldsymbol{p_n}-\boldsymbol{o_n}| = \langle|\boldsymbol{P}-\boldsymbol{O}|\rangle \quad (2)$$

$$\mathbf{MSE}(\boldsymbol{P},\boldsymbol{O}) = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{n}=\mathbf{1}}^{\boldsymbol{N}}(\boldsymbol{p_n}-\boldsymbol{o_n})^{\mathbf{2}} = \langle(\boldsymbol{P}-\boldsymbol{O})^{\mathbf{2}}\rangle \quad (3)$$

It is not possible to determine if a model over- or under-predicts just from the MAE and for this reason it is often suggested that bias and MAE (or MSE) be reported together. Since the MSE is the second moment of the error, it gives information about the variance of a parameter and its bias.

B. Clustering parameter to score basin homogeneity

In the field of data mining, data are often sorted into clusters to distinguish between groups with similarities of particular variables. One of the most popular methods, *K*-means clustering, sorts data into *K* clusters, where *K* has been chosen by the user [21]. This approach has been used by [22] to create representative sea states from field data and to create a quality metric based on the MAE that represented the ‘compactness’ of the variables. This method of looking at the compactness of a group was essentially based on the residuals from the group mean and the inclusion of different variable types.

Rather than consider a number of distinct degrees of freedom, we propose to quantify the homogeneity of wave production by examining certain wave parameters across a group of wave gauges, using a two-dimensional clustering parameter. The method uses the mean, $\overline{x_g}$, and the variance, v_g, of the parameter of choice, for example wave height or period, as calculated from the wave gauge data to assess how closely the data match across gauges. As this is a measure of the homogeneity of the basin, the method does not make reference to the input values.

The mean value for all *G* gauges, x_G, is calculated using (5). Lastly, the difference between $\overline{x_g}$ and x_G is calculated for all gauges (6), which is the residual of the sample, r_g. The squares of the residuals are then analogous to the mean square error (MSE) calculated in (3).

$$\overline{x_g} = \frac{1}{W}\sum_{w=1}^{W} x_{w,g} \quad (4)$$

$$x_G = \frac{1}{G}\sum_{g=1}^{G}\overline{x_g} \quad (5)$$

$$r_g = \left(\overline{x_g} - x_G\right) \quad (6)$$

To examine the homogenity, the residuals, the r_g values calculated in (6), can be plotted against the standard deviation of the values on a particular gauge. To quantify the homogeneity, a clustering parameter, c_p, is calculated as the mean vector distance between the origin and all the points in the group:

$$c_p = \sqrt{r_g^2 + v_g} \quad (7)$$

thus low variance on each wave gauge and between each wave gauge combine to give low scores. Expressing the data in this way means that the data have similar ranges in both dimensions and the units are consistent.

C. Comparative performance and skill

For simple analysis of two groups, it may be appropriate to compare the mean values. This can be made more sophisticated by calculating whether a change in experimental set-up had a significant, and quantifiable, effect [23]. For data that has come from multiple experiments with a two-dimensional parameter space, however, such simple analysis is not appropriate. Moreover, a quantification of the comparative performance of two groups that makes no reference to the desired or ideal performance is of limited use.

Skill may be defined as the overall performance of a prediction based on observations with reference to a baseline. Meteorology commonly uses skill scores to assess the performance of forecast models and these have been adopted more recently by coastal morphodynamicists to quantify the skill of predictive sediment transport models [20].

We propose that skill scores can be used to provide a method to systematically and consistently benchmark the quality of waves in a wave basin system as a whole as it undergoes commissioning, maintenance and upgrading. The computer-to-waves system used in a basin should be entirely deterministic but variations in local conditions, basin construction and measurement mean that there will not be perfect agreement between the input and the output. In this case, the predicted values are the system inputs, the observed values are recent wave gauge measurements and the baseline values are those observed prior to some change in hardware or software. Some of the possible modifications to the entire system path-way are shown in Figure 1, with a modification to the paddle hardware given as an example. Two versions of the skill score can be computed for the parameters space. The mean squared error skill score (MSESS), given in Equation (8) and also called the Brier skill score [24], gives the skill of a new system, *M*1, over an old system, *M*0, with reference to the input values, *I*.

$$\begin{aligned}\mathbf{MSESS} &= \mathbf{1} - \frac{\mathbf{MSE}_{\boldsymbol{new\ system}}}{\mathbf{MSE}_{\boldsymbol{old\ system}}} \\ &= \mathbf{1} - \frac{\langle(\boldsymbol{M1}-\boldsymbol{I})^{\mathbf{2}}\rangle}{\langle(\boldsymbol{M0}-\boldsymbol{I})^{\mathbf{2}}\rangle}\end{aligned} \quad (8)$$

This can be defined to be a measure of accuracy skill and can be used on both the height and the

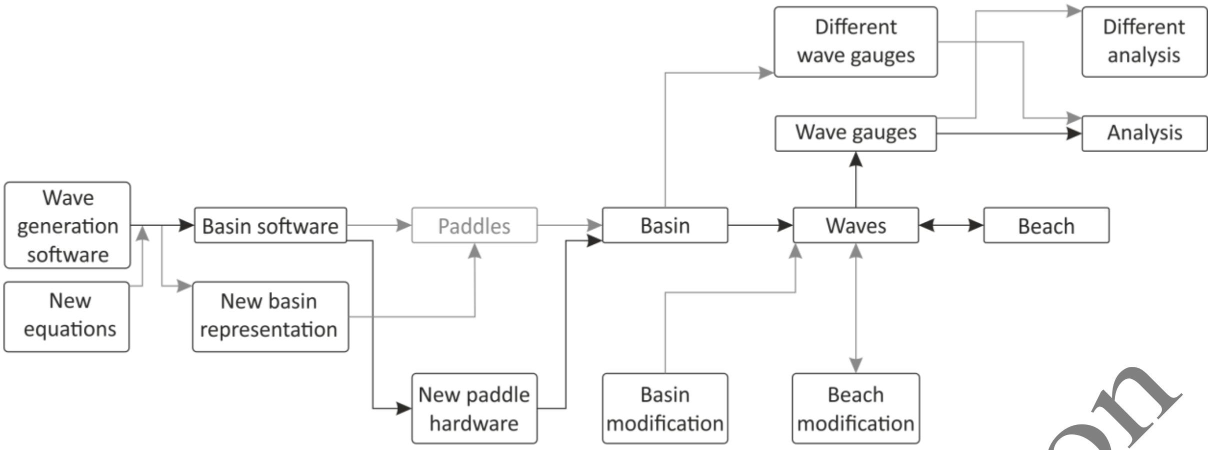

Figure 1 – Diagram showing the possible modifications to the whole system. Each deviation from the path (indicated in black) could lead to the calculation of a skill score. Here a modification to the paddle hardware (such as different gussets between the paddles or a change in the pivot point) is used to illustrate deviations from the typical path.

period of the waves, for example. The MSESS returns a value of 1 if the new system perfectly models that of the input and a value of zero if the new system is the same as the old system. If the new system performs worse than the old system, a negative value of MSESS is returned. Owing to the way in which the MSESS is calculated, for wave gauge data that is averaged over time, the values of skill are point measurements and the shortcomings of the MSESS are discussed in Section IV.

The idea of a skill score comparing a new and old system can be extended to take the clustering parameter as an input, as (9) shows. This results in a measure of skill related to the homogeneity, thus precision of the basin. These two measures can be used in conjunction to assess the quality of the basin after any hardware, software or analysis change, such as those shown in Figure 1.

$$CSS = 1 - \frac{c_{p1}}{c_{p0}} \qquad (9)$$

III. Case Study: UoP Ocean Basin

This paper investigates the application of the skill score analysis methodology to assess the quality of wave field in a deep-water basin. The UoP Ocean Basin, opened in autumn 2012, is a deep-water facility that is able to produce waves and currents at the same time. It has nominal dimensions of 35 m long by 15.5 m across. The depth of the central movable floor section, shown in Figure 2, may be up to 3.0 m but owing to a slight slope at the leading edge of the floor mechanism, the water depth in front of the paddles is 3.6 m. The floor can be raised to above the still water level, shown by the grey line in Figure 2, allowing models to be secured to the floor. The waves are produced by 24 dry-backed, 2 m hinge-depth, flap-type paddles produced by Edinburgh Designs Ltd. At the far end of the basin (+*y*), a convex parabolic beach structure covered in a porous wire mesh attenuates the incoming wave energy.

A proposed measure of accuracy of the UoP Ocean Basin is that wave heights should be within ±5% of the target value, although calibration can readily ensure that accuracy is within ±1%. While the waves in the basin are always measured during experiments, discrepancy between the target wave height and the produced wave height will have implications for experiments concerning power conversion efficiency since the power per unit of wave crest is proportional to the square of the wave height. More important is consistency, both in time (stationarity) and in space (homogeneity) as that measured changes in the wave conditions can then be attributed to the experimental device.

A. *Methods*

In this paper, the results of regular wave experiments are considered. Methods for evaluating spectra will be discussed later in the paper. In the simplest case of regular monochromatic waves, no reflected wave energy would contaminate the time series, and it has been suggested as the ideal situation for an experiment [25]. However, with a real basin it is not practical to run experiments that have no reflections present.

1) *Chosen parameter space*

Three data sets were compared. The first experimental session was held during 2014 after modifications to the baseline transfer functions to improve wave height accuracy had been made, but before major software updates had been installed. This data set provided a benchmark for comparison and gave data for the whole of the parameter space. Waves were run in the UoP Ocean Basin with the floor at its maximum depth of 3.0 m. Regular monochromatic, long-crested waves were run in batches sorted according to height. A selection of heights and frequencies were chosen to cover the full-operating parameter space, Figure 3, and chosen frequencies were alternated as the height increased. Waves that were considered outside of the paddle limits according to the Biésel transfer function, which quantifies the limits of the paddles themselves, were not included in the parameter space but waves that were predicted to be on the limit of breaking according to the Miche breaking criterion were used.

The second experimental session was run in 2016 and covered half the parameter space as shown in Figure 3. During the second experimental session, the chosen frequencies were not alternated as the height increased leading to greater in-fill of the parameter space. Figure 3 shows the parameter space and the number of experiments at each combination of frequency and amplitude.[1]

Finally, a third data set was taken during 2016-2017. For these waves, the gauge layout was substantially different and the area of the basin covered was much smaller. They are included because the floor depth was 2.0 m which allows the influence of water depth to be discussed.

2) Experimental conditions

Waves were programmed in the Edinburgh Designs Ltd wave synthesiser program; v1.2 was used in 2014 with an old version of the basin geometry set-up file and v1.3 was used in 2016 with an updated geometry file. The synthesiser program included the option to set a focal point along the basin for the waves and this was left at the default value (zero).

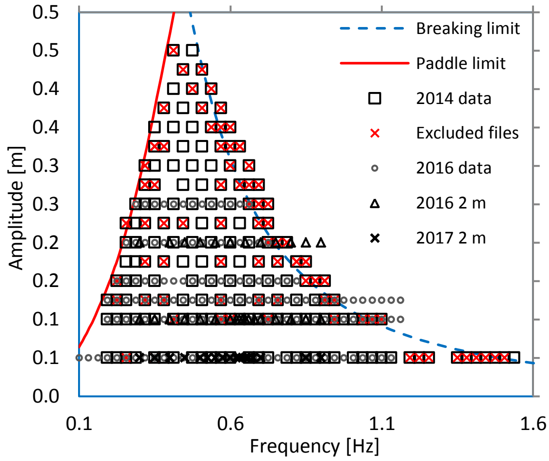

Figure 3 – The experimental parameter space was determined based on the theoretical capabilites of the basin.

