Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Site-independent prediction of field vane undrained shear strength of Shanghai clay: machine-learning models versus a log-normal multivariate baseline under leave-one-site-out validation

Seyed Mahdi Naseri Baygi

civil.engineers85@yahoo.com

Islamic Azad University, Mashhad Branch https://orcid.org/0009-0007-4414-251X

Research Article

Keywords: Undrained shear strength, Field vane test, Machine learning, Leave-one-site-out validation, Conformal prediction, Shanghai clay

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11653232/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Site-independent prediction of field vane undrained shear strength of Shanghai clay: machine-learning models versus a log-normal multivariate baseline under leave-one-site-out validation

Seyed Mahdi Naseri Baygi (corresponding author) · Department of Civil Engineering, Islamic Azad University, Mashhad Branch, Mashhad, Iran · seyedmahdi.naseribaygi@iau.ac.ir · ORCID: https://orcid.org/0009-0007-4414-251X

Preprint; not peer reviewed.

Abstract

Undrained shear strength (S_u) is often estimated from index properties, but reported machine-learning (ML) accuracies frequently rest on random splits that place rows from the same site in both training and test sets. Using the public SH-CLAY/11/4051 database of Shanghai clay, we evaluate eight regression models (ridge regression, support vector regression, Gaussian process regression, random forest, XGBoost, LightGBM, a perceptron and a perceptron ensemble) for predicting field vane S_u from depth, plasticity index, void ratio and effective vertical stress (377 complete rows from 28 sites). Models are assessed by nested leave-one-site-out cross-validation, against log-normal multivariate (ln-MVN) and linear mixed-effects baselines using identical inputs, with site-cluster bootstrap intervals. All ML models outperform the ln-MVN baseline: pooled out-of-site root-mean-square error (RMSE) is 5.15 to 5.34 kPa for the nonlinear models, 5.64 kPa for ridge regression and 6.09 to 6.10 kPa for the ln-MVN baselines (paired differences 0.4 to 0.9 kPa, 95% intervals excluding zero). Differences among the nonlinear models are small; only the ANN ensemble and random forest clearly outperform the mixed-effects baseline, and adding depth improves no model. Split-conformal intervals reach 0.88 to 0.91 coverage at the 90% level, but coverage varies strongly between sites. Effective vertical stress dominates out-of-site permutation importance. Monotonicity constraints did not change accuracy; an offset estimated from five local measurements lowered the remaining error by 0.1 to 0.4 kPa. The work applies existing methods to an existing database under deployment-consistent validation; it has no external validation set.

Keywords: Undrained shear strength; Field vane test; Machine learning; Leave-one-site-out validation; Conformal prediction; Shanghai clay

1 Introduction

Undrained shear strength (S_u) is a primary input to the design of foundations, embankments and excavations in soft clay. Direct determination by laboratory testing or by the field vane test (VST) is costly and rarely dense enough to characterise a whole site, so practice relies on correlations with index properties, stress state and in-situ measurements, and on probabilistic descriptions that carry the uncertainty of such correlations into design. Ching and Phoon [1] built a multivariate distribution of clay parameters from a global database, and Zhang et al. [2] compiled the SH-CLAY/11/4051 database of Shanghai clay and fitted a multivariate normal model to it, which supports conditional prediction of one parameter from the others.

Machine-learning (ML) models are now widely applied to the same task. Most reported studies use data sets of modest size from a single region: 230 observations from Turkey for S_u from standard penetration test results and consistency indices [3], 127 samples from a Vietnamese highway project [4], 216 sensitive-clay samples from southern Finland modelled with an adaptive neuro-fuzzy system [5], and 384 data points from 24 Finnish locations modelled with ensemble methods [6]; Elsawy et al. [7] modelled S_u of sensitive alluvial soft clay with ML regressors. Larger collections have also been used. Zhang et al. [8] predicted S_u of soft clays from the TC304 database with XGBoost and random forest under five-fold cross-validation; Xiong et al. [9] addressed missing data in the multi-site CLAY/10/7490 database by imputation and developed probabilistic neural-network and

boosting models, validated by stratified five-fold partitioning; and Zhang et al. [10] applied Bayesian neural networks to S_u and the compression index and reported low reliability where data were sparse. Where the validation scheme is described, it is a train–test split [3, 7] or k-fold cross-validation [8, 9]; none of the descriptions we examined forms folds by site.

This matters because geotechnical data are hierarchical. Rows from one site share geology, sampling practice and test procedures and are therefore not independent. If rows are split at random, information from the held-out site reaches the training set, and the resulting accuracy describes interpolation among known sites rather than prediction for a new one [11]. Taherdangkoo et al. [12] argue that claims of ML performance in geotechnics should be supported by deployment-consistent validation, calibrated uncertainty, physical admissibility and explicit applicability limits, and show with a multi-site clay database that apparent performance and its engineering interpretation depend strongly on the validation scheme; in their case study, leave-one-site-out evaluation of a Bayesian ridge model gave a strongly negative fold-mean coefficient of determination for the logarithm of a normalised mobilised strength. Ching and Wang [13] likewise used leave-one-site-out testing in predicting foundation capacity from a load-test database.

The present study addresses three questions for field-vane S_u of Shanghai clay. (RQ1) Under leave-one-site-out validation, do ML models outperform a multivariate log-normal model that uses the same inputs? (RQ2) Do the nonlinear ML families differ from one another, or from a linear model, by more than the sampling uncertainty of the sites, and does adding depth to three index properties help? (RQ3) Can prediction intervals calibrated on out-of-site residuals reach their nominal coverage, and how does coverage vary between sites? To our knowledge, a leave-one-site-out comparison of several ML families against a multivariate log-normal baseline with identical inputs has not been reported for the SH-CLAY database; this statement rests on a screening of the titles and abstracts of the 61 records citing Zhang et al. [2] in Scopus (export of 5 November 2026) and on the literature we examined; a full-text systematic search has not been performed. The contribution is therefore an application of existing methods to an existing database under deployment-consistent validation, not a new algorithm. Specifically, we (i) evaluate eight ML models and five baselines with nested leave-one-site-out cross-validation; (ii) compare ML models and the ln-MVN baseline on identical input sets, with and without depth; (iii) quantify uncertainty with site-cluster bootstrap intervals for the metrics and split-conformal prediction intervals for the predictions; (iv) examine site-level and regime-level errors, the dependence of accuracy on the number of training sites, and the physical behaviour of the fitted models; and (v) test whether physics-guided constraints (monotonicity in effective stress, a strength-ratio target) or an empirical-Bayes site offset based on a few local measurements, in the spirit of quasi-site-specific prediction [13], improve out-of-site accuracy.

2 Data

2.1 The SH-CLAY/11/4051 database

SH-CLAY/11/4051 contains 4051 rows from 50 sites in Shanghai with 11 clay parameters and was compiled by Zhang et al. [2]; it is openly available [14]. A row is a composite of boreholes within a 35 m radius at the same depth, and the vane test is not necessarily co-located with the boreholes that supplied the index properties, so the inputs and the target of one row do not always originate from one specimen [2]. Missingness is structural: not every parameter was measured at every site.

2.2 Analysis subset and variables

We analysed the complete-case subset in which depth (z), plasticity index (PI), void ratio (e), effective vertical stress (σ'_v) and field vane S_u are all available. This subset has 377 rows from 28 sites (3 to 41 rows per site; the largest site contributes 10.9% of rows) with no missing values and no exact duplicate rows. No row was removed on the basis of target values or model error. Effective vertical stress is not measured independently; it is derived

from depth and an effective unit weight [2], which makes it strongly collinear with depth (Spearman's $\rho \approx 0.98$ in this subset). Table 1 lists the variables. Whether the vane strengths in SH-CLAY are corrected for rate and anisotropy (for example with a Bjerrum-type factor) is not stated in the source text available to us; S_u is analysed as provided. Fig. 1 shows the distribution of S_u, the number of rows per site and the rank correlations among the variables, and Fig. 2 shows S_u against each input.

Table 1 Data dictionary and descriptive statistics of the analysis subset (n = 377 rows, 28 sites)

Variable	Meaning	Unit	Role	Min	Median	Max	Mean	SD
depth	Sampling depth z	m	Input	2.50	15.00	28.00	14.74	5.53
PI	Plasticity index	%	Input	10.50	18.60	25.50	18.70	2.94
e	Void ratio	–	Input	0.85	1.30	1.86	1.28	0.17
sigv	Effective vertical stress σ′v (depth × effective unit weight)	kPa	Input	29.20	133.20	243.40	129.80	43.31
Su_VST	Undrained shear strength from field vane test	kPa	Target	20.50	39.10	76.60	40.51	9.40

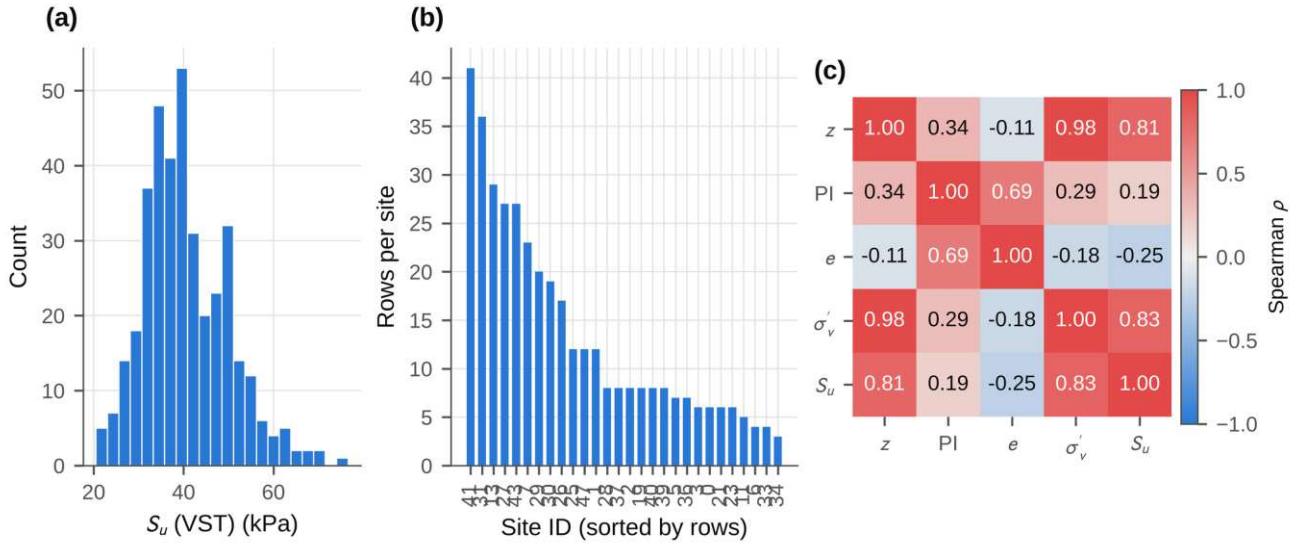


Fig. 1 Data overview: (a) histogram of S_u (VST); (b) number of rows per site; (c) Spearman rank correlations among depth, PI, e, σ'_v and S_u

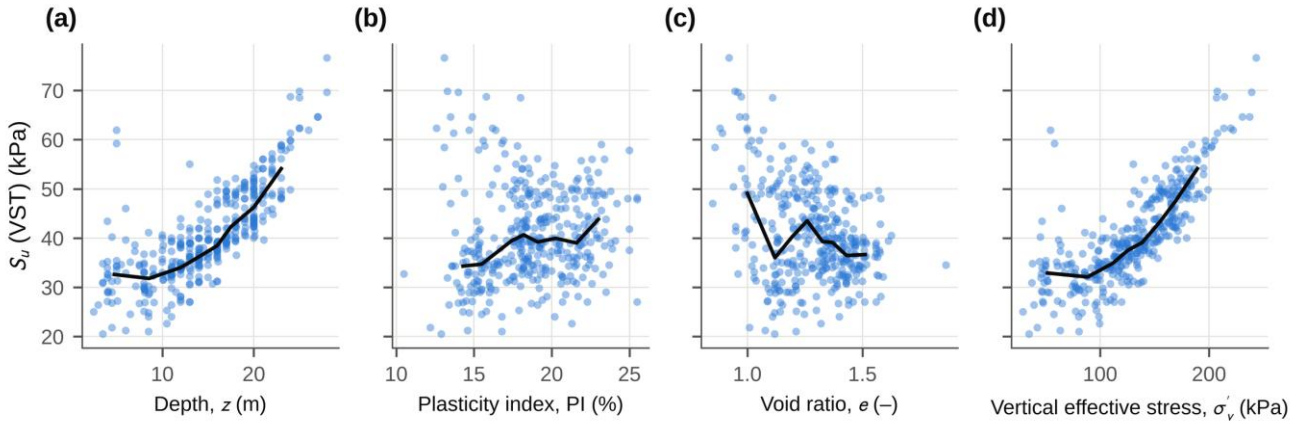


Fig. 2 S_u (VST) versus (a) depth, (b) plasticity index, (c) void ratio and (d) effective vertical stress; the line is a binned running mean

3 Methods

3.1 Validation design

The target use is prediction for a site that was not used for model development. The outer loop is therefore leave-one-site-out (LOSO) cross-validation with 28 folds: all rows of one site form the test set, and the remaining 27 sites form the training set. Within each training set, hyperparameters are selected by grid search with inner GroupKFold (five folds, grouped by site) using RMSE as the criterion, so no test-site information reaches model selection. Feature scaling and, where used, target scaling are fitted inside the training data only. The code asserts that no site appears in both the training and test sets of any fold and that every row is tested exactly once. Predictions from all folds are pooled, and all metrics are computed on the pooled out-of-site predictions. Fold-mean coefficients of determination are not used because small sites make them unstable [12]; per-site error distributions are reported instead. Random ten-fold cross-validation, which mixes rows of the same site across training and test sets, is run only as a diagnostic of validation optimism and is not used for any claim about new sites. Fig. 3 summarises the workflow.

