Evaluation of a Cone Penetration Test Thin-Layer Correction Procedure in the Context of Global Liquefaction Model Performance

Mertcan Geyin[1] and Brett W. Maurer[1]

Abstract: Engineering geologists routinely perform liquefaction hazard assessments using data from the cone penetration test (CPT). However, the volume of soil mobilized by the CPT acts as a low-pass filter on the true stratigraphy, potentially removing information such as the data defining a thin layer of soil or the interface between two dissimilar soils. The Boulanger and DeJong (2018) CPT inversion procedure, which aims to correct these effects, is herein evaluated in the context of CPT-based liquefaction model performance. Using over 15,000 case-histories from 24 earthquakes parsed into 2 datasets, 18 different liquefaction models are studied, resulting in 36 performance trials. In 1 of these trials, the CPT inversion procedure increases model efficiency to a statistically significant degree, but in 23 others it significantly decreases efficiency. This decline in performance tends to grow as profiles become more stratified. To explore remedies, a liquefaction triggering curve is rederived from inverted CPT data, such that its training and forward implementation are made consistent. Nonetheless, this exacerbates the decline in prediction efficiency. Ultimately, the results of this study are not a direct assessment of the pioneering Boulanger and DeJong (2018) procedure. However, the results do provide evidence that this procedure – when applied to existing CPT-based liquefaction models – may provide no demonstrable performance benefit.

Keywords: cone penetration test; thin layer corrections; liquefaction model

1. Introduction

Cone Penetration Test (CPT) measurements, which principally include cone-tip resistance (q_c) and sleeve friction (f_s), may be used to infer various soil properties and behaviors. The CPT, for example, has significant benefits over other tests on which liquefaction models have been based (NRC 2016). In this regard, engineering geologists routinely carry out liquefaction hazard studies, both locally and regionally, using CPT data (e.g., among many, Juang et al. 2009; Heidari and Andrus, 2010; Khoshnevisan et al., 2015; Zhang et al. 2016, 2018; Chen et al., 2016; Gheibi et al., 2016; Hasek and Gassman, 2019; Bastin et al., 2020; Norini et al. 2021). However, because the CPT is an intermediate-to-large-strain penetration test, its measurements are still potentially obfuscated by the volume of mobilized soil. That is, q_c and f_s are influenced by soil conditions both relatively proximal to, and distal from, the CPT sensors. While CPT

[1] Department of Civil and Environmental Engineering, University of Washington, Seattle USA.

measurements are recorded at discrete depths – typically 1-2 cm intervals – they sample a mobilized zone of influence that may extend ~10-30 cone diameters from the q_c sensor (Boulanger and DeJong 2018), with the zone growing and shrinking as a function of soil properties. For the industry standard 10 cm^2 cone, this equates to a zone 0.35-1.05 m thick. This mobilized zone effectively acts as a "low-pass filter" on the true soil stratigraphy, filtering information from the low spatial wavelengths, such as the data defining a thin layer of soil or the interface between two dissimilar soils. These spatial smoothing effects, which are often called "thin layer" and "transition" effects, have been investigated by many authors (e.g., van der Linden 2018; Ching et al. 2015; Robertson 2011; Ahmadi and Robertson 2005; Lunne et al. 1997; Treadwell 1976). Although chart-based methods have been proposed to manually correct CPT data for these effects, Boulanger and DeJong (2018) proposed what may be the first fully automated approach. This "inverse filtering and interface detection" procedure attempts to correct CPT measurements to their "true" values. Since the direct measurements reflect conditions averaged over a volume, rather than at discrete points, their correction would invariably improve CPT site characterization. As an example, CPT data is shown in Fig. 1, with and without correction by the Boulanger and DeJong (2018) procedure.

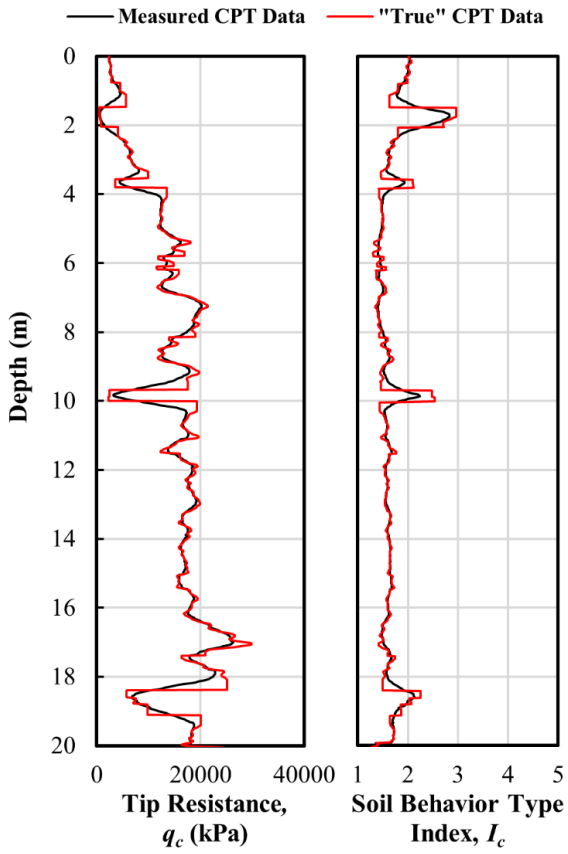

Fig. 1. CPT data with (i.e. "true") and without (i.e. measured) inverse-filtering and interface correction per Boulanger and DeJong (2018), as implemented in the software *Horizon* (Geyin and Maurer 2020a).

The efficacy of the Boulanger and DeJong (2018) procedure has yet to be rigorously studied in the literature, but its adopted can change the computed liquefaction hazard at a site, with the magnitude and direction of change dependent on many factors. Moreover, this procedure was recently incorporated into software programs that predict liquefaction and its consequences (e.g., *CLiq* by Geologismiki 2020; *Horizon* by Geyin and Maurer 2020a). Accordingly, the objective of this paper is to assess the Boulanger and DeJong (2018) procedure in the context of CPT-based liquefaction model performance. That is, to determine whether use of the procedure improves prediction-model efficiency. Using field case histories from 24 earthquakes parsed into 2 datasets, 18 different liquefaction models will be studied, resulting in 36 performance trials. Uncertainty due to finite sampling will be accounted for and used to establish statistical significance. Ultimately, the results of these trials spur further inquiries that expand on the initial objective, and which are introduced later in the paper.

In the following, the Boulanger and DeJong (2018) procedure is first succinctly summarized. The 18 liquefaction models in which this procedure will be evaluated, and the field case-history data to which it will be applied, are then identified. Lastly, the methodologies for evaluating predictive performance are described and trial results are presented and discussed.

2. Summary of the Boulanger and DeJong (2018) Procedure

The correction of CPT data can be idealized as the conversion of measured q_c and f_s values, or q^m and f^m, to the "true" values q^t and f^t, which would be obtained if measurements reflected conditions at discrete points. This can be viewed as an inverse problem, with the goal of determining the "true" values via inversion of what was measured. As illustrated in Fig. 2 (in this case for tip resistance), q^m may deviate from q^t near layer interfaces because the measurements are influenced by materials both above and below the interface, even though the measurement is recorded at a discrete depth. In the case shown in Fig. 2, q^m is artificially increased in the upper soft clay as the cone approaches, and is influenced by, the underlying dense sand. The opposite occurs in the dense sand as the cone approaches the underlying soft clay. That is, q^m is artificially decreased when the cone begins to sense the softer material. These errors are commonly referred to as "transistion" or "interface" effects. It can also be seen in Figure 2 that should the dense sand be insufficiently thick for the cone to receive no influence, at least momentarily, from the soft clay either above or below it, then q^m in the dense sand will never reach its true value. Of course, the opposite would occur in a soft, loose stratum sandwhiched by denser strata. While these latter errors have the same root cause as transition effects (i.e., measurement over a volume) they are often separetely referred to as "thin layer" effects.

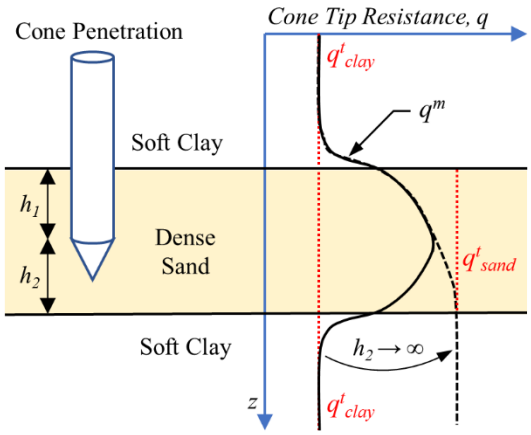

Fig. 2. Conceptual schematic showing cone penetration in layered soil. The measured cone-tip resistance (q^m) deviates from the "true" resistance (q^t) that would be measured in each material if measurements reflected conditions at discrete points (modified from Mo et al. 2017).

Nomenclature aside, the Boulanger and DeJong (2018) procedure aims to correct these collective effects by viewing q^m as the convolution of q^t with a physical low-pass filter (w_c) over a zone of influence (a depth window 60 times that of the cone diameter (d_c), centered at the cone tip). With this approach, q^m is computed as:

$$q^m(z) = q^t(z) * w_c(z) \quad (1)$$

where q^m, q^t, and w_c are a function of depth, z, and the asterisk indicates convolution, which is the integral of the point-wise multiplication of $q^t(z)$ and $w_c(z)$, as a function of the amount that one of the functions is shifted relative to the other. While complete details will not be given here, $w_c(z)$ is a function of two other terms: w_1, which decreases the relative influence of any soil as its distance from the cone tip increases; and w_2, which weights the influence of soil above or below the cone tip based on whether those soils are stronger or weaker than that immediately at the tip. Specifically, it is assumed that q^m receives more influence from the soil immediately near the tip when that soil is relatively weaker than the surrounding soil. Conversely, it is assumed that q^m receives less influence from the soil immediately near the tip when that soil is relatively stronger than the surrounding soil. In essence, the Boulanger and DeJong (2018) technique identifies the q^t that, when convolved with w_c, best predicts q^m. As part of this procedure, additional filters are used to smooth high frequency noise considering the CPT sampling interval, thereby increasing the speed and likelihood of convergence to an optimal solution.

Following estimation of q^t via inversion of q^m, Boulanger and DeJong (2018) propose, at least for the time being, that f^t be estimated from q^t. Specifically, the proposed "inversion" of f^m follows the

assumption that both inverted and measured pairs of normalized tip resistance (Q) and normalized sleeve friction ratio (F) lie along the same radial line originating from the origin of the Soil Behavior Type Index, I_c, proposed by Robertson and Wride (1998), which maps in Q-F space. In effect, this approach changes I_c near interfaces, and in-turn the inferred soil type (e.g., the inferred susceptibility to liquefaction) but results in relatively minimal changes to I_c otherwise (i.e., away from interfaces). Following the initial development of q^t and f^t profiles, a separate procedure detects and corrects interfaces based on the rate of change of q^t with respect to depth. In total, the Boulanger and DeJong (2018) procedure has five parameters, for which recommended "baseline" values were provided: $z'_{50,ref} = 4.2$; $m_{z50} = 0.5$; $m_z = 3$; $m_q = 2$; and $m_t = 0.1$. While complete details are provided in Boulanger and DeJong (2018), these baseline parameters were herein adopted to compute "true" CPT data. It is plausible that these values could be calibrated at the site-specific level (e.g., using high resolution borings adjacent to a CPT). This would change the magnitude of correction and the sensing and development distances, thereby potentially improve identification of thin layers. However, the information compiled for this study was generally insufficient to attempt calibration and, when available, provided insufficient statistical support to justify it. As part of the CPT processing methodology, statistical cross-correlation (Buck et al. 2002) was used to align tip- and sleeve-measurements, both for the measured and "true" CPT data. All CPT processing, including implementation of the Boulanger and DeJong (2018) procedure, was completed using the open-source software *Horizon* (Geyin and Maurer 2020a).