For the 3 m deep experiments, waves were run for a period of approximately three or five minutes, depending on the wave length, to cover the majority of the chosen parameter space. The time between runs was manually adjusted at the beginning of the experimental session and was later set to be 3 minutes (2014 data) or 5 minutes (2016 data). For waves that were predicted to cause cross-waves in the basin, the gap between runs was manually extended to approximately 10 minutes for the 2014 data set only. For the 2 m deep experiments, 100 waves were run at each frequency and so the experimental time varied.

3) Surface elevation measurement

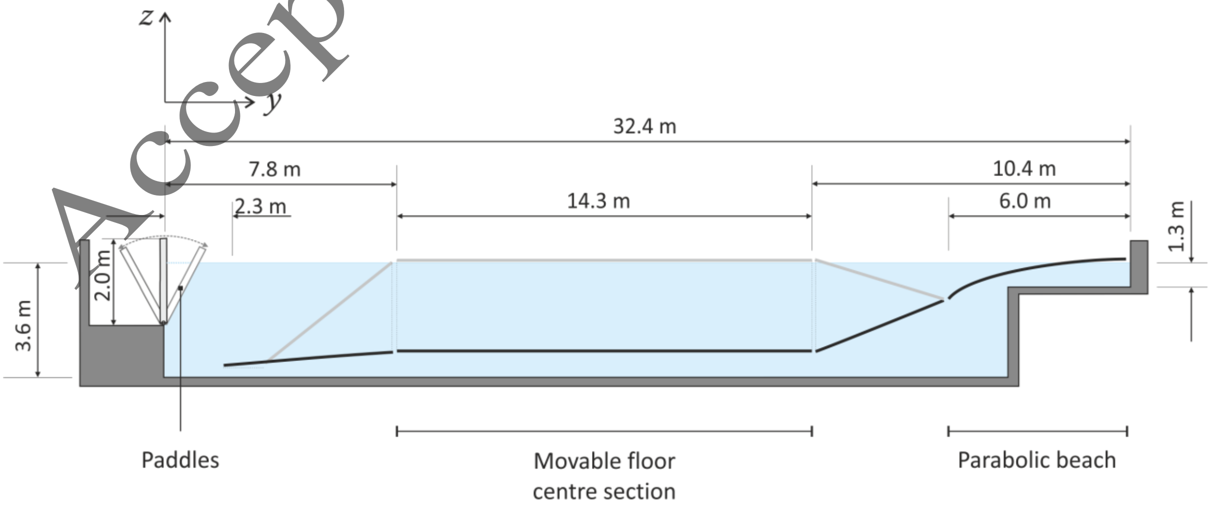

Figure 2 – Side view of the Plymouth Ocean Basin showing the movable floor at maximum depth (black line) and at minimum depth (grey line). All measurements are approximate.

[1] Frequency and amplitude are used in preference to height and period in this section since these are the inputs to the Edinburgh Designs Ltd. software.

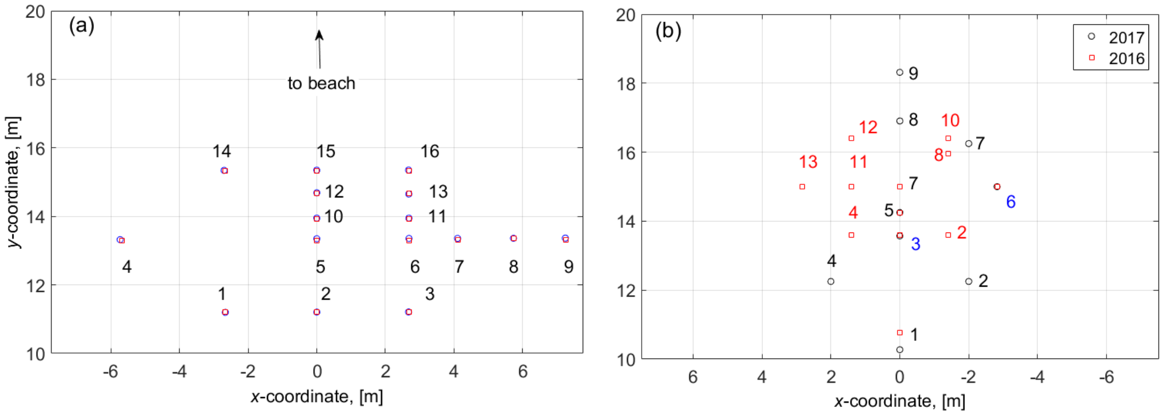

Figure 4 – Wave gauges were placed around the central section of the 35 m long basin for the 3 m deep experiments (a) and for the 2 m deep experiments (b), for which there were two gauge layouts.

To measure the wave height, twin-wire resistance wave gauges were placed in the basin as shown in Figure 4. The positions of the wave gauges in the Ocean Basin in 2014 were measured using a Leica Flexline TS06 plus EDM with the front face of the paddles used as a reference point. In 2016, a reference point on the floor was used to replicate the 2014 wave gauge pattern and Figure 4 shows that there was good agreement between the locations. Owing to an electrical fault Gauges 7 and 9 were not used in the 3 m 2016 data set.

4) Preliminary data analysis

A series of bespoke MATLAB programs was created to manage, collate and analyse the data from the viable experiments. The main functions that manipulated the data or were used to calculate derived parameters were:

- Low-pass filtering of the data (<20Hz)
- Calculation of wavelength, λ, and celerity, *c*, from input parameters using dispersion relation
- Section delimitation based on theoretical group velocity
- Zero-crossing analysis

Since all experimental basins have to operate under physical constraints, the waves generated in the basins will not be stationary throughout the entire time series. If a stationary signal is desirable, the goal is to choose the portion of the time series that can be said to be most stationary, which for a physical basin should be when the operating conditions are not changing.

The time series from the regular wave experiments were divided into five sections according to their presumed operating conditions: called run-up, Stages 1, 2 and 3 and run-down. The run-up period accounted for the first 8 s and was set in the basin control software. Stage 1 describes the portion of the time series starting at the end of the run-up and finishing when reflected waves reach a particular wave gauge, now travelling in the –*y* direction. This stage should only contain waves that are unaffected by reflections and so is referred to as the clean (generation) section. Stage 2 describes the segment of the time series in which any reflected wave energy reaches the paddles and may be re-reflected but has not yet reached a particular gauge. In this segment, the paddles begin to account for any reflected waves and so operating conditions are not stationary. In Stage 3, the “steady state”, any re-reflected wave energy will have reached a particular wave gauge and operating conditions should be constant. This stage ends when the run-down period commences. Of course, any re-reflected waves present in Stages 2 and 3 have the possibility of being reflected once again by the beach structure. The likelihood and implications of this occurring are discussed in Section III.C.

The timing of each stage was dependent on the group speed at which the waves were travelling (dependent on the input parameters of the wave and the dispersion relation being applicable) and the distance travelled in each stage (dependent on the physical characteristics of the basin and the set-up of the wave gauges). The distance travelled used the full length of the basin cavity, rather than trying to estimate which portion of the beach was responsible for reflecting the waves, and the distance of the wave gauges in relation to face of the paddles at rest. This is, of course, a simplification and its implications are discussed in Section III.C.

For the experiments conducted in 2 m water depth, the experiments were run for 100 waves rather than for a fixed duration. This meant that the steady state section was deemed to end when 100 waves had been measured by the wave gauges. This was an underestimate of the actual duration since it did not take into account the time that the 100th wave would take to travel to the gauge of interest.

A zero down-crossing method (as recommended by the IAHR [26]) was used to calculate the wave heights and periods for the full wave record and each of the wave stages. In order not to exclude waves that crossed the boundaries of the stages, waves were deemed to belong to a certain stage if the final crossing point occurred in that wave stage, even if the wave began in a different stage and this is discussed further in Section III.C.

B. Results

The measures presented in Section II can be used to answer questions relating to the quality of a generated wave field but equally can be applied to any three-dimensional system for which an assessment of the average parameters is required. In this Section, we present and discuss the results in the context of the UoP Ocean Basin. The goal is not to demonstrate absolute quality but rather to show the functioning of the methods previously discussed and developed, therefore many of the results only consider the 2014 data set as the parameter space was larger. It was found that the period was much less variable than the height and so the results focus on the measured wave heights as they better demonstrate the methods. The skill scores developed are used to compare the clean and steady state generation sections and the old and the new data sets. The 2 m deep experiments are only used to compare the CSS.

Figure 5 shows the time series as recorded by Gauge 2 during one of the experiments from the middle of the parameter space ($f = 0.50625$ Hz, $a = 0.175$ m). The vertical lines represent the section delimitations calculated according to the method outlined in the previous Section. From Figure 5, features common to all the experiments can be highlighted. The peak in the period during the run up phase is an artefact of the first waves generated, which are not yet full-height waves. The last wave in the run up section typically had a lower period and height than the subsequent waves. The first wave in Section 1 (the clean waves section) typically had a larger magnitude than the subsequent waves and this may be larger than the demand value as well. In the UoP Ocean Basin, a small degree of breaking of the first wave was observed at the leading edge of the working floor area, Figure 2. Figure 5 demonstrates that the variation in the zero-crossing heights was greater than in the zero-crossing periods, in particular during the steady state portion of the experiment.

For most of the 3 m deep experimental parameter space, the duration of wave generation was fixed to approximately 300 . This means that for the slower waves, the steady state portion of the experiment was smaller as a proportion of the total experiment time. The clean generation and the transition phases are both quite short compared to typical experiment lengths. The length of these stages is inversely proportional to the group celerity, therefore proportional to the frequency of the waves, but also affected by the position of the gauge. For the wave shown in Figure 5, only 15 waves were present in the clean generation section. For the fastest waves in the parameter space, the clean phase comprised only two waves, compared to 78 waves for the slowest travelling waves.