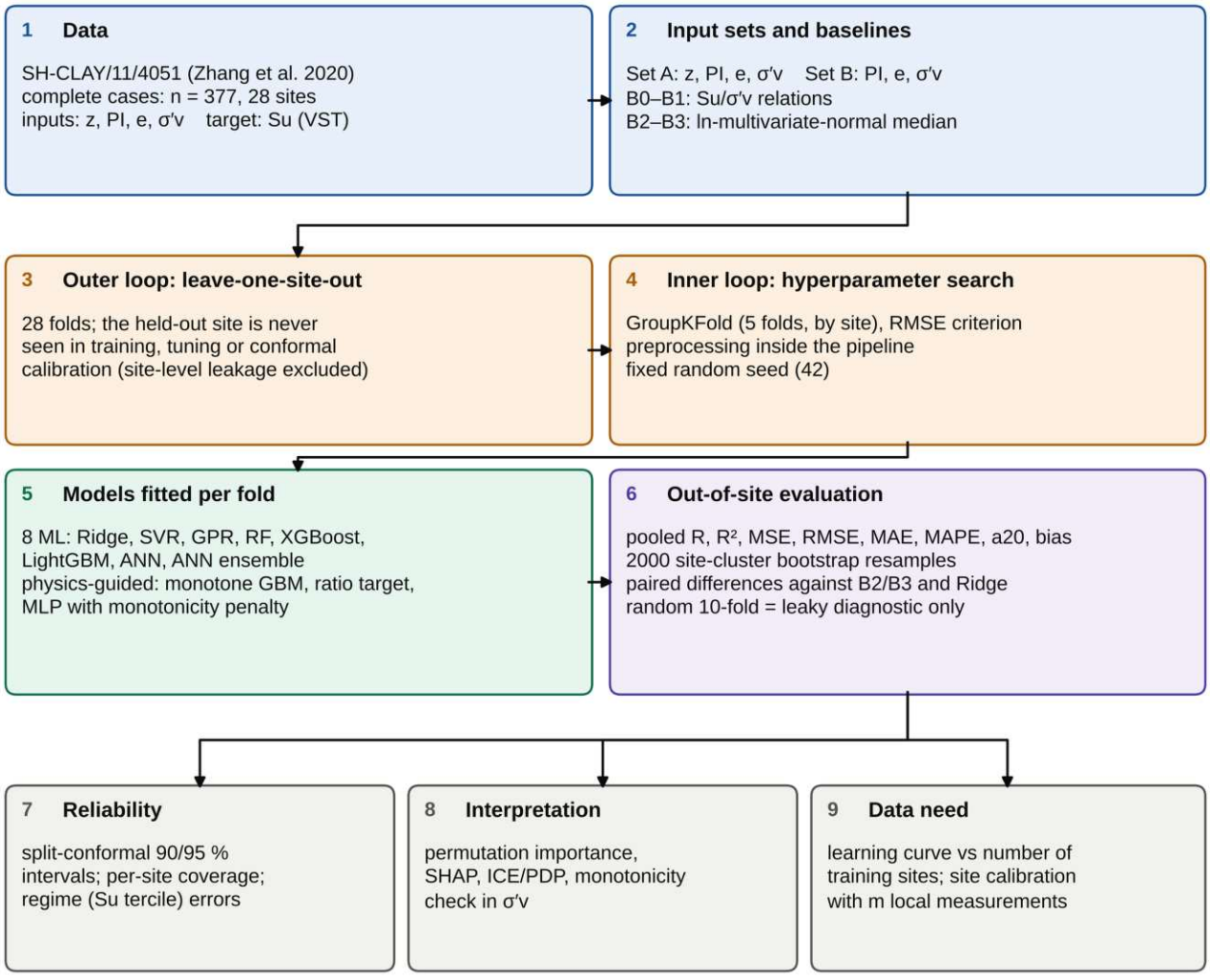


Fig. 3 Workflow of the study: data, nested leave-one-site-out validation, models, out-of-site evaluation and diagnostics (numbers refer to the sections of the analysis)

3.2 Models

Eight ML models were compared: ridge regression; support vector regression with an RBF kernel (SVR); Gaussian process regression with an anisotropic (ARD) Matérn-5/2 kernel plus a white-noise term (GPR) [15]; random forest (RF) [16]; XGBoost [17]; LightGBM [18]; a single multilayer perceptron (ANN; one hidden layer,

tanh activation, L-BFGS optimiser, L2 penalty); and an ensemble of ten such perceptrons, each fitted on a bootstrap resample of the training rows, whose mean is the prediction (ANN ensemble; 16 hidden neurons and L2 penalty 3.0, fixed rather than tuned in the inner loop). These ensemble settings were chosen by the author with knowledge of earlier cross-validation runs on the same data, so the ensemble result may be slightly optimistic relative to the nested-tuned models. Implementations used scikit-learn [19], XGBoost 3.4.1 and LightGBM 4.7.0. The hyperparameter grids are listed in Online Resource 1; GPR kernel parameters are optimised by marginal likelihood. All stochastic components use fixed seeds (seed 42).

3.3 Baselines

$$\ln Su = \beta 0 + \beta 1 \ln PI + \beta 2 \ln e + \beta 3 \ln \sigma' v + \varepsilon \qquad (1)$$

A sixth baseline, B4, is a linear mixed-effects model with a random intercept for each site, $S_u = \beta 0 + \beta 1\ PI + \beta 2\ e + \beta 3\ \sigma'_v + u_{site} + \varepsilon$, fitted by restricted maximum likelihood on standardised inputs of set B in each training fold. Because the site effect of a new site is unknown, the population-level prediction (fixed effects only) is used. Variants with a random slope in σ'_v and with all variables on the logarithmic scale (back-transformed to the median) were also evaluated. This model family is the standard formulation for grouped data and is the natural benchmark for any claim about site-to-site variability.

Five baselines were fitted inside each training fold. The constant baseline predicts the training mean. B0 is $S_u = a \cdot \sigma'_v$ and B1 is $S_u = \sigma'_v(a + b \cdot PI)$; both are data-fitted linear models, not published empirical relations. B2 is a simplified log-normal multivariate (ln-MVN) model: ordinary least squares on the natural logarithms of S_u and of PI, e and σ'_v, with prediction of the conditional median exp(·) (Eq. 1). This corresponds to the conditional median of a multivariate normal model in log space, in the spirit of Ching and Phoon [1] and Zhang et al. [2], but without their Johnson-system marginal transformations, so it is a simplified version of the published approach. B3 is B2 with depth added, to match the four-input ML models. Because B2 does not use depth, ML models were fitted on two input sets: set A (z, PI, e, σ'_v) is compared with B3, and set B (PI, e, σ'_v) with B2.

3.4 Performance metrics and statistical comparison

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

$$RMSE = \sqrt{\frac{1}{n}\sum(yi - \hat{y}i)^2} \qquad (3)$$

For measured values y_i and predictions $\hat{y}_i$ (i = 1…n, kPa), we report the Pearson correlation coefficient R; R^2 (Eq. 2); the mean squared error (MSE); the root-mean-square error (RMSE, Eq. 3); the mean absolute error (MAE); the mean absolute percentage error (MAPE, %); a20, the fraction of predictions within ±20% of the measured value; and the mean bias, mean($\hat{y}_i - y_i$). These are standard definitions computed on pooled out-of-site predictions. Uncertainty of each metric is estimated by a site-cluster bootstrap (2000 resamples of sites with replacement, percentile 95% intervals). Differences between two models are computed on the same bootstrap resamples (paired). We report the paired difference with its percentile interval and, descriptively, the bootstrap fraction in which the first model is better; this fraction is not a p-value. An interval that includes zero is treated as inconclusive, and no model is declared best on the basis of R^2 alone. For the interpretation analyses a single model is selected by the lowest pooled out-of-site RMSE; this selection uses the same predictions it is then reported on, so the label "lowest RMSE" is slightly optimistic.

3.5 Prediction intervals

$$\hat{y} \pm q,\ q = r(k),\ k = \lceil (n+1)(1-\alpha) \rceil \qquad (4)$$

$$IS = (u - l) + \frac{2}{\alpha}(l - y) \cdot 1\{y < l\} + \frac{2}{\alpha}(y - u) \cdot 1\{y > u\} \qquad (5)$$

Split-conformal intervals [20] were built separately within every outer fold. The absolute residuals of the training sites were obtained by inner out-of-site cross-validation (GroupKFold, up to five folds), and the finite-sample quantile of order $\lceil (n + 1)(1 - \alpha) \rceil / n$ of these residuals gives a symmetric half-width q, so that the interval is $\hat{y} \pm q$ (Eq. 4), where r(k) is the k-th smallest residual. Because the inner models see about 80% of the training sites, the residuals are slightly larger than those of the final model, which makes the intervals somewhat conservative. Intervals were evaluated for $\alpha = 0.2$, 0.1 and 0.05 by empirical coverage, mean width, the interval score (Eq. 5; penalty $2/\alpha$, with l and u the interval limits and y the measured value) and per-site coverage. The conformal guarantee assumes exchangeability of calibration and test residuals, which site-level heterogeneity can violate; coverage is therefore checked per site rather than assumed. For GPR and the ANN ensemble the model's own predictive standard deviation was also evaluated as a Gaussian 90% interval.

3.6 Error diagnostics and model interpretation

Per-site RMSE was computed for each held-out site. Regime errors were computed in terciles of the measured S_u and of σ'_v. For the model with the lowest pooled RMSE we computed out-of-site permutation importance (increase in out-of-site RMSE after permuting one input in the test site, repeated), Shapley additive explanations (SHAP) values [21] for the final model fitted to all data (Online Resource 1), individual conditional expectation (ICE) and partial dependence (PDP) curves between the 5th and 95th percentiles of each input, and a one-at-a-time sensitivity analysis in which each input (and depth with σ'_v together) is changed by ±10% and ±20% and the mean relative change of the prediction is recorded. Variance-based (Sobol) indices were not used because they assume independent inputs, which is violated by z and σ'_v. For every model fitted to all data we checked that predictions were positive and counted the fraction of ICE curves that were non-decreasing in σ'_v and in z (tolerance 1% of the standard deviation of S_u). Because z and σ'_v are collinear, ICE, PDP and permutation procedures evaluate combinations of inputs that rarely occur in the data; their results are interpreted as properties of the fitted functions, not as causal effects. Further diagnostic displays (a Taylor diagram [22], radar chart, residual plots and SHAP views) are given in Online Resource 1. In addition, the dependence of out-of-site accuracy on the number of training sites was examined for ridge regression, SVR and random forest with fixed, untuned hyperparameters (ridge penalty 1; SVR with C = 10, gamma = 0.1, epsilon = 0.1; random forest with 200 trees, maximum depth 3 and minimum leaf size 5): for each held-out site, k sites (k = 4, 8, 12, 16, 20, 24, 27) were drawn at random from the remaining sites, the model was fitted on their rows, and the pooled out-of-site RMSE was computed; draws were repeated eight times with a fixed seed (at k = 27 all remaining sites are used, so there is no sampling variability).

3.7 Physics-guided variants and quasi-site-specific calibration

Two kinds of physical knowledge were encoded and evaluated on input set B under the same nested LOSO design. (i) Monotonicity: S_u is expected not to decrease with σ'_v at fixed PI and e. XGBoost and LightGBM were refitted with a monotone constraint in σ'_v (small grids given in the released code). The perceptron ensemble was refitted with a soft penalty $\lambda \cdot \mathrm{mean}(\max(-\partial S_u/\partial \sigma'_v, 0)^2)$, evaluated on the training rows and on 200 collocation points drawn uniformly from the observed range of the inputs, with $\lambda = 0$ (no penalty) and $\lambda = 10$; the gradients of this implementation were verified against finite differences. This is a constraint-based, PINN-style regularisation, not a physics-informed neural network in the sense of a partial-differential-equation residual, because no governing equation of the vane strength is available. (ii) Normalisation: XGBoost, SVR and ridge regression were fitted to the strength ratio S_u/σ'_v and the prediction was multiplied by σ'_v. The ratio-target models and the penalty weight λ were not tuned. For calibration with local data, the held-out site contributes m randomly chosen rows (m = 1, 2, 3, 5; 30 random draws per site), and an additive offset equal to the shrunken mean residual, $m \cdot \tau^2/(m \cdot \tau^2 + \sigma^2)$ times the mean residual of these rows, is applied to the remaining rows of that site. The between-site variance τ^2 and within-site variance σ^2 of the out-of-site residuals were estimated by one-way random-effects analysis of variance from the other 27 sites; the intraclass correlation is $\tau^2/(\tau^2 + \sigma^2)$. For the mixed model the

same shrinkage corresponds to the best linear unbiased prediction of the site effect, with the variance components of the training sites. Sites with m or fewer rows are skipped for that m, and every offset result is compared with the uncorrected RMSE on the same rows.

3.8 Software and reproducibility

The analysis was run with Python 3.13, NumPy 2.5.3, pandas 3.0.5, SciPy 1.18.1, scikit-learn 1.9.1, XGBoost 3.4.1, LightGBM 4.7.0, SHAP 0.54.1 and Matplotlib 3.11.2 and statsmodels 0.15.0 (mixed models). The code includes unit tests for the metric definitions, the absence of site leakage, recovery of a known signal, a negative control, reproducibility of the bootstrap and the conformal coverage on synthetic data. The SHA-256 hash of the analysis file, all arguments and all library versions are stored with the results. Out-of-site predictions and the hyperparameters selected in each fold are provided in the public repository described under Data availability.

3.9 Use of large language models

Claude (Anthropic) was used as an assistant for writing and checking the Python analysis code and for drafting and editing the manuscript text. The author defined the research question and the validation design, supplied the data and the literature, ran the code, verified the results and references against the original sources, and takes full responsibility for the content.

4 Results

4.1 Out-of-site accuracy

Table 2 and Fig. 4 summarise pooled LOSO performance. The nonlinear ML models reached RMSE between 5.15 kPa (SVR) and 5.34 kPa (single ANN) and R^2 between 0.68 and 0.70; ridge regression reached RMSE 5.64 kPa (R^2 = 0.64). The ln-MVN baselines B2 and B3 reached RMSE 6.09 and 6.10 kPa (R^2 = 0.58), whereas the linear stress-ratio baselines B0 and B1 (R^2 = 0.25) and the constant baseline (R^2 = −0.05) were clearly worse. The 95% site-cluster bootstrap intervals of the individual metrics are wide (for example 4.35 to 6.25 kPa for the RMSE of SVR) and overlap strongly among ML models. All models showed a mean bias between −0.18 and +0.13 kPa except the baselines B0 to B3 (−1.81 to −0.53 kPa). Fig. 5 gives the metrics with their bootstrap intervals, Fig. 6 a radar chart of the normalised scores, Fig. 7 a Taylor diagram [22] of the predictions and Fig. 8 diagnostics of the residuals.

Table 2 Pooled leave-one-site-out performance (n = 377, 28 sites). Square brackets: 95% site-cluster bootstrap interval. RMSE, MAE and bias in kPa; MAPE in %; a20 is the fraction within ±20%

Model	R	R^2 [95% CI]	RMSE [95% CI]	MAE	MAPE	a20	Bias
Ridge	0.80	0.64 [0.42, 0.74]	5.64 [4.77, 6.75]	4.23	10.9	0.87	-0.07
SVR	0.84	0.70 [0.50, 0.80]	5.15 [4.35, 6.25]	3.83	10.0	0.88	-0.18
GPR	0.83	0.69 [0.51, 0.78]	5.25 [4.42, 6.21]	3.92	10.2	0.87	0.06
RF	0.83	0.69 [0.51, 0.79]	5.19 [4.37, 6.15]	3.85	10.1	0.88	-0.01
XGBoost	0.83	0.69 [0.52, 0.78]	5.24 [4.40, 6.16]	3.93	10.3	0.88	0.05
LightGBM	0.83	0.69 [0.53, 0.78]	5.24 [4.39, 6.16]	3.89	10.2	0.88	-0.02
ANN (single MLP)	0.83	0.68 [0.50, 0.77]	5.34 [4.51, 6.35]	3.95	10.3	0.88	0.13
ANN ensemble	0.84	0.70 [0.51, 0.80]	5.16 [4.28, 6.22]	3.75	9.9	0.88	-0.06
B2: ln-MVN (PI, e, σ′v)	0.77	0.58 [0.38, 0.67]	6.09 [5.07, 7.20]	4.52	11.5	0.85	-0.53
B3: ln-MVN (z, PI, e, σ′v)	0.76	0.58 [0.38, 0.67]	6.10 [5.09, 7.22]	4.55	11.6	0.85	-0.53
B1: Su = σ′v(a + b·PI)	0.80	0.25 [-0.33, 0.55]	8.11 [6.84, 9.76]	5.85	15.4	0.77	-1.73
B0: Su = a·σ′v	0.79	0.25 [-0.37, 0.58]	8.14 [6.67, 9.83]	5.74	15.4	0.76	-1.81
Mean (constant)	-0.52	-0.05 [-0.16, -0.02]	9.60 [7.80, 11.34]	7.57	19.6	0.61	0.03