3. Liquefaction Models

The Boulanger and DeJong (2018) procedure will be evaluated in the context of CPT-based liquefaction model performance. That is, whether use of the procedure improves or worsens their prediction efficiency. Towards that end, six triggering models based on the so-called "simplified stress-based" framework, first envisioned by Whitman (1971) and Seed and Idriss (1971), are herein adopted: Green et al. (2019), Boulanger and Idriss (2014), Idriss and Boulanger (2008), Moss et al. (2006), Architectural Institute of Japan (2001), and Robertson and Wride (1998). However, because these models predict triggering at specific depths below ground, a true evaluation of their performance requires subsurface exploration or instrumentation (i.e., to assess whether predicted and actual responses agree). This could potentially be achieved in advance of an earthquake using buried sensors (e.g., Holzer et al. 2007) or after an earthquake using vision penetrometers (Raschke and Hryciw 1997) or geoslicers (Nakata and Shimazaki 1997). Yet, case histories with such data are exceedingly rare and still may not result in definitive interpretations of what did, and did not, liquefy (e.g., Takada and Atwater 2004). As a result,

nearly all existing case-histories document only whether liquefaction manifestations were observed at the ground surface. Accordingly, to compare subsurface predictions of liquefaction against surface observations, the predictions from each triggering model will be input to three different models that predict surficial manifestations of liquefaction: Maurer et al. (2015a), van Ballegooy et al. (2014), and Iwasaki et al. (1978), who developed models termed *LPI*, *LSN*, and LPI_{ISH}, respectively. In this study, "liquefaction model" therefore refers to the use of two models in series: one triggering model and one manifestation model. The independent performance of these models can simply not be assessed via any practical, objective means. A summary of the 18 models to be used (6 triggering models x 3 manifestation models), and the symbols that will be used to identify them, is provided in Table 1.

Table 1. Summary of Liquefaction Triggering and Manifestation Models used in this Study.

Model	Type	Symbol
Robertson & Wride (1998)	Triggering	RW98
Arch. Institute Japan (2001)	Triggering	AIJ01
Moss et al. (2006)	Triggering	Mea06
Idriss & Boulanger (2008)	Triggering	IB08
Boulanger & Idriss (2014)	Triggering	BI14
Green et al. (2019)	Triggering	Gea19
Iwasaki et al. (1978)	Manifestation	*LPI*
van Ballegooy et al. (2014)	Manifestation	*LSN*
Maurer et al. (2015)	Manifestation	LPI_{ISH}

4. Liquefaction Case-History Data

A total of 15,223 liquefaction case histories from 24 earthquakes will be analyzed, as listed in Table 2. However, because most of these cases are from three earthquakes in New Zealand's Canterbury province, analyses will be carried out separately for these and the remaining 21 events, henceforth referred to as the "Canterbury dataset" and "global dataset," respectively. Each case history includes estimates of groundwater depth and peak ground acceleration (*PGA*), CPT data, and observations of the presence or absence of surficial liquefaction manifestations.

Table 2. Summary of Liquefaction Case-Histories Analyzed (after Geyin et al. 2020b)

Date	Event	Country	Magnitude (M_w)	Number of Case Histories
16/6/1964	Niigata	Japan	7.60	3
9/2/1971	San Fernando	USA	6.60	2
4/2/1975	Haicheng	China	7.00	2
27/7/1976	Tangshan	China	7.60	10
15/10/1979	Imperial Valley	USA	6.53	7
9/6/1980	Victora (Mexicali)	Mexico	6.33	5
26/4/1981	Westmorland	USA	5.90	9
26/5/1983	Nihonkai-Chubu	Japan	7.70	2
28/10/1983	Borah Peak	USA	6.88	3
2/3/1987	Edgecumbe	New Zealand	6.60	23
24/11/1987	Elmore Ranch	USA	6.22	2
24/11/1987	Superstition Hills	USA	6.54	8
18/10/1989	Loma Prieta	USA	6.93	67
17/1/1994	Northridge	USA	6.69	3
16/1/1995	Hyogoken-Nambu	Japan	6.90	21
17/8/1999	Kocaeli	Turkey	7.51	16
20/9/1999	Chi-Chi	Taiwan	7.62	34
8/6/2008	Achaia-Ilia	Greece	6.40	2
4/4/2010	Baja	Mexico	7.20	3
11/3/2011	Tohoku	Japan	9.00	7
20/5/2012	Emilia	Italy	6.10	46
4/10/2010	Darfield	New Zealand	7.10	5371
22/2/2011	Christchurch	New Zealand	6.20	4806
14/2/2016	Christchurch	New Zealand	5.70	4771

The Canterbury data was sourced from Geyin et al. (2020a, 2021), who compiled liquefaction case-histories from the: (i) M_w7.1, 4 Sept. 2010 Darfield earthquake; (ii) M_w6.2, 22 Feb. 2011 Christchurch earthquake; and (iii) M_w5.7, 14 Feb. 2016 Christchurch earthquake. The Geyin et al. (2020a, 2021) database built upon earlier compilations from Canterbury (Maurer et al. 2014, 2015b), resulting in 15,890 case histories. Of those, 14,948 were ultimately selected for analysis in the present study. Cases were excluded when: (i) the predominant manifestation of liquefaction was lateral spreading, since the manifestation models adopted herein are not intended to predict it; (ii) the depth of CPT pre-drill significantly exceeded that of the groundwater, since CPT data in the pre-drill zone must be estimated rather than measured (e.g., by extrapolating upward from just below the pre-drill); or (iii) the estimated median *PGA* was less than 0.075 g, since such cases may not provide meaningful tests of prediction efficiency, given that the absence of liquefaction is easily predicted by judgement. With respect to manifestations, Geyin et al. (2020a, 2021) classified each case history as "none," "marginal," "moderate," "severe," "lateral spreading," or "severe lateral spreading" using criteria from Green et al. (2014), wherein classifications were based on a circular sample centered on each CPT, with approximate radius of 10 m. In this study, liquefaction-model performance will be judged on the ability to predict surficial

manifestations of liquefaction on free-field level ground. Cases with other expressions of liquefaction, such as lateral spreading, foundation movements, or evidence from ground-motions are removed because the adopted liquefaction models are not designed to predict these expressions. To facilitate model assessment, the Geyin et al. (2020a, 2021) case histories are binomially classified as "No Manifestation" and "Manifestation," where the latter are cases with either "minor," "moderate," or "severe" manifestations. Of the resulting cases assembled from Canterbury, 65% are "No Manifestation" and 35% are "Manifestation." The reader is referred to Geyin et al. (2020a, 2021) for further details relevant to data collection and processing, and where the complete Canterbury dataset may be obtained.

To compare results from Canterbury with regions elsewhere, 275 case histories from 21 global earthquakes will be analyzed in parallel. These cases were sourced from Geyin and Maurer (2021), who compiled case-history data from the literature. Whereas documentation of liquefaction in Canterbury was aided by remote sensing, case histories elsewhere are often preserved in less detail, occasionally with few details about the nature of manifestation. Thus, while manifestations were again classified binomially using the Green et al. (2014) criteria, uncertainty is unavoidably present. Among these 275 case histories, 42% are "No Manifestation and 58% are "Manifestation." The reader is referred to Geyin et al. (2021) for further details, and where the complete global dataset may be obtained. The Canterbury and Global datasets were previously studied by Geyin et al. (2020b), who compared the efficacies of various geospatial and geotechnical liquefaction models, and by Geyin and Maurer (2020b), who developed fragility functions for predicting the probability of liquefaction-induced ground failure. In the current study, these datasets are used to rigorously evaluate whether the Boulanger and DeJong (2018) procedure improves CPT-based liquefaction model performance.

5. Liquefaction Model Methodology

All calculations described in this section were performed with the software *Horizon* (Geyin and Maurer 2020a), which has been used in previous research (e.g., Geyin et al. 2020b). The six triggering models listed in Table 1 were each used to compute the factor-of-safety against liquefaction (FS_{liq}) vs. depth. As part of this process, liquefaction susceptibility was first inferred using the CPT soil-behavior-type index (I_c) (Robertson and Wride 1998). Using a criterion developed from lab and field data in Canterbury (Maurer et al. 2019), soils with $I_c > 2.5$ were assumed not susceptible. However, because this value is within the range of common, generic thresholds (e.g., 2.4-2.6) (Youd et al. 2001), it was also applied to the global dataset. Ultimately, the most salient findings of this study are found to be independent of this criterion, as will be further discussed. For soils deemed susceptible, the IB08, BI14, and Gea19 models

(see Table 1) consider the influence of fines-content (FC) on liquefaction resistance. Accordingly, an $I_c - FC$ model specific to Canterbury (Maurer et al. 2019) was used for the Canterbury dataset while a generic $I_c - FC$ correlation (Boulanger and Idriss 2014) was used for the global dataset. The six triggering models were otherwise implemented as proposed by the respective publications in Table 1.

The predictions of FS_{liq} vs. depth made by each triggering model were then input to LPI, LPI_{ISH}, and LSN, which all have the same objective - to characterize the cumulative response of a soil profile in the free field, thereby predicting manifestations of liquefaction at the ground surface. For brevity, the formulae describing LPI, LSN, and LPI_{ISH} are not given here but are provided in Geyin et al. (2020c). Their implementation is exactly as described in Geyin et al. (2020c) and mirrors that in popular practice.

6. Performance Evaluation

Receiver-operating-characteristic (ROC) analyses are a popular diagnostic tool to evaluate models (e.g., Fawcett 2006; Zou 2007) and are widely used in geoscience and geoengineering (e.g., among many, Lin et al. 2021; Upadhyaya et al. 2021; Ju et al. 2020; Sarma et al. 2020). In this study, ROC analyses will be used to: (i) quantify the efficiency of liquefaction models; and (ii) assess whether "true" CPT data improves that efficiency to a statistically significant degree. In all classification problems (e.g., predicting whether sites have observations of liquefaction), "positive" and "negative" observations overlap when plotted as a function of a diagnostic model index (e.g., LPI, LSN, etc.). As an example, two distributions are plotted in Figure 3a. ROC curves plot the true-positive prediction rate (R_{TP}) (i.e., the rate at which positives, or liquefaction manifestations, are correctly predicted) versus the false-positive prediction rate (R_{FP}) (i.e., the rate at which negatives, or a lack of liquefaction manifestations, are correctly predicted) for a range of classification "thresholds," which are used to predict outcomes. Values above a threshold predict positives and those below a threshold predict negatives. Figure 3b depicts the relationship between the observations, thresholds, and ROC curve.

As a model segregates the distributions of positive and negative outcomes more efficiently (i.e., the distributions have less overlap), a corresponding ROC curve trends toward a point at the coordinates (0,1) in ROC space, indicating the existence of a threshold value that perfectly separates the distributions (i.e., the model is perfectly efficient). Conversely, random guessing appears as a 1:1 line in ROC space, in which case the model has no utility and the positive and negative distributions perfectly overlap. For this reason, the area under a ROC curve (AUC) is widely adopted to characterize model efficiency (e.g., Fawcett 2006). AUC also has statistical significance. In this case, it is the likelihood that sites with manifestations have larger model values than sites without manifestations. In this regard, it is equivalent

to the nonparametric Wilcoxon statistic (Hanley and Mc Neil 1982). In the case of a perfectly efficient model, AUC = 1.0 (or 100%), whereas with random guessing AUC = 0.5 (or 50%). Better prediction models thus have higher AUC. With this approach, false-positive predictions (liquefaction is predicted but is not observed) and false-negative predictions (liquefaction is not predicted but is observed) are given equal importance. In other words, AUC reflects the overall misprediction rate, rather than treating either false positives or false negatives as being more important. Another desirable feature of AUC is its relative insensitivity to class imbalance. Suppose a hypothetical dataset includes 1 positive case and 99 negative cases. For this dataset, a hypothetical model that predicts negative outcomes 100% of the time would be 99% accurate, even though the model is objectively useless. For this reason, model accuracy – while commonly reported – is a poor metric unless the positive and negative classes are equal in size. Other oft-reported performance metrics that similarly focus on only positive or negative predictions, such as sensitivity and specificity (e.g., Powers 2011), would be similarly inappropriate for the aims of this study. AUC, however, would appropriately characterize this model's lack of utility by finding a value near 0.5. Accordingly, AUC will be used to quantify model performance in each of the 36 performance trials.

