Proceedings of the 21st ICSMGE, Vienna, Austria, 14 – 19 June 2026. Pistrol, Adam & Schweiger (eds.)
Published by: ÖGG, Austrian Society for Geomechanics, Salzburg, Austria, ISBN 978-3-9503898-4-5
https://doi.org/10.53243/ICSMGE2026-1531

Application of a quantitative goodness-of-fit criterion for pore pressure simulations during liquefaction behaviour

Catherine Jiang, Andy O'Sullivan
Andy O'Sullivan Geotechnical Engineering, New Zealand, catherine@aosullivan.co.nz

Liam Wotherspoon
The University of Auckland, New Zealand

ABSTRACT: Studies focusing on the numerical modelling of liquefaction will generally involve the analysis of pore water pressure responses and require comparisons between measured and modelled responses in order to validate the modelling assumptions or to select a best-fit model. Early studies typically based this comparison on qualitative judgement that can introduce potential bias and may vary from person to person. A recent study proposed a quantitative criterion to evaluate the goodness-of-fit between two pore pressure response time series based on a set of quality metrics. This paper presents an application of the proposed criterion to quantify the goodness-of-fit of a previously collected dataset, which consists of pairs of pore pressure records from dynamic centrifuge tests and numerical simulations. A separately conducted survey collected qualitative judgements of this dataset, and the survey data was analyzed against the criterion evaluation. A total of 25 respondents were been involved in the survey to score eighteen pairs of records. Analysis of the survey data revealed considerable variability in the subjective evaluations, with individual scores ranging from 1 (poor match) to 10 (perfect match) for one pair of records. The average survey scores were then calculated and compared to the criterion scores. The results indicated that the average survey scores exhibited a reasonably high correlation with the criterion evaluation, including the weighted average of the criterion and most of the individual quality metrics. Results suggest the criterionoffers a more systematic, reproducible and unbiased framework for performance evaluation.

KEYWORDS: pore pressure simulation, goodness-of-fit, quantitative criterion.

1 INTRODUCTION

A range of different approaches have been used to assess the goodness-of-fit (GOF) of the time-series response of model simulations against experimental data or recorded data from the field. Early studies typically used qualitative approaches to evaluate the GOF of models based on the match of the overall shape of a response and/or the peak intensity (Sabetta, Pugliese 1996, Hartzell 1982). The results based on subjective judgement are generally less convincing because people may perceive different levels of similarity based on the same data. Instead, quantitative criteria are usually preferred to evaluate the GOF between time series, as it is generally objective and ensures a consistent scale.

Over recent decades, numerous quantitative criteria have been developed to assess the goodness-of-fit (GOF) between ground acceleration time series. In contrast, limited efforts have been directed toward establishing equivalent criteria for evaluating excess pore pressure time histories. A recent study introduced a quantitative GOF metric specifically for pore pressure records (Jiang, Wotherspoon 2024). In the present study, this criterion is applied to a previously collected dataset that consists of paired pore pressure simulations and the corresponding centrifuge test records. The performance of the proposed GOF criterion is compared with qualitative assessments obtained from a survey conducted among postgraduate students in the Department of Civil and Environmental Engineering at the University of Auckland.

This paper aims to evaluate the effectiveness of the proposed quantitative criterion and highlight the benefits of numerical assessment methods over traditional qualitative evaluations, which often rely on visual interpretation.

2 METHODOLOGY

2.1 *Criterion Introduction*

2.1.1 *Score equation*

The proposed criterion consists of five individual quality metrics, each of which is evaluated by a score equation. The generic score equation is a negative exponential function defined in Equation (1).

$$S(n_1, n_2) = 10exp\left\{-\left|\frac{|n_1 - n_2|}{mean(|n_1|, |n_2|)}\right|\right\} \quad (1)$$

where S is the parameter GOF score and n_1 and n_2 are any proposed quality metric from the two records. The function $exp\ (-|z|)$ ranges between 0 and 1, and the factor 10 sets the range to vary from 0 to 10. A score lower than 4 is a poor fit, a score between 4 and 6 is a fair fit, a score between 6 and 8 is a good fit and a score higher than 8 is an excellent fit. Figure 1 shows the relationship between the metric ratios ($\frac{n_1}{n_2}$) and the GOF scores with colors indicating the grade scale.