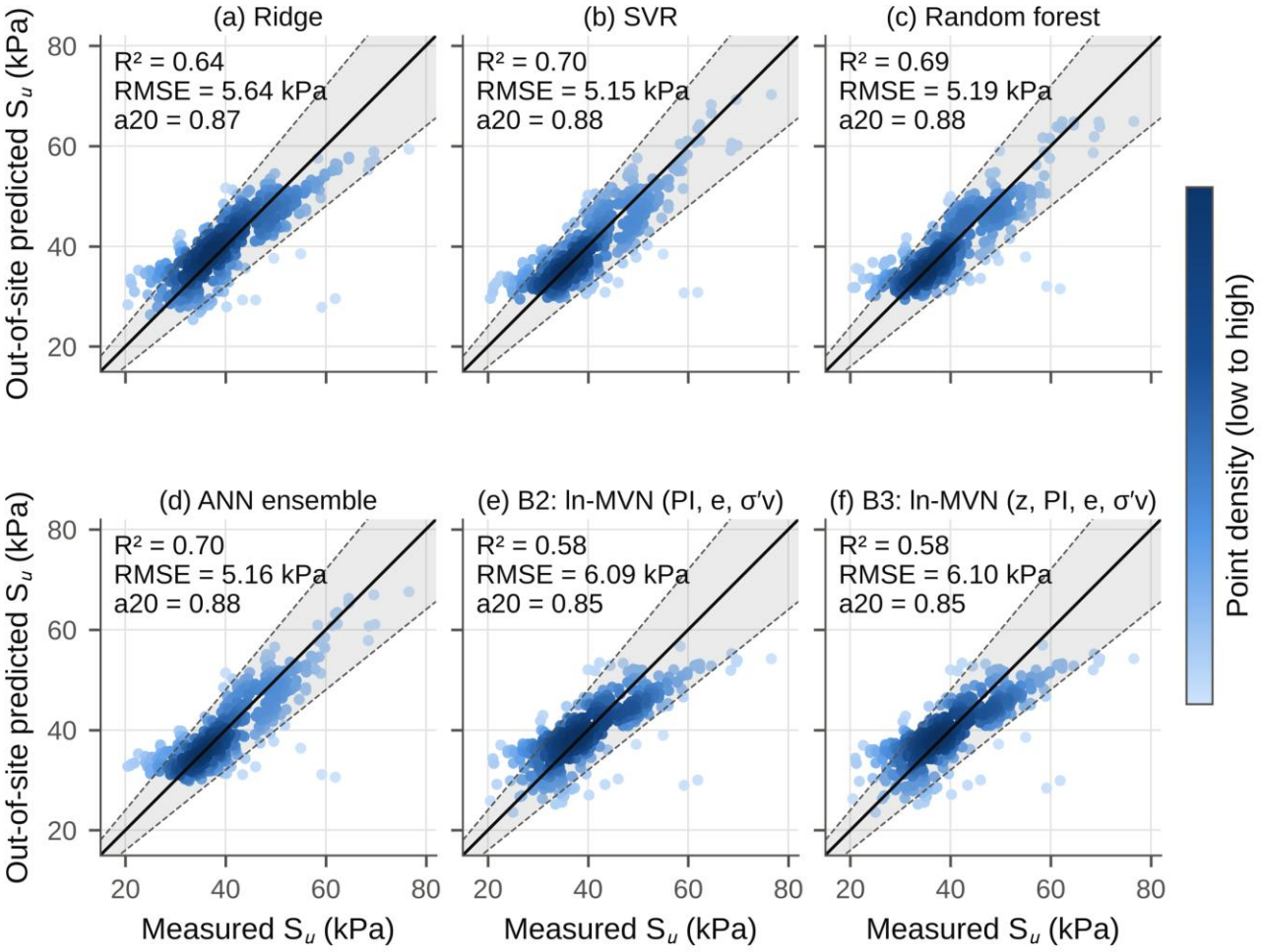


Fig. 4 Measured versus out-of-site predicted S_u for ridge regression, SVR, random forest, the ANN ensemble and the baselines B2 and B3; point colour shows the local density of points, the solid line is the 1:1 line and the dashed lines and shading mark ±20%

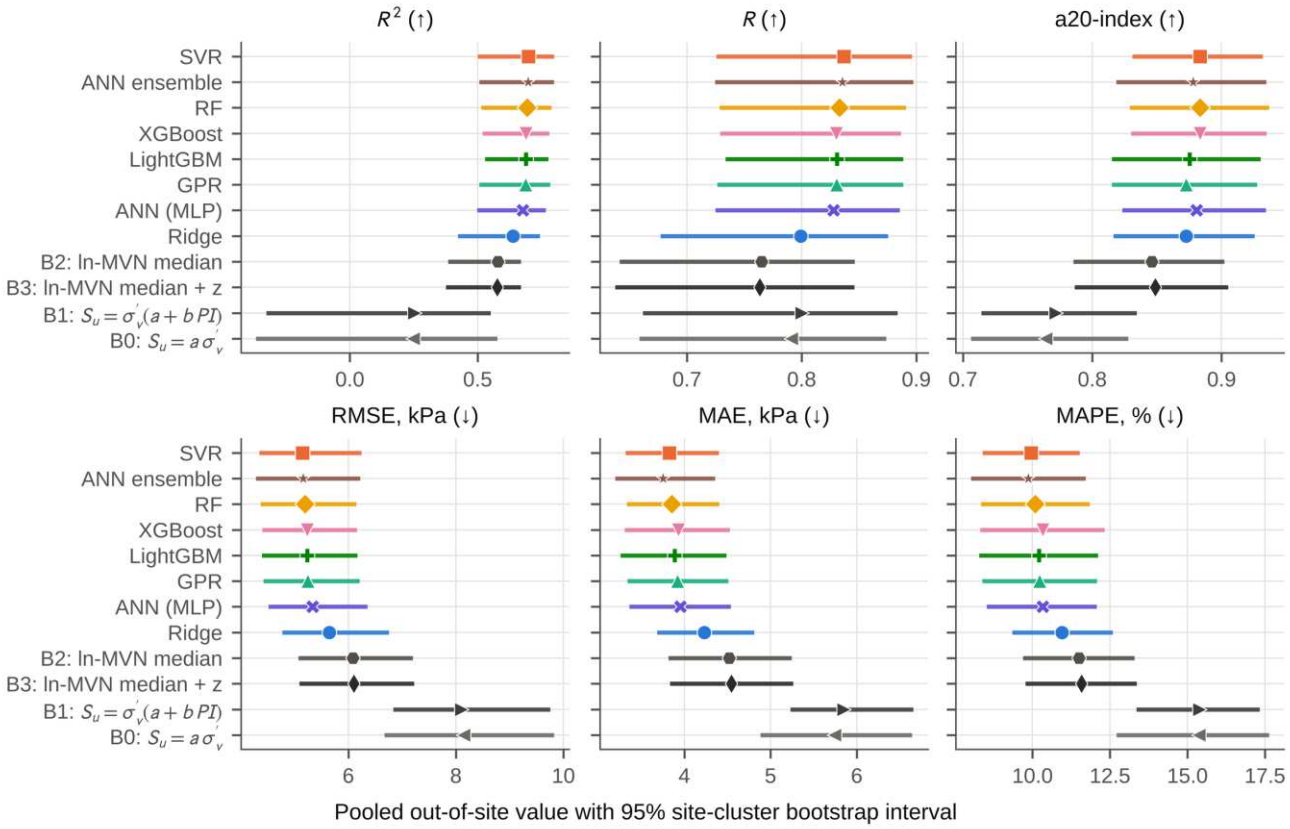


Fig. 5 Pooled out-of-site metrics with 95% site-cluster bootstrap intervals

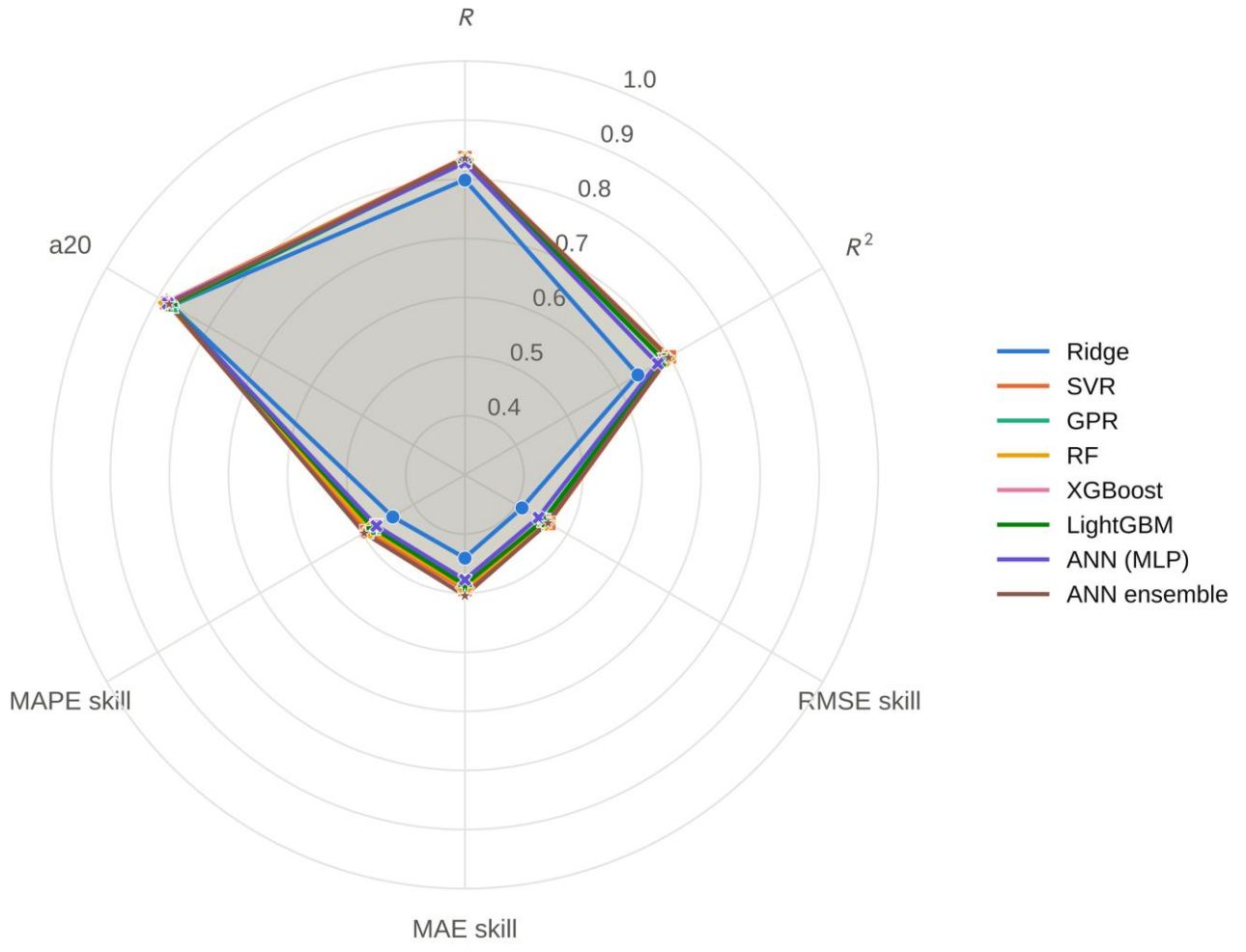


Fig. 6 Radar chart of normalised out-of-site scores (radial axis starts above zero)

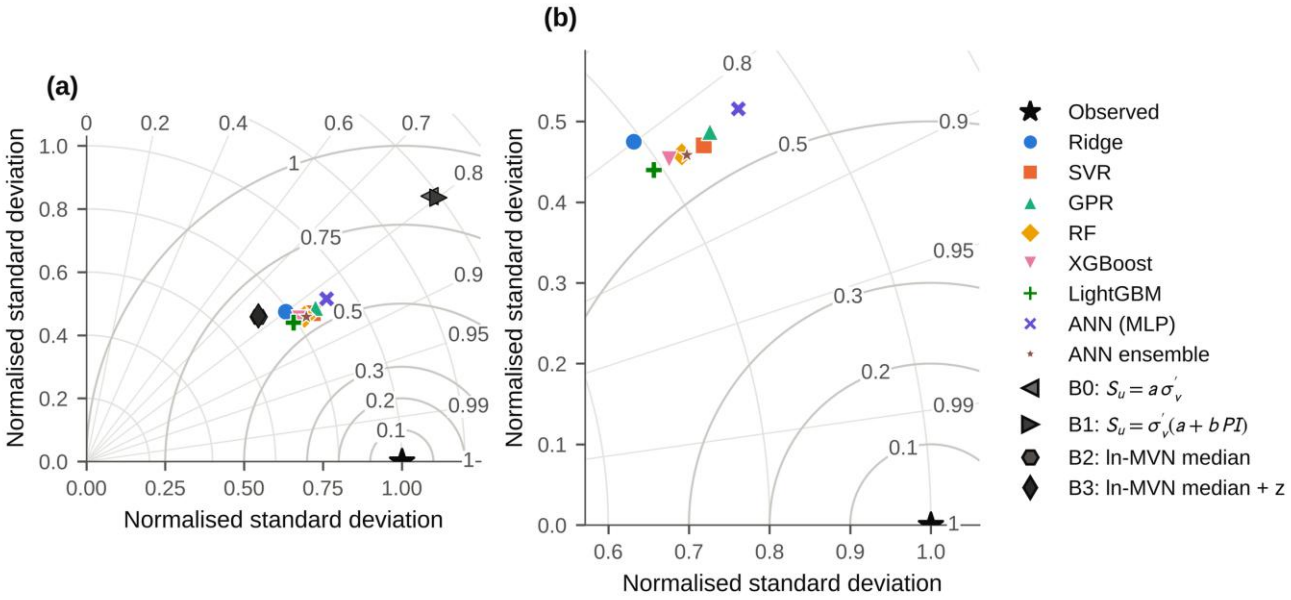


Fig. 7 Taylor diagram [22] of the out-of-site predictions

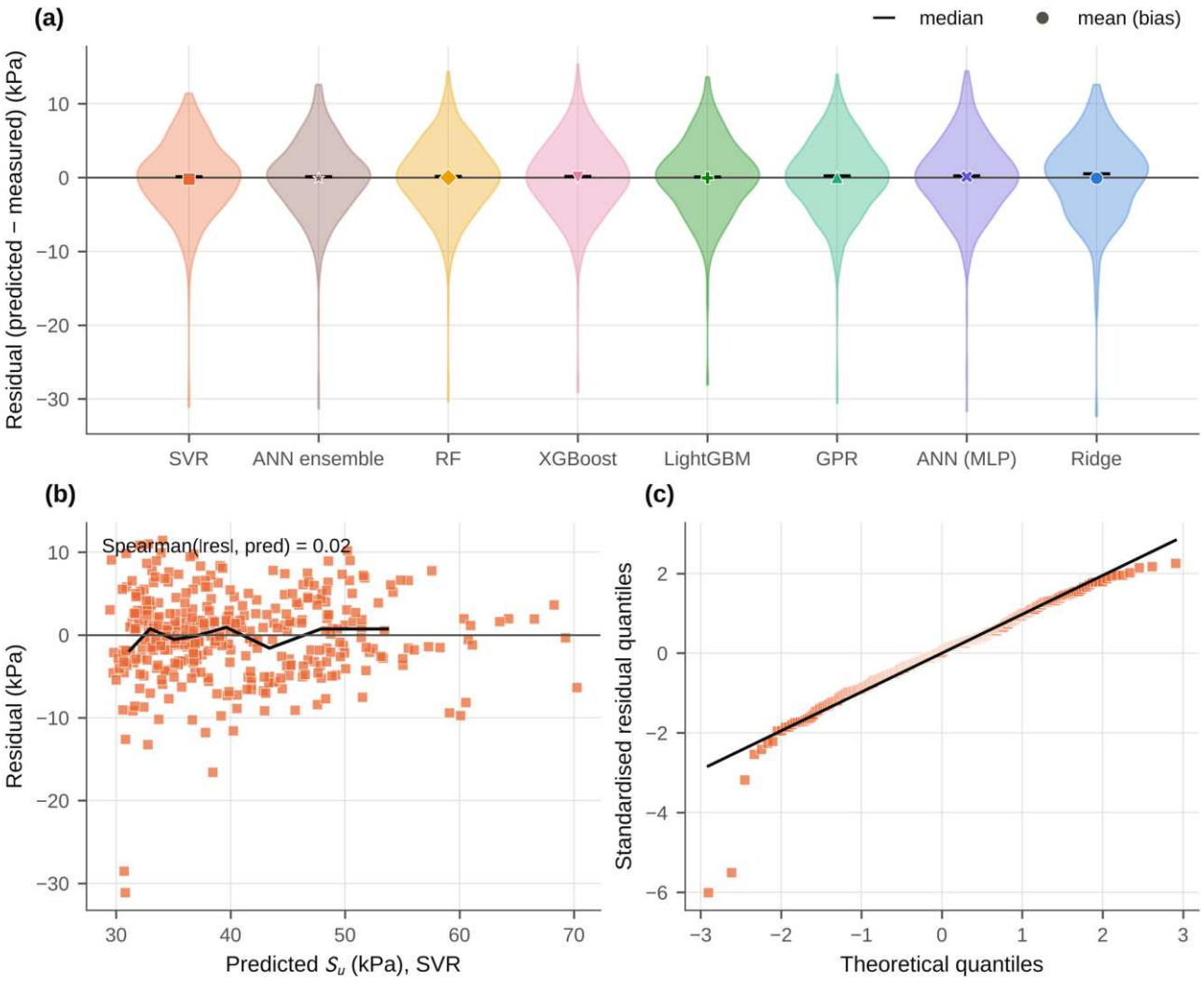


Fig. 8 Residual diagnostics of the out-of-site predictions: (a) distribution of residuals (predicted − measured) for all models; (b) residuals versus predicted S_u for SVR; (c) normal quantile–quantile plot of standardised SVR residuals

4.2 Like-for-like comparison with the ln-MVN baseline and effect of depth

Table 3 gives paired RMSE differences. With the four inputs, every ML model had a lower RMSE than B3 (differences −0.45 to −0.94 kPa), and the 95% intervals excluded zero for all eight models. With the three index inputs only, every ML model again had a lower RMSE than B2 (−0.39 to −0.92 kPa), again with intervals excluding zero. For set A the margin reached 0.94 kPa (about 15% of the baseline RMSE) for SVR and the ANN ensemble and was 0.76 to 0.90 kPa for the other nonlinear models. Relative to ridge regression, only SVR (−0.49 kPa [−0.80, −0.13]) and the ANN ensemble (−0.49 kPa [−0.83, −0.11]) had intervals that excluded zero; for GPR, RF, XGBoost, LightGBM and the single ANN the intervals included zero. Adding depth to the three index inputs changed the RMSE of the models by −0.24 to +0.18 kPa; the interval excluded zero only for the single ANN, for which depth increased the error (+0.18 kPa [+0.01, +0.41]).