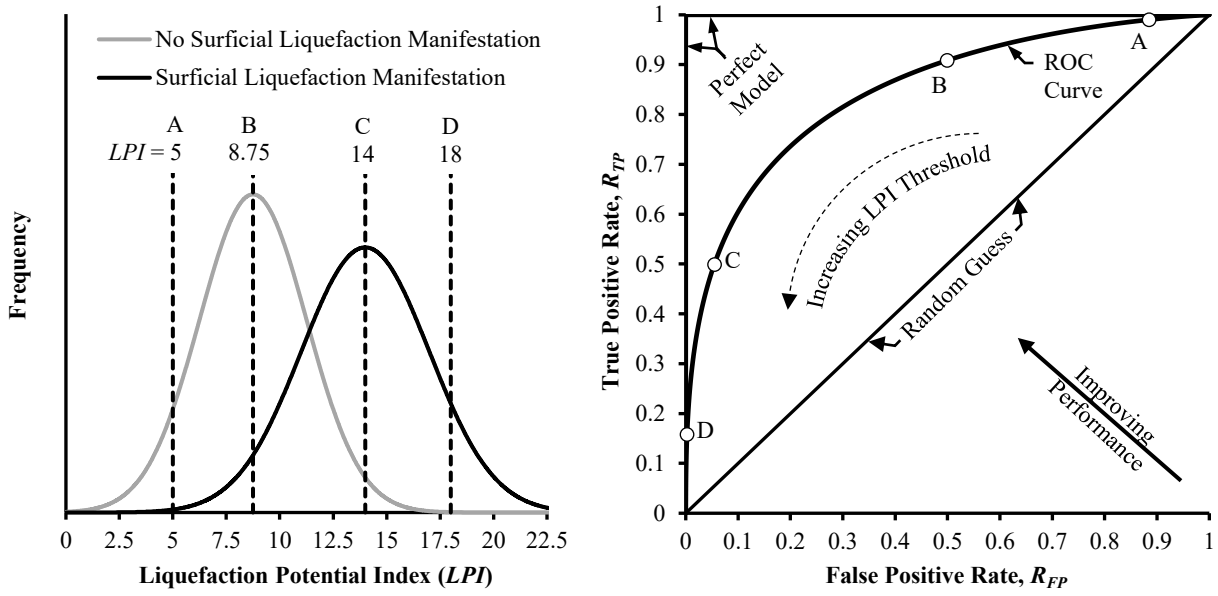

Fig. 3. ROC analyses: (a) positive and negative observations vs. computed LPI; (b) derivative ROC curve, and depiction of how the efficiency of a diagnostic test is assessed via AUC (after Geyin et al. 2020b).

To account for the finite availability of case histories within each dataset (i.e., the Canterbury and global datasets), bootstrap simulations (e.g., Diaconis and Efron 1983) will be used to quantify the finite-sample uncertainty of AUC for each model. This will characterize the sensitivity of results to the data

chosen for study and be used to assess whether differences in *AUC* could arise from chance (i.e., due to finite samples) and not because one model is truly better. Tests of significance will be carried out via the ROC-specific DeLong et al. (1988) methodology. This approach computes the P-values, or probabilities, that two *AUC* samples could have come from the same distribution. In each of the 36 trials, model performance using measured CPT data will be compared to that using "true" CPT data to determine which, if any, is statistically better. Since this method is predicated on *AUC* normality, both Anderson-Darling and Lilliefors tests (Anderson and Darling 1952; Lilliefors 1967) were used to confirm, with 95% confidence, that all *AUC* samples came from a normally distributed family.

7. Results and Discussion

Using the data and methodology above, liquefaction manifestations were predicted for CPT-based case histories by 18 models, both with and without correction by the Boulanger and DeJong (2018) procedure. To illustrate how model performance will be evaluated using ROC analysis, results will first be shown for one model, following which summary statistics from all models are presented. In Figures 4a and 4b, results using the Gea19 – *LPI* model (i.e., the Gea19 triggering model and *LPI* manifestation model) are shown for the global dataset using measured and "true" CPT data, respectively. In each figure, 95%-confidence intervals (CIs) were computed from a total of 10,000 bootstrapped samples. The median ROC curve is the same as that resulting from an analysis of all data without resampling. It can be seen, for this model and dataset, that use of the Boulanger and DeJong (2018) correction procedure decreases the median *AUC* by 1.62% (i.e., predictions are less efficient), counter to what ideally should occur.

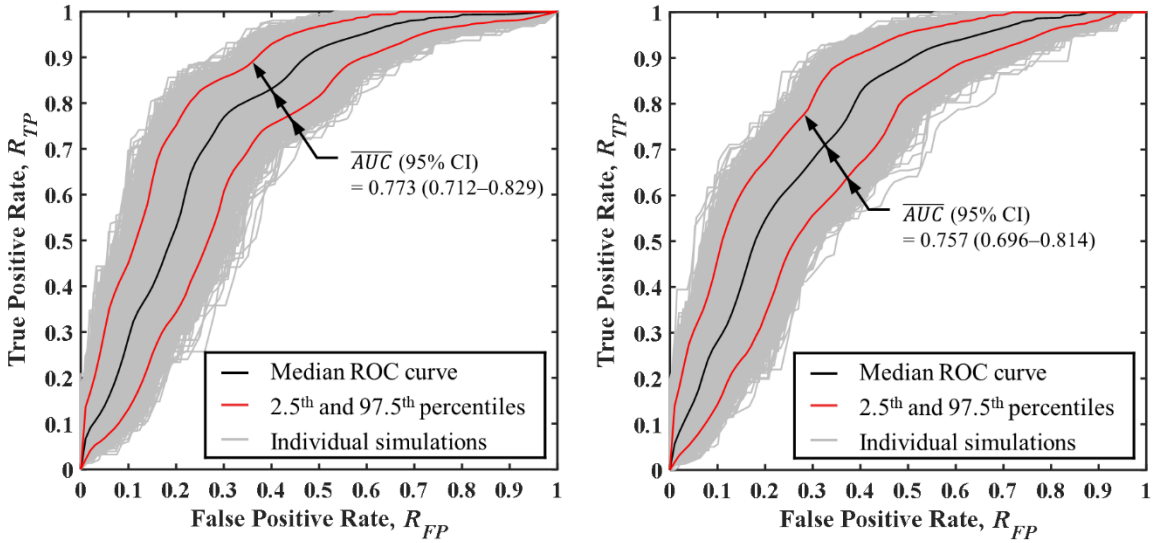

Fig. 4. ROC analysis of Gea19-*LPI* model performance in predicting surficial manifestations of liquefaction for the global dataset: (a) measured CPT data; and (b) "true" CPT data.

From identical ROC analyses of all 18 models, summary statistics are compiled in Figures 5a and 5b for the global and Canterbury datasets, respectively, wherein the resulting shift in *AUC* from using the Boulanger and DeJong (2018) procedure is plotted. Shown for each of 36 trials is the median shift in *AUC* and the 95% CI on that shift, sorted by the year each model was proposed.

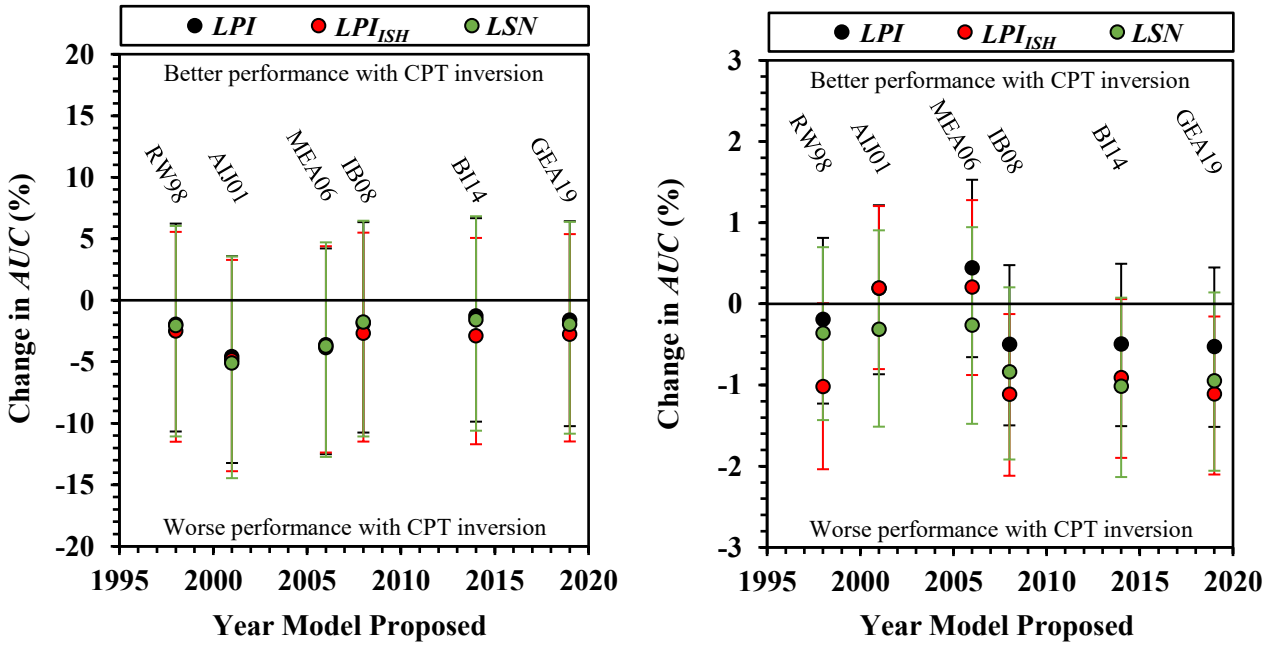

Fig. 5. Summary of change in liquefaction-model performance – quantified by *AUC* – for 18 models to which the Boulanger and DeJong (2018) procedure was applied, ordered by year proposed: (a) global dataset; and (b) Canterbury dataset. Markers denote median shift in *AUC*; bars are 95% confidence intervals on that shift; all model acronyms are identified in Table 1.

As shown in Figure 5, the behavior exhibited by Gea19-*LPI* in Figure 4 (i.e., the decline in median *AUC* using "true" CPT data) is also true of all 18 models when tested on the global dataset, and true of 14 models when tested on the Canterbury dataset. Considering all models, the average changes in *AUC* owed to using "true" CPT data are respectively -2.8% and -0.47% for the global and Canterbury datasets; the average 95% CIs on these changes are respectively -11.50% to 5.29% and -1.51% to 0.53%. The larger finite-sample uncertainty of the global dataset is expected given its much smaller size, all else being equal. Its larger uncertainty may also be further augmented by the relative geomorphic and seismologic diversity of the global dataset and/or because the data collection methods (e.g., ground-motion estimates, CPTs) varied somewhat with time and place.