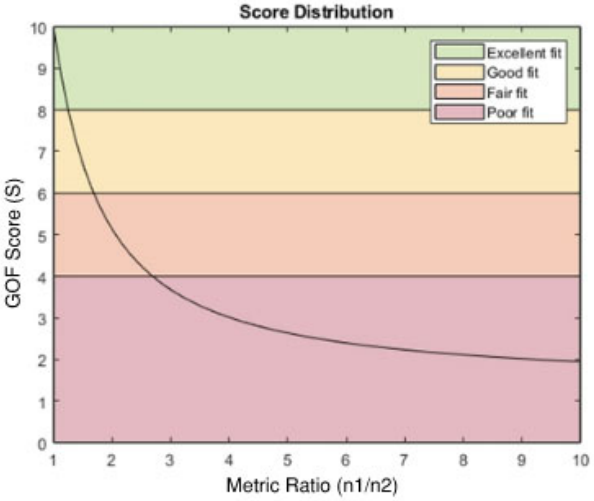

Figure 1. GOF scores as a function of metric ratio ($\frac{n_1}{n_2}$) and the grade scale for different GOF scores.

2.1.2 *Quality metrics*

During the liquefaction process, the excess pore pressure develops over time with a similar general trend. Figure 2 shows

a pair of representative pore pressure records from soils that have liquefied, with one record from a centrifuge test and the other from numerical modelling. Both records consist of a rapid increase in pore pressure followed by a roughly constant period, with significant fluctuations observed throughout the records. The termination of the records may depend on the duration of the ground motion input. Pore pressure records may involve a gradual decrease in magnitude if the ground motion includes post-earthquake response, whereas a sudden cutoff may occur if the post-earthquake response is excluded.

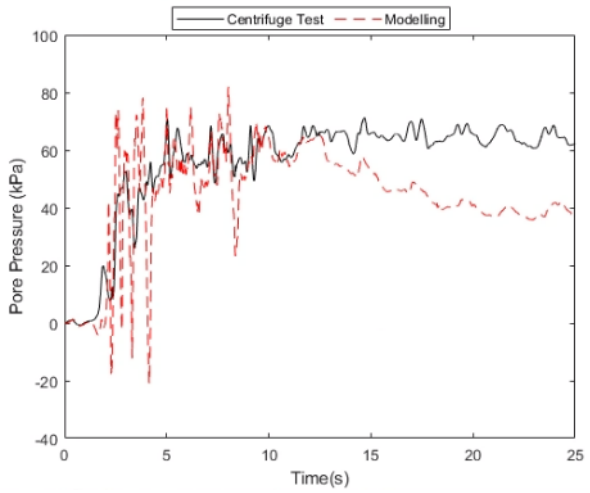

Figure 2. Example pore pressure records extracted from a centrifuge experiment and the corresponding numerical modelling record.

In this study, a total of five criteria quality metrics were proposed to ensure a comprehensive comparison between the two records. Table 1 summarizes the five proposed individual quality metrics to compare two pore pressure time series, $p_1(t)$ and $p_2(t)$.

An overall goodness-of-fit (GOF) score can be derived by aggregating the five proposed quality metrics using either an arithmetic mean or a weighted average. The arithmetic mean provides a simple and uniform assessment by treating all metrics with equal importance. In contrast, the weighted average incorporates engineering judgment by assigning relative weights to each metric according to its perceived significance in evaluating the performance of pore pressure simulations. A weighted average is recommended in this study and Table 1 summarizes the proposed weight combination.

Table 1. Proposed pore pressure GOF criteria quality metrics and the weights used in this study.