Table 3 Paired differences in pooled out-of-site RMSE (kPa) with 95% site-cluster bootstrap intervals; negative values favour the first model of each comparison. Set A: z, PI, e, σ'_v; set B: PI, e, σ'_v

Model	Set A vs B3	Set B vs B2	Set A vs set B (effect of depth)	Set A vs ridge (set A)
Ridge	-0.45 [-0.83, -0.12]	-0.39 [-0.79, -0.07]	-0.04 [-0.14, +0.04]	–
SVR	-0.94 [-1.60, -0.36]	-0.68 [-1.05, -0.27]	-0.24 [-0.66, +0.05]	-0.49 [-0.80, -0.13]
GPR	-0.85 [-1.47, -0.22]	-0.87 [-1.49, -0.26]	+0.04 [-0.01, +0.08]	-0.40 [-0.79, +0.05]
RF	-0.90 [-1.55, -0.26]	-0.89 [-1.53, -0.26]	+0.01 [-0.07, +0.09]	-0.45 [-0.84, +0.01]
XGBoost	-0.86 [-1.47, -0.22]	-0.87 [-1.50, -0.23]	+0.03 [-0.05, +0.12]	-0.41 [-0.82, +0.08]
LightGBM	-0.86 [-1.43, -0.25]	-0.81 [-1.41, -0.17]	-0.04 [-0.17, +0.07]	-0.41 [-0.86, +0.05]
ANN (single MLP)	-0.76 [-1.25, -0.25]	-0.92 [-1.58, -0.33]	+0.18 [+0.01, +0.41]	-0.31 [-0.65, +0.06]
ANN ensemble	-0.94 [-1.61, -0.35]	-0.92 [-1.59, -0.37]	+0.00 [-0.10, +0.10]	-0.49 [-0.83, -0.11]

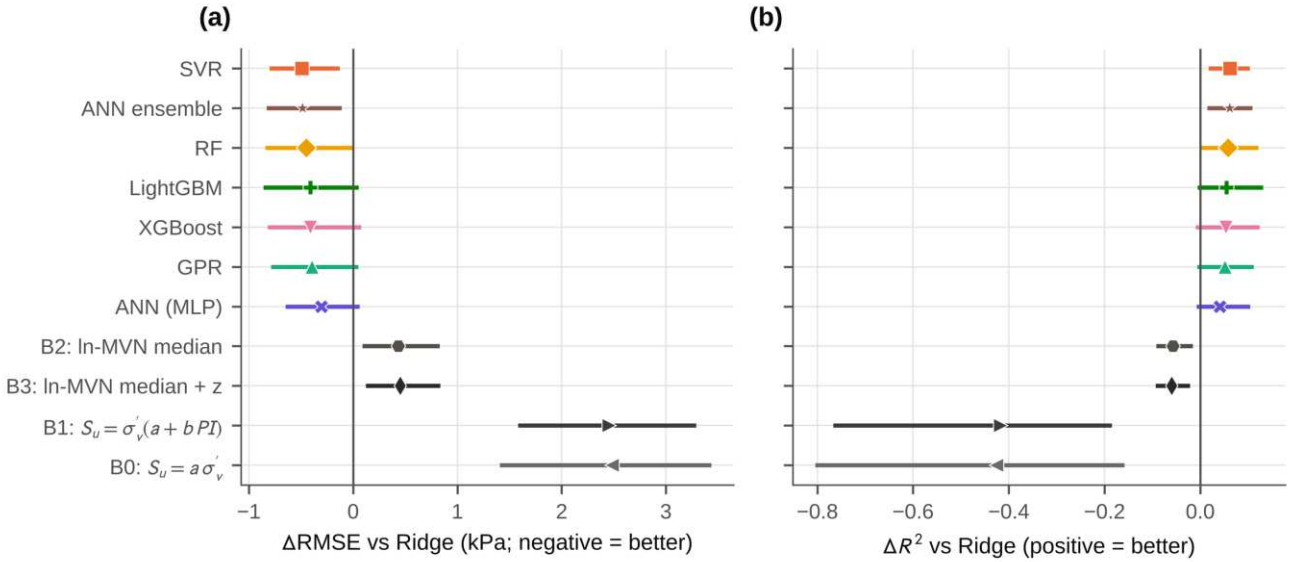


Fig. 9 Paired differences from ridge regression in pooled out-of-site RMSE (a) and R^2 (b) with 95% site-cluster bootstrap intervals (set A inputs). Intervals that do not cross zero indicate a difference supported by the resampled sites

4.3 Validation optimism

Random ten-fold cross-validation, which allows rows of the same site in training and test sets, gave R^2 values that were higher than the LOSO values by 0.013 (XGBoost) to 0.038 (single ANN) and RMSE values lower by 0.11 to 0.32 kPa for the ML models (Online Resource 1, Table S1). SVR and the ANN ensemble had the highest R^2 under both schemes, whereas the order of the remaining models changed. For the ln-MVN baselines the difference in R^2 was 0.037 to 0.039. Fig. 10 compares the two validation schemes.

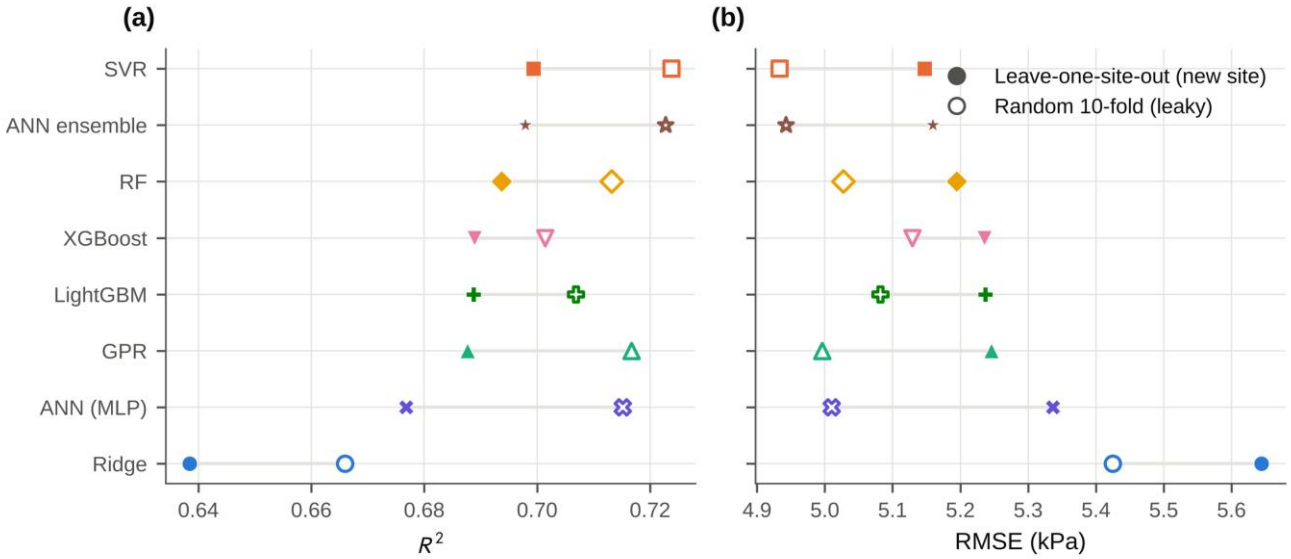


Fig. 10 Effect of the validation strategy on (a) R^2 and (b) RMSE: leave-one-site-out (filled markers, new site) versus random ten-fold (open markers, leaky diagnostic)

4.4 Site-level and regime-level errors

Errors varied strongly between sites (Fig. 11). Per-site RMSE ranged from 2.0 to 11.9 kPa for RF (median 4.7 kPa), from 2.1 to 12.3 kPa for the ANN ensemble (median 4.8 kPa) and from 2.6 to 13.1 kPa for B3 (median 5.8 kPa). RF had a lower per-site RMSE than B3 at 19 of 28 sites and the ANN ensemble at 21 of 28 sites; the sites with the largest errors had few rows (for example four rows at site 6) or, in one case, 12 rows (site 25), so per-site values are themselves uncertain (Fig. 12 gives the error of every model at every site, and Fig. 13 the residuals of the SVR model against the inputs). In terciles of measured S_u, the lowest tercile (20.5 to 35.6 kPa) was over-predicted on average (bias +2.8 kPa for RF) and the highest tercile (43.2 to 76.6 kPa) was under-predicted (−2.9 kPa); in the lowest tercile 24% (RF) and 25% (ANN ensemble) of predictions exceeded the measured value by more than 20% (27% for B3). By σ'_v tercile, the RMSE of RF was 6.27 kPa at $\sigma'_v \leq 113$ kPa and 4.51 to 4.63 kPa above it. Fig. 14 shows measured and predicted depth profiles at the six sites with the most rows: at Site 41 the predictions lie above the measurements between about 10 and 17 m, and at Site 7 both models lie below them between about 10 and 18 m. Fig. 15 gives, for six models, the share of predictions that deviate from the measured S_u by more than 20% in each tercile.

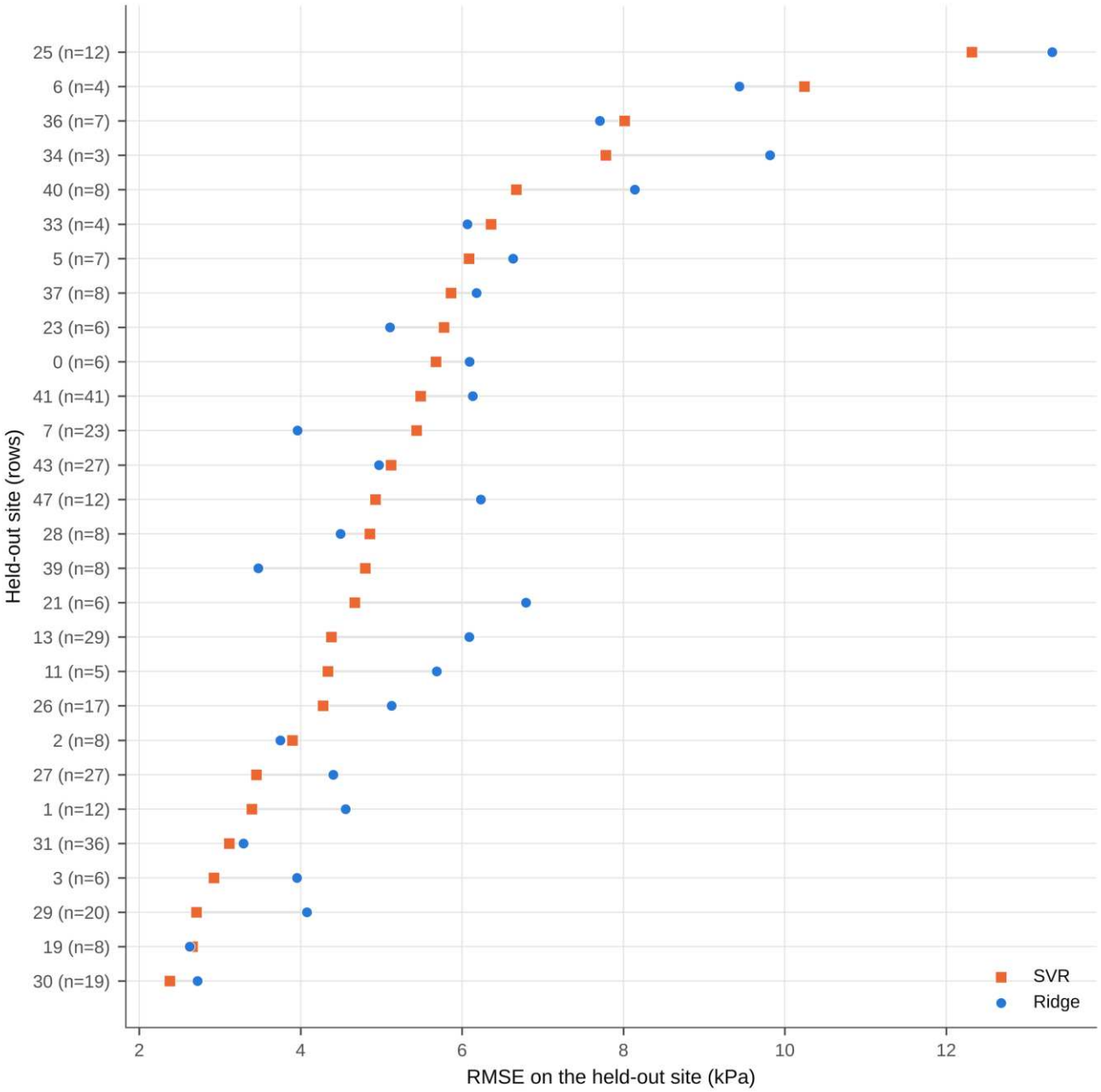


Fig. 11 RMSE on each held-out site for SVR and ridge regression (set A). Sites are ordered by the SVR error; n is the number of rows at the site

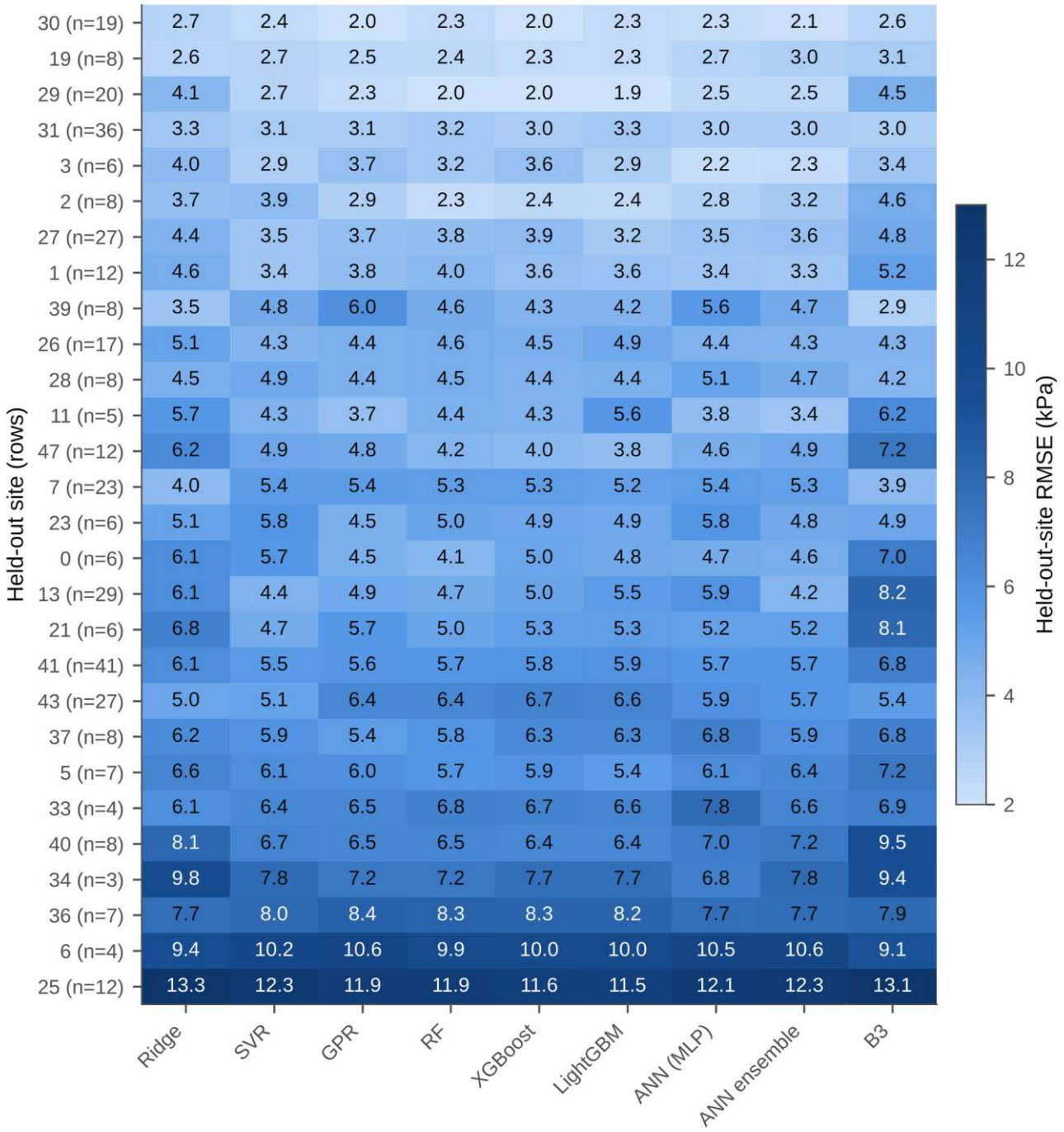


Fig. 12 Held-out-site RMSE of each model; rows are sites ordered by mean RMSE over models (number of rows in parentheses) and the numbers are printed in the cells

