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.

Groundwater chemistry inference, transferability and physical consistency evaluation with machine learning across global monitoring records

Md Rakib Hasan

rakibhhridoy@fermium.systems

University of Dhaka https://orcid.org/0009-0002-4007-7590

Shoumik Zubyer

University of Dhaka https://orcid.org/0009-0006-4085-1493

Anwar Zahid

University of Dhaka https://orcid.org/0000-0002-0456-6760

Research Article

Keywords: groundwater quality, machine learning, spatial transfer, conformal prediction, foundation model, arsenic, quantum machine learning

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11578171/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.

Article

Preprint

Groundwater chemistry inference, transferability and physical consistency evaluation with machine learning across global monitoring records

Md Rakib Hasan [1,3] ✉ Shoumik Zubyer [1,3] Anwar Zahid [1,2]

Machine learning is increasingly used to fill gaps in groundwater quality records, yet models are rarely tested in regions they were not trained on. Here nine methods, from logistic regression and raster-based hazard models to a generic tabular foundation model, a groundwater-pretrained Transformer and a simulated quantum kernel, are compared on identical splits drawn from 2.1 million samples. Field parameters and major ions recovered the major-ion chemistry of unseen wells (R^2 0.83–0.95), but arsenic reached at most 0.53. Across 47 leave-one-region-out tests, measured chemistry ranked contaminant exceedances with AUC 0.71–0.76, whereas environmental rasters alone stayed near chance (0.58). Out of region, exceedance coverage fell below its nominal 90% for every method, and the best-ranking model missed the most exceedances (35%). Domain pretraining matched a generic tabular model, and the quantum kernel matched its classical counterpart. Ranking accuracy and trustworthy uncertainty therefore separate under geographic shift, and redox-sensitive contaminants still require direct measurement.

[1]Department of Soil, Water and Environment, University of Dhaka, Dhaka 1000, Bangladesh. [2]Institute of Water Modelling, Dhaka, Bangladesh. [3]Fermium Systems, Dhaka 1207, Bangladesh. ✉e-mail: rakibhhridoy@fermium.systems

Introduction

Groundwater supplies drinking water to over half the world's population and irrigates 40% of global cropland [1], yet its chemical quality is measured far less completely than its use would warrant. Arsenic, fluoride and manganese occur naturally above health-based guidelines in aquifers on every inhabited continent [2–6], and the Bengal Basin alone exposes tens of millions of people to drinking water above the WHO arsenic guideline of 10 μg/L [7,8]. Comprehensive multi-parameter databases exist in fewer than 30 countries, concentrated in Europe and North America [9], while the regions with the most severe geogenic contamination in South and Southeast Asia, sub-Saharan Africa and Latin America hold the sparsest records. Most samples that do exist report a handful of routine parameters, not the full suite needed to judge redox state or contaminant mobility. The practical question for a water manager is therefore which unmeasured parameters can be inferred from the ones in hand, and how far that inference can be trusted.

Machine learning has become the standard answer to that question in three forms. Global and national hazard maps predict the probability that a contaminant exceeds its guideline from gridded climate, soil and geology layers, trained on point measurements and extrapolated to unsampled land [2–4,10]. Transfer learning carries models from data-rich regions or basins to data-scarce ones [11–14]. Pretrained foundation models, which learn general representations from large unlabelled corpora before being adapted to a task [15], now extend to tabular data through generic models such as TabPFN [16] as well as through domain-specific pretraining. All three promise skill where monitoring is thin, and all three are usually evaluated with random or within-region cross-validation, which measures interpolation between neighbouring samples rather than prediction in a new place. For large-scale ecological maps, random cross-validation has reported most of the variance explained where spatial validation found almost none [17]. That gap motivated an explicit area of applicability for spatial models [18] and a continuing debate on how map accuracy should be assessed [19,20]. For geogenic groundwater contaminants, the one direct test of transferability found that random forests and neural networks transferred fluoride predictions no better than logistic regression [21]. That result is consistent with the broader observation that tree and linear models remain hard to beat on tabular scientific data [22,23].

Four questions therefore remain open for groundwater chemistry, and each needs a different test. The first is inferability, meaning which parameters routine measurements can recover and which must be measured, a question that sets the minimum sampling suite for a monitoring programme. The second is transfer, whether predictions of guideline exceedance hold in a region that contributed no training data. The third, and arguably the most consequential for practice, is calibration, because a monitoring agency acts on a stated risk level and a model that ranks wells well but understates its uncertainty sends field teams to the wrong wells with misplaced confidence. Conformal prediction gives such risk levels a finite-sample coverage guarantee when calibration and test data are exchangeable [24,25], and so offers a direct test of whether the guarantee survives a geographic shift. Calibrated coverage has, to the authors' knowledge, not been reported for groundwater contamination models. The fourth is physical consistency, whether a reconstructed water still balances its charges and sits at plausible mineral saturation, since a reconstruction that violates either cannot be trusted for geochemical interpretation however accurate its individual values appear.