Quality Metrics	Definition	Weight
P_D	Average difference	20
P_M	Magnitude of steady pore pressure	40
P_T	Time when steady pore pressure starts	20
P_I	Pore pressure integral	10
P_{DTW}	Dynamic time warping	10

P_D provided a measure of the average difference between two records by evaluating and averaging the point-to-point GOF for the entire time series, as described in Equation (2).

$$P_D = mean[S(p_1(t), p_2(t))] \quad (2)$$

P_M and P_T are the coordinates (magnitude and time, respectively) of the turning point when the pore pressure buildup process stops and reaches a stable state, as described in Equation (3) and (4).

$$P_M = S\left(P_1^{x\%}, P_2^{x\%}\right) \quad (3)$$

$$P_T = S\left(T_1^{x\%}, T_2^{x\%}\right) \quad (4)$$

where $P^{x\%}$ and $T^{x\%}$ are defined as the magnitude and time when $p(t)$ first reaches a set percentage ($x\%$) of its maximum, which is defined as 85% in this study.

The pore pressure integral (P_I), is defined as the integral of the pore pressure record above $x\%$ of the maximum smoothed pore pressure record $p(t)$. This metric is used to describe the time intervals with the pore pressure higher than the threshold. In addition, the metric can capture the fluctuation that occurred above the threshold value. The integral equation and corresponding score equation are defined in Equation (5) and (6).

$$I^{x\%} = \int [p(t) - y(t)]\, dt \quad (5)$$

$$P_I = S\left(I_1^{x\%}, I_2^{x\%}\right) \quad (6)$$

where the threshold pore pressure $y(t) = x\%\ of\ max\ p(t)$.

Dynamic time warping, often used in speech recognition (Rabiner 1993), computes the optimal distance between two records with a non-linear time scale. It can be used to match two shapes regardless of lags or duration differences, but it is sensitive to time step and magnitude. P_{DTW} employs this concept to evaluate the spike matches within a certain time window w, where w = 1 s for this study. Within each time window, both records are calibrated to start from 0 and the *DTW* is used to verify if there are similar spike shapes within the window. The score equation is defined as Equation (7).

$$P_{DTW} = dtw(p_1(t^w)), dtw(p_2(t^w)) \quad (7)$$

2.2 *Survey Data*

A voluntary and anonymous survey was conducted over a one-month period from 23rd January 2024 to 22nd February 2024 to collect people's perceptions about the similarity between pairs of data. Any postgraduate student who was in the Department of Civil and Environmental Engineering at the University of Auckland was invited to participate in this research. Participants had the option to leave their email address to receive a summary of the survey findings upon completion. The survey responses were stored securely in the university's database where only the research team will have access to. The data will be stored for 6 years and then automatically erased.

The participants finished an online questionnaire and used their own judgement to score a dataset. The questionnaire includes a total of 18 pairs of pore pressure simulation records from a previously performed numerical modelling study as shown in Figure 3Figure 1 (Part 1) and Figure 4Figure 2 (Part 2), including the centrifuge test record in blue and the numerical modelling records in red. Both figures show how pore water pressure changes during an input excitation. The participant was asked to use consistent judgement to compare each data pair and score the similarity between experimental and numerical data. A score of 1 was a poor match and a score of 10 was an excellent match.

3 RESULTS AND DISCUSSION

Three pore pressure records from Figure 3 – subplots 1, 4, and 7 – are selected to demonstrate the analysis. The criterion scores are compared against the collected survey scores to correlate the qualitative judgement and quantitative evaluation.