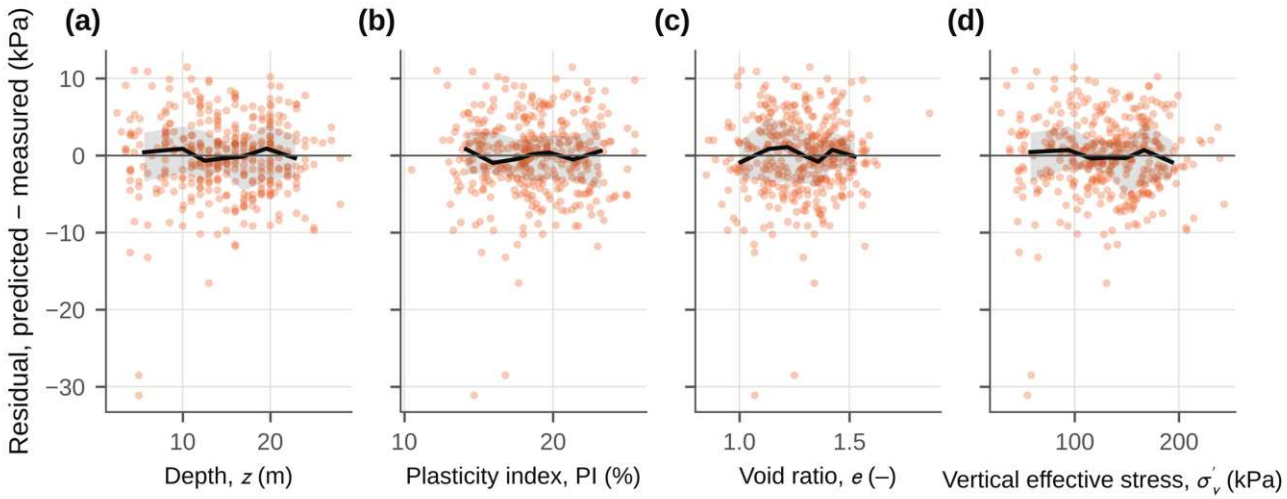


Fig. 13 Residuals (predicted − measured) of the SVR out-of-site predictions versus (a) depth, (b) plasticity index, (c) void ratio and (d) effective vertical stress; the line is the median and the band the interquartile range in seven quantile bins

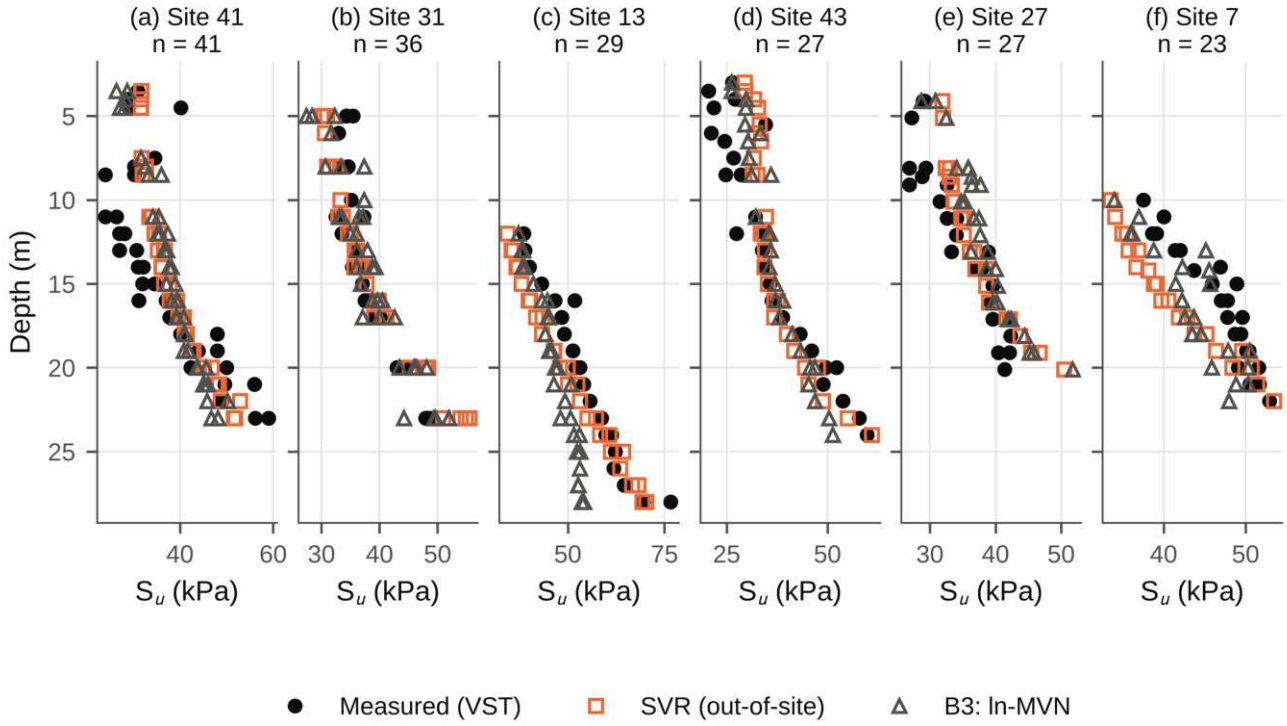


Fig. 14 Depth profiles of measured S_u (VST) and out-of-site predictions of SVR and B3 at the six sites with the most rows

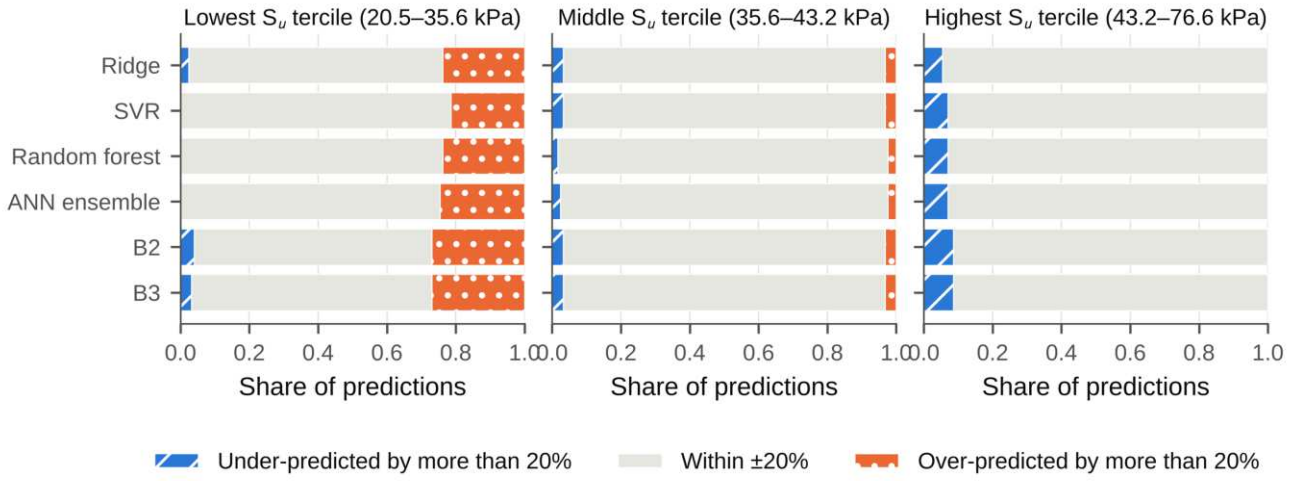


Fig. 15 Share of out-of-site predictions that are within ±20% of, or more than 20% below or above, the measured S_u in each tercile of measured S_u

4.5 Prediction intervals

Table 4 gives the conformal results. Pooled empirical coverage at the nominal 90% level was 0.875 to 0.912 across the eight models and at the nominal 95% level 0.936 to 0.960, with mean 90% interval widths of 15.5 to 16.7 kPa. Coverage was not uniform across sites: the lowest coverage of a single site at the 90% level was 0.43 for every model, and between 7% and 18% of sites (depending on the model) had coverage more than 0.2 below the nominal level. The predictive standard deviation of GPR, used as a Gaussian interval, reached 0.902 coverage at the nominal 90% level (mean width 16.2 kPa); the spread of the ANN ensemble members alone reached 0.276 (mean width 2.6 kPa). Fig. 16 shows the coverage at each site for four models. Fig. 17 shows the calibration of the predictive uncertainty of the Gaussian process model.

Table 4 Split-conformal prediction intervals under leave-one-site-out validation. Coverage and mean width pooled over all rows; the interval score uses penalty $2/\alpha$ ($\alpha = 0.1$). "Min site cov." is the lowest coverage of a single site; "Share of sites < nominal − 0.2" is the fraction of sites with coverage below 0.70 at the 90% level

Model	Cov. 90%	Width 90% (kPa)	Interval score 90%	Min site cov.	Share of sites < nominal − 0.2	Cov. 95%	Width 95% (kPa)
Ridge	0.902	16.7	24.0	0.43	0.11	0.952	21.2
SVR	0.894	15.5	22.1	0.43	0.18	0.960	19.8
GPR	0.910	16.3	22.6	0.43	0.14	0.958	20.3
RF	0.902	16.2	22.8	0.43	0.11	0.955	19.7
XGBoost	0.912	16.5	22.7	0.43	0.07	0.955	20.1
LightGBM	0.910	16.5	22.4	0.43	0.07	0.958	20.4
ANN (single MLP)	0.875	15.7	23.6	0.43	0.18	0.936	19.7
ANN ensemble	0.897	16.1	22.6	0.43	0.11	0.955	20.2

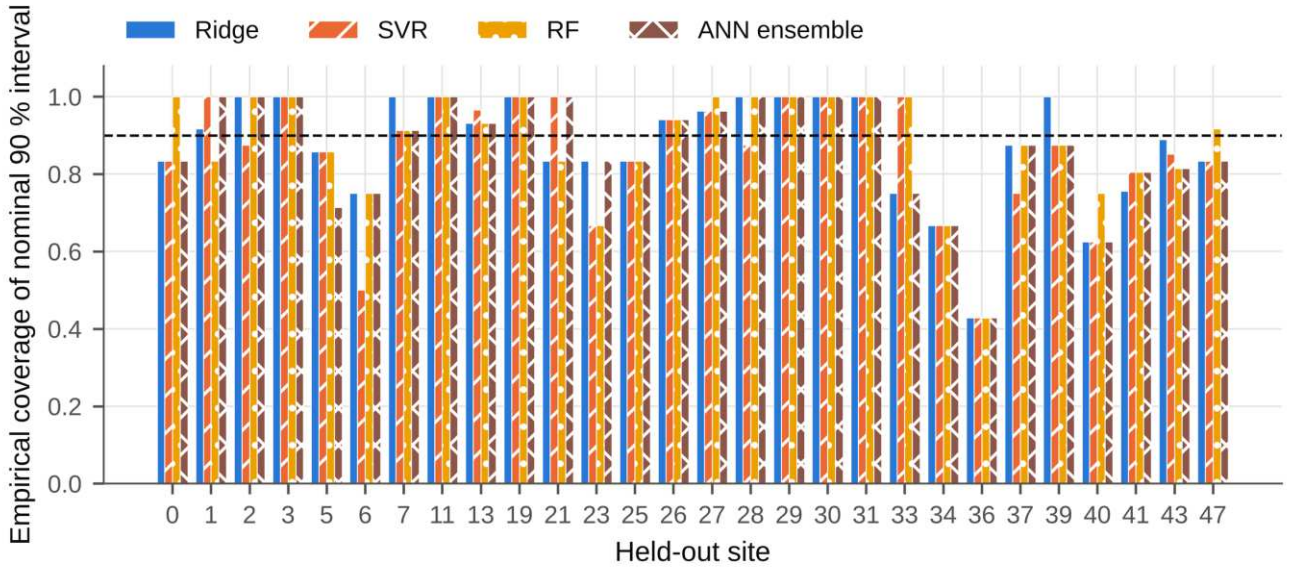


Fig. 16 Empirical coverage of the nominal 90% split-conformal interval on each held-out site for ridge regression, SVR, random forest and the ANN ensemble; the dashed line marks the nominal level

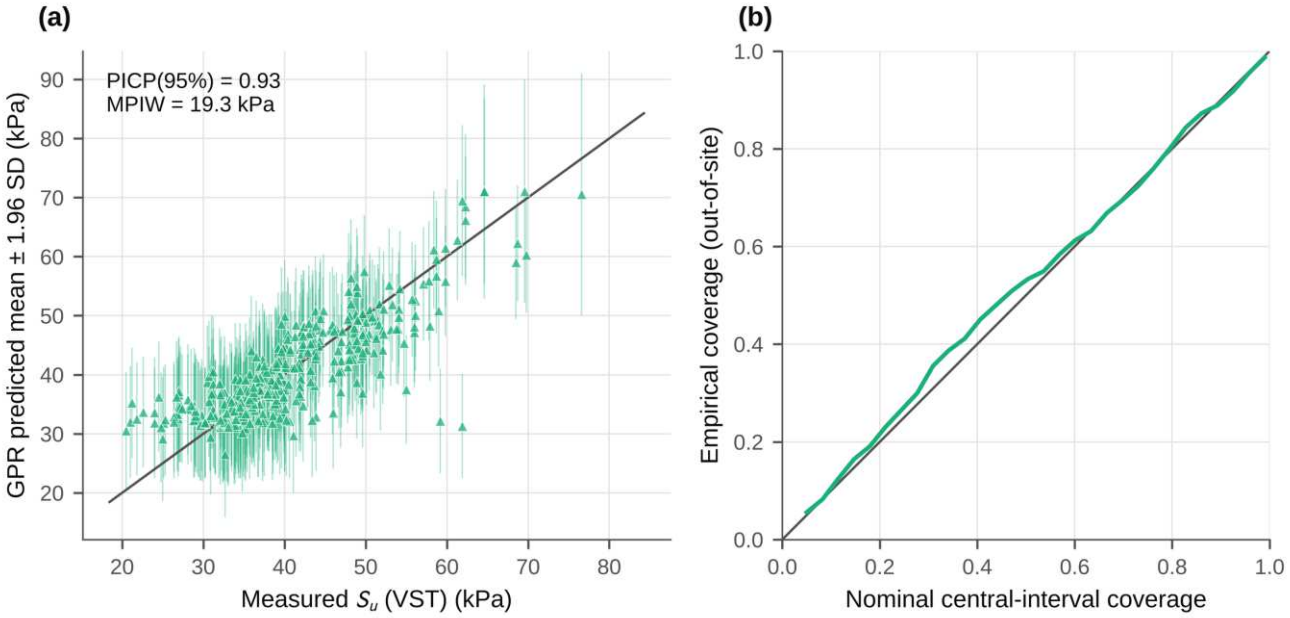


Fig. 17 GPR uncertainty under leave-one-site-out validation: (a) measured versus predicted mean ± 1.96 predictive standard deviations; (b) empirical versus nominal coverage of central intervals

4.6 Input importance and fitted behaviour

Out-of-site permutation importance (Fig. 18) ranked σ'_v first for seven of the eight models (RMSE increase 3.33 to 5.91 kPa; 1.55 kPa for ridge regression, for which depth ranked first with 2.30 kPa); depth was generally second (for example 1.83 kPa for SVR and 1.24 kPa for RF). The increase for PI was close to zero for ridge regression, GPR, XGBoost and LightGBM (−0.03 to 0.05 kPa) and 0.15 to 0.61 kPa for the other models; the increase for e ranged from 0.02 to 0.86 kPa. Because depth and σ'_v are collinear, the split of importance between them depends on the model (for GPR, 5.91 kPa for σ'_v and 0.22 kPa for depth). SHAP values of the SVR model fitted to all data (Fig. 19) gave the same ordering of mean absolute contributions (Fig. 20): σ'_v (3.57 kPa), depth (2.08 kPa), void ratio (1.14 kPa) and PI (0.88 kPa). The dependence plots (Fig. 21) and the SHAP heatmap (Fig. 22) show how the contribution of each input varies between rows. The partial dependence of SVR on σ'_v was flat to slightly decreasing below about 80 kPa and increased above it (Fig. 23); the dependence on PI was weak and that on e was weakly decreasing. The one-at-a-time sensitivity of the SVR prediction to changes of ±10% and

±20% in each input is shown in Fig. 24. In the one-at-a-time analysis of SVR, ±20% in σ'_v changed the mean prediction by −7.0% and +7.7%, and the coupled change of z and σ'_v by −8.7% and +9.5%.