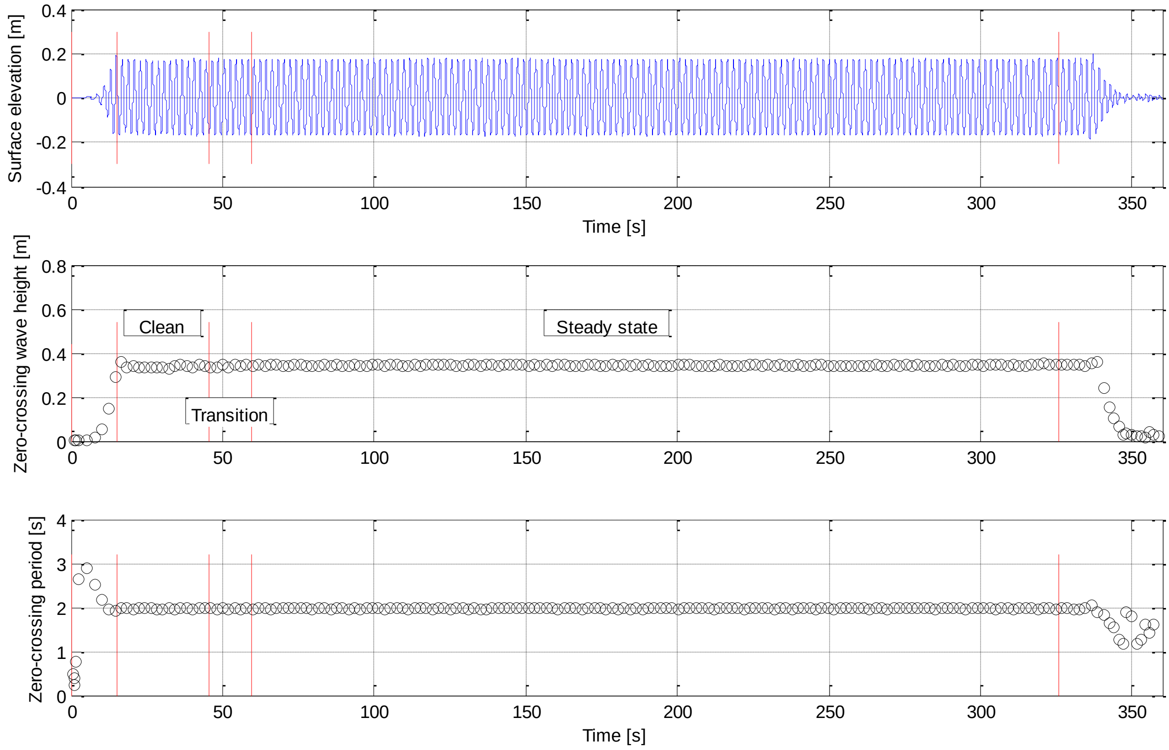

Figure 5 – Time series and wave heights as calculated by zero-crossing analysis from the old data set. Data shown come from Gauge 2 for a wave with 0.175 m amplitude and frequency, $f = 0.50625$ Hz ($T = 1.9753$ s).

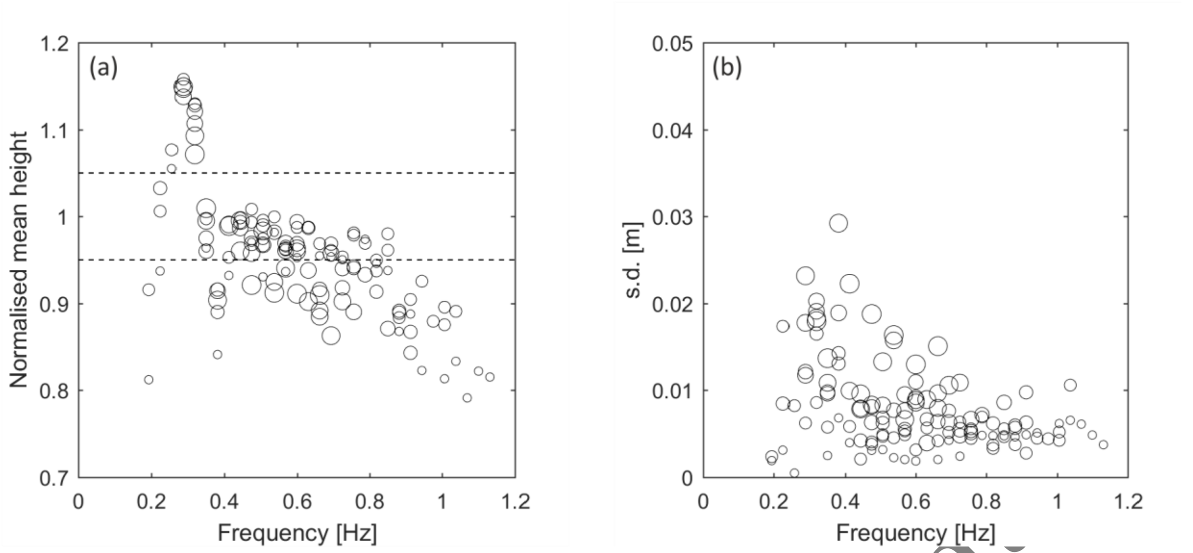

Figure 6 – Normalised mean wave height (a) and standard deviation, s.d. (b), as a function of frequency for all wave heights for Gauge 2 in the clean generation section. Marker size is proportional to the target wave height.

1) Accuracy and precision

The mean wave height as measured on Gauge 2 and normalised by the target value and the standard deviation of height are shown in Figure 6. Without further calibration effort, the normalised mean wave height varied throughout the parameter space but there was a significant frequency effect on the wave height: around $f = 0.30$ Hz waves were up to 15% larger than the target value. Since Figure 6 shows the clean generation section, reflected wave energy was not the cause. At frequencies either side of this point, normalised wave height dropped considerably. The peak around $f = 0.30$ Hz only affected the smaller waves (indicated by the smaller marker size) because the larger waves were out of the theoretical paddle limits at this frequency, *cf.* Figure 3. As the frequency increased, the normalised wave height fell from unity to 0.85 for the larger waves and 0.80 for the smaller waves. With the number of data available, it is not possible to ascertain whether the decline of the normalised wave height was significantly steeper for the larger waves compared to the smaller waves

Figure 6(b) indicates that the larger waves had a larger standard deviation than the smaller waves, which is expected in an absolute measure, although there is still some variation. At the higher end of the frequency axis, more waves were present in each wave recording so the value of s.d. is a better approximation for the true population standard deviation, σ, however the values of s.d. may still be influenced by the frequency. For example, when the normalised wave heights were greater than one, the standard deviations were larger compared to other data with the same input wave height. This suggests that a physical effect of the basin was more likely the cause than the paddle transfer

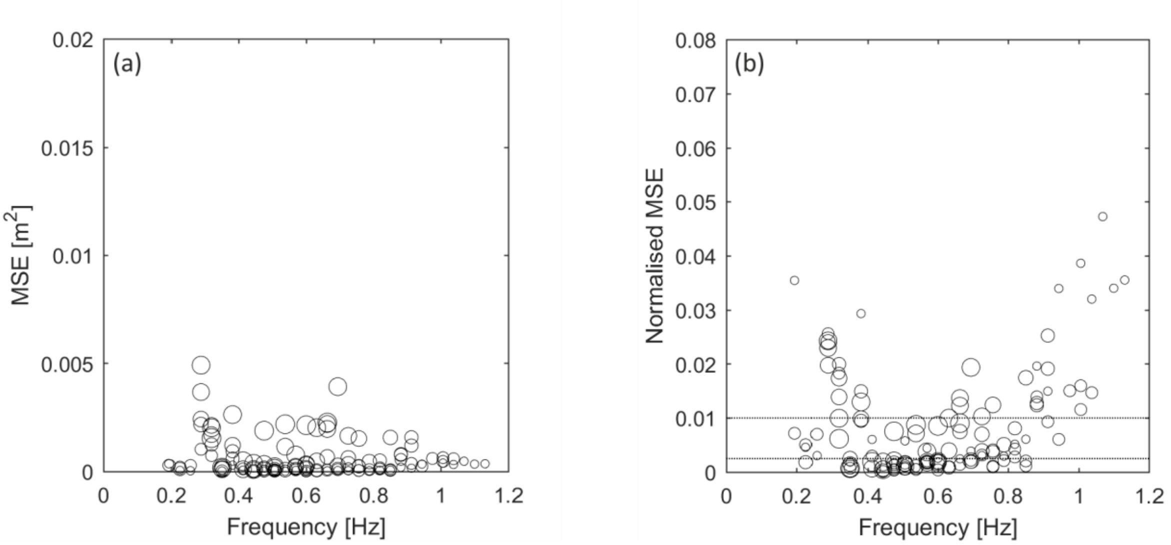

Figure 7 – MSE (a) and normalised MSE (b) of the wave height as a function of frequency for the clean generation section for Gauge 2.

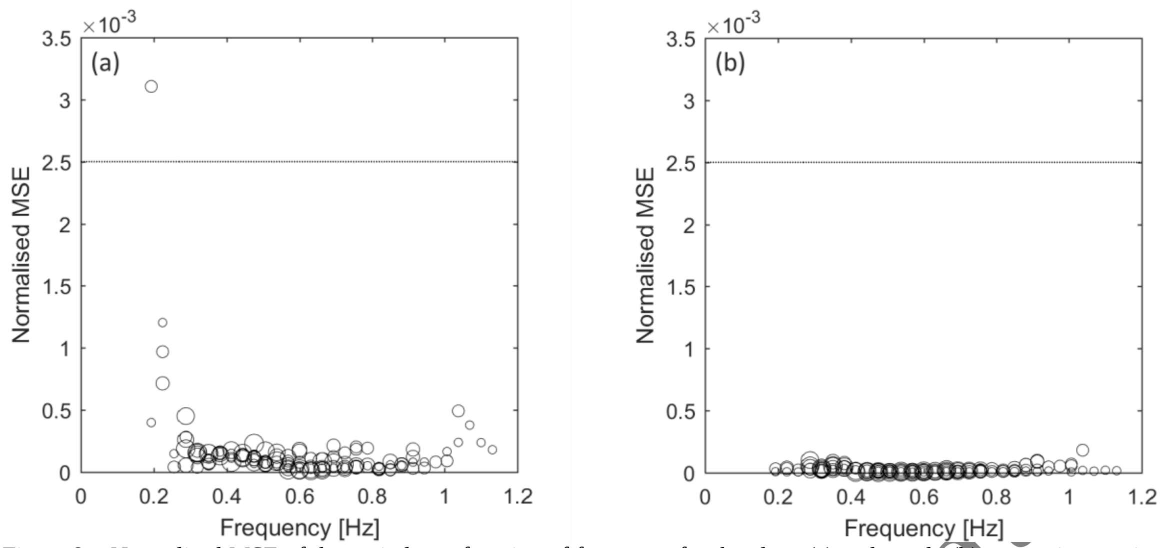

Figure 9 – Normalised MSE of the period as a function of frequency for the clean (a) and steady (b) generation section for Gauge 2.

function.

Further evidence that the paddle transfer function was not the most likely cause of the large values of normalised wave height is presented in Figure 8, which shows the normalised wave height and standard deviation for Gauge 15, which was located furthest away from the paddles, *q.v.* Figure 4. The frequency effect around $f = 0.30$ Hz reduced the normalised wave height considerably on Gauge 15 with a spike in values at the higher frequencies. Since the effect that this particular narrow band of frequencies had on the wave height was constructive or destructive depending on the location in the physical space of the basin, it is almost certainly not a paddle transfer function effect.

For most basins, the limit of the deep-water wave generation depends on the basin geometry itself rather than on the wavemakers. The limit is defined by the ratio of basin water depth, *d*, to the wavelength, λ, calculated using the wave dispersion equation. A ratio of d/λ greater than 0.5 indicates the waves are deep water waves which in the case of the UoP Ocean Basin gives a minimum frequency of $f = 0.51$ Hz at 3 m water depth. This is unlikely to be the cause of all the frequency-dependent effects shown in Figure 6 and Figure 8 but it will introduce compounding factors.