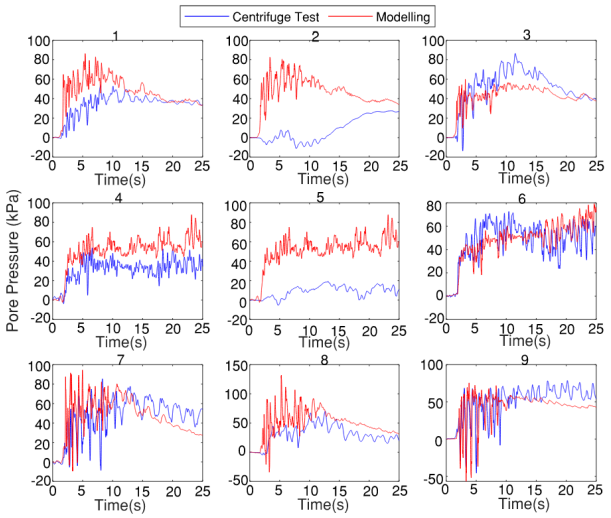

Figure 3. Surveyed pore pressure records in pairs (Part 1) – centrifuge test records in blue and numerical modelling records in red.

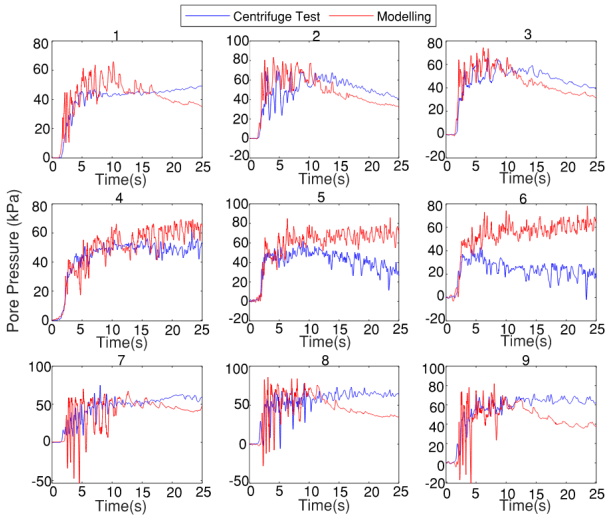

Figure 4. Surveyed pore pressure records in pairs (Part 2) – centrifuge test records in blue and numerical modelling records in red.

3.1 *Survey Data Analysis*

Figure 5 presents the three selected records to demonstrate the relationship between the criteria scores and survey scores. For each pair of pore pressure records, three subplots are used to illustrate the results:

(i) A pair of pore pressure records plotted against time,
(ii) A histogram of the 25 survey scores with a vertical dashed line showing the criterion weighted average score (WA), and
(iii) A scatter of the criteria metric scores with a horizontal dashed line showing the average survey score.

Figure 5 a(i) and b(i) both illustrate the scenarios where the numerical simulation overestimates the measured pore pressure response; however, the characteristics of these discrepancies differ. A preliminary visual observation of Figure 5 a(i) indicates that the simulated pore pressure record is nearly twice the experimental record during 2-10s before converging to similar amplitudes after 12s. In contrast, Figure 5 b(i) shows a consistent overestimation across the whole 25s duration. Figure 5 c(i) presents a simulation that more closely aligns with the experimental data throughout the excitation.

These varying levels of agreement lead to high variation of qualitative judgements, which is reflected in the histograms of survey responses shown in Figure 5 a(ii) and b(ii). The black dashed line denotes the criterion-derived WA. For both cases, survey scores are widely distributed across the full 1–10 range, indicating a high variability in qualitative judgment. Figure 5 c(ii) shows a left-skewed distribution, suggesting general agreement among respondents on the higher quality of this match.

The average survey scores for the first two data pairs are approximately 6, which are shown as the red dashed lines in Figure 5 a(iii) and b(iii). The criterion WA in each case is broadly consistent with the survey averages, although the individual metric scores may deviate from these lines. Similarly, Figure 5 c(iii) shows a close alignment between the average survey rating and the criterion WA.