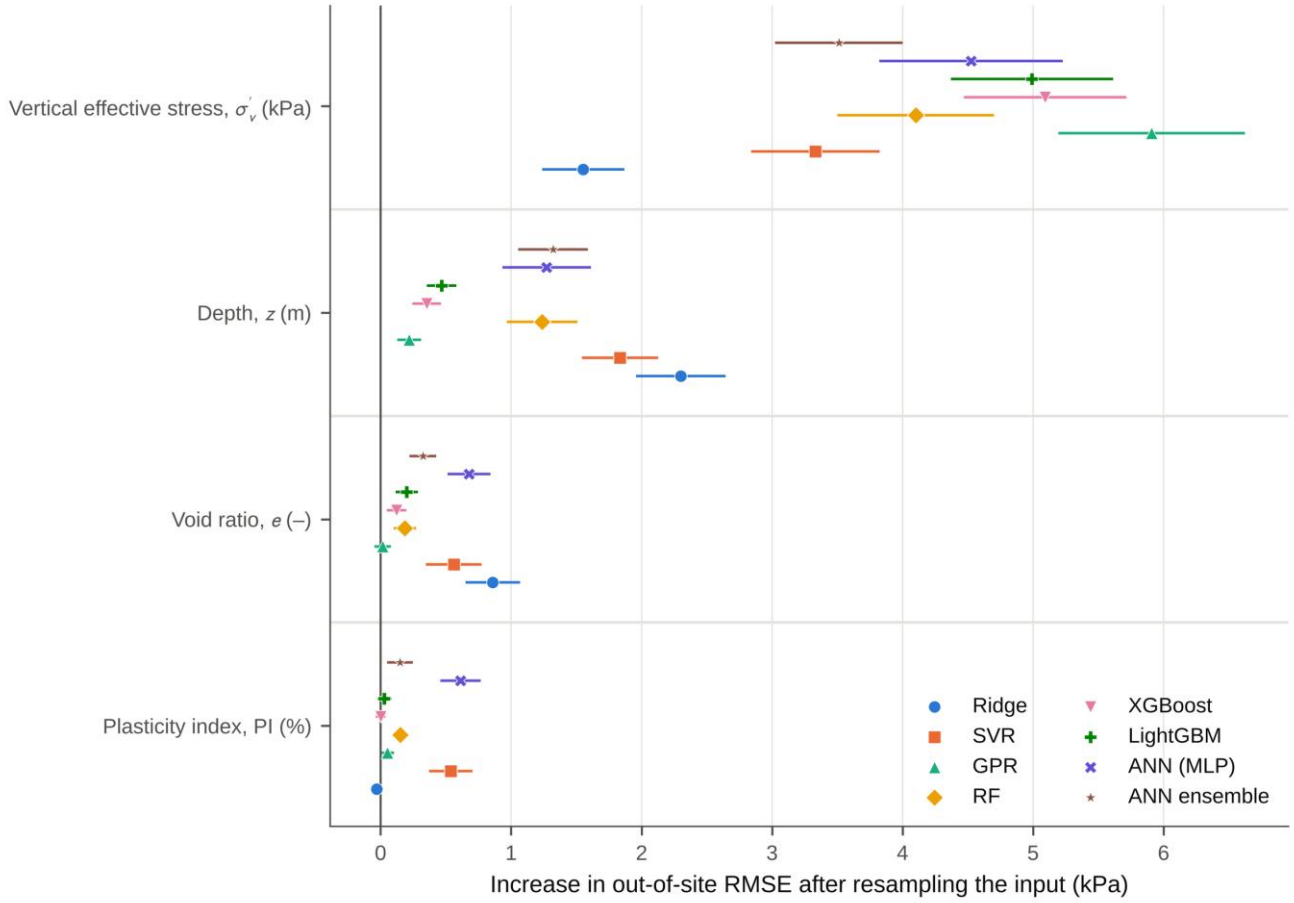


Fig. 18 Out-of-site permutation importance of the inputs for all models (set A): increase in out-of-site RMSE after resampling one input in the held-out site. Horizontal bars show the variation over repeats

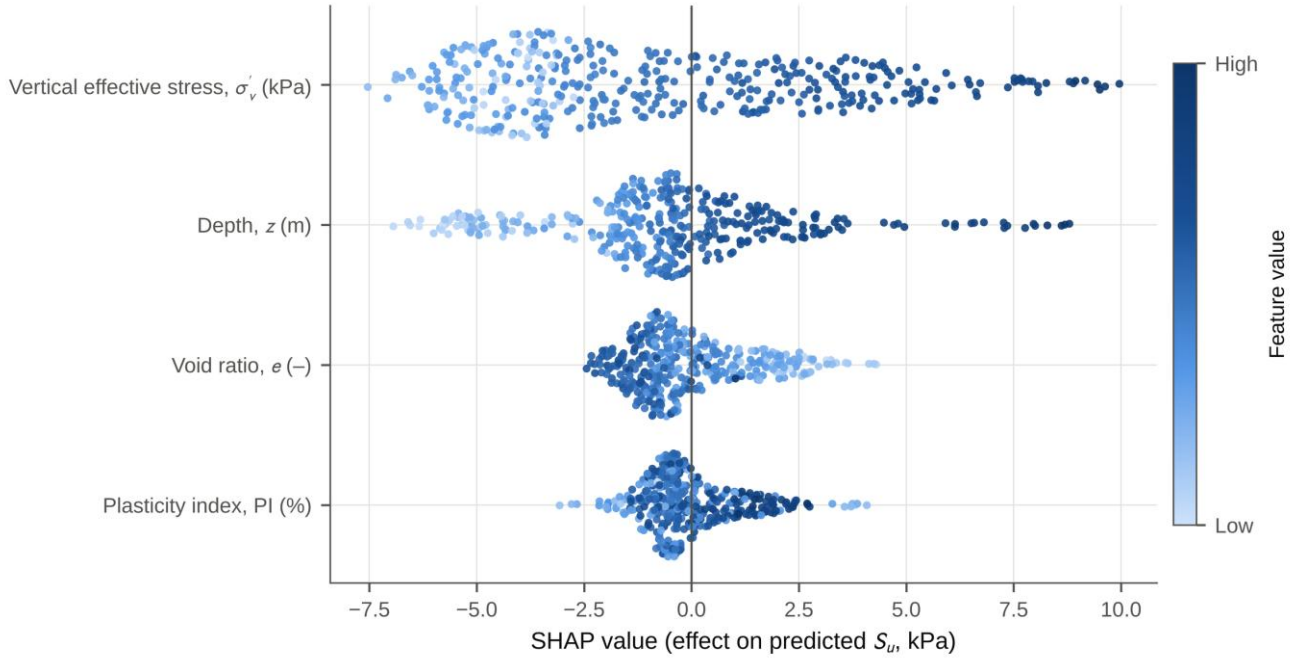


Fig. 19 SHAP values of the SVR model fitted to all data (beeswarm); each point is one row, coloured by the value of the input

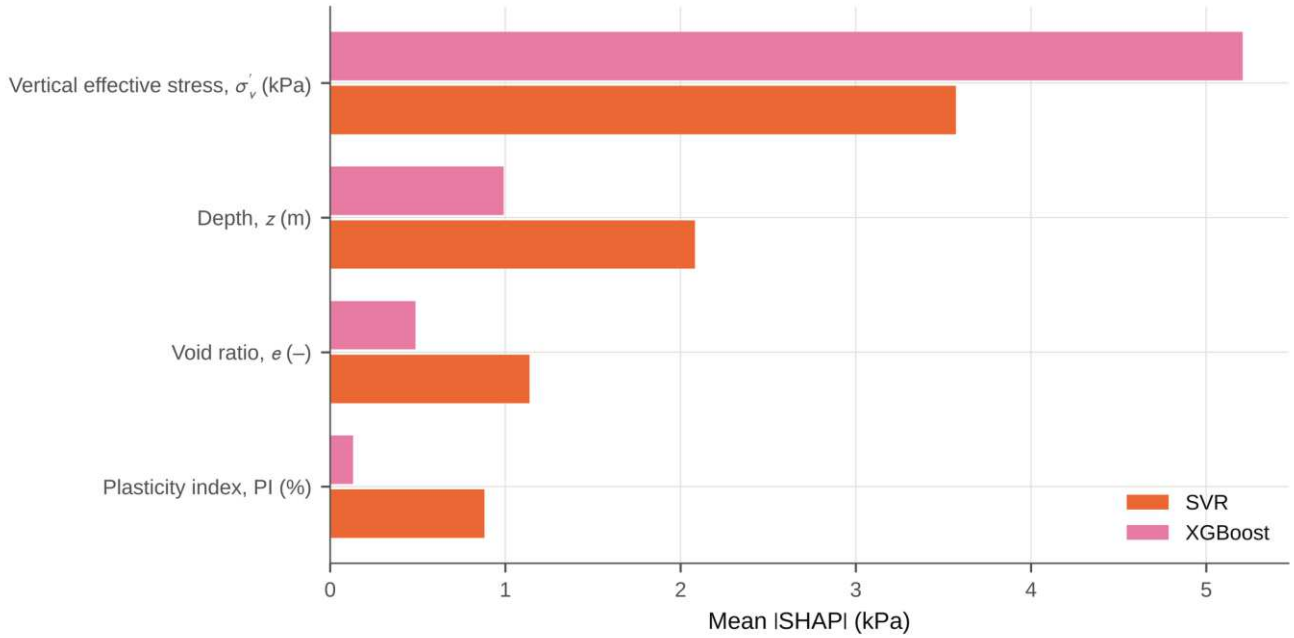


Fig. 20 Mean absolute SHAP values (SVR, fitted to all data)

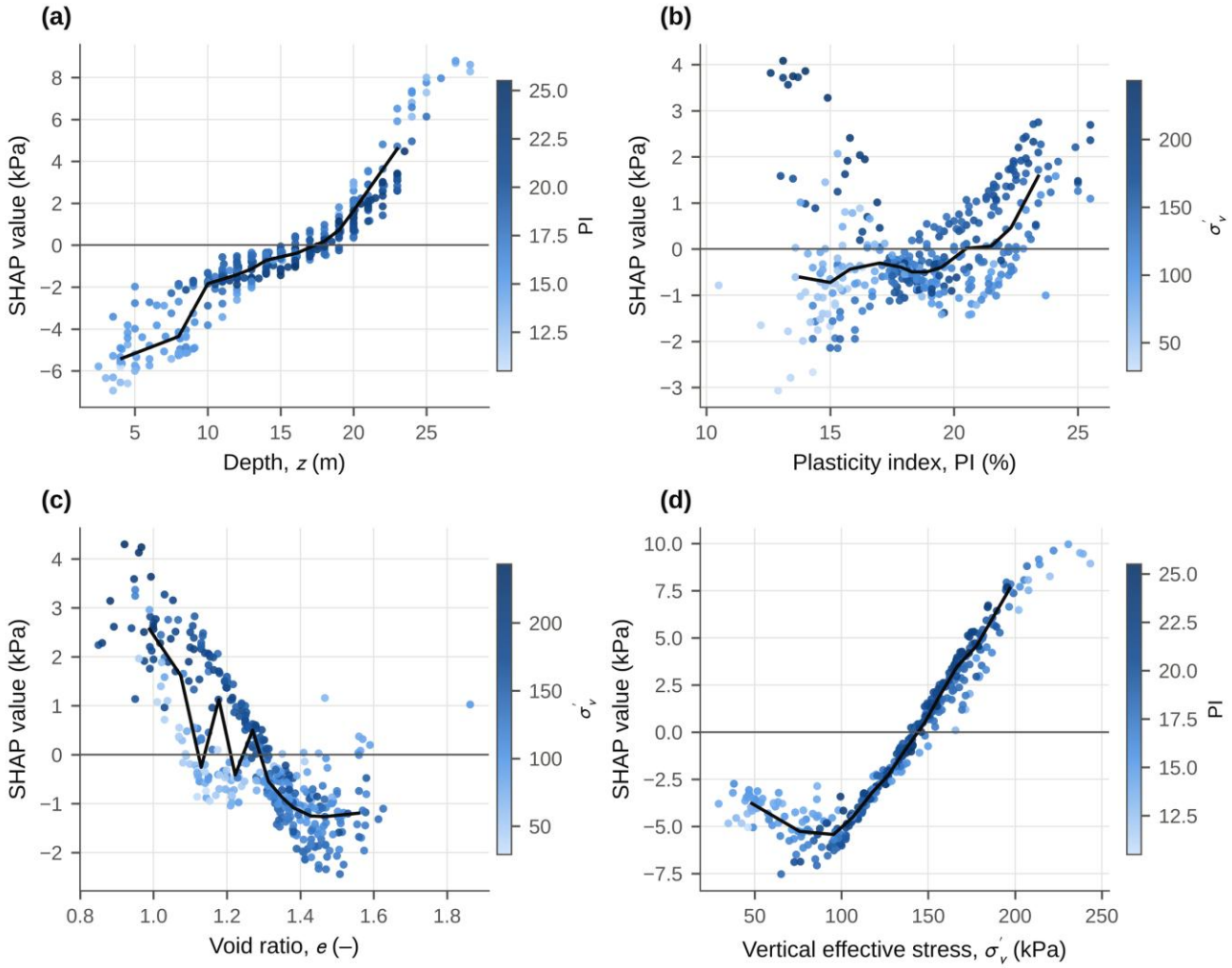


Fig. 21 SHAP dependence plots (SVR, fitted to all data)