In Section II.A, bias and mean square error, MSE, were presented as ways to assess the accuracy of a system. It is recommended that bias and MSE be presented together since bias of equal value but opposite sign may cancel out, and as such it is possible to have a low bias and a large MSE, but not the opposite. Since the normalised mean heights have been presented, bias is not. Figure 7 presents the MSE for Gauge 2 in the clean generation section (a) and the MSE normalised by

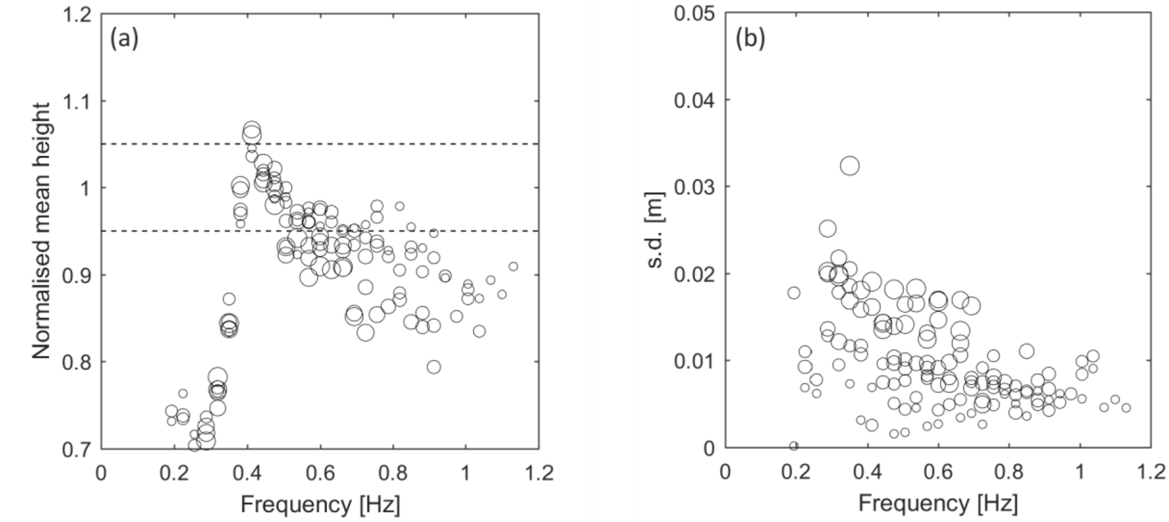

Figure 8 – Normalised mean wave height (a) and standard deviation, s.d. (b), as a function of frequency for all wave heights for Gauge 15 in the clean generation section. Marker size is proportional to the target wave height.

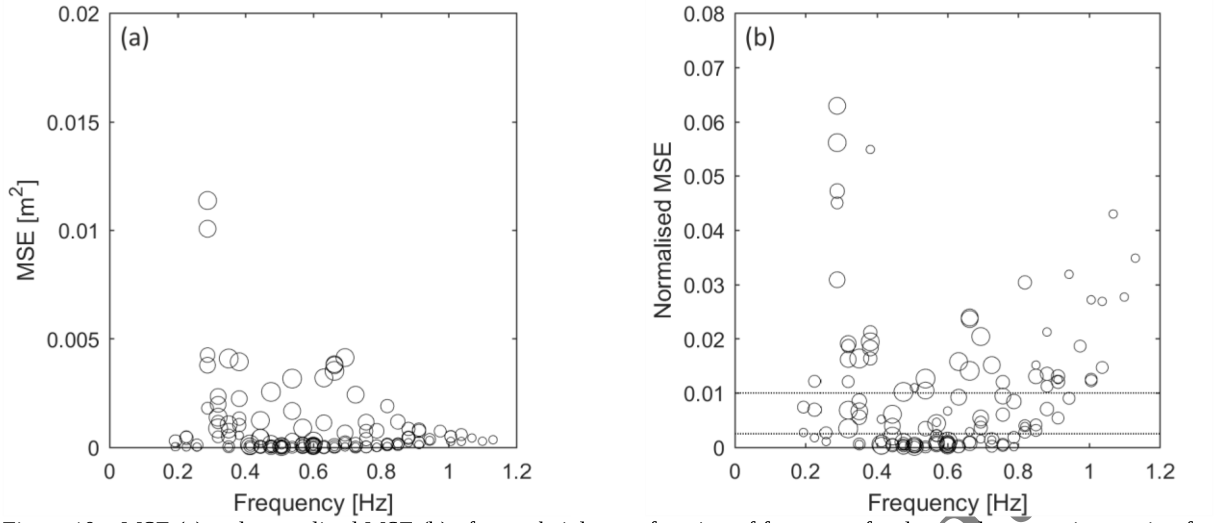

Figure 10 – MSE (a) and normalised MSE (b) of wave height as a function of frequency for the steady generation section for Gauge 2.

the square of the target wave height (b).

By normalising the MSE by the target wave height, we can carry forward the requirement for the mean wave height to fall within ±5% of the target value and impose a stricter requirement. If every wave in the sample were to fall within the ±5% bound, it can be shown that the normalised MSE should be less than 0.0025. The lower dashed line in Figure 7(b) shows this boundary, with the upper dashed line describing the less strict requirement of every wave falling within ±10%. It is interesting that at the high frequency end of the parameter space, the smallest waves had the largest values of normalised MSE and referring back to Figure 8 shows that these waves were typically 85% – 90% of the target wave height.

As with the standard deviation, the MSE tended to be larger for larger waves. The peak in the MSE corresponds to *f* = 0.475 Hz, *H* = 0.8 m; the highest wave in the parameter space. The form of the s.d. plot and the MSE plot are broadly similar, but there is lower correlation between MSE and height than there is between s.d. and height.

Figure 10 shows the MSE and the normalised MSE of the wave height in the steady generation section for Gauge 2. Comparing Figure 7 and Figure 10, the values of MSE and normalised MSE are similar and change throughout the parameter space in a similar way. There appears to be little overall difference between the MSE values for the wave height throughout the parameter space. However, these plots mask differences at discrete points in the parameter space; these are discussed in terms of the MSESS in the next Section.

The proportion of waves meeting the stricter requirement (that all waves should fall within ±5% of the target value) is roughly the same in the clean (Figure 7) and steady (Figure 10) generation sections for the wave height. Contrast this with the values for the period in the clean and steady generation sections, revealed by the normalised MSE plots in Figure 9. The normalised MSE of the period in the clean generation section is very low and all but one datum is lower than the imposed threshold. Despite the low values, there still appears to be a frequency effect – at either end of the frequency range there is a small rise in the normalised MSE.

The steady state generation section plot, Figure 9(b), shows that the normalised MSE of the period was very low for the entire parameter space. This demonstrates that the period was very consistent, not only in terms of the target value but also throughout each wave record.

2) Point comparisons and skill

In Section II, the MSE skill score (MSESS) was introduced as a method to compare predictions and observations against a standard. The MSESS is a way of quantifying the differences seen in the MSE, as demonstrated by Figure 7 and Figure 10, for example. For the wave data, it allows us to compare the measured variables, such as, but not exclusively, wave height and period, to the target values. The skill of a certain condition can then be compared with that of another; here, the clean generation section is compared to the steady generation section for the 2014 data set.

Figure 12 shows how the MSESS for both height and period from one gauge (Gauge 2) varied throughout the parameter space with the marker size indicative of the input wave height and the colour indicating the MSESS value. The MSESS was calculated using (8) and has an upper limit of one. Two things are immediately apparent from Figure 12: more points have negative values of MSESS and the range of values is much greater for the height than for the period. A negative value indicates that the steady generation stage performed worse compared to the clean generation section with respect to the target value. For Gauge 2, there

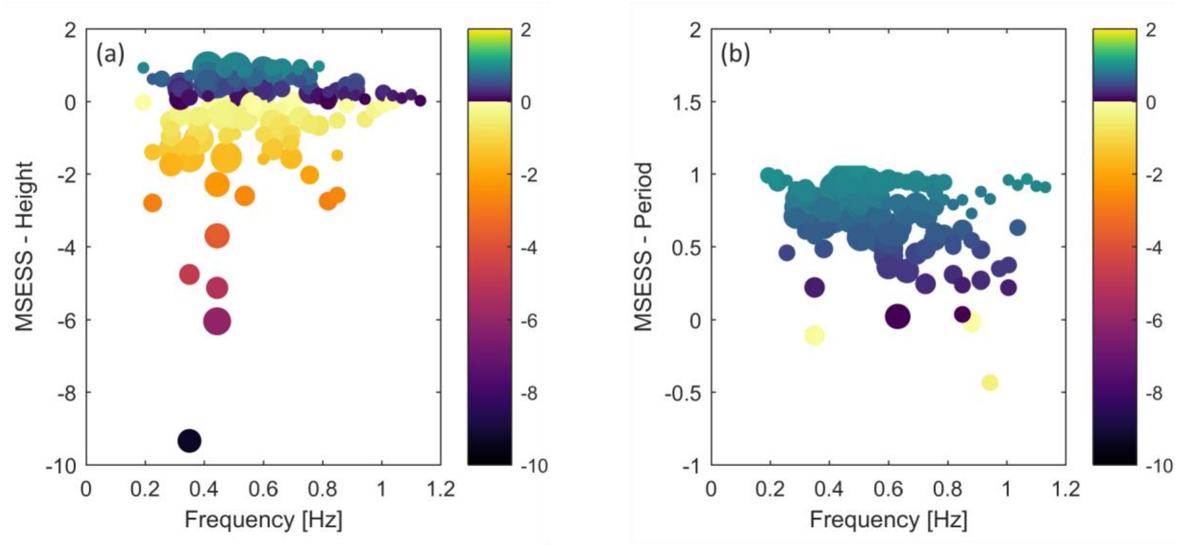

Figure 12 – MSESS throughout the parameter space for both wave height (a) and period (b) as calculated on Gauge 2 of the 2014 data set.

was no pattern linking the negative values of height MSESS to the input height, although there is a concentration of negative values at approximately $f = 0.4$ Hz – 0.5 Hz. This is in part owing to the extension of the parameter space at this point: there were more values of height tested at this range of frequencies because the theoretical parameter space (see Figure 3) allowed larger heights to be produced. The negative values of MSESS shown in Figure 12(a) indicate that for much of the parameter space, the clean generation section performed better than the steady generation section in terms of meeting the target wave height. In contrast, there were only three points in the parameter space where this was true for the period, as shown by Figure 12(b).

If all gauges are considered, the differences between the height and the period are more remarkable, as shown in Figure 11. For the height MSESS, there were many data throughout the parameter space (906 of 2400 data) that had negative values. The ordinate of Figure 11 has been truncated at -10 for clarity, concealing 21 data with values between -10 and -55. For the period MSESS, the negative values only appeared at the edges of the parameter space and there were fewer negative points (109 of 2400 data).

Many of the very low height MSESS values occurred between $f = 0.3187$ Hz and $f = 0.4437$ Hz and this range corresponds to the points at which much larger waves were seen in the normalised measured height in both the clean and the steady state generation section (*cf.* Figure 6 and Figure 8). The low period MSESS values tended to fall at either end of the frequency space, where the normalised MSE was also highest in the clean generation section as indicated in Figure 9.