When assessing the apparent, general decline in performance using "true" CPT data, it should be noted that some of the global cases used to test performance in Figure 5a were also previously used to train

triggering models when they were originally developed. That is, the models were previously trained using the same measured CPT data that is now being using to test them, which might reasonably result in bias that favors the measured data and disfavors the "true" data obtained from Boulanger and DeJong (2018). In this regard, the percentage of test data previously used in training varies from 0% (e.g., RW98) to ~75% (e.g., Gea19). However, the matter of bias is more complicated. *First*, the developers of triggering models, to the degree then possible, manually corrected for interface and thin-layer effects using chart-based solutions or judgement, mitigating the possibility of the aforementioned bias. *Second*, while a case history may have been used to train a triggering model, this training was performed using a judgement-based manifestation model (i.e., an analyst selected the so-called "critical layer" using their judgement to analyze an observation at the ground surface; they did not use *LPI*, *LSN*, or LPI_{ISH} in-reverse to select it). Similarly, *LPI*, *LSN*, and LPI_{ISH} were not formulated or optimized using case-history data, but rather, were developed heuristically and then retrospectively shown to provide useful predictions on field data. In this respect, the training and test datasets used herein might be considered wholly different. Complications aside, it can be seen in Figure 5a that the 18 models perform relatively similarly on the global dataset, despite the large variability in possible bias. Regarding the Canterbury data, the BI14 and Gea19 triggering models, when originally proposed, were trained on a dataset of which 20% was from Canterbury, whereas all other models were trained independent of Canterbury data. It can be seen in Figure 5b that while the BI14 and Gea19 models do exhibit worse performance with CPT inversion, relative to others, these models perform very similar to the unbiased IB08 model, to which they owe their analytical provenance. Thus, there is no readily apparent difference between models with and without conceivable bias. Nonetheless, the possibility that models may perform better if retrained on "true" CPT data will be explored later.

It was found that in 32 of the 36 trials performed, *AUC* decreased because of CPT inversion. To determine whether these changes in *AUC* are statistically significant, P-values were computed per DeLong et al. (1988) to compare the performance of each model with and without CPT inversion. These values are given in Table 3 and are the probabilities that two *AUC* samples (i.e., one model with and without CPT inversion) come from the same parent distribution. Thus, when P-values are small, an observed difference in *AUC* is more likely the result of CPT inversion and less likely a consequence of finite-sample uncertainty. Small P-values occur when (i) two *AUC* values are dissimilar; and/or (ii) the uncertainties of those *AUC* values are small. The common significance level of 0.05 is adopted for these analyses, meaning that two models are classified as "significantly" different if the difference is at least 95% probable. Using this approach, Table 3 summarizes, for each of 36 trials, whether CPT inversion made the efficiency of a liquefaction model better or worse and whether that change was significant.

299 Notable observations from Table 3 are: (i) of the 32 out of 36 trials in which performance decreased
300 using "true" data, those decreases were significant in 23 trials and insignificant in the remaining 9; (ii) of
301 the 4 out of 36 trials in which performance increased, those increases were significant in 1 trial and
302 insignificant in the remaining 3. While these findings would invariably change if the adopted significance
303 threshold were changed (say, from 0.05 to 0.10), the findings would nonetheless suggest a lack of
304 demonstrable improvement using the Boulanger and DeJong (2018) procedure, as implemented herein.
305 Additional inquiries will be investigated in the following.

Table 3. P-Value Matrix to Compare Model Performance With and Without CPT Inversion.

Triggering Model	**Manifestation Model**	***AUC* Better or Worse with CPT Inversion? (p-value)[1]**	
		Global Dataset	**Canterbury Dataset**
RW98	*LPI*	Worse (0.0765)	Worse (0.0424)
	LPI_{ISH}	Worse (0.0333)	Worse (0.0001)
	LSN	Worse (0.0804)	Worse (0.0003)
AIJ01	*LPI*	Worse (0.0004)	Better (0.0511)
	LPI_{ISH}	Worse (0.0005)	Better (0.0972)
	LSN	Worse (0.0002)	Worse 0.0121)
MEA06	*LPI*	Worse (0.0044)	Better (0.0007)
	LPI_{ISH}	Worse (0.0087)	Better (0.1823)
	LSN	Worse (0.0077)	Worse (0.0772)
IB08	*LPI*	Worse (0.077)	Worse (0.0001)
	LPI_{ISH}	Worse (0.025)	Worse (0.0001)
	LSN	Worse (0.1534)	Worse (0.0001)
BI14	*LPI*	Worse (0.2565)	Worse (0.0001)
	LPI_{ISH}	Worse (0.0123)	Worse (0.0001)
	LSN	Worse (0.2111)	Worse (0.0001)
GEA19	*LPI*	Worse (0.1386)	Worse (0.0001)

	LPI_{ISH}	Worse (0.0165)	Worse (0.0001)
	LSN	Worse (0.109)	Worse (0.0001)

[1]Probability that *AUC* using measured CPT data and *AUC* using "true" CPT data could be from the same distribution. Values less than 0.05 are classified as statistically "significant", in which case significance is indicated with shading.

7.1 General Shifts in Perceived Hazard

Towards identifying conditions under which "true" CPT data changes the efficiency of liquefaction models, the general causes of shifting predictions are next investigated. Plotted in Figure 6 are Gea19-*LPI* values computed using measured CPT data versus "true" data, for both the global and Canterbury datasets. The Boulanger and DeJong (2018) procedure, on average, minimally alters the computed Gea19-*LPI* value, with average changes of +0.04 and -0.76 for the global and Canterbury datasets, respectively. For a minority of cases, however, the computed value changes by as much as ±5 or ±10. Of course, profiles inferred to be relatively homogenous from CPT data plot near the 1:1 line in Figure 6, since the Boulanger and DeJong (2018) procedure will minimally alter such data. By corollary, profiles plotting well above or below the 1:1 line tend to be interbedded. From an investigation of all such outliers, two representative cases are next highlighted.

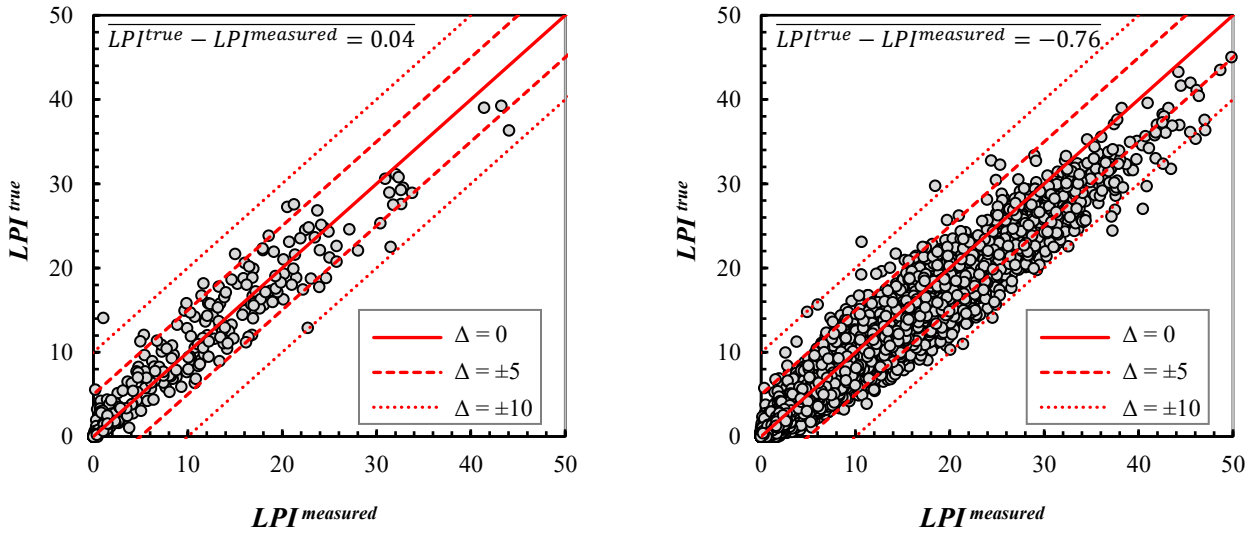

Fig. 6. Computed Gea19-*LPI* values using measured CPT data versus "true" CPT data corrected by the Boulanger and DeJong (2018) procedure: (a) global dataset; and (b) Canterbury dataset.

Plotted in Figure 7 is a case history from Canterbury in which the computed Gea19-*LPI* value increased from 5.7 to 16.8 following CPT inversion. The Boulanger and DeJong (2018) procedure both decreased and increased the measured q_c in thin layers, albeit by relatively small margins. However, the perceived hazard increased substantially because of a downward I_c shift in some layers. Specifically, where the procedure perceives that a thin layer of susceptible sand is sandwiched between softer unsusceptible soils, which may artificially decrease measured q_c and increase measured I_c, the procedure attempts to correct these effects. Thus, while the corrective increase in q_c increases liquefaction resistance, the decrease in I_c changes the inferred soil type to one that is susceptible to liquefaction. It stands to reason, then, that CPT inversion may increase the perceived hazard for one level of seismic loading (e.g., that in Figure 7) but decrease the perceived hazard for a different, lesser level of seismic loading.

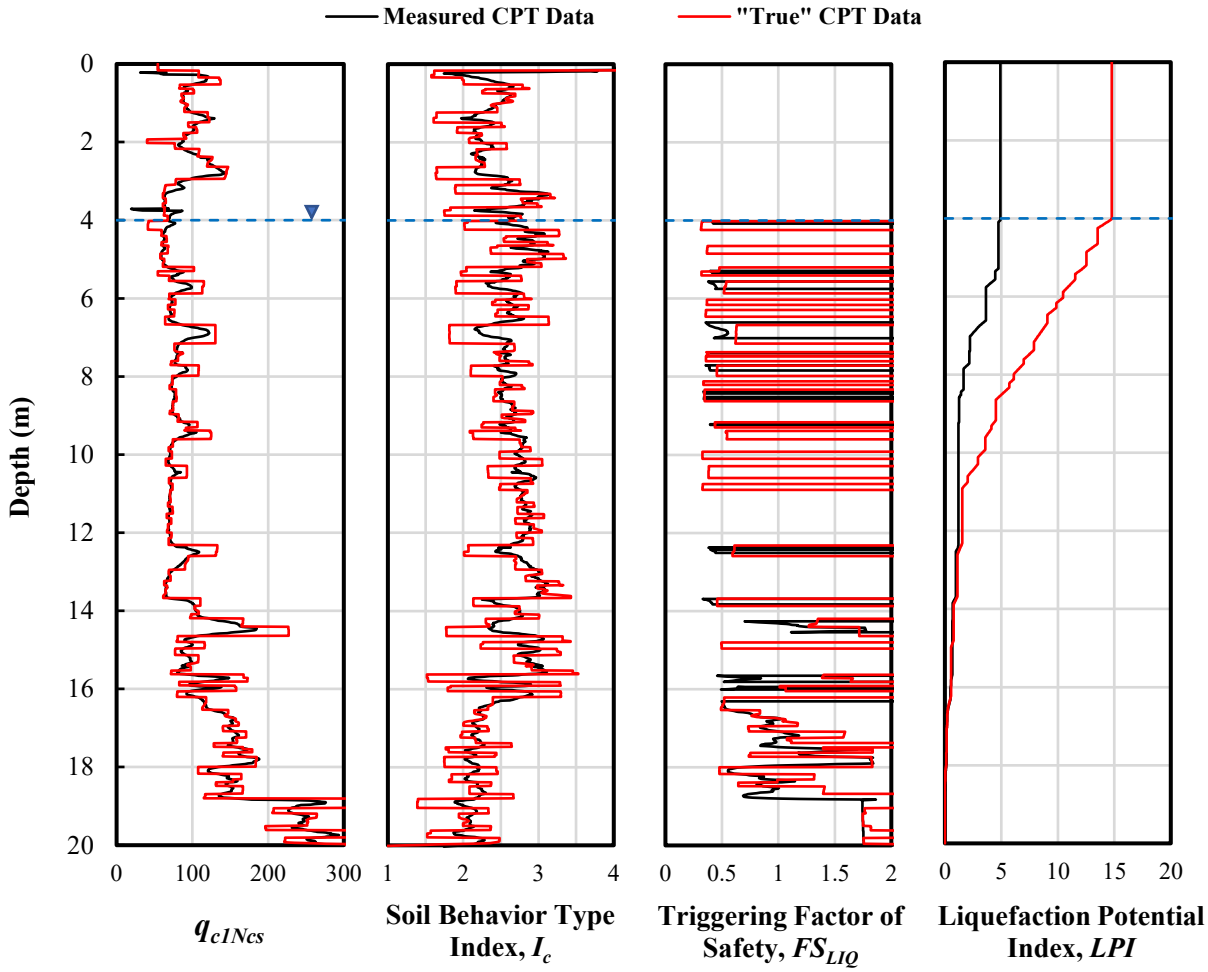

Fig. 7. Computed Gea19-*LPI* values using measured CPT data versus "true" CPT data corrected by the Boulanger and DeJong (2018) procedure [Canterbury case history #5510 from Geyin et al. (2020a, 2021)].