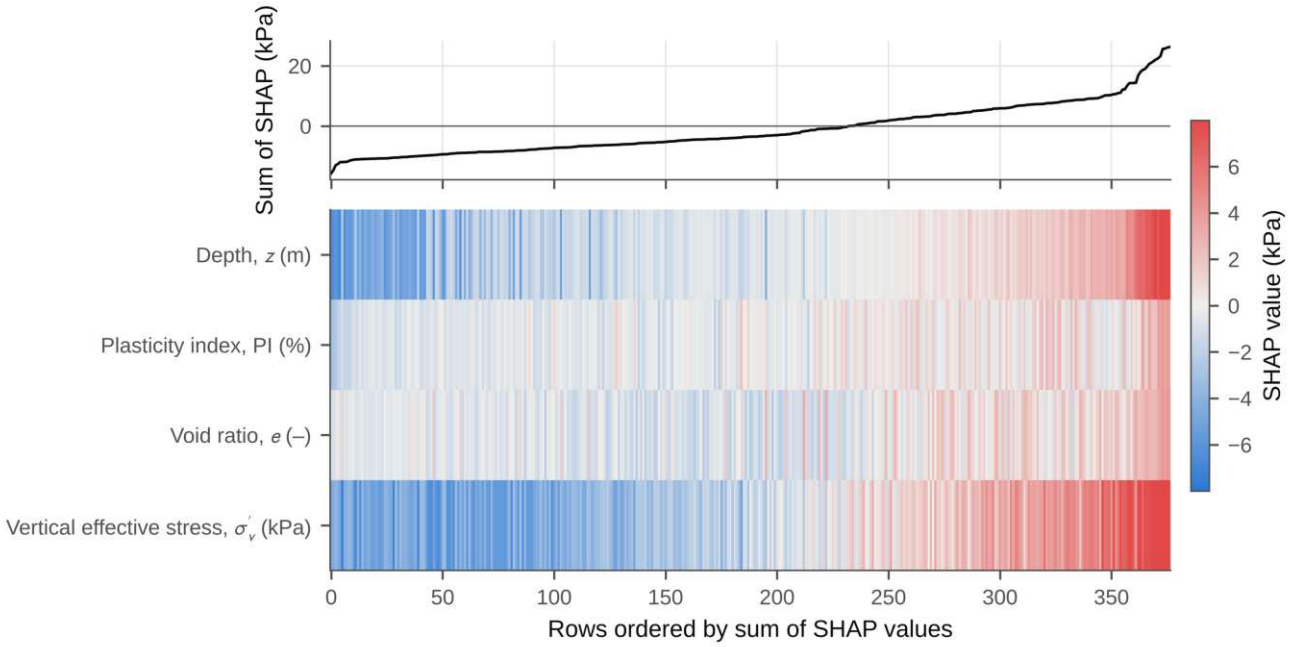


Fig. 22 SHAP heatmap of the SVR model fitted to all data: rows of the data ordered by the sum of their SHAP values (top: the sum; below: the contribution of each input)

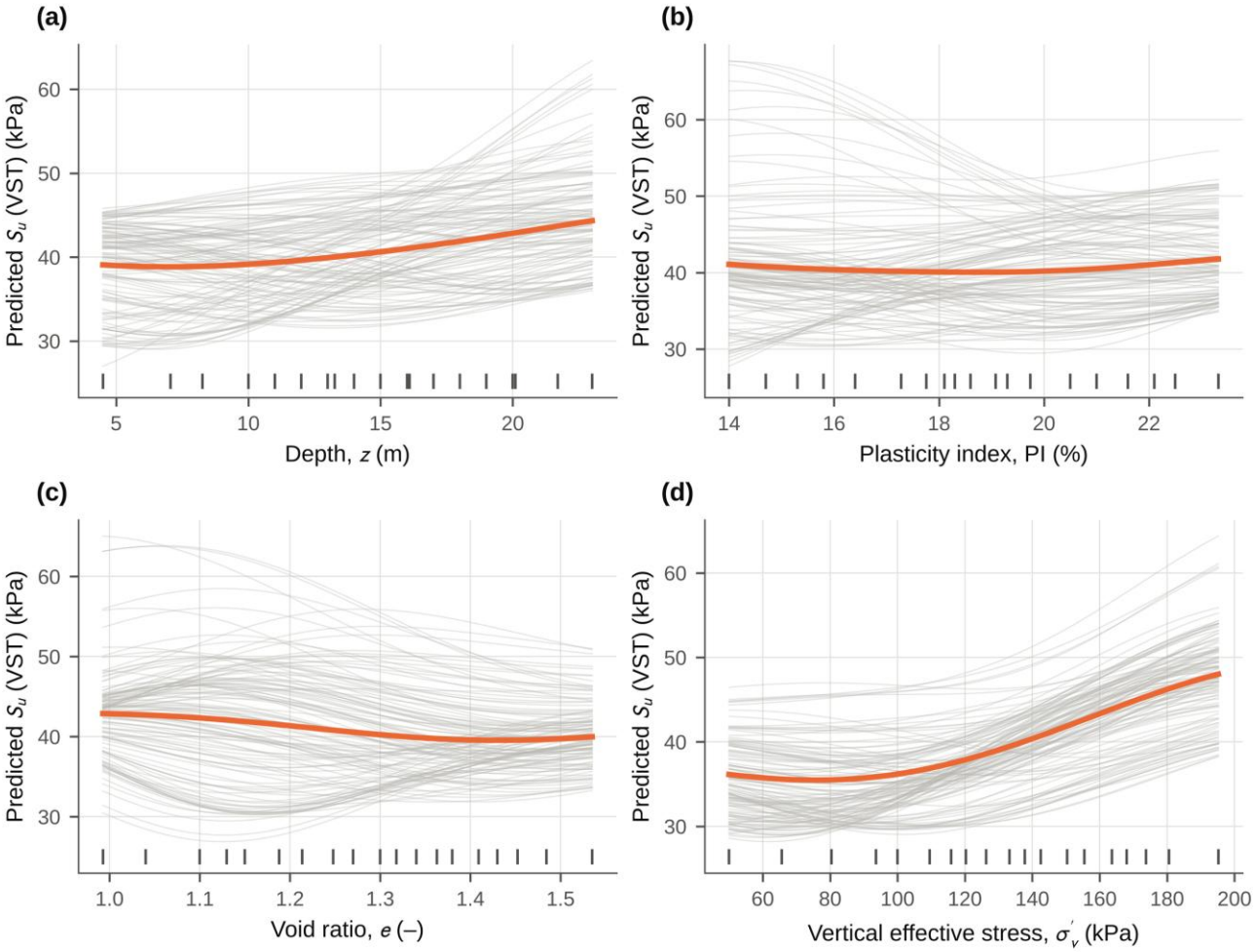


Fig. 23 Individual conditional expectation curves (grey) and partial dependence (bold) of the SVR model fitted to all data, between the 5th and 95th percentiles of each input: (a) depth, (b) plasticity index, (c) void ratio, (d) effective vertical stress. Ticks on the abscissa mark deciles of the data

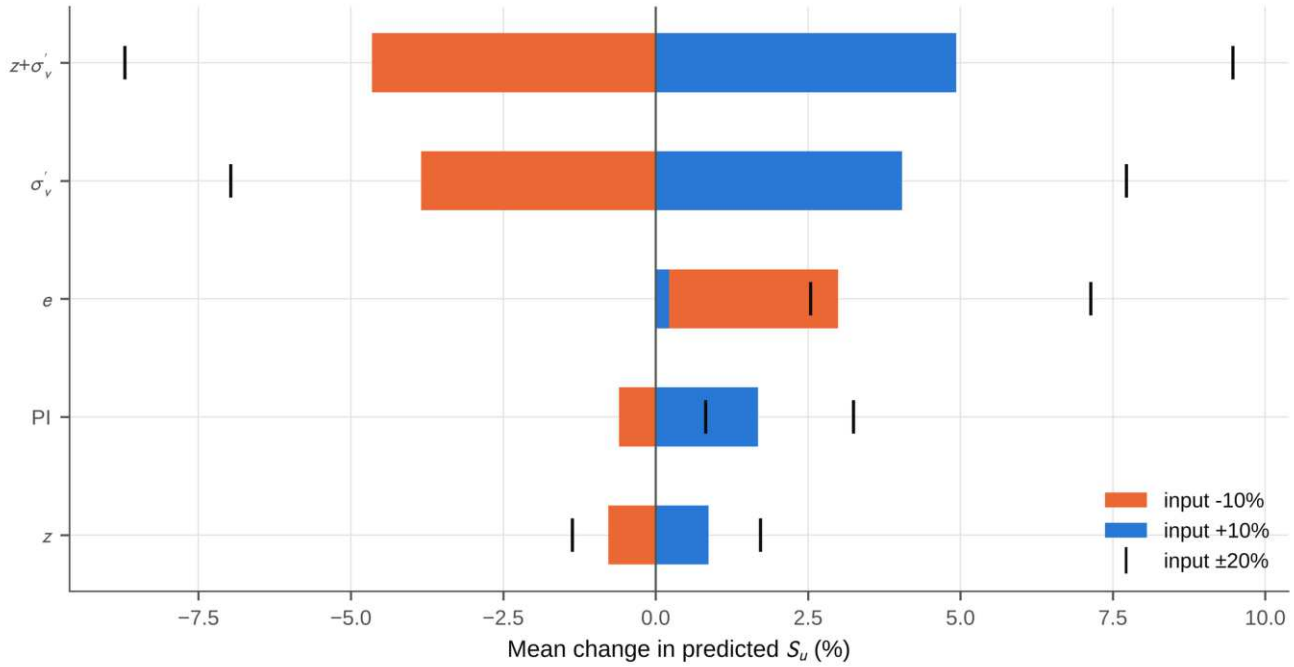


Fig. 24 One-at-a-time sensitivity of the SVR prediction: mean relative change for ±10% and ±20% change of each input, and of depth with σ'_v together

4.7 Physical admissibility checks

All predictions of all models fitted to all data were positive (minimum 25.95 to 31.24 kPa). The fraction of ICE curves that were non-decreasing in σ'_v was 1.00 for ridge regression, XGBoost, LightGBM, the single ANN and the ANN ensemble, 0.23 for SVR, 0.02 for GPR and 0.00 for RF. For depth the fractions were 1.00 for ridge regression, GPR and XGBoost, 0.85 for the ANN ensemble, 0.49 for the single ANN, 0.33 for SVR, 0.03 for RF and 0.00 for LightGBM. No monotonicity constraint was imposed on any model in the main comparison; constrained variants are examined in Section 4.9.

4.8 Dependence on the number of training sites

With fixed, untuned hyperparameters, the pooled out-of-site RMSE decreased as more sites were used for training; the decrease slowed beyond about 20 sites for SVR (5.33 kPa at 20 sites, 5.29 at 24 and 5.28 at 27) but continued for random forest (5.38, 5.29 and 5.25 kPa) (Fig. 25). With four training sites, ridge regression (mean RMSE 6.09 kPa) was better than SVR (6.37 kPa) and random forest (6.42 kPa). SVR was better than ridge regression from eight sites onward and random forest from twelve sites onward, and with all 27 remaining sites the RMSE was 5.67 kPa for ridge regression, 5.28 kPa for SVR and 5.25 kPa for random forest. Between 16 and 27 sites the mean RMSE of SVR decreased from 5.47 to 5.28 kPa and that of random forest from 5.58 to 5.25 kPa, whereas that of ridge regression decreased from 5.80 to 5.67 kPa. The values at 27 sites are higher than the tuned results of Table 2 (for example 5.28 against 5.15 kPa for SVR) because the hyperparameters were not tuned.

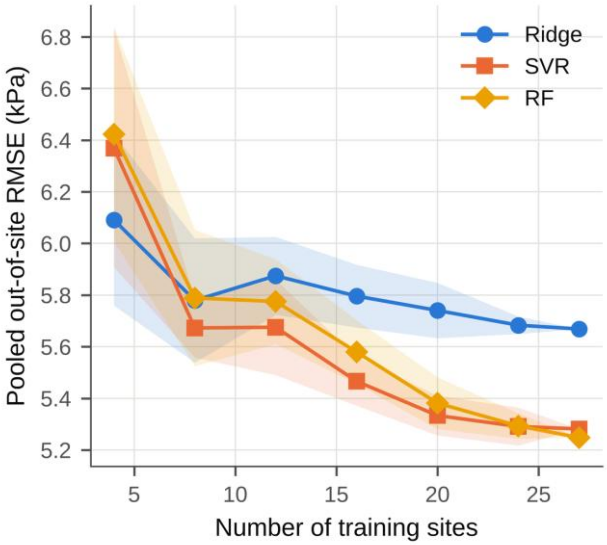


Fig. 25 Pooled out-of-site RMSE versus the number of training sites for ridge regression, SVR and random forest with fixed, untuned hyperparameters. Lines are means over eight random draws of training sites; bands are ± one standard deviation over the draws

4.9 Mixed-effects baseline, physics-guided variants and calibration with local measurements

The linear mixed model with a random site intercept (B4) gave a pooled out-of-site RMSE of 5.62 kPa [4.72, 6.71] (R^2 = 0.641) on input set B, practically equal to ridge regression (paired difference −0.07 kPa [−0.17, +0.03]) and 0.45 kPa lower than the ln-MVN baseline B2 (paired difference −0.45 kPa [−0.95, −0.11]). The random-slope variant gave 5.63 kPa, and the logarithmic variant 6.04 kPa, which is 0.05 kPa below B2 (interval [−0.14, +0.04]) and above ridge regression. Relative to B4, the ANN ensemble was lower by 0.47 kPa [−0.76, −0.14] and the random forest by 0.44 kPa [−0.83, −0.01], whereas SVR was not distinguishable (−0.23 kPa [−0.62, +0.24]). The fitted model attributes a between-site variance of 6.8 kPa^2 and a within-site variance of 23.9 kPa^2 (conditional intraclass correlation 0.22) to the data after the fixed effects (Fig. 26a). Using m rows of the new site to predict its site effect lowered the RMSE on the remaining rows from 5.57 to 5.53 kPa (m = 1), from 5.51 to 5.37 kPa (m = 2), from 5.44 to 5.21 kPa (m = 3) and from 5.37 to 5.01 kPa (m = 5).

On input set B, the pooled out-of-site RMSE of the monotone XGBoost and LightGBM models was 5.17 and 5.23 kPa, compared with 5.21 and 5.27 kPa for the unconstrained models (paired differences −0.04 kPa [−0.13, +0.03] and −0.05 kPa [−0.18, +0.06]). The perceptron ensemble with the monotonicity penalty (λ = 10) had RMSE 5.171 kPa and without it (λ = 0) 5.169 kPa; the paired difference was +0.002 kPa [−0.004, +0.008]. The unconstrained ensemble already produced non-decreasing ICE curves in σ'_v for all rows examined, and so did the monotone boosting models (fraction 1.00 for all four), so the constraint was effectively inactive. The strength-ratio target did not help: RMSE was 5.29 kPa for XGBoost, 5.75 kPa for SVR and 7.73 kPa for ridge regression, and the ICE curves of the ratio models were non-decreasing in σ'_v for a fraction of 0.00 (XGBoost), 0.63 (SVR) and 0.00 (ridge regression) of rows. All physics-guided variants except the ridge ratio model had lower RMSE than the ln-MVN baseline B2 (6.09 kPa), with paired intervals excluding zero for all except the SVR ratio model (−0.33 kPa [−0.70, +0.04]).

The out-of-site residuals showed a modest between-site component: the intraclass correlation was 0.14 for SVR (τ^2 = 4.1 kPa^2, σ^2 = 25.3 kPa^2) and 0.24 for ridge regression (τ^2 = 7.9 kPa^2, σ^2 = 25.0 kPa^2). Using m local measurements to estimate a site offset lowered the RMSE on the remaining rows of the site (Fig. 26b). For SVR the reduction relative to the uncorrected predictions on the same rows was negligible at m = 1 (5.34 against 5.33 kPa), 0.03 kPa at m = 2, 0.07 kPa at m = 3 and 0.11 kPa at m = 5 (5.01 against 5.12 kPa); for ridge regression the corresponding reductions were 0.05, 0.17, 0.27 and 0.41 kPa (5.03 against 5.44 kPa at m = 5). After correction with five measurements, ridge regression and SVR had nearly the same RMSE.