The MSESS highlights the differences between

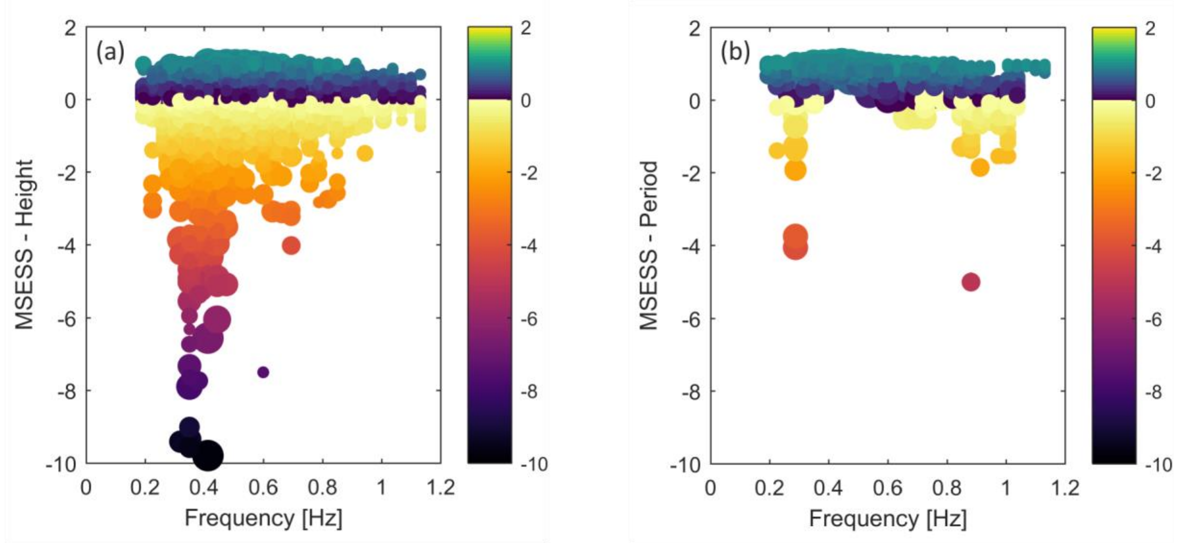

Figure 11 – MSESS throughout the parameter space for both wave height (a) and period (b) as calculated for all gauges in the 2014 data set.

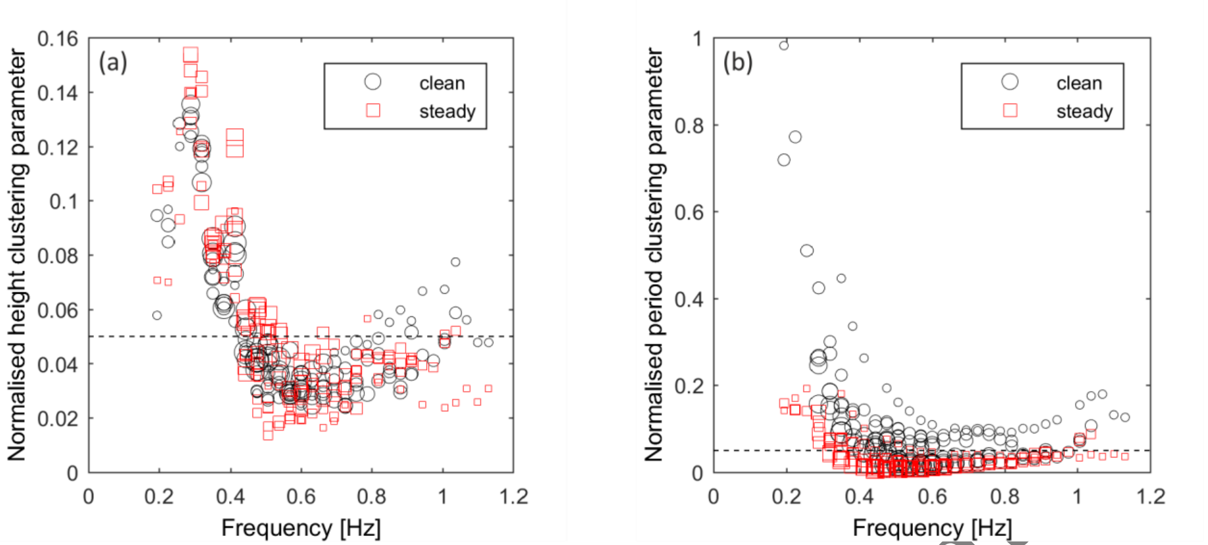

Figure 13 – Clustering parameter, c_p, based on the wave height (a) and period (b) throughout the parameter space of the 2014 data set. Marker size is indicative of input wave height.

the clean and the steady generation sections but caution must be exercised when making comparisons; since it is a comparison between two different conditions or states, it cannot be used to say how good either of those conditions were, only which was better. This is especially true if we consider two values of MSESS representing two different parameters; it is not possible to say which parameter was more like the input value, only which performed better compared to a baseline value.

Threshold levels of MSESS were proposed by [27] and used by [20] for qualification of prediction models for sediment transport. For example, a MSESS of 0.80 would always be considered excellent using the criteria and MSESS values calculated. To put that into context, if we take the threshold values for normalised MSE associated with every wave falling within ±10% of the target value as the old value, and the value for the ±5% boundary of normalised MSE as the new value, the MSESS threshold would be 0.75. It is noted by [20] that different disciplines will have different thresholds of MSESS that are classified as 'useful' and this must be taken into consideration when proscribing qualitative labels for MSESS values.

3) Cp and skill for homogeneity assessment

Figure 13 shows the clustering parameter, c_p, as a function of input frequency for wave height, normalised with respect to the input height and period normalised by the reciprocal of the input frequency. For both the clean generation and the steady section of the wave record, the clustering parameter followed the same trend with increasing frequency: a rapid increase and a peak at approximately $f = 0.3$ Hz, a rapid decrease to a local minimum between $f = 0.5$ Hz and $f = 0.6$ Hz and a final slow rise towards the end of the

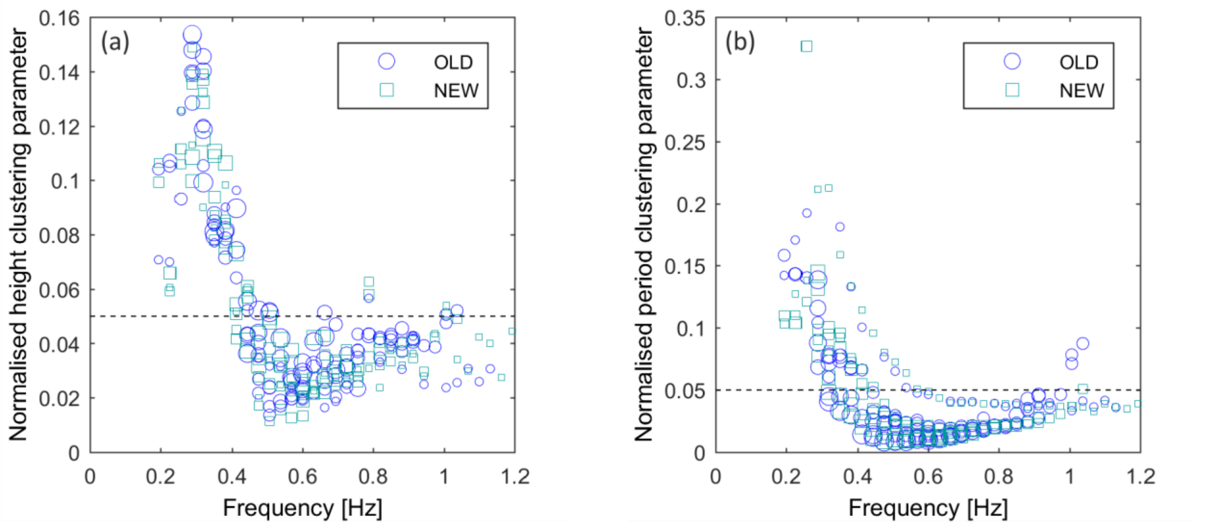

Figure 14 – Clustering parameter, c_p, based on the wave height (a) and period (b) in the steady state throughout the overlapping parameter space for both (2014 and 2016) data sets. Marker size is indicative of input wave height.

parameter space.

Figure 13(a) indicates that overall, the height clustering parameter, hence the homogeneity of the wave height, had similar values in the clean and the steady generation section. As with the MSE plots presented, this does not highlight the differences at each point in the parameter space, but does indicate that frequency had more of an effect on the c_p than the choice of wave record section did. The point at which the clustering parameter switched from decreasing to increasing values approximately coincides with the deep water wave limit of the basin. After this minimum, the clustering parameter showed a small increase with the input height across the parameter space.

As the frequency increased, there was a similar overall trend for the period clustering parameter, Figure 13(b), as for that of the height, especially at the lower frequencies, although the values for the period clustering parameter were several times larger than the values for the height clustering parameter for the clean section. Figure 13(b) has had its ordinate truncated at one to better show the data, obscuring two data from the clean section. In general, the steady state section had lower values than the clean section for clustering of the period data.

The cause for such high values of clustering parameter are not discernible from Figure 13(b) and may be surprising considering the high quality of the period data presented so far. Further investigation into the data reveals several contributing factors. Figure 15 shows that for a particular experiment in the parameter space, the clean section data had larger values of both group residual and s.d., with one gauge considered an outlier (Gauge 14).

At the low frequencies, such as the one shown in Figure 15, for which the speed of the wave is high, there are as few as three data contributing to the mean measurement in the clean section. This means that not only are the mean and s.d. less reliable as measures of the true values, but the influence of the section delimitation is higher. It was already shown in Figure 5 that waves preceding the start of a clean section can have parameters different to target values and it may be that the period suffers more from this than the height for this point in the parameter space. It is not necessary however, that the lower homogeneity of the height data in the clean and the steady state sections be caused by the same physical phenomenon. Reflected wave energy will certainly play a role in the steady state at the lower frequencies. An assessment of the reflections in the UoP Ocean Basin showed that at frequencies lower than 0.3 Hz, the reflections were greater than 20% of the input height [28].

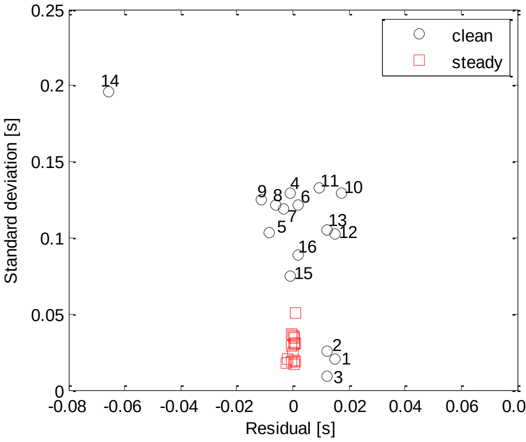

Figure 15 – Clustering of the gauge data for period in the s.d.–residuals space for one file, with $H = 0.2$ m, $f = 0.25625$ Hz. Data are numbered for the gauges for the clean section only, numbering as in Figure 4(a).

The clustering skill score, CSS, introduced in Section II, is presented in Table 1 to compare the clean portion versus the steady state data for both the 2014 and 2016 data, and to compare the old and new data delineated into both clean and steady state sections. The CSS quantifies the clustering

Table 1 – Values of clustering parameter skill score, *CSS*, for height and period comparing both clean and steady generation sections for both data sets with 3 m water depth. Values in parentheses denote the *CSS* from the comparison subset of the old (2014) data.