Plotted in Figure 8 is a case history from the global dataset in which the computed Gea19-*LPI* value decreased from 34.5 to 22.51 following CPT inversion. In this case the predominant effect of the Boulanger and DeJong (2018) procedure was to increase measured q_c in some thin layers by a significant margin (much more than in Figure 7). This may be due to the inferred interbedded layers in Figure 8 being relatively thinner and surrounded by relatively softer material. While these corrections are conceptually reasonable, their accuracy cannot be directly assessed without "true" CPT data from calibration chamber tests or numerical simulations. Nonetheless, the preceding examples demonstrate that large upward or downward shifts in the perceived hazard can occur in highly interbedded profiles, particularly when the involved soils are transitional in nature with I_c values near the threshold for discriminating susceptibility.

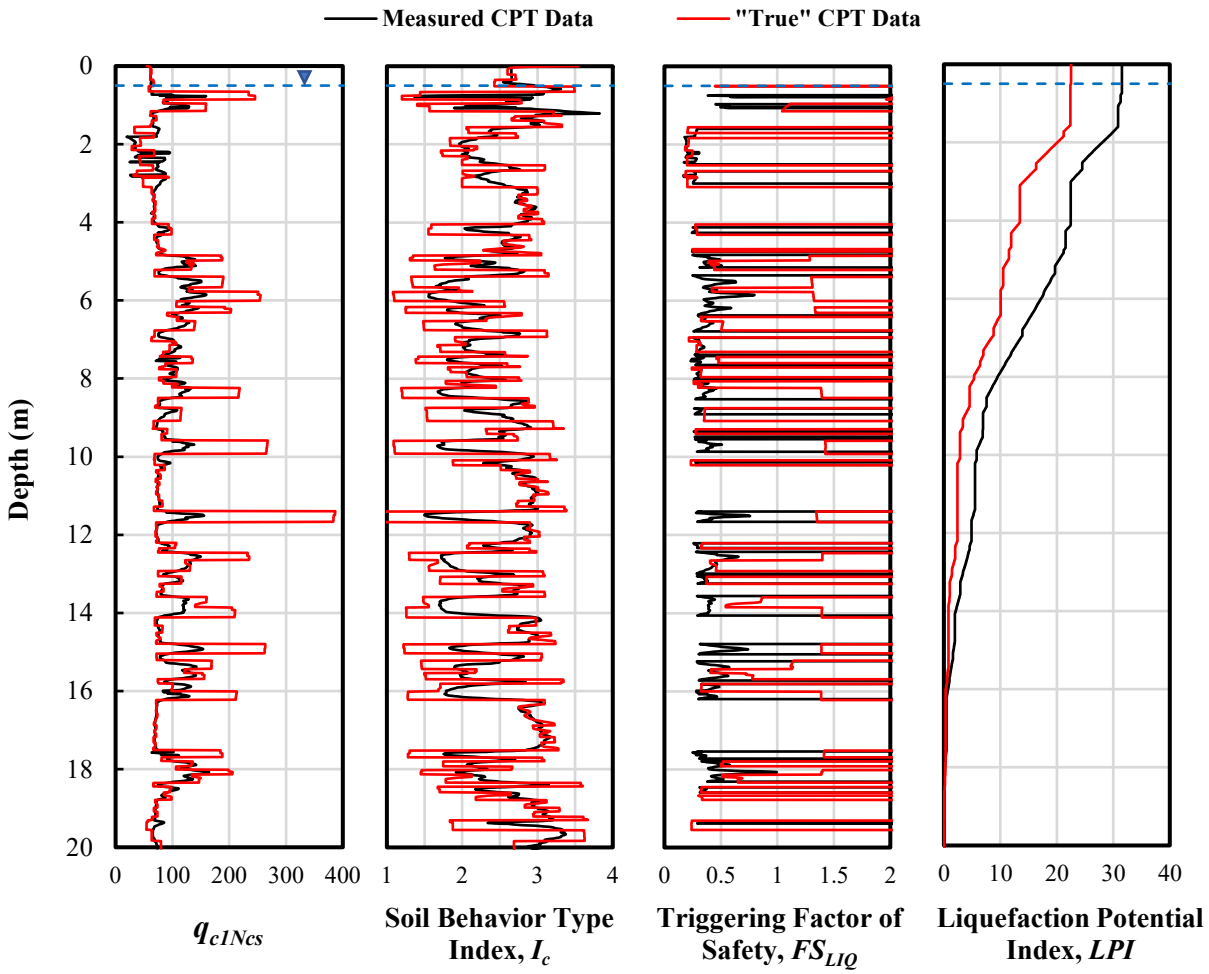

Fig. 8. Computed Gea19-*LPI* values using measured CPT data versus "true" CPT data corrected by the Boulanger and DeJong (2018) procedure [global case history #95 from Geyin and Maurer 2021)].

7.2 Dependence of Results on the I_c Cutoff Used to Infer Liquefaction Susceptibility

Given that large shifts in the perceived hazard are often associated with I_c values moving above or below the $I_c = 2.5$ threshold for discriminating susceptibility, the sensitivity of previous results to this threshold should be assessed. In Figure 7, for example, Gea19-*LPI* increased significantly because I_c values initially just above 2.5 (and thus inferred to be unsusceptible) were adjusted by the Boulanger and DeJong (2018) procedure to just below 2.5. Had the $I_c = 2.5$ threshold been increased in this case and others (e.g., to account for uncertainty in whether sleeve friction is inverted correctly), then CPT inversion could result in less drastic changes to the perceived hazard. To investigate whether the results summarized in Figure 5 and Table 3 would change if a different I_c cutoff were used, all previous analyses were repeated using cutoffs ranging from 2.5 to 3.5 in increments of 0.1. The results of these analyses, which are summarized in Figure 9, indicate that prior findings were insensitive to the I_c cutoff. That is, CPT inversion tends to slightly decrease model efficiency independent of the cutoff. While CPT inversion *does*, for some of the 18 models, slightly improve efficiency when very high cutoffs (e.g., $I_c > 3.2$) are used, those efficiencies are initially much less than optimal. Thus, while CPT inversion may technically increase model performance in this range of the I_c-cutoff domain, there is no compelling reason to employ these high cutoffs, or to use CPT inversion with the liquefaction models assessed herein.

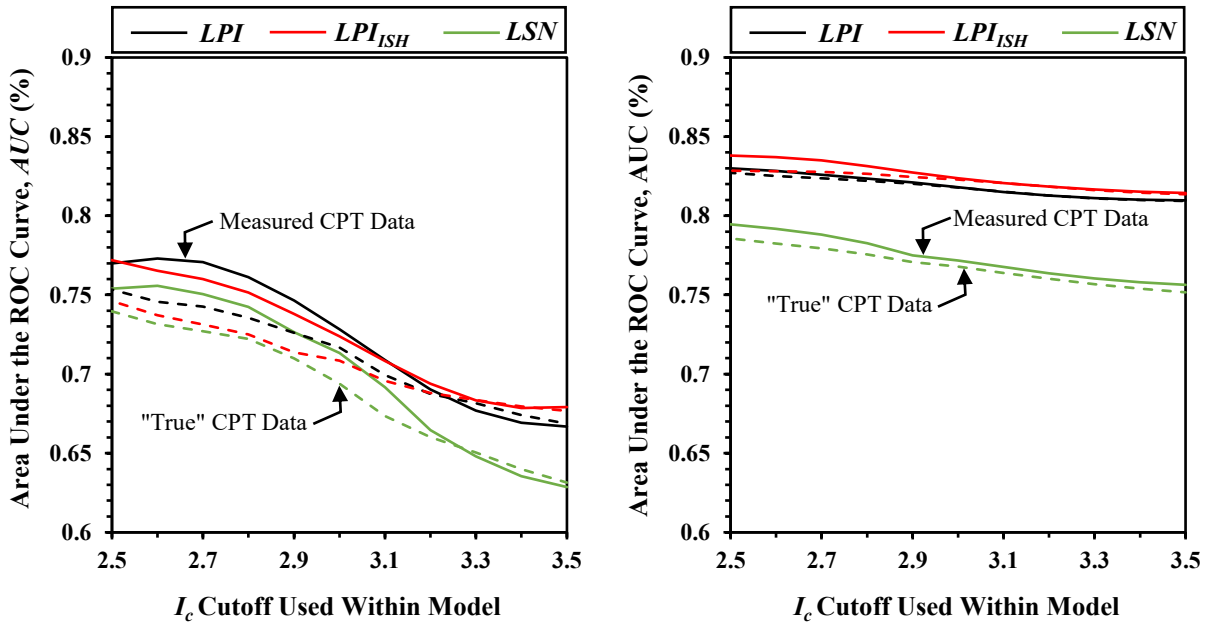

Fig. 9. Liquefaction-model performance – quantified by *AUC* – with and without CPT inversion, plotted as a function of the I_c cutoff used to infer liquefaction susceptibility: (a) global dataset; and (b) Canterbury dataset. Solid lines denote the average *AUC* across all six triggering models using measured CPT data; dotted lines denote the same average using "true" CPT data.

7.3 Correlations Between Geomorphology and Shifts in Model Efficiency

While large shifts in the perceived hazard may occur in profiles inferred to be highly interbedded, it is not yet clear whether the observed, general decline in liquefaction-model efficiency is directly connected to such profiles. For this investigation we focus on the Canterbury dataset, which provides case histories from different geologic units, with differing degrees of "interbeddedness," having experienced similar ground motions. Figure 10a shows the expected, surficial geologic units in the vicinity of Christhurch as mapped by Brown (1975) and Brown and Weeber (1992), and the locations of CPT soundings in the Canterbury dataset. The units in Figure 10a are: (A) Alluvial sand and silt of overbank deposits; (B) Peat swamps now drained; (C) Fixed dune sand and beach sand deposits; (D) Saline sand, silt and peat of drained lagoons and estuaries; and (E) Fluviatile gravel, sand, and silt of historic flood channels.