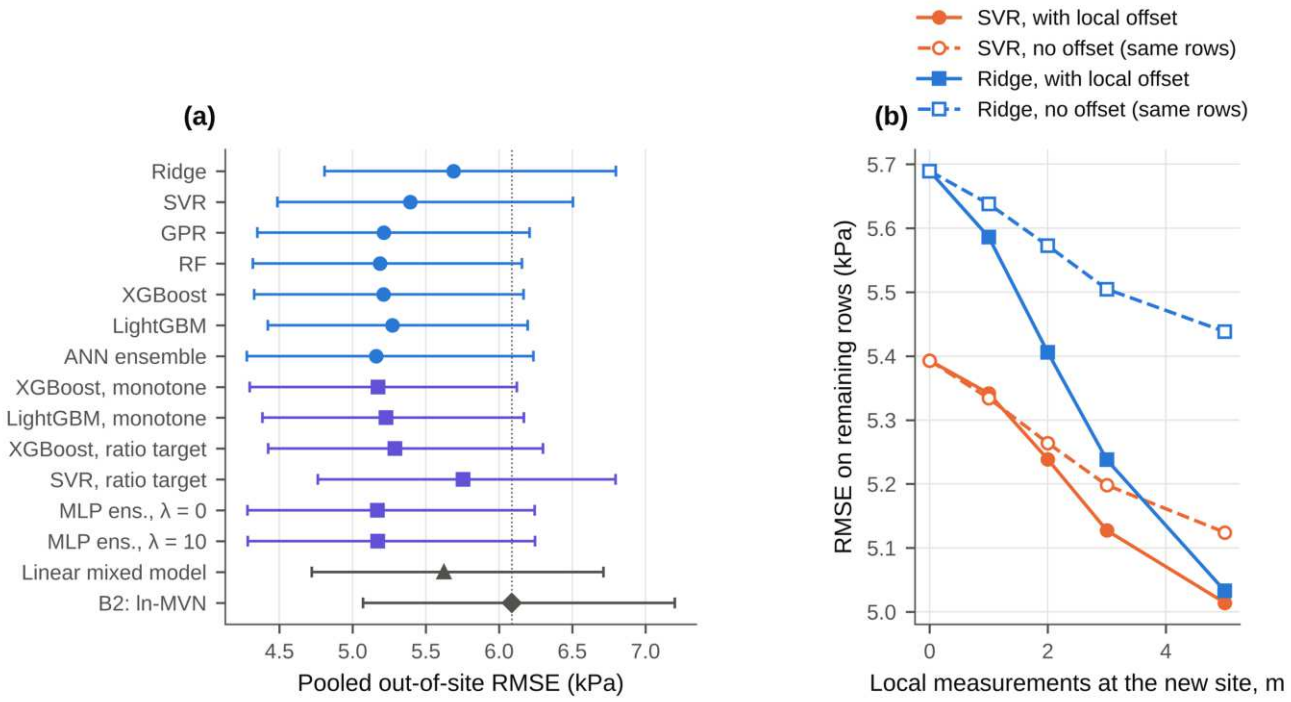


Fig. 26 (a) Pooled out-of-site RMSE with 95% site-cluster bootstrap intervals on input set B: standard models (circles), physics-guided variants (squares), linear mixed model B4 (triangle) and the ln-MVN baseline B2 (diamond; dotted line); (b) RMSE on the remaining rows of a new site after an empirical-Bayes offset estimated from m local measurements (filled markers) and without the offset on the same rows (open markers)

5 Discussion

5.1 Machine learning versus the ln-MVN baseline

On identical inputs, ML models reduced the pooled out-of-site RMSE by 0.4 to 0.9 kPa relative to a log-linear (ln-MVN) baseline, and the paired intervals excluded zero in both input configurations. This supports the statement that, for this database and under LOSO validation, the baseline's log-linear structure leaves some systematic structure unexplained. The data do not identify the reason. Candidate explanations are nonlinear or non-lognormal relations between the inputs and S_u and heteroscedasticity; none was tested here. The margin is modest in absolute terms: the standard deviation of S_u in the subset is 9.4 kPa, and even the best RMSE (5.15 kPa) is more than half of it. The result also holds only against a simplified ln-MVN model without the Johnson-system marginal transformations of the published approach [1, 2]; the comparison with the full published model remains to be done.

The mixed-effects benchmark (Section 4.9) sharpens this reading. A linear mixed model on the original scale performs like ridge regression, and the gap to the ln-MVN baseline is therefore at least partly a gap between a model on the original scale and a log-linear model, not necessarily a gain from nonlinear flexibility. Only the ANN ensemble and the random forest were separated from the mixed model by the bootstrap intervals (the random forest marginally), and SVR was not. The benefit of nonlinear ML over a properly specified linear grouped-data model is thus about 0.4 to 0.5 kPa at best in this database, smaller than the 0.4 to 0.9 kPa obtained against the ln-MVN baseline.

5.2 Model ranking is not resolved

The nonlinear models lie within 0.19 kPa of RMSE of each other, and their bootstrap intervals overlap almost completely; only SVR and the ANN ensemble separate from ridge regression in the paired comparison. The data therefore do not support naming a best algorithm. The finding that the single ANN was ranked lowest among the nonlinear models, whereas the ensemble of the same architecture ranked among the highest, suggests that the

variance of a single network fitted on 300-odd rows contributes to the difference; this is consistent with, but not proven by, our results. The choice among the nonlinear models may reasonably rest on other criteria, such as the availability of an uncertainty estimate (GPR), the ability to impose monotonicity (Section 4.9 shows that doing so did not change accuracy), or computational cost.

5.3 Depth, effective stress and collinearity

Depth did not improve any model on average, and for the single ANN it increased the error. Since σ'_v is derived from depth, the two carry nearly the same information, and a model with both has more freedom to fit site-specific noise. The permutation and SHAP results should therefore not be read as evidence that depth has an independent physical effect: permuting one of two collinear inputs creates input combinations that do not occur in the field. The dominance of σ'_v is consistent with the established use of the ratio S_u/σ'_v as a normalised strength measure; Zhang et al. [2] reported a negative relation between PI and S_u/σ'_v from the vane test in Shanghai clay, whereas in our models PI contributed little once σ'_v and e were available. We did not test the S_u/σ'_v formulation as a target.

5.4 Validation optimism

Random cross-validation was optimistic compared with LOSO, but only slightly (ΔR^2 of 0.013 to 0.038), and SVR and the ANN ensemble led under both schemes. This is smaller than the strong validation dependence described in the multi-site study of Taherdangkoo et al. [12], which used a different target (the logarithm of a normalised mobilised strength), a different database and different models, so the numbers are not comparable. A possible reason for the small optimism here is that four index inputs carry little site-specific information that a flexible model could exploit; we did not test this explanation. The small optimism does not make random splitting appropriate: it is specific to this database and input set.

5.5 Uncertainty and decision relevance

Split-conformal intervals achieved near-nominal pooled coverage without distributional assumptions, but the guarantee is marginal. Site-level coverage was highly uneven, so a user predicting for one new site cannot assume 90% coverage for that site. The ensemble spread alone is far too narrow to be used as an interval (0.28 coverage at nominal 90%), because it measures only the variation among members; GPR's predictive standard deviation was adequately calibrated in pooled terms (0.90). The regime analysis shows regression toward the mean: low strengths are over-predicted and high strengths under-predicted. Over-prediction of low strengths is the unconservative direction for design: about a quarter of the predictions in the lowest tercile exceeded the measured value by more than 20%, and the ln-MVN baseline did not do better (27%). We did not evaluate classification thresholds or factors of safety, so the decision-level consequences remain unquantified.

5.6 Position relative to previous work and evidence level

Previous ML studies of S_u used train–test splits on small single-region data sets (111 samples in Elsawy et al. [7], 230 in Mbarak et al. [3], 127 in Pham et al. [4], 216 in Tran et al. [5]) or five-fold validation on larger databases [8, 9], and several reported high accuracy on the validation data. Those results cannot be compared directly with ours, because the targets, inputs, data sets and validation schemes differ, and the studies did not use site-wise folds. The present study keeps to the direction proposed by Taherdangkoo et al. [12] on one database and one target: Table 5 summarises, theme by theme, the evidence provided and what is missing. The most important gap is the absence of an external validation set from outside Shanghai. The learning curve (Fig. 25) adds a practical message: with 27 training sites the accuracy of random forest was still improving and that of SVR only slowly, so these data cannot show whether a ceiling has been reached; the curves suggest that additional sites would help at least as much as more elaborate models, but this was not tested.

Table 5 Evidence summary following the themes of the minimum-evidence framework of Taherdangkoo et al. [12]

Theme	Evidence in this study	Status
Leakage control	Nested LOSO; preprocessing inside training folds; asserted site disjointness; negative-control and signal-recovery tests	Provided
Deployment-consistent validation	Pooled out-of-site metrics, site-cluster bootstrap, per-site errors; random CV only as diagnostic	Provided for new sites within SH-CLAY
Uncertainty calibration	Conformal intervals with pooled and per-site coverage	Partly (site-level coverage uneven)
Decision relevance	Over-prediction share by strength tercile; no thresholds or factors of safety	Partly
Physical admissibility	Positivity satisfied; monotonicity not satisfied by several models; enforced variants tested and accuracy unchanged	Partly
Applicability limits	Input ranges (Table 1); one city; composite rows; vane correction unknown	Stated; no external test
Auditability	Code, seeds, file hash, library versions, fold-level hyperparameters and predictions	Provided

5.7 Physics-guided learning, site calibration and recent architectures

The physics-guided variants of Section 4.9 do not support a claim that physical constraints improve prediction here: the unconstrained flexible models were already monotone in σ'_v over the observed range, and imposing the constraint changed the RMSE by less than the resampling uncertainty. This is a null result for this database and a single, simple constraint; it does not exclude benefits for sparser data or for constraints with more physical content. Site calibration with a few local measurements reduced the error by an amount that grew with m, but the gain was small for SVR (about 2% at m = 5) and larger for ridge regression (about 8%), and the variance components came from the same residuals that were then corrected, so the benefit may be somewhat optimistic. Operator-learning methods (physics-informed neural operators, DeepONet, Fourier neural operators) learn maps between functions and need field data or a governing equation, and sequence models (Transformers, state-space models such as Mamba) need long ordered sequences such as continuous depth profiles; the present data are 377 independent rows with four scalar inputs and one scalar target, so these architectures were not applied, and no claim is made about their performance. Pretrained tabular transformers designed for small data sets may be worth testing in future work under the same site-wise validation.

5.8 Limitations

(1) The data set is small (377 rows, 28 sites) and sites differ in size; site-level conclusions are uncertain. (2) Each row is a composite of boreholes within 35 m, and the vane test is not necessarily co-located with the index tests, so part of the unexplained variance is measurement mismatch, not model error [2]. (3) σ'_v is derived from depth and is collinear with it. (4) The correction applied to the vane strengths is not stated in the source text available to us. (5) The ln-MVN baseline is simplified (no Johnson transformations). (6) No external validation set was used; results do not transfer to other regions without testing. (7) Complete-case selection may be biased by the structural, possibly non-random missingness; the results apply to rows where all four inputs were measured. (8) Hyperparameter search reached the edge of the grid for several parameters (for example the minimum leaf size of RF and the minimum child samples of LightGBM), so the grids may still be restrictive. (9) The conformal and bootstrap procedures treat sites as exchangeable. (10) The model used for interpretation was chosen by the lowest pooled RMSE, a mildly optimistic selection. (11) The settings of the ANN ensemble were not tuned inside the nested loop (Section 3.2). (12) The learning curve used fixed hyperparameters and random draws of sites, so it is indicative only. (13) The physics-guided variants and the calibration were added after the main comparison and were not tuned; the variance components of the calibration were estimated from the out-of-site residuals of the other sites, whose models were trained with the calibrated site, which can bias the result slightly.

6 Conclusions

Under leave-one-site-out validation on the SH-CLAY/11/4051 database (377 rows, 28 sites), nonlinear ML models predicted field vane S_u from depth, PI, e and σ'_v with pooled out-of-site RMSE of 5.15 to 5.34 kPa (R^2 0.68 to 0.70), compared with 6.09 to 6.10 kPa (R^2 0.58) for an ln-MVN baseline with identical inputs. The 95% intervals of the paired RMSE differences excluded zero for all ML models and both input sets. The nonlinear models could not be ranked against each other, and depth added no benefit to the three index inputs. Conformal intervals reached near-nominal pooled coverage, but coverage differed greatly between sites, and low strengths tended to be over-predicted. Effective vertical stress dominated the fitted models. A linear mixed-effects model matched ridge regression and was outperformed only by the ANN ensemble and random forest. Monotonicity constraints and a strength-ratio target did not improve accuracy, whereas five local measurements at a new site reduced the error on its remaining rows by 0.1 to 0.4 kPa. These findings apply to Shanghai clay and the stated data subset; external validation, a comparison with the full published multivariate model and a treatment of the vane-correction question are the next steps.

Statements and Declarations

Funding. The author declares that no funds, grants or other support were received during the preparation of this manuscript.

Competing interests. The author has no relevant financial or non-financial interests to disclose.

Data availability

The SH-CLAY/11/4051 database is available from Mendeley Data [14] (DOI 10.17632/f3v73r2wn9.1). The code, a script that rebuilds the 377-row analysis subset from the original database, the out-of-site predictions and the per-fold hyperparameters will be deposited in a public repository with a persistent identifier [DOI TO BE INSERTED BY THE AUTHOR AFTER DEPOSIT]. Further tables are provided as Online Resource 1.

References

1. Ching J, Phoon KK (2014) Correlations among some clay parameters—the multivariate distribution. Can Geotech J 51(6):686–704. https://doi.org/10.1139/cgj-2013-0353
2. Zhang D, Zhou Y, Phoon KK, Huang H (2020) Multivariate probability distribution of Shanghai clay properties. Eng Geol 273:105675. https://doi.org/10.1016/j.enggeo.2020.105675
3. Mbarak WK, Cinicioglu EN, Cinicioglu O (2020) SPT based determination of undrained shear strength: regression models and machine learning. Front Struct Civ Eng 14(1):185–198. https://doi.org/10.1007/s11709-019-0591-x
4. Pham BT, Qi C, Ho LS et al (2020) A novel hybrid soft computing model using random forest and particle swarm optimization for estimation of undrained shear strength of soil. Sustainability 12(6):2218. https://doi.org/10.3390/su12062218
5. Tran QA, Ho LS, Le HV, Prakash I, Pham BT (2022) Estimation of the undrained shear strength of sensitive clays using optimized inference intelligence system. Neural Comput Appl 34(10):7835–7849. https://doi.org/10.1007/s00521-022-06891-5
6. Bherde V, Koushik PMV, Balunaini U (2024) Application of ensemble-based methods for prediction of undrained shear strength of soft sensitive clays. In: Geotechnical Special Publication GSP 352, pp 52–61. https://doi.org/10.1061/9780784485347.006
7. Elsawy MBD, Alsharekh MF, Shaban M (2022) Modeling undrained shear strength of sensitive alluvial soft clay using machine learning approach. Appl Sci 12(19):10177. https://doi.org/10.3390/app121910177
8. Zhang W, Wu C, Zhong H, Li Y, Wang L (2021) Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. Geosci Front 12(1):469–477. https://doi.org/10.1016/j.gsf.2020.03.007
9. Xiong H, Dai S, Lan P, He X, Tong C, Zhang S, Sheng D (2025) Probabilistic indirect models for undrained shear strength: addressing significant data missing and variability with advanced imputation and machine learning techniques. Comput Geotech 187:107488. https://doi.org/10.1016/j.compgeo.2025.107488

10. Zhang P, Yin Z-Y, Jin Y-F (2022) Bayesian neural network-based uncertainty modelling: application to soil compressibility and undrained shear strength prediction. Can Geotech J 59(4):546–557. https://doi.org/10.1139/cgj-2020-0751
11. Roberts DR, Bahn V, Ciuti S et al (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913–929. https://doi.org/10.1111/ecog.02881
12. Taherdangkoo R, Arab A, Shakiba S, Doulati Ardejani F, Nagel T (2026) What does “good performance” mean in geotechnical machine learning? Geomech Energy Environ 47:100865. https://doi.org/10.1016/j.gete.2026.100865
13. Ching J, Wang J-S (2025) Quasi-site-specific prediction for foundation capacity based on load-test database. Comput Geotech 184:107305. https://doi.org/10.1016/j.compgeo.2025.107305
14. Zhang D, Zhou Y, Phoon KK, Huang H (2020) Data for: Multivariate probability distribution of Shanghai clay properties. Mendeley Data, V1. https://doi.org/10.17632/f3v73r2wn9.1
15. Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. MIT Press, Cambridge, MA. ISBN 978-0-262-18253-9
16. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
17. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16), pp 785–794. https://doi.org/10.1145/2939672.2939785
18. Ke G, Meng Q, Finley T et al (2017) LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30:3146–3154
19. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html
20. Lei J, G’Sell M, Rinaldo A, Tibshirani RJ, Wasserman L (2018) Distribution-free predictive inference for regression. J Am Stat Assoc 113(523):1094–1111. https://doi.org/10.1080/01621459.2017.1307116
21. Lundberg SM, Erion G, Chen H et al (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9
22. Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res 106(D7):7183–7192. https://doi.org/10.1029/2000JD900719

Further declarations

Funding: None. **Competing interests:** None. **Ethics approval and consent:** Not applicable (secondary analysis of a published database). **Author contributions (CRediT):** Seyed Mahdi Naseri Baygi: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft, writing – review and editing, visualization. **Use of large language models:** see Section 3.9.