	Clean vs Steady		Old vs New	
	OLD 2014	NEW 2016	CLEAN	STEADY
Height	-0.0879 (-0.0203)	-0.1424	0.1553	0.0542
	Figure 13(a)	--	--	Figure 14(a)
Conclusion	Clean generation section produced more homogenous heights	Clean generation section produced more homogenous heights	New height data showed more homogeneity	New height data showed more homogeneity
Period	0.6825 (0.6802)	0.3402	0.4930	-0.0459
	Figure 13(b)	--	--	Figure 14(b)
Conclusion	Steady generation section produced more homogenous periods	Steady generation section produced more homogenous periods	New period data showed more homogeneity	Old period data showed more homogeneity
…across 16 gauges for the whole parameter space				

parameter results, such as from Figure 13, and the CSS from comparisons not presented graphically are also given. The CSS values given in Table 1 associated with Figure 13 show that the clean section produced more homogenous wave heights than the steady state section for the old and new data, although this was a small effect. There was a bigger effect associated with the period homogeneity measured in the clean section and the steady state: the CSS was >0.68 for the old data set and 0.34 for the new data set.

Figure 14 shows the clustering parameter comparing the steady state of the 2014 (old) and the 2016 (new) 3 m data sets. In order to compare similar situations, a subset of both data sets was used such that data were from the same values of frequency and amplitude. The behaviour of the clustering parameter is very similar for both the 2014 and the 2016 data in the steady state, indicating that the factors affecting the wave have a strong frequency component. From this, it is not unreasonable to conclude that a significant proportion of the variability in the wave data, particularly at the low frequencies, may be due to basin characteristics.

With the water depth equal to 3 m, the deep water wave limit is $f_{\min}$ = 0.51 Hz and so the low-frequency end of the parameter space, *qv*. Figure 3, contains transitional waves, not deep-water waves. This means that waves in this part of the parameter space will feel the presence of at least some portion of the slope in front of the main floor section (Figure 2) and the main floor section itself.

If the presence of the floor adversely affects the values recorded on the wave gauges, then for the experiments conducted at a water depth of 2 m, more of the parameter space will have higher values of the quality measures presented so far. Figure 16 shows the clean and steady state clustering parameters for the 2 m deep basin experiments. Comparing Figure 16 with Figure 14, it can be seen that the minimum clustering parameter for the wave height is approximately coincident with the onset of deep water waves, now $f_{\min}$ = 0.62 Hz. The period clustering parameter had a similar behaviour with frequency in both the 3 m and 2 m experimental cases, although the latter provided much larger values of clustering parameter. This indicates that the homogeneity of the period when the water depth was 2 m was lower than when the floor was at 3 m. This does not indicate which water depth provided values closer to the input values, although from the results presented so far, it is reasonable to assume that the period was very close to the input in both cases.

C. *Discussion of the case study*

The use of the clean and the steady state as comparative data sets was not to demonstrate the quality of the waves in either section but to allow the analysis methods and the wave quality to be discussed while teasing out issues such as reflections and data paucity. It also served to demonstrate that some artefacts in the data are likely physical basin effects and therefore unavoidable in the context of experimental design.

The stand-out result of the UoP Basin case study is that there was a distinct frequency effect visible in nearly all of the results presented. This typically took the form of larger normalised mean and s.d. of the target parameters at frequencies around f = 0.3 Hz. Most interestingly, this phenomenon was apparent in both the clean and the steady section so the absence, or presence, of any reflections is not the sole cause.

The fact that all gauges seem to be affected by the frequency of the waves (as seen in Figure 11) and in both sections (see, for example, Figure 13) points to a physical effect of the wave paddle-basin system. A simple explanation would be that the

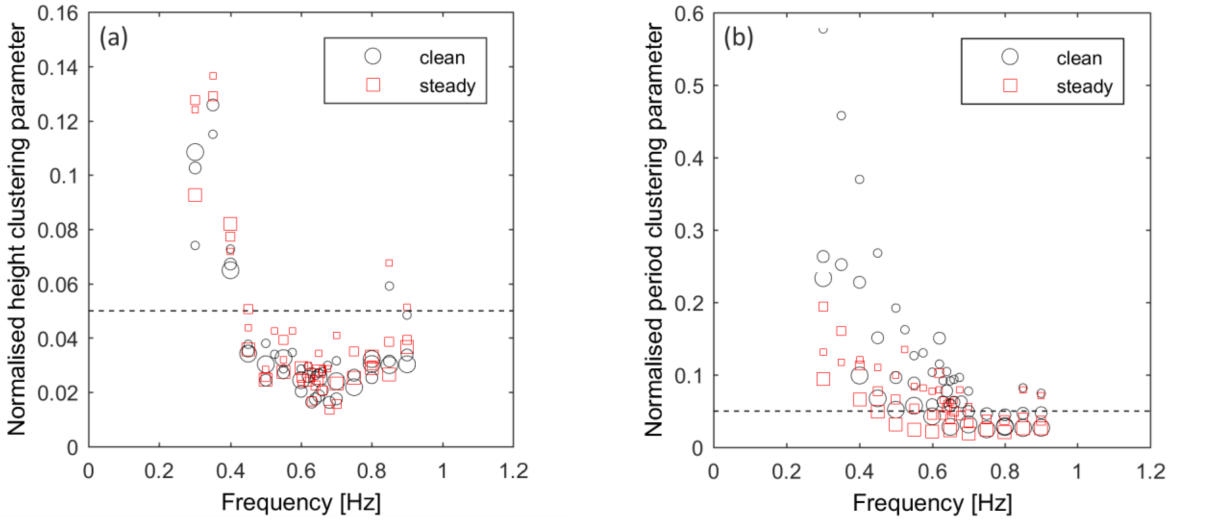

Figure 16 – Clustering parameter, c_p, based on the wave height (a) and period (b) for the 2 m deep basin experiments comparing the clean and the steady states. Marker size is indicative of input wave height.

gain of the paddle transfer functions was too low at very low frequencies and too high at $f = 0.3$ Hz, but this is unlikely to be the primary issue since the variation in wave height and period also increased when the normalised values were high (e.g. Figure 6 and Figure 8).

1) Experimental inefficiency

Data collection for the whole parameter space was a time-intensive task that limited the periods in which data could be collected. Others have published work in which packets of waves containing multiple frequencies were used to calibrate the transfer functions and examine the wave climate in basins [18, 29]. Whilst this dramatically reduces experiment time and confers other advantages as described in [29], this method relies on the multiple waves not interacting. Furthermore, it provides no data for testing stationarity, which then has to be assumed. In future benchmarking operations, it will be interesting to use different data collection strategies.

2) Data paucity and section delimitation

A significant influence on the results is the number of data that make up each mean gauge value. As previously mentioned, the finite length of the time series coupled with a certain wave celerity meant that at the very low frequency end of the parameter space, the number of waves that could be classed as belonging to the clean generation section was very small. At these points in the parameter space, the clean section record length, and hence the number of waves was small; however, during the steady state portion of the record, the number of waves was much larger.

The number of data available obviously impacts on the validity of mean and standard deviation measurements but there are two factors to consider. First, without a change to basin geometry, there is no way to extend the clean generation period in the wave record. Whilst it would be informative to compare these results with those from a longer basin, the comparison might confuse the issue when different basin effects were present. Secondly, the effects of frequency on the values can be seen in both the clean and the steady generation sections and so it is reasonable to conclude that the frequency effect in the clean generation section is not only an artefact of data paucity.

The section delimitation according to the group celerity introduced some uncertainty to the results, in particular those from the clean generation section which were more likely to be contaminated by the first waves, and low-frequency waves where the influence of the first waves would be bigger. The section delimitation was an approximation based on the theoretical value of group celerity and an overestimate of the distance travelled, owing to using the basin cavity length as the full-basin length. Cross-correlation of the signals arriving at different wave gauges to determine a more accurate measure of the group celerity was investigated but was not found to be a reliable method. By reducing the length of the basin in the calculation of the section delimitation, the first (transient) waves, *q.v.* Figure 5, were often captured within the clean generation section. A better method of dividing the time series into sections may be possible but given the other problems with relying on the clean section, this may be moot.

3) Non-deep water waves

The presence of the floor introduces a component of variability to all the non-deep water wave results that is difficult to address. For example if the floor is considered to be perfectly planar but is not, or has local high points, the water depth will not be homogenous around the basin thus it is not surprising that the results should not be homogenous either. It was found by O'Boyle et al [16] that (unreported) variations in the measured bathymetry of the QUB Portaferry basin caused observable spatial variation in spectral wave height in their numerical model, although with a water depth of only 0.5 m, that floor would have had more of an effect. From that, we can conclude that the variability in the wave height and period across the basin in the non-deep water end of the parameter space could be reduced by ensuring a homogenous floor bathymetry. In practice this is difficult to achieve and the cost/benefit ratio would likely render it unfeasible. Later work at the QUB basin reported the side- and back-wall reflections and diffraction as the main sources of wave climate variability [30].

For any basin (using linear wave theory to generate waves), the limit of the deep-water wave generation ultimately depends on the basin geometry itself rather than on the wavemakers. Owing to the hyperbolic tangent function in the dispersion relation, the relationship between the depth and the maximum period (minimum frequency) is not linear. For example, a basin with $d = 2.0$ m can produce deep water waves with a maximum period of $T = 1.60$ s (minimum $f = 0.62$ Hz), whereas for a basin with $d = 3.0$ m, the maximum deep water wave period is 1.96 s (minimum $f = 0.51$ Hz).

4) Reflections

Section III.B reported that reflections in the basin were previously found to be as high as 20% of the input height at the very low frequency end of the parameter space. The beach, a parabolic glass-fibre structure covered with metal meshing, was not designed to attenuate all incident waves but to reduce those from the likely working area of the parameter space the most, thus reflections will be present in much of the steady-state. If the steady state is to be used, then it makes sense to adjust the basin settings such that the wave height is best represented in the steady state, taking into account the potential reflected wave energy. However the facility manager and the experimentalist must ensure that this is agreed before an experimental campaign.

Another potential physical effect could be cross waves in the basin, which are frequency (and basin geometry) dependent. Cross waves can be predicted using the ratio of basin width to the half wave length. If the number of half wave lengths that can be accommodated in the basin width is given by n, the remainder, as calculated by mod(n,1), can be used to predict the presence of cross waves. As the remainder approaches these zero or one, the closer the number of half wave lengths is to an integer value, indicating the likelihood of cross waves. Figure 17 shows mod(n,1) as calculated for the parameter space with the red dotted lines and the grey dashed lines designating 10% and 15% from zero and one.

Comparison of Figure 17 with Figure 6 shows that a cross wave was likely present at $f = 0.2875$ Hz, which is the same frequency at which there was an increase in the normalised wave height on Gauge 2. At this frequency, the wavelength was 15.73 m; close to the width of the basin. The subsequent frequencies at which there may have been cross waves (f = [0.38125, 0.44375, 0.63125, 0.97500] Hz) do not correspond to high points in the normalised height (Figure 6), though $f = 0.38125$ Hz corresponds to the peak in standard deviation on Gauge 2. It may be that by examining the harmonic content of the time series the presence of cross waves can be established, but this is out of the scope of this work. It is noted that some schemes for analysis of reflections can also deal with cross waves [9]. These trends are replicated for the normalised mean height in the steady generation section for Gauge 2. Figure 8(a) shows that the peak in the normalised mean height coincided with $f = 0.41250$ Hz, which Figure 17 indicates was not a frequency at which cross waves were likely to be present.