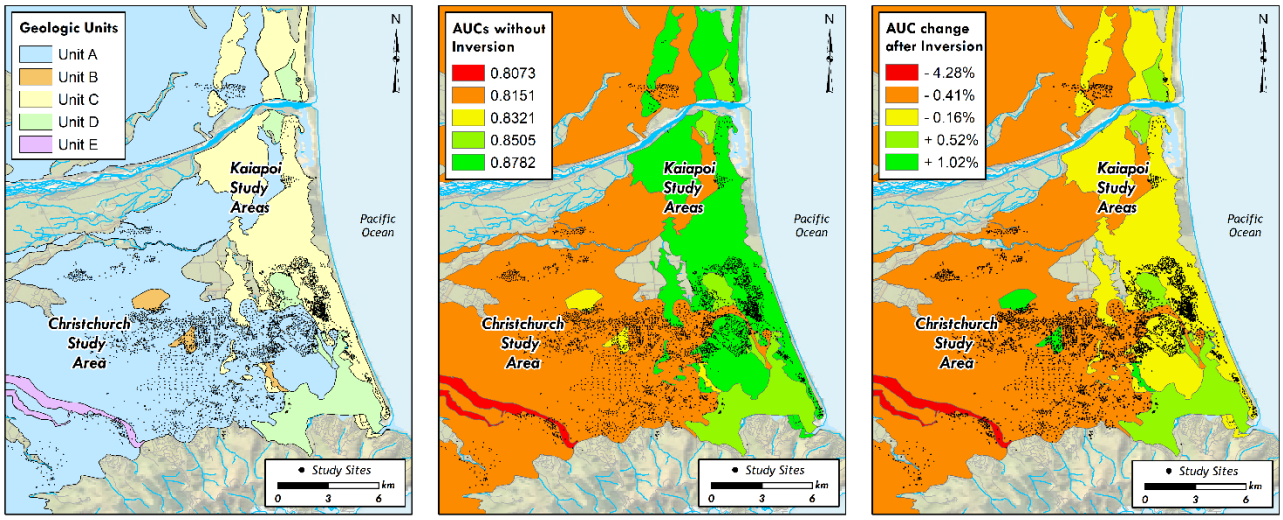

Fig. 10. (a) Mapped geologic units of the Canterbury database, as described in Table 4, and CPT locations; (b) Intra-unit *AUC*s computed from Gea19-*LPI* predictions using measured CPT data; (c) the change to intra-unit *AUC* values due to CPT inversion by the Boulanger and DeJong (2018) procedure.

Intra-unit ROC analyses were performed on the five geologic units in Table 4, resulting in *AUC* values for each model within each unit. These values convey the efficiency with which a model separates cases with and without manifestations of liquefaction, independent of cases in all other units. Results for the Gea19-*LPI* model using measured CPT data are mapped in Figure 10b and are representative of all other models. A spatial dependence may be observed, such that *AUC* values are highest in the east (e.g., dune and beach deposits) and 6-7% lower in the west (e.g., alluvial sand and silt overbank deposits). The less efficient performance of liquefaction models in western Christchurch was previously noted by other investigators (e.g., Beyzaei et al. 2017; McLaughlin 2017; Boulanger et al. 2019). Mapped in Figure 10c

are the shifts in *AUC*, within each unit, resulting from CPT inversion. While a statistically significant overall decrease in *AUC* was previously computed for the study area, performance actually locally increased in some units (where performance initially tended to be better) and locally decreased in others (where performance initially tended to be worse).

In the search for traits by which study sites in Canterbury may be further segregated, a quantitative "interbeddedness index," or the number of inferred soil type changes per meter, is proposed. Here, we adopt the following I_c boundaries proposed by Robertson and Wride (1998), which segregate soils into different behavior types: I_c = 1.31, 2.05, 2.6, 2.95, and 3.6. As an example, Robertson and Wride (1998) proposed that I_c = 2.05 separates silty sand from sandy silt. We limit this index to the upper 10 m of each CPT, given that *LPI*, *LSN*, and LPI_{ISH} assume that surface expression is largely a result of response in the upper 10 m of a profile. Adopting this index, the Canterbury dataset was binned based on inferred soil type changes per meter, wherein a moving bin width of 0.5 changes per meter, and a moving bin increment of 0.25 changes per meter, were adopted. As was done previously, 10,000 bootstrap simulations were performed in each bin to quantify finite-sample uncertainty.

Using the Gea19-*LPI* model as a representative example, *AUC* values are shown in Figure 11a as a function of interbeddedness. It can be seen that as the number of inferred soil type changes per meter increases from zero to seven, the median *AUC* decreases by ~10%. Thus, prior to CPT inversion, liquefaction models tend to perform worse as profiles become more interbedded. In this respect, it is well documented in lab, numerical, and field research that interbedded low-permeability soils can influence the triggering and surface expression of liquefaction (e.g., Fiegel and Kutter 1994; Ozutsumi et al. 2002; Brennan and Madabhushi 2005; Juang et al. 2005; Özener, et al. 2008; Maurer et al. 2015c; Cubrinovski et al. 2019). Despite this, none of the triggering or manifestation models evaluated herein explicitly consider this influence. As such, the performance exhibited in Figure 11a could be considered unsurprising but could also be partly ameliorated via correction of thin layer effects. With respect to this possibility, Figure 11b shows the resulting change in *AUC* due to CPT inversion. As could be expected, there is little to no change in model efficiency for profiles without inferred soil type changes. However, in highly interbedded profiles, inversion tends to exacerbate the discrepancy seen in Figure 11a. That is, when many soil type changes are inferred, CPT inversion decreases liquefaction model efficiency, counter to what ideally would occur. As could be expected, and as shown in Figure 12, there exists in the Canterbury dataset a strong correlation between mapped surficial geology and the inferred interbeddedness of soil profiles. Where interbeddedness is greatest, liquefaction models tend to perform worse initially and tend to be made worse by CPT inversion.

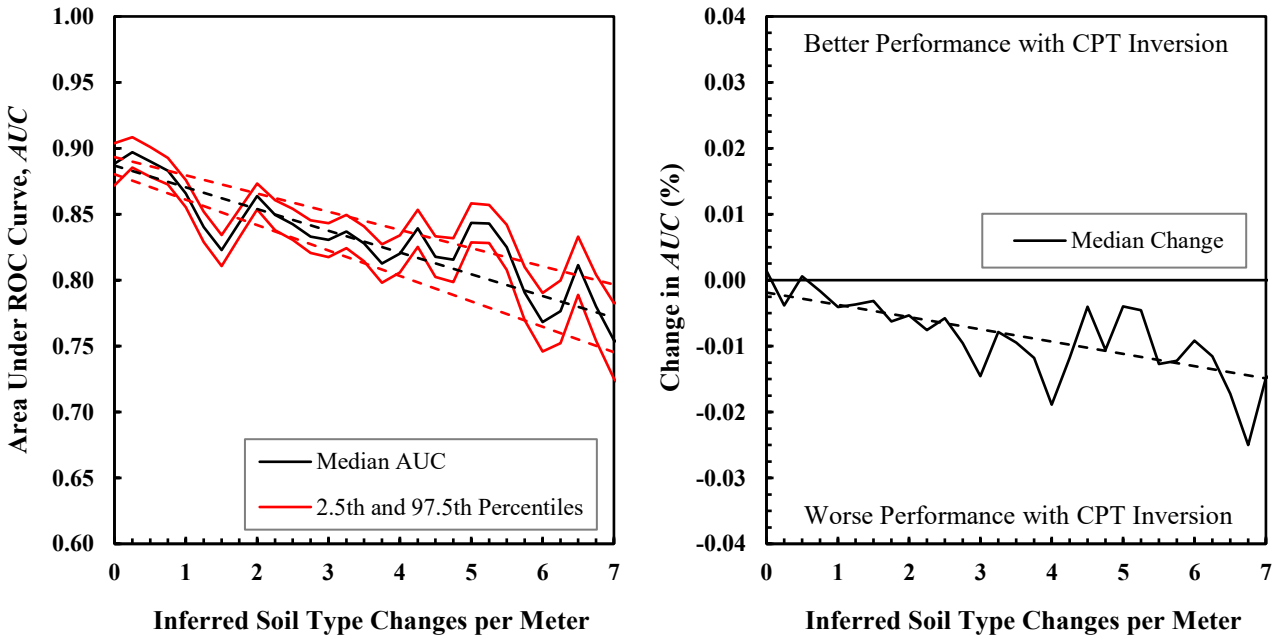

Fig. 11. Gea19-*LPI* model performance as a function of profile interbeddedness (i.e., inferred soil type changes per meter): (a) *AUC* prior to CPT inversion; (b) Change in *AUC* due to CPT inversion.

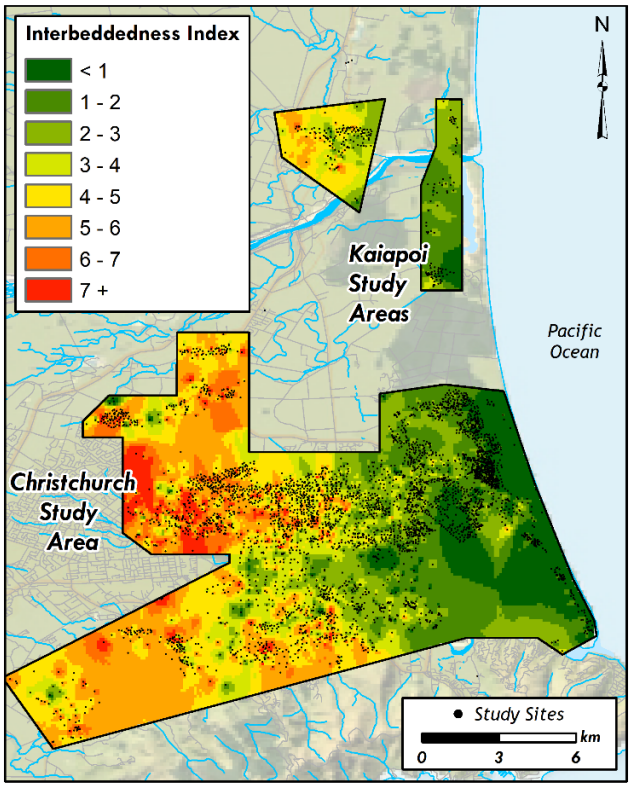

Fig. 12. Profile interbeddedness (i.e., inferred soil type changes per meter, as described in the text) across the Canterbury database study area.

7.4 Development and Assessment of a Liquefaction Model Trained on "True" CPT Data

As discussed, the results in Figure 5 (i.e., changes in *AUC* due to CPT inversion, considering all 18 models) do not suggest testing bias against "true" CPT data. Nonetheless the possibility persists that liquefaction models could perform better if they were both trained and tested on such data. To explore this possibility, the BI14 triggering model is herein retrained on case history data to which CPT inversion has been applied. Boulanger and Idriss (2014) define their limit-state triggering curve as:

$$CRR_{M=7.5,\sigma'_v=1atm} = \exp\left(\frac{q_{c1Ncs}}{113} + \left(\frac{q_{c1Ncs}}{1000}\right)^2 - \left(\frac{q_{c1Ncs}}{140}\right)^3 + \left(\frac{q_{c1Ncs}}{137}\right)^4 - C_o + \varepsilon_{\ln(R)}\right) \quad (5)$$

Where C_o is a fitting parameter that serves to scale the relationship and which has a recommended value of 2.60, and where $\varepsilon_{\ln(R)}$ is normally distributed with a mean of 0.0 and recommended standard deviation of $\sigma_{\ln(R)} = 0.20$. Due to historical precedent, Boulanger and Idriss (2014) proposed that their deterministic triggering model (which was herein evaluated as BI14) correspond to a triggering probability (P_L) of ~16%. As such, their deterministic model is defined by the equation above when $\sigma_{\ln(R)} = -0.20$ and is plotted in Figure 13 as a black line (note that by removing the $\varepsilon_{\ln(R)}$ term in Eq. (5), the deterministic curve is defined using $C_o = 2.8$). Adopting the optimization/training routine of BI14 exactly as prescribed therein, the BI14 triggering model was first retrained with measured CPT data using 81% of the BI14 dataset (the authors were unable to obtain the raw CPT data for the remaining 19%). Owing to this difference, an optimal C_o of 2.87 was found (as compared to the value of 2.80 proposed by BI14), creating a new baseline for comparison. This triggering curve in shown in Figure 13 as a dashed orange line.