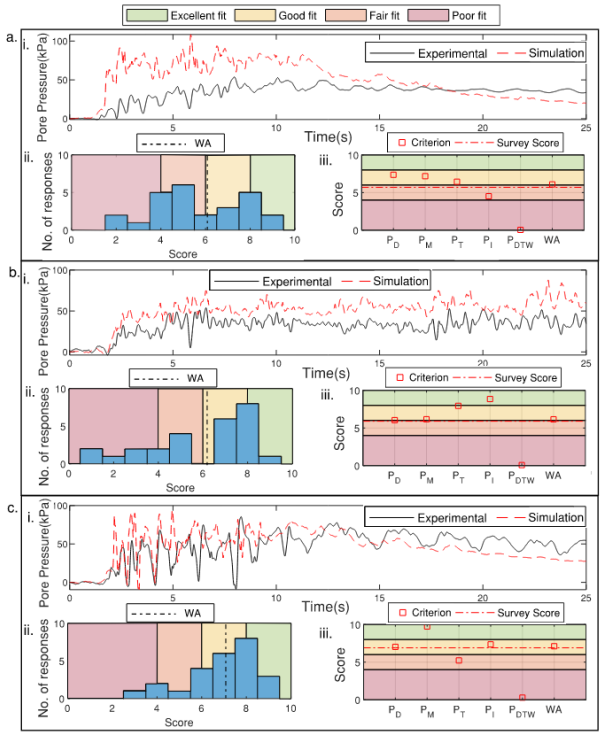

Figure 5. The comparison between pore pressure criterion metric scores and survey data of data pairs 1, 4 and 7. Subplot (i). the pore pressure records; subplot (ii) the histogram of the 25 survey results with a dashed line showing *WA*; subplot (iii) the scatter of metric scores with a dashed line showing the corresponding average survey score.

The rest of the survey scores are analyzed similarly, and the results are generally consistent with the observations above. Overall, the comparative analysis emphasizes the enhanced consistency, reproducibility, and analytical rigor provided by the quantitative criterion evaluation.

3.2 *Statistical Correlation Analysis*

A criterion correlation summary is performed to assess the correlations between the survey average scores and the criteria evaluated WA. A programming language for statistical computing and graphics, RStudio IDE, is adopted to generate the pairwise scatter plots with histograms and correlations.

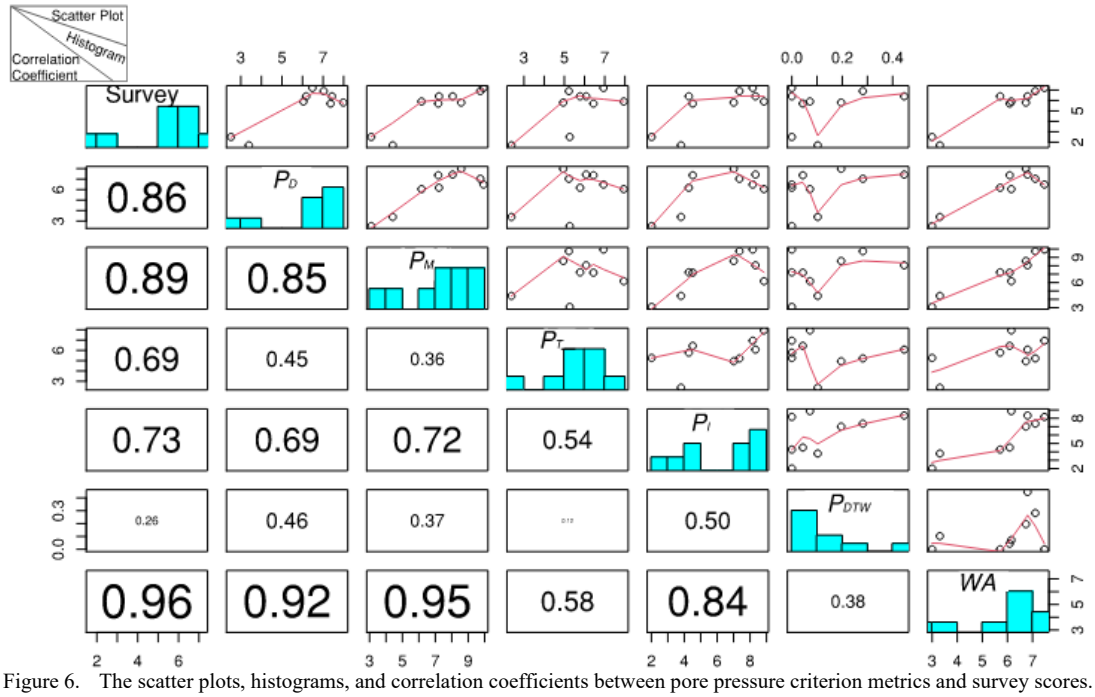

Figure 6. The scatter plots, histograms, and correlation coefficients between pore pressure criterion metrics and survey scores.