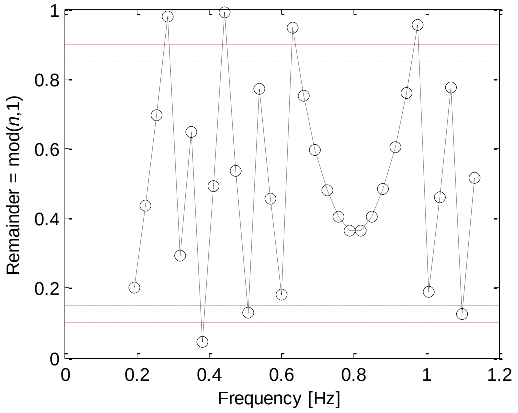

Figure 17 – Remainder values close to zero or one indicate that a particular frequency produced waves with a half wavelength close to an integer factor of the basin width.

5) Selecting a portion of the wave record

As previously mentioned, there is no real consensus on the number of waves that need to be analysed nor whether they should be free from reflections or recorded in the steady state of the basin in question.

For regular waves, a guideline figure of 50 – 100 waves is suggested when testing a device/structure to allow resonance effects or instabilities to be revealed [15] although for these effects to be measurable, the length of the signal has to be longer if the effect is small in magnitude. Other advice is to select the time interval for which the length is a multiple of the period, starting after the first transient sequence and ending before the apparition of reflections and the last transient sequence [25].

To allow 50 waves to pass by a point in a basin before reflections from the far end can be detected requires either high frequency waves, since they travel more slowly than low frequency waves, or a very long basin. For example, with a basin 35 m long such as the UoP Ocean Basin, with a wave gauge situated in the middle, the lowest wave frequency that would conform to these restrictions would be $f = 1.06$ Hz. Comparing this limit to the theoretical parameter space of the basin, Figure 3, reveals that wave amplitudes would have to be kept small to be able to produce these waves, but also that most of the parameter space would be inaccessible if the condition of 50 clean waves were imposed. To allow fifty 0.5 Hz waves to pass by a point 15 m along a basin would require a basin 95 m long. It seems then that the ability to have large numbers of clean waves is something only flumes and towing tanks with wave-making capability can reasonably expect to achieve. In doing so, however, the ability to examine 3D effects would be compromised.

In light of the physical limitations of the basin and measurement requirements, it is reasonable to conclude that the steady state portion of a regular wave time series should be used for experimentation in the UoP Ocean Basin. With a finite-length basin, reflected wave energy is unavoidable but may not be a barrier to good experimentation. For devices operating in heave, for example, the direction of wave movement is not as important as the local wave height, so reflections need not be a concern [16].

D. Conclusions of the case study

Given the imposed necessity to record at least 50 steady-state waves, a logical question to ask is how the choice of the steady state affects the quality of the waves. The normalised mean wave height data for the gauge closest to the paddles, Figure 7(b) and Figure 10(b), suggest that wave height was less accurate throughout the parameter space during the steady state than during the clean section. However, the accuracy is dependent on the gauge position and the frequency and height of the wave. The plot of MSE skill score, Figure 11 shows that the majority of data had positive MSESS, indicating that the steady state was more accurate than the clean generation section. In terms of homogeneity, the clustering parameter results summarised in Table 1 show that for the both the old and new data, the clean section had better homogeneity of wave heights.

For the measured wave periods, nearly all of the presented results suggest that the steady state delivers higher quality waves than the clean generation section. The exceptions are the MSESS values on certain gauges at the edges of the parameter space, Figure 11, however it is noted that the periods fell well under the quality thresholds imposed by a ±5% accuracy requirement; see for example Figure 9.

IV. Discussion of the quality metrics methods proposed

The goal of this paper was to propose a methodology by which the quality of a wave field in a basin could be assessed in terms of the accuracy and precision of the wave parameters, the basin homogeneity and by a quantification of alterations to the basin hardware, software or the analysis procedures used. In Section III, the selection of graphs presented indicated that a portion of the waves in the parameter space fell within the ±5% of target height and period but that there were also frequency effects believed to be related to basin geometry. The MSE and normalised MSE were used to demonstrate that a stricter quality measure could be imposed that had implications for the consistency (precision) of the basin. The MSESS was used to show that the steady generation section was overall much better than the clean generation section for the wave period but not necessarily for the wave height and again frequency played a large part. Finally, the clustering parameter and the CSS were used to show that the homogeneity of the basin was greatly improved in the steady generation section for the period but not for the height.

A. Accuracy and precision measures

The mean squared error, MSE, was used to provide a measure of accuracy but only at one point in the physical space of the basin. MSE gives more importance to outliers than the MAE (mean absolute error) does owing to the squaring operation but the threshold value that guarantees all waves fall within the boundary can be calculated in the same way. Since the MSE does not give an indication of the sign of the error it is often presented with bias, although here it was presented with the normalised mean values.

The MSE can be expressed as the sum of the variance and the bias squared, which gives another calculation method. If the bias and the MSE are equal, this implies that the variance is zero and that perhaps the basin gain should be adjusted to reduce the wave height bias. This also implies that the MSE and the variance are equivalent if there is no bias. Since we have introduced a quality threshold for the normalised MSE, the threshold value can be assigned to the variance and the bias. For example, if the mean height is to be within ±5% of the target value and every wave should fall within that bound too, the normalised variance cannot be larger than the threshold of the normalised MSE. .

The main drawback with the normalised wave variables and MSE values is that they are per gauge measurements. This means that there would be one status plot for each gauge, and for each further condition (for example floor depth). Given that spatial measurements can be almost infinite in number, this is too much information to deal with. This also makes quantification of the quality of a basin hard to achieve.

B. MSESS critique

Whilst it is possible to use the MSESS for both spatial and temporal data, the MSESS cannot be extended to accommodate both. The application of the score to sediment transport models in [20] compared a baseline prediction and a more sophisticated model of sand movement. Despite the modelling process being iterative, this is an example of spatial data as the intermediate iterations of the model are of no concern in the MSESS. This is in contrast to the way in which we propose to apply the MSESS, in which the accuracy of each wave is important, not only the accuracy of the final wave. Thus the need to define a separate measure of homogeneity.

Much of the criticism levelled at using the MSESS concerns its use with probabilistic models, such as those used in [20] and many of these are reviewed and discussed by [24]. One of the principal concerns is the choice of the baseline: [24] notes that for meteorological forecast skill, the baseline is chosen to be an unskilful but not unreasonable forecast, yet in [20] the baseline used is generally the initial (sediment bed) formation. In the application of the MSESS to wave basin data to benchmark quality, the baseline is the current or former condition of the basin before any upgrades or changes have been made, in a similar way to the use of recently observed values in meteorology. The MSESS is not assessing the skill of a prediction method but quantifying the wave accuracy before and after a ‘treatment’.

C. Clustering parameter critique

The clustering parameter was developed to amalgamate precision and homogeneity information to provide a whole-basin measure of quality using the variance from each gauge and the residuals based on the gauge-group mean. The clustering parameter is essentially a form of the root mean squared error (RMSE) that uses residuals rather than prediction errors, i.e. an RMSE that does not reference the target value. If, however, we apply the same threshold technique as for the normalised MSE and introduce the requirement that the gauge-group residuals should be within ±5% of the group mean, the limit of the normalised clustering parameter is also 5% or 0.05. The graph of the normalised clustering parameter, Figure 13, reveals that for the height clustering parameter, most data at frequencies higher than 0.45 Hz fall within this boundary; for both the clean and the steady state section. For the period clustering parameter, all but one of the data are within this boundary.

The advantage of ignoring the target value in the clustering parameter is that it allows a measure of how close the values are to the others in the gauge group, thereby providing a measure of the homogeneity. It can be argued that homogeneity of the basin is more important than achieving the target value if the size of the experimental area is large. For the UoP Ocean Basin, array tests covering much of the basin width are often performed. For early-stage array tests, in which monochromatic waves are used, each device is expecting the same wave climate (at least within the first row). Spatial variation in wave height due to basin reflections and non-homogeneity has been found to obscure array effects, which might be small compared to the parameter being measured [16]. In addition, non-homogeneity may also be an issue for numerical model validation if it is not replicated by the numerical model [17].

The clustering parameter is a broad measure and, like the other methods here, does not highlight the causes of the results seen. It has the advantage that it is not possible to have a low clustering parameter value with bad homogeneity; good and bad values do not cancel as they might in bias calculations. As the clustering parameter is nominally in the same units as the parameter in question, care must be taken when comparing two values. In III.B, the clustering parameters were normalised by the input values allowing them to be compared across the parameter space. Once normalised, it is possible to compare the clustering of the data with respect to two different input parameters, although these should not be conflated.

The clustering parameter could be used as a metric to define the working area of the basin with enough wave gauge data from around the basin. For example, the working area could be defined as the area covered by wave gauges whose clustering parameters fell below a certain threshold. However, this relies on a certain density of spatial data and assumes reproducibility and stationarity. When the results presented in the previous Section are considered, it seems likely that this interpretation of what constitutes a working area would have to be defined at each point throughout the parameter space, including at different water depths. For experimentalists planning array tests, it is a much better strategy to run empty basin tests with wave gauges at the locations of interest and to work with the facility to minimise the clustering parameter through calibration.

D. Analysis artefacts

In this paper we have presented all the analysis as the result of a time domain-based, zero-crossing analysis of the wave heights and periods. From Figure 1 it is clear that the analysis forms part of the whole system and introduces its own simplifications and errors. The MSE can be expressed as the sum of the variance and the squared bias, which can both be derived from the spectral moments, so it is possible to calculate the MSE, MSESS and the CSS in the frequency domain, although this will introduce a different set of assumptions and errors. However, it is worth considering so that the methods used here can be extended to irregular waves for which a zero-crossing method would not be appropriate. In this case it would be appropriate to measure height and period and to formulate a metric to quantify adherence to the desired spectral shape. This could be done either as a deviation from the shape itself or by using each of the spectral moments as quantifiers.

E. What cannot be discerned from these measures?

What the methods developed in this paper do not cover is the cause or source of any deviation in quality. Since the measures take into account the whole process from wave file creation to analysis, as described in Figure 1, an incorrect wave file, basin variation or analysis artefacts could all adversely affect the results but it is not possible to attribute variation to any of these without further investigation. This is not an atypical situation though; these methods just serve to quantify the accuracy, precision and homogeneity of the data. A non-exhaustive summary of the artefacts seen in the results presented and their potential causes is presented in the Appendix in the form of a troubleshooting guide.

It has been noted that the MSE and the MAE, and by extension the MSESS and CSS, imply that all error is due to the real-life variations of the data and not attributable to measurement errors [20]. Our consideration of the full system recognises that measurement plays a role in the quality scores without specifically quantifying the associated errors. The next step for further investigating the data would be a full consideration of the measurement errors and how the metrics should be interpreted in light of these. A good discussion of error and uncertainty can be found in the EquiMar project deliverables [14].