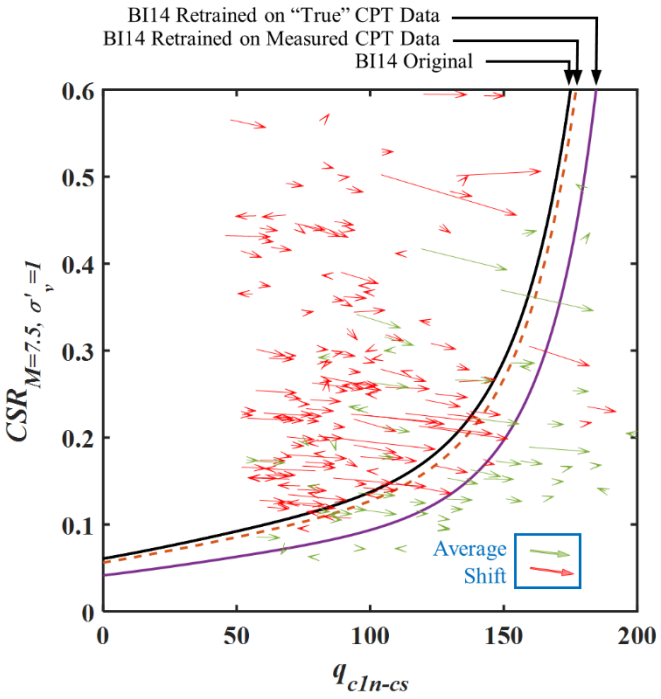

Fig. 13. Retraining of the Boulanger and Idriss (2014) triggering curve using "true" CPT data from, as described in the text. Vectors indicate the change to case-history datapoints resulting from CPT inversion.

Next, the Boulanger and DeJong (2018) inversion procedure was applied to the partial BI14 database, and "critical layers" proposed by BI14 for those cases were resampled. The resulting shift of each case-history point in triggering space is shown in Figure 13, wherein red vectors indicate shifts for cases with manifestations of liquefaction and green vectors indicate shifts for cases without manifestations of liquefaction. While CPT inversion both decreased and increased q_{c1Ncs} (cone tip resistance, normalized for overburden pressure and adjusted for fines content), the average changes were +16.24 and +14.67 for cases with and without manifestations, respectively. The corresponding changes in $CSR_{M=7.5,\ \sigma'_V=1}$ (cyclic stress ratio, normalized for overburden pressure and adjusted for earthquake magnitude) were respectively -0.007 and -0.005. This lesser change in *CSR* is unsurprising, given the lesser dependence of *CSR* on soil properties, relative to q_{c1Ncs}. Repeating the BI14 optimization procedure, but training on "true" CPT data, the optimal C_o was 3.18 (compared to 2.87 using measured data). Given the shift of this curve, which is shown in Figure 13 as a solid purple line, it appears that "true" data is not optimally compatible with a model trained on measured data. Accordingly, using the version of BI14 newly trained on "true" data, the preceding ROC analyses of case-histories were repeated on the global and Canterbury "true" datasets. However, relative to the *AUC*s obtained by applying the original BI14 model to measured data, those

obtained by applying the new BI14 model to "true" data were less. Using BI14-*LPI* as an example, *AUC* was 1.2% less in both the global and Canterbury datasets.

The final question, then, is whether any triggering curve can be found to improve predictions of liquefaction manifestations when inverted CPT data is used? To investigate, the C_0 parameter in Eq. (5) was varied from 1.5 to 5.0 in increments of 0.1, which is analogous to changing the probability associated with liquefaction triggering from 16% to various other values. For each C_0, BI14-*LPI* predictions were recomputed for each case history using "true" data, after which *AUC* values were computed for the global and Canterbury datasets. The results of these analyses, which are summarized in Figure 14, show that while *AUC*s computed from "true" data may be slightly improved using C_0 values different from the BI14 default of 2.8, these *AUC*s are still less than those computed from measured data. Thus, there exists no triggering curve obtainable via C_0 recalibration for which CPT inversion demonstrably improves the efficiency of liquefaction predictions.

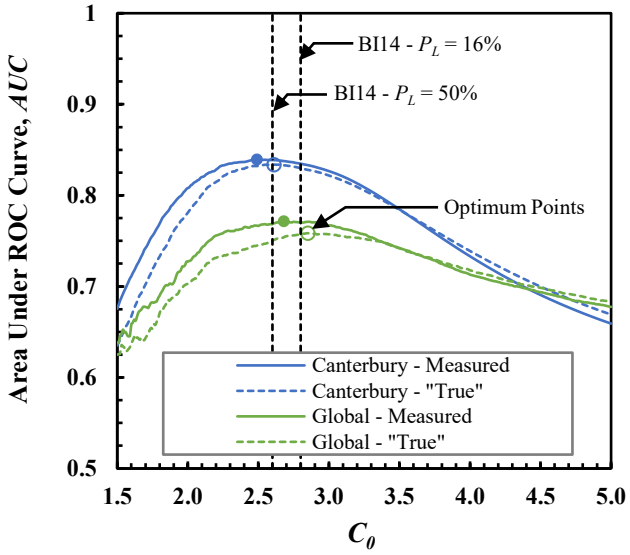

Fig. 14. BI14-*LPI AUC* values with and without inversion, as a function of the BI14 parameter C_0.

8. Caveats and Limitations

The presented findings are inherently tied to the datasets studied herein. The applicability of these findings to other datasets (e.g., soils of unusual age, minerology, composition, etc.) – or to other methodologies, is unknown. The phrases "true CPT data" and "CPT inversion" refer specifically to the Boulanger and DeJong (2018) inversion procedure using "baseline" parameters, as implemented in the software *Horizon*. As previously discussed, it is plausible these parameters could be calibrated at the site-

specific level to potentially improve identification and correction of thin layers. As an example, it is inherently a challenge to distinguish graded strata (i.e., fining upwards or downwards) from relatively distinct material interfaces, since both appear as a rate of change in CPT data, and since both naturally occur. While the procedure includes a baseline parameter for flagging material interfaces, its selection invariably includes subjectivity that ideally could be confirmed by borehole sampling or knowledge of local geology. In the present study, all globally available liquefaction case histories are studied. As a result, site-specific calibration was not undertaken, but could conceivably improve performance. In addition, it should be emphasized that the Boulanger and DeJong (2018) procedure was not directly evaluated, and as such, nothing can be directly concluded about its efficacy. That is, its ability to accurately correct CPT data for multiple thin-layer effects was not assessed. This would require CPT calibration chamber data or numerical simulations of such data, both in uniform and layered deposits. When analyzing field case histories, as done in this study, only the combined performance of a CPT inversion procedure *and* a liquefaction model can be quantified. In this regard, the Boulanger and DeJong (2018) procedure might improve liquefaction predictions using some triggering and/or manifestation models other than those utilized in this study. Similarly, the procedure might provide utility in other geotechnical applications. That is, the lack of utility observed herein could be due to fundamental limitations in the liquefaction models, rather than to limitations of the procedure itself. However, given that model performance at-best increased insignificantly (but most often, decreased significantly), it is unlikely that minor adjustments to liquefaction models would alter this outcome. Ultimately, additional data will confirm or update the findings presented herein and summarized below.

9. Conclusions

The Boulanger and DeJong (2018) CPT inversion procedure was evaluated in the context of CPT-based liquefaction model performance. Using field case-histories parsed into 2 datasets, 18 different liquefaction models were studied, resulting in 36 performance trials. In only 1 trial did the CPT inversion procedure increase model efficiency to a statistically significant degree, while in 23 others it significantly decreased efficiency. This decline in performance, which was independent of the I_c cutoff used to infer liquefaction susceptibility, tended to grow as profiles became more stratified, opposite of what ideally should occur. To explore possible remedies, a liquefaction triggering curve was rederived from inverted CPT data, such that its training and forward implementation were made consistent. Nonetheless, this exacerbated the decline in prediction efficiency when applied to field case histories. Moreover, no readily conceivable triggering curve could be found to improve predictions of liquefaction when inverted CPT

data was used. Ultimately, the results of this study are not a direct assessment of the pioneering Boulanger and DeJong (2018) procedure. However, the results do provide strong evidence that this procedure may provide no demonstrable performance benefit when applied to existing CPT-based liquefaction models. This conclusion should be weighed against caveats and limitations discussed in the preceding section.

10. Acknowledgements

This study is based on work supported by the National Science Foundation (NSF), US Geological Survey (USGS), and Pacific Earthquake Engineering Research Center (PEER) under Grant Nos. CMMI-1751216, G18AP-00006, and 1132-NCTRBM, respectively. The authors also wish to acknowledge the numerous researchers who contributed to the data studied herein. Much of this data was collected under the sponsorship of either the New Zealand Earthquake Commission (EQC) or one of the three organizations above. However, any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF, USGS, PEER, or EQC.

11. Data Availability

All data analyzed in this study is publicly available. In addition, many calculations performed as part of the study, including CPT processing and liquefaction modelling, were carried out using *Horizon*, a freely available program developed by the authors.

12. References

Ahmadi MM and Robertson PK (2005) Thin-layer effects on the CPT q_c measurement. *Canadian Geotechnical Journal* 42(5): 1302-1317.

Anderson TW and Darling DA (1952) Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics* 23(2): 193-212.

Architectural Institute of Japan (2001) *Recommendations for design of building foundations*, 486 p.

Bastin, S., van Ballegooy, S., Mellsop, N., & Wotherspoon, L. (2020). Liquefaction case histories from the 1987 Edgecumbe earthquake, New Zealand–Insights from an extensive CPT dataset and paleo-liquefaction trenching. *Engineering Geology*, 271: 105404.

Beyzaei CZ, Bray JD, van Ballegooy S, Cubrinovski M, & Bastin S (2018) Depositional environment effects on observed liquefaction performance in silt swamps during the Canterbury earthquake sequence. *Soil Dynamics and Earthquake Engineering* 107: 303-321.

Boulanger RW and Idriss IM (2014) *CPT and SPT Based Liquefaction Triggering Procedures, Report No. UCD/CGM-14/01*, Center for Geotechnical Modeling, University of California, Davis, CA.

Boulanger RW and DeJong JT (2018) Inverse filtering procedure to correct cone penetration data for thin-layer and transition effects. *Cone Penetration Testing 2018*, Hicks, Pisano, and Peuchen, eds., Delft University of Technology, The Netherlands: 25-44.

Boulanger RW, Khosravi M, Cox BR and DeJong JT (2019) Liquefaction Evaluation for an Interbedded Soil Deposit: St. Teresa's School, Christchurch, New Zealand. IACGE 2018: Geotechnical and Seismic Research and Practices for Sustainability: 686-704.