This study answers the four questions with one controlled comparison rather than by proposing a new model. Eight methods are evaluated on identical rows and splits, namely logistic regression, spatial k-nearest-neighbour interpolation, a raster-only random forest of the kind used in global hazard maps, random forest, gradient boosting, TabPFN, and the GENESIS Transformer both frozen and fine-tuned, and a simulated quantum kernel enters as a ninth. GENESIS was pretrained by masked reconstruction on 2,087,970 harmonized samples from four public monitoring databases (GEMStat, USGS NWIS, French ADES and EEA Waterbase), so it enters as the domain-pretrained candidate and not as the object of the study. Inferability is measured on sites absent from training, transfer and calibration on 47 leave-one-region-out cells covering seven contaminants in 15 regions, and physical consistency through charge balance and the saturation indices of calcite, gypsum and fluorite. A Bangladesh field dataset, held out from every stage of training, serves as a stress test for strongly reducing aquifers. The harmonized corpus, the fixed splits and the per-sample predictions of every method are released so that later methods can be scored on the same footing.

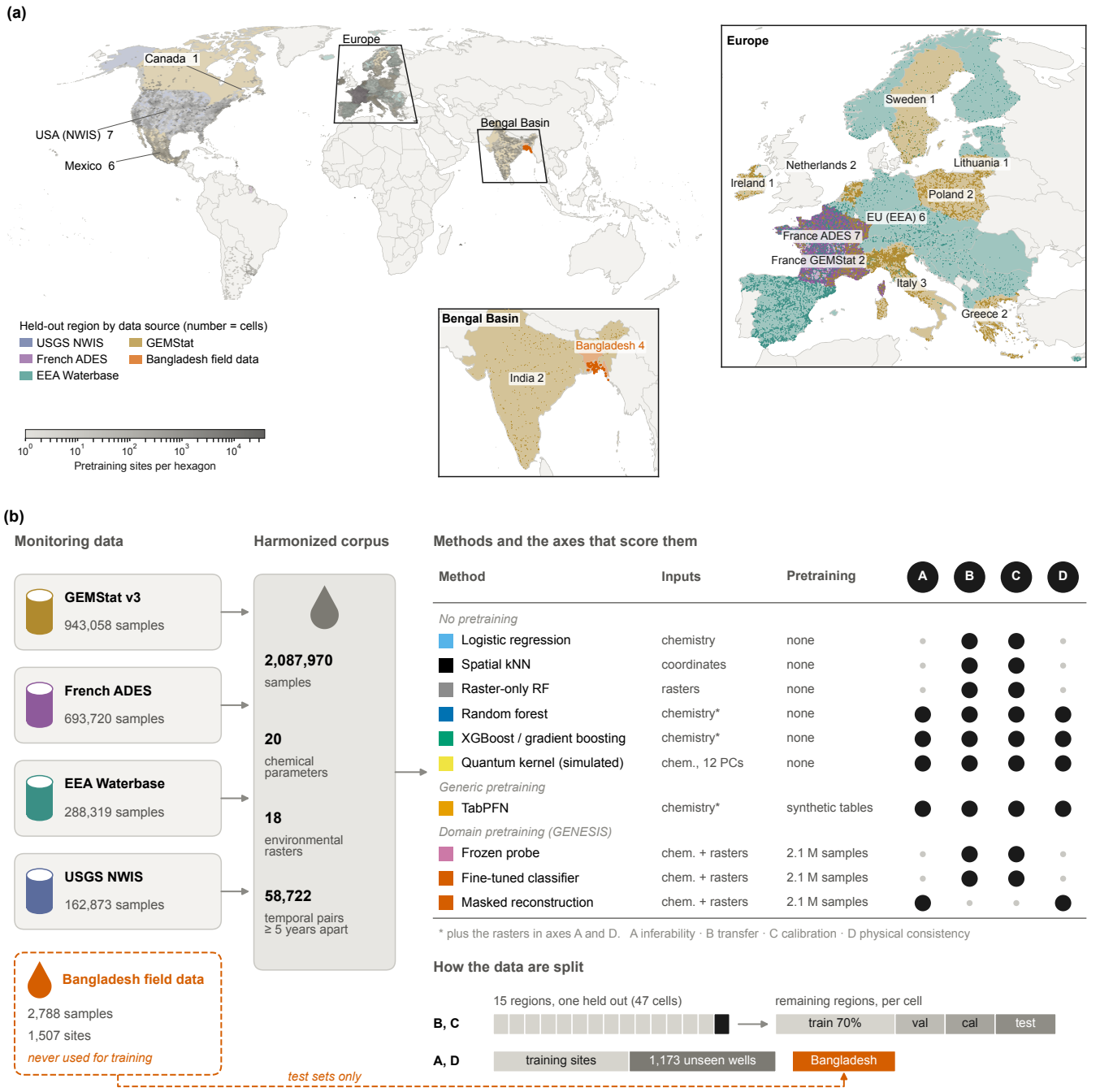


Fig. 1. Data coverage and study design. (a) Global distribution of pretraining sites (grey hexagons, log density) and of the 15 regions held out in turn, with each held-out region's countries tinted by its data source and its number of leave-one-region-out cells given after its name. Insets enlarge Europe and the Bengal Basin, with temporal-pair sites coloured by source and the Bangladesh field data, which no model was trained on, in vermillion. (b) The four databases are harmonized into one corpus, the matrix shows the inputs and pretraining of each method and which axes score it, and the bars show how the data are split for each axis. PCs, principal components.

Results

Inferability from routine measurements

Routine measurements fixed the major-ion composition of unseen groundwater but not its redox chemistry (Fig. 2). With pH, EC and the major ions supplied (level L2), GENESIS reconstructed Ca, Mg, Na, Cl, HCO_3, SO_4, EC and TDS at site-disjoint test wells with R^2 of 0.83–0.95, while K (0.73) and pH (0.61) lagged. The redox-sensitive parameters needed the full suite, Fe rising from 0.53 at L2 to 0.71 at L4, Mn from 0.47 to 0.68 and U from 0.41 to 0.64. Arsenic was the extreme case, at 0.22 with the major ions and 0.53 with every other parameter supplied. The ceiling for arsenic was shared by every method, since the best baseline at L4 reached 0.51, which places its limit in the data rather than in any one model.