Just as the root causes cannot be determined with a few metrics without further investigation, any necessary corrective action is not implied either. Modifications to the system that lead to reduced MSE values or clustering parameters may not be cost effective if they produce little or mixed improvements. However, with an understanding of the skill scores, it is possible to perform a cost-benefit analysis; the EquiMar project [14] discussed the use of repeated measures for decreasing the precision limits and how it is up to the Facility Manager to decide on the cost/benefit.

F. *Extension of the methods*

Essentially any metric of the quality is a trade-off between aggregated parameters and detail. The clustering parameter could easily be extended to consider multiple parameters, such as period and steepness, resulting in one metric but this would be less useful in the interpretation of the data and in the consideration of further action. It would also be possible to apply a weighting to the components that make up the clustering parameter, such that group residual or standard deviation could be penalised more heavily than the other depending on the objective of the assessment.

V. Conclusions

In this paper, we have devised and presented a series of novel metrics for the quantitative evaluation of a wave basin. Not only can these metrics be used to quantify accuracy and precision as an absolute measure, but we have developed a skill score that allows the relative comparison of wave field quality, thus the benchmarking and evolution of a basin can be quantified.

These methods have applications for all those with an interest in wave basins. Initially, the methods of evaluation are interesting to Facilities Managers during the commissioning process of a new facility or during demonstration of capacity.

We have shown how the homogeneity of wave height and period can be quantified. These methods can be extended and applied to all types of basins to allow potential basin users to determine whether a basin is suitable for their needs. For example a basin suitable for a single device may not be as good for testing an array of devices. The benchmarking of the basin is also useful to users of the basin who may be testing several months, or even years, apart. A user may opt to change the experiment design in light of an update to the basin and uncertainty in the measurements may be more fully discussed with an accurate picture of the wave quality.

Finally, as noted in [17], it is often taken for granted during numerical model validation that the waves in physical basins are accurate, precise and homogenous, which can lead to errors in the validation process. By understanding the limits of physical basin wave quality, better agreement between physical and numerical modelling can be achieved. This paper presents a first step towards the quantification of wave quality and it is hoped that these methods are adopted to allow greater understanding of the wave generation facilities available to end-users.

VI. References and Acknowledgements

KMC gratefully acknowledges the support from HEFCE funding for the overall research through the School Engineering (formerly the School of Marine Science and Engineering) at the University of Plymouth. Colleagues in the School of Engineering are thanked for their comments on the manuscript.

1. Munk, W.H. *ORIGIN AND GENERATION OF WAVES*. in *Coastal Engineering*. 1950. Long Beach, California, October.
2. Cavaleri, L., J.H.G.M. Alves, F. Ardhuin, A. Babanin, M. Banner, K. Belibassakis, M. Benoit, M. Donelan, J. Groeneweg, T.H.C. Herbers, P. Hwang, P.A.E.M. Janssen, T. Janssen, I.V. Lavrenov, R. Magne, J. Monbaliu, M. Onorato, V. Polnikov, D. Resio, W.E. Rogers, A. Sheremet, J.M. Smith, H.L. Tolman, G. van Vledder, J. Wolf, and I. Young, *Wave modelling - The state of the art.* Progress in Oceanography, 2007. **75**(4): p. 603-674.
3. Havelock, T.H., *LIX. Forced surface-waves on water.* The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1929. **8**(51): p. 569-576.
4. Frigaard, P., M. Høgedal, and M. Christensen, *Wave Generation Theory*. 1993: Hydraulics & Coastal Engineering Laboratory, Department of Civil Engineering, Aalborg University.
5. Spinneken, J. and C. Swan, *The operation of a 3D wave basin in force control.* Ocean Engineering, 2012. **55**: p. 88-100.
6. U.S. Army Corps of Engineers, *Water wave mechanics*, in Coastal Engineering Manual. EM1110-2-1100, 2008
7. Suquet, F. and F. Biésel, *Laboratory wave-gnerating apparatus*. 1953, Laboratoire Dauphinois d'Hydraulique: University of Minnesota. p. 133.
8. Brossard, J., A. Hemon, and E. Rovoalen, *Mesure de houle par une ou deux sondes mobiles*, in *6e Journées de l'hydrodynamique*. 1997: Nantes.

9. Andersen, H., J. Grønbech, and T. Jensen, *Determination of incident waves using reflection analysis.* 1995.
10. Schäffer, H.A., *Second-order wavemaker theory for irregular waves.* Ocean Engineering, 1996. **23**(1): p. 47-88.
11. Spinneken, J. and C. Swan, *Second-order wave maker theory using force-feedback control. Part I: A new theory for regular wave generation.* Ocean Engineering, 2009. **36**(8): p. 539-548.
12. Spinneken, J. and C. Swan, *Second-order wave maker theory using force-feedback control. Part II: An experimental verification of regular wave generation.* Ocean Engineering, 2009. **36**(8): p. 549-555.
13. *Final Report and Recommendations to the 23rd ITTC*, in The Specialist Committee on Waves. 2002
14. *D3.4 Best practice for tank testing of small marine energy devices*, in EquiMar. 2010
15. Holmes, B., *Tank Testing of Wave Energy Conversion Systems*. Marine Renewable Energy Guides. 2009, London: BSI.
16. O'Boyle, L., B. Elsaesser, M. Folley, and T. Whittaker, *Assessment of wave basin homogeneity for wave energy converter array studies*, in *9th European Wave and Tidal Energy Conference*. 2011: Southampton, UK.
17. Lamont-Kane, P., M. Folley, and T. Whittaker, *Investigating uncertainties in physical testing of wave energy converter arrays*, in *10th European Wave and Tidal Energy Conference* 2013: Aalborg, Denmark.
18. O'Boyle, L., B. Elsäßer, and T. Whittaker, *Methods to enhance the performance of a 3D coastal wave basin.* Ocean Engineering, 2017. **135**: p. 158-169.
19. Nurmi, P., *Recommendations on the verification of local weather forecasts*, in *Technical Memorandum*. 2003, ECMWF: Shinfield Park, Reading. p. 19.
20. Sutherland, J., A.H. Peet, and R.L. Soulsby, *Evaluating the performance of morphological models.* Coastal Engineering, 2004. **51**(8-9): p. 917-939.
21. Berry, M.J.A. and G.S. Linoff, *DataMining Techniques*. Second ed. 2004: Wiley.
22. Draycott, S., T. Davey, D.M. Ingram, J. Lawrence, A. Day, L. Johanning, J. Steynor, and D. Noble, *Applying site specific resource assessment: methodologies for replicating real seas in the FloWave facility*, in *International Conference on Ocean Energy*. 2014: Halifax, Nova Scotia.
23. Coe, R., *It's the effect size, Stupid: what 'effect size' is and why it is important*, in *Annual Conference of the British Educational Research Association*. 2002: Exeter, UK.
24. Bosboom, J., A.J.H.M. Reniers, and A.P. Luijendijk, *On the perception of morphodynamic model skill.* Coastal Engineering, 2014. **94**: p. 112-125.
25. Le Boulluec, M., *Data analysis*, in *HYDRALAB Young researchers workshop*. 2013, IFREMER.
26. Darras, M. *IAHR list of sea state parameters: a presentation*. in *22nd IAHR World Congress: Wave analysis and generation in laboratory basins*. 1987. Lausanne.
27. van Rijn, L.C., D.J.R. Walstra, B. Grasmeijer, J. Sutherland, S. Pan, and J.P. Sierra, *The predictability of cross-shore bed evolution of sandy beaches at the time scale of storms and seasons using process-based Profile models.* Coastal Engineering, 2003. **47**(3): p. 295-327.
28. Collins, K.M., G. Iglesias, D. Greaves, A. Toffoli, and S. Stripling. *The new COAST laboratory of Plymouth University: A world-class facility for marine energy*. in *ICE Coasts, Marine Structures and Breakwaters Conference*. 2013. Edinburgh.
29. Masterton, S.R. and C. Swan, *On the accurate and efficient calibration of a 3D wave basin.* Ocean Engineering, 2008. **35**(8-9): p. 763-773.
30. O'Boyle, L., *Wave Fields around Wave Energy Converter Arrays*. 2013, Queen's University Belfast.

APPENDIX – A non-exhaustive guide to establishing the performance metrics of a basin.

PARAMETER	POSSIBLE VALUES				NOTES/CAUTIONS
CSS	<0	0	0<1	1	
Comparison of the clustering parameters from two data sets (new and old)	New Cp values higher than old values of Cp	New and old values of Cp exactly the same	New values of Cp are smaller than old values	New value of Cp is equal to zero, which in this context implies a perfect wave on every gauge (residuals and s.d. values equal to zero), but not actually defined by the (theoretical) input.	If old value of Cp is zero (i.e. perfect wave at all points) then the CSS would be -∞.
Investigative actions	Plot Cp values to look for outliers/inconsistencies Consider number of data involved in measurement Work out if reflections, cross waves or depth effects are playing a role. Consider if normalised Cp is under target threshold.				
Cp			0<∞		Full basin measurement.
Quantification of basin homogeneity			Check residuals and s.d. values in a cluster plot (Figure 15)		
	If all large residuals		Indicative of bad basin homogeneity. May be affected by reflections, cross waves, depth effects.		
	If one large residual		Check gauge time series/raw data (eg Figure 5).		
	If generally large values of s.d.		Indicative of bad basin stationarity. May be affected by reflections, cross waves or depth effects.		
	If one large s.d		Check gauge time series/raw data (eg Figure 5).		
MSESS	<0	0	0<1	1	
Comparison of two sets (new, old) of MSE values	New MSE values higher than old values of MSE, i.e. more error relative to input in new set compared to old set.	New and old MSE values exactly the same.	New MSE values are smaller than old MSE values, i.e. less error relative to input in new set compared to old set.	New MSE values have no error, i.e. exactly match input.	Point measurement so affected by position in basin. Both sets could have very large or very small MSE values so only relative accuracy is assessed. MSESS may be meaningless if MSE values are smaller than measurement error.
Investigative actions	Check MSE values (and subsequent actions).				
MSE			0<∞		
Quantification of error of values wrt to input (target)			Larger values indicate larger deviations from input.		Point measurement so affected by position in basin.
Investigative actions	Check normalised MSE for input parameter effect. Check range of values similar at points around the basin. Check for presence of cross waves. Check normalised values against threshold for quality assessment. Check for depth effects. Check for reflection effects.				

Investigative actions in more detail:

If you suspect…	**Cause**	**Action**
…a depth effect	Influence of floor induces variance in measurements	Check if operating in deep water Consider conducting experiment in deeper water/at smaller scale Consider not using that part of parameter space
…deflections	Reflected wave energy from beach (and perhaps paddles) that affects homogeneity and stationarity. May introduce non-linearities/higher-order effects.	Do a full reflection analysis to quantify influence of reflections at that water depth Check wave gauges are not at (anti)nodes in basin Consider running in clean only section (not recommended) Investigate modifications to beach structure Consider not using that part of the parameter space. Check for evidence of non-linearities/higher-order modes.
…cross waves	Half wave lengths that are factors of basin width cause cross waves that affect homogeneity and stationarity. May introduce non-linearities/higher-order effects.	Work out if cross waves are likely (see Figure 17) Consider separating out cross waves with a suitable reflection analysis e.g. [9] Investigate modifications to sidewalls to attenuate cross waves Consider not using that part of the parameter space. Check for evidence of non-linearities/higher-order modes.