Brennan AJ and Madabhushi, SP (2005) Liquefaction and drainage in stratified soil. *JGGE* 131(7): 876-885.
Brown LJ (1975) Water well data. Sheet 576/7-8, Belfast-Styx. New Zealand Geological Survey, Report 72.
Brown LJ and Weeber JH (1992) Geology of the Christchurch urban area: Institute of Geological and Nuclear Sciences Geological Map 1. Institute of Geological and Nuclear Sciences Limited, Lower Hutt, New Zealand, scale, 1(25,000), 1.
Buck JR, Daniel MM, and Singer AC (2002) *Computer Explorations in Signals and Systems Using MATLAB®,* 2nd Edition. Upper Saddle River, NJ: Prentice Hall.
Chen, Q., Wang, C., & Juang, C. H. (2016). Probabilistic and spatial assessment of liquefaction-induced settlements through multiscale random field models. *Engineering Geology*, *211*, 135-149.
Ching J, Wang JS, Juang CH, & Ku CS (2015) Cone penetration test (CPT)-based stratigraphic profiling using the wavelet transform modulus maxima method. *Canadian Geotechnical Journal 52*(12): 1993-2007.
Cubrinovski M, Rhodes A, Ntritsos N, and van Ballegooy S, (2019). System response of liquefiable deposits. *SDEE* 124: 212-229.
DeLong ER, DeLong DM and Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44: 837-845.
Diaconis, P and Efron B (1983) Computer intensive methods in statistics. *Sci. Amer*, 248(5): 116–130.
Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognition Letters* 27(8): 861–874.
Fiegel GL and Kutter BL, (1994). Liquefaction mechanism for layered soils. *JGGE* 120(4): 737-755.
Gheibi, E., & Gassman, S. L. (2016). Application of GMPEs to estimate the minimum magnitude and peak ground acceleration of prehistoric earthquakes at Hollywood, SC. *Engineering Geology*, *214*, 60-66.
Geologismiki (2020). CLiq v3, CPT soil liquefaction software, https://geologismiki.gr/products/cliq/. Geologismiki Geotechnical Software, Serres, Greece.
Geyin M, and Maurer BW (2019). An analysis of liquefaction-induced free-field ground settlement using 1,000+ case-histories: observations vs. state-of-practice predictions. Geotechnical Special Publication 308: 489-498.
Geyin M and Maurer BW (2020a) Horizon: CPT-based liquefaction risk assessment and decision software. DesignSafe-CI, doi: 10.17603/ds2-2fky-tm46.
Geyin M and Maurer BW (2020b) Fragility functions for liquefaction-induced ground failure. *Journal of Geotechnical and Geoenvironmental Engineering* 146(12): 04020142.
Geyin M and Maurer BW (2021) CPT-Based Liquefaction Case Histories from Global Earthquakes: A Digital Dataset (Version 1). DesignSafe-CI. https://doi.org/10.17603/
Geyin M, Maurer BW, Bradley BA, Green RA, and van Ballegooy S. (2020a) CPT-Based Liquefaction Case Histories Resulting from the 2010-2016 Canterbury, New Zealand, Earthquakes: A Curated Digital Dataset (Version 2). DesignSafe-CI. https://doi.org/10.17603/ds2-tygh-ht91.
Geyin M, Baird AJ and Maurer BW (2020b) Field assessment of liquefaction prediction models based on geotechnical vs. geospatial data, with lessons for each. Earthquake Spectra 36(3): 1386–1411.
Geyin M, Maurer BW, Bradley BA, Green RA, and van Ballegooy S (2021) CPT-based liquefaction case histories compiled from three earthquakes in Canterbury, New Zealand. *Earthquake Spectra*, In Press.
Green RA, Cubrinovski M, Cox B, Wood C, Wotherspoon L, Bradley B and Maurer B (2014) Select Liquefaction Case Histories from the 2010-2011 Canterbury Earthquake Sequence. *Earthquake Spectra* 30(1): 131-153.
Green RA, Bommer JJ, Rodriguez-Marek A, Maurer BW, Stafford PJ, Edwards B, Kruiver PP, De Lange G and Van Elk J (2019) Addressing limitations in existing 'simplified' liquefaction triggering evaluation procedures: application to induced seismicity in the Groningen gas field. *Bulletin of Earthquake Eng* 17(8): 4539-4557.
Hanley JA & McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology 143*(1): 29-36.
Hasek, M. J., & Gassman, S. L. (2019). Cyclic resistance ratio of pleistocene-age sands from the South Carolina coastal plain. *Engineering Geology*, *251*, 158-171.

Heidari, T., & Andrus, R. D. (2010). Mapping liquefaction potential of aged soil deposits in Mount Pleasant, South Carolina. *Engineering Geology*, *112*(1-4), 1-12.

Holzer TL and Youd TL (2007) Liquefaction, ground oscillation, and soil deformation at the wildlife array, California. *Bulletin of the Seismological Society of America* 97(3): 961-976.

Idriss IM and Boulanger RW (2008). Soil liquefaction during earthquakes. *Monograph MNO-12* 2008; Earthquake Engineering Research Institute, Oakland, CA, 261 pp.

Iwasaki T, Tatsuoka F, Tokida K, and Yasuda S (1978) A practical method for assessing soil liquefaction potential based on case studies at various sites in Japan. 2nd Intl Conf. Microzonation.

Ju, N., Huang, J., He, C., Van Asch, T. W. J., Huang, R., Fan, X., ... & Wang, J. (2020). Landslide early warning, case studies from Southwest China. *Engineering Geology*, *279*: 105917.

Juang, C. H., Lu, C. C., & Hwang, J. H. (2009). Assessing probability of surface manifestation of liquefaction at a given site in a given exposure time using CPTU. *Engineering Geology*, 104(3-4): 223-231.

Khoshnevisan, S., Juang, H., Zhou, Y. G., & Gong, W. (2015). Probabilistic assessment of liquefaction-induced lateral spreads using CPT—focusing on the 2010–2011 Canterbury earthquake sequence. *Engineering Geology*, *192*, 113-128.

Lilliefors HW (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association 62*(318): 399-402.

Lin, A., Wotherspoon, L., Bradley, B., & Motha, J. (2021). Evaluation and modification of geospatial liquefaction models using land damage observational data from the 2010–2011 Canterbury Earthquake Sequence. *Engineering Geology*, *287*: 106099.

Lunne T, Robertson PK and Powell JM (1997) *Cone Penetration Testing in Geotechnical Practice*. Blackie Academic & Professional, London, U.K.

Maurer BW, Green RA, Cubrinovski M and Bradley BA (2014) Evaluation of the liquefaction potential index for assessing liquefaction hazard in Christchurch, New Zealand. *JGGE* 140(7): 04014032.

Maurer BW, Green RA and Taylor ODS (2015a) Moving towards an improved index for assessing liquefaction hazard: lessons from historical data. *Soils and Foundations* 55(4): 778-787.

Maurer BW, Green RA, Cubrinovski M and Bradley B (2015b) Assessment of CPT-based methods for liquefaction evaluation in a liquefaction potential index framework. Géotechnique 65(5): 328-336.

Maurer BW, Green RA, Cubrinovski M and Bradley BA (2015c) Fines-content effects on liquefaction hazard evaluation for infrastructure during the 2010-2011 Canterbury, New Zealand earthquake sequence. *Soil Dynamics and Earthquake Engineering* 76: 58-68.

Maurer BW, Green, RA, van Ballegooy S, and Wotherspoon L (2019). Development of region-specific soil behavior type index correlations for evaluating liquefaction hazard in Christchurch, New Zealand. *SDEE* 117: 96-105.

McLaughlin K (2017) Investigation of false-positive liquefaction case history sites in Christchurch, New Zealand. M.S. Thesis. The University of Texas at Austin, Austin, TX.

Mo PQ, Marshall AM and Yu HS (2017). Interpretation of cone penetration test data in layered soils using cavity expansion analysis. *JGGE 143*(1): 04016084.

Moss RES, Seed RB, Kayen RE, Stewart JP, Der Kiureghian A and Cetin KO (2006) CPT-based probabilistic and deterministic assessment of in situ seismic soil liquefaction potential. *JGGE* 132(8): 1032-1051.

Nakata T and Shimazaki K (1997) Geo-slicer, a newly invented soil sampler, for high-resolution active fault studies. *Journal of Geography* 106(1): 59-69 (in Japanese).

National Research Council (NRC) (2016). *State of the Art and Practice in the Assessment of Earthquake-Induced Soil Liquefaction and its Consequences*, Committee on Earthquake Induced Soil Liquefaction Assessment (Edward Kavazanjian, Jr., Chair, Jose E. Andrade, Kandian "Arul" Arulmoli, Brian F. Atwater, John T. Christian, Russell A. Green, Steven L. Kramer, Lelio Mejia, James K. Mitchell, Ellen Rathje, James R. Rice, and Yumie Wang), The National Academies Press, Washington, DC.

Norini, G., Aghib, F. S., Di Capua, A., Facciorusso, J., Castaldini, D., Marchetti, M., ... & Piccin, A. (2021). Assessment of liquefaction potential in the central Po plain from integrated geomorphological, stratigraphic and geotechnical analysis. *Engineering Geology*, 282: 105997.

Özener P, Özaydin K, and Berilgen M (2008). Numerical and Physical Modeling of Liquefaction Mechanisms in Layered Sands. *Geotechnical Earthquake Engineering and Soil Dynamics* IV, 1-12.

Ozutsumi O, Sawada S, Iai S, Takeshima Y, Sugiyama W, & Shimazu T (2002). Effective stress analyses of liquefaction-induced deformation in river dikes. *SDEE* **22**(9): 1075-1082.

Powers, D.M.W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies* (1): 37–63.

Raschke SA, and Hryciw RD (1997). Vision cone penetrometer for direct subsurface soil observation. *JGGE* **123**(11): 1074-1076.

Robertson PK (2011) Automated detection of CPT transition zones. *Geotechnical News*, June: 35-38.

Robertson PK and Wride CE (1998). Evaluating cyclic liquefaction potential using cone penetration test. *Canadian Geotechnical Journal* 35(3): 442-459.

Sarma, C. P., Dey, A., & Krishna, A. M. (2020). Influence of digital elevation models on the simulation of rainfall-induced landslides in the hillslopes of Guwahati, India. *Engineering Geology*, *268*: 105523.

Seed HB & Idriss IM (1971). Simplified procedure for evaluating soil liquefaction potential. *ASCE Journal of Soil Mechanics and Foundations Division* 97(9): 1249–1273.

Takada K and Atwater BF (2004). Evidence for liquefaction identified in peeled slices of Holocene deposits along the lower Columbia River, Washington. *BSSA* 94(2): 550-575.

Treadwell DD (1976) The influence of gravity, prestress, compressibility, and layering on soil resistance to static penetration. PhD Dissertation, Univ. of California, Berkeley, CA.

Upadhyaya, S., Maurer, B. W., Green, R. A., & Rodriguez-Marek, A. (2021). Selecting the optimal factor of safety or probability of liquefaction triggering for engineering projects based on misprediction costs. *Journal of Geotechnical and Geoenvironmental Engineering*, *147*(6): 04021026.

van Ballegooy S, Malan P, Lacrosse V, Jacka ME, Cubrinovski M, Bray JD, O'Rourke TD, Crawford SA, and Cowan H, (2014). Assessment of liquefaction-induced land damage for residential Christchurch. *Earthquake Spectra* 30(1): 31-55.

van der Linden TI, De Lange DA and Korff M (2018) Cone penetration testing in thinly inter-layered soils. *Proceedings of the Institution of Civil Engineers-Geotechnical Engineering* 171(3): 215-231.

Whitman RV, (1971) Resistance of soil to liquefaction and settlement, *Soils & Foundations* 11(4): 59–68.

Youd TL, Idriss IM, Andrus RD, Arango I, Castro G, Christian JT, Dobry R, Finn WDL, Harder LF, Hynes ME et al. (2001) Liquefaction resistance of soils: summary report from the 1996 NCEER and 1998 NCEER/NSF workshops on evaluation of liquefaction resistance of soils. *JGGE* 127: 817–833.

Zhang, J., Chen, F.Y., Juang, C.H., & Chen, Q. (2018). Developing joint distribution of amax and Mw of seismic loading for performance-based assessment of liquefaction induced structural damage. *Engineering Geology*, 232: 1-11.

Zhang, J., Juang, C.H., Martin, J.R., & Huang, H.W. (2016). Inter-region variability of Robertson and Wride method for liquefaction hazard analysis. *Engineering Geology*, 203: 191-203.

Zou KH (2007) Receiver operating characteristic (ROC) literature research. *On-line bibliography*, http://www.spl.harvard.edu/archive/spl-pre2007/pages/ppl/zou/roc.html