The ordering of methods depended on how much chemistry was available (Fig. 2a). With rasters alone (L0), GENESIS had a median R^2 of 0.13 against 0.38–0.45 for gradient boosting, random forest and TabPFN, and it trailed a baseline on all 19 targets, as expected of a model whose pretraining always included chemistry. Once the major ions were present the order reversed, and at L4 GENESIS reached a median R^2 of 0.80 against 0.71–0.73 for the baselines (Wilcoxon $p \leq 0.0014$ against each). Its lead was widest on Mn (0.68 against a best baseline of 0.49), NO_3 (0.75 against 0.58) and Fe (0.71 against 0.63) (Fig. 2b), and it trailed a baseline on four targets only, U, DOC, EC and TDS. That advantage did not survive the move to Bangladesh, where no method reached a positive median R^2 at L4 (GENESIS −0.15, gradient boosting −0.08, random forest −0.01, TabPFN −0.06, with GENESIS no

different from any of them, $p \geq 0.32$). The redox-sensitive species failed there for GENESIS without exception (As −0.25, Fe −0.76, Mn −0.21, PO_4 −0.62), as did nitrate (−1.65). Na, Cl, EC and TDS stayed above 0.69, so the failure is specific to the parameters that the strongly reducing aquifer controls rather than a general loss of skill.

Greedy selection named the measurements that most help each contaminant, and its choices follow known geochemistry. Fe was chosen first for Mn and Mn first for Fe, HCO_3 first for U and F, and Fe first for As. The rule did not rescue arsenic, whose site-disjoint R^2 stayed between 0.31 and 0.38 for gradient boosting and between 0.09 and 0.13 for GENESIS over the four additions, so arsenic has to be measured. With these sparse inputs gradient boosting outperformed GENESIS for As, Fe, U, F and PO_4, consistent with the L0 and L1 result that the pretrained model is weakest when few parameters are observed.

Transfer to unseen regions

Chemistry transferred to unseen regions and location did not (Fig. 3a). Across the 47 leave-one-region-out cells, the methods using chemistry ranked exceedances with mean AUC between 0.708 (the simulated quantum kernel) and 0.760 (TabPFN), whereas the raster-only random forest reached 0.583 and spatial interpolation 0.532, close to chance. The raster-only model lost 0.148 AUC against random forest on the same rows ($p < 10^{-6}$), which is the gap between a hazard map built from environmental layers and a model given the chemistry of the well.