The plot consists of three components:

(i) Along the diagonal subplots: univariate histograms of each criterion, which is a high-level summary of the criteria scores.

(ii) Subplots at upper right corner: scatter plots with locally estimated scatterplot smoothing (LOESS) smoothers between any two criteria to visualize the correlation between two criteria.

(iii) Subplots at bottom left corner: the correlation coefficient between any two criteria, and the size of the font of each number is proportional to its magnitude, i.e., the correlation coefficients have large fonts when the two criteria are highly correlated with each other and vice versa.

Figure 6 shows a statistical correlation summary of the 9 pairs of pore pressure records (Survey data Part 1). The results indicate a strong positive correlation between the survey evaluation and criterion WA with a correlation coefficient of 0.96. The high degree of correlation suggests that the criterion WA effectively captures the overall trends reflected in the averaged representation of the subjective human judgments.

In terms of individual metrics, the survey scores exhibit relatively higher correlations with P_D and P_M (0.86 and 0.89, respectively). This observation aligns with the previous conclusions and further proves the effectiveness of the quantitative criterion, especially in capturing features that are perceptually important to human evaluators.

It is notable that not all the criterion metrics are closely correlated with each other. The lack of strong internal correlation suggests that each metric captures distinct characteristics of the pore pressure time histories. Therefore, the inclusion of multiple metrics enables a more comprehensive evaluation of GOF.

4 CONCLUSIONS

In this study, the recently proposed quantitative GOF criterion for excess pore pressure time series was applied to a previously collected dataset to evaluate the agreement between simulated and experimental records. The resulting quantitative assessments were compared against qualitative evaluations obtained from a questionnaire survey conducted within the Department of Civil and Environmental Engineering at the University of Auckland. A total of 25 responses were collected over a one-month period from postgraduate students and researchers.

Analysis of the survey data revealed considerable variability in the subjective evaluations, with individual scores ranging from 1 (poor match) to 10 (perfect match). This wide distribution suggests that individual expertise might influence the interpretation of perceived GOF, emphasizing the inherent subjectivity of visual or experience-based assessments.

Despite this high variability, the averaged trends observed in the survey responses are reasonably consistent with those derived from the proposed quantitative criterion. This consistency highlights the robustness and reliability of the criterion and suggests that it successfully captures the key features that are commonly recognized by human perceptions. While qualitative assessments may be sensitive to subtle features in the data, the quantitative approach offers a systematic, reproducible and unbiased framework for performance evaluation. As such, the proposed criterion provides a sound basis for comparative analysis across different cases, enhancing objectivity and credibility in the assessment of pore pressure simulation quality.

5 ACKNOWLEDGEMENT

The authors gratefully acknowledge Andy O'Sullivan Geotechnical Engineering for technical and financial support.

6 REFERENCES

HARTZELL, S., 1982. Simulation of ground accelerations for the May 1980 Mammoth Lakes, California, earthquakes. Bulletin of the Seismological Society of America, 72(6A), pp. 2381–2387.

JIANG, C. and WOTHERSPOON, L., 2024. Quantitative criterion for the goodness-of-fit of synthetic pore pressure time series, 8th International Conference on Earthquake Geotechnical Engineering, 7-10 May 2024.

SABETTA, F. and PUGLIESE, A., 1996. Estimation of response spectra and simulation of nonstationary earthquake ground motions. Bulletin of the Seismological Society of America, 86(2), pp. 337–352.