Among the chemistry models the two pretrained ones ranked best, with TabPFN (0.760) and fine-tuned GENESIS (0.755) statistically indistinguishable ($p = 0.30$, and $p = 0.81$ on the 34 redox cells), and both outperformed random forest (0.731, $p = 0.0002$ for TabPFN and $p = 0.022$ for GENESIS) and XGBoost (0.732). Every chemistry model ranked the conservative contaminants best (AUC 0.81–0.88), the redox-sensitive ones next (0.69–0.74) and uranium worst (0.54–0.64, four cells), so the ordering of methods held within each group (Fig. 3b). The frozen GENESIS probe (0.723) did no better than random forest ($p = 0.47$) or logistic regression ($p = 0.40$), so groundwater-specific pretraining contributed nothing until the encoder was tuned on the task (fine-tuned against frozen, $p < 10^{-5}$). A generic tabular model pretrained only on synthetic data, with a context of 10,000 rows against roughly four times as many for the other methods, therefore matched the domain-pretrained model on transfer.

The audit of raster-only models explains where hazard maps lose their skill (Fig. 3c). Under random cross-validation the raster-only forest scored AUC 0.74–0.89 for all 20 parameters, with a mean of 0.81, and that in-distribution skill was unrelated to how well each parameter could be reconstructed from chemistry (Spearman $\rho = 0.20$, permutation $p = 0.40$). Held out by region, the same models fell to 0.37–0.71 (mean 0.59), now in close order of reconstruction difficulty ($\rho = 0.76$, $p = 0.0001$, and 0.72 without Eh). Cl, Na and TDS kept the most skill (0.68–0.71), while Mn (0.54), DOC (0.48) and Eh (0.37) lost it entirely. The order held within single held-out regions, with ρ of 0.66–0.78 in the USA, France, the EEA set and Bangladesh, but it was weak in Mexico (0.03) and Greece (0.14, six parameters). Within one data source it held for the French network ($\rho = 0.77$, $p = 0.0001$) but not for the US network (0.31, $p = 0.19$), so part of the cross-region gradient may reflect differences in sampling protocol rather than geochemistry alone. Detection limits were ruled out as the cause, since the labelling threshold lay above every source's reporting limit for all 20 parameters.

Calibration under geographic shift

Every method kept its stated coverage in distribution and lost it out of region (Fig. 4). On held-aside test pairs from the training regions, mean exceedance coverage was 0.91 for all nine methods, as the conformal guarantee requires. In the held-out region it fell to 0.65–0.81 ($p \leq 0.006$ for every method except the raster-only forest), and more than half of the cells dropped below 0.85 for every method except logistic regression (45%) and the raster-only forest (34%). Exceedance prevalence rose from 0.08 in calibration to 0.16 in the held-out regions, but class-conditional calibration is insensitive to prevalence, so the loss reflects a change in the chemistry of exceeding samples. The loss was not confined to redox species, coverage for random forest being 0.81 on redox cells, 0.73 on conservative cells and 0.48 on uranium cells, and the redox cells were no worse than the conservative ones ($p = 0.79$). For every chemistry model the miss rate was lowest on redox cells and highest on uranium cells (Fig. 4c).

Sharpness and safety traded against each other across methods (Fig. 4a), with fine-tuned GENESIS returning the fewest ambiguous sets (12%) and missing the most exceedances (35%). Its coverage was below that of random forest overall (0.65 against 0.77, $p < 10^{-4}$) and on redox cells (0.70 against 0.81, $p = 0.001$). TabPFN and random forest missed 23–24% of exceedances with 21–22% ambiguous sets. The methods with the lowest miss rates bought them with ambiguity, logistic regression missing 20% with 38% ambiguous sets and the raster-only forest 20% with 55%, which is coverage achieved by declining to decide. The best-ranking chemistry model was therefore the least reliable about its own uncertainty, being confidently wrong more often than the forest it outranked. Recalibration on labelled pairs from the held-out region restored coverage to about 0.95, but only through sets containing both classes for 67–88% of samples at $k = 200$, so it offered no practical remedy.

Physical consistency of reconstructed waters

On unseen sites, reconstructed waters obeyed mineral equilibrium and charge balance about as well as observed ones, for every accurate regressor (Fig. 5). Saturation indices computed from reconstructed waters agreed with those from observed waters with R^2 of 0.87–0.91 for calcite and 0.89–0.94 for gypsum, against shuffled nulls near −1, and the saturation state of calcite was classified correctly for 83–85% of samples. Median absolute charge-balance error was 4.2–6.1% against 3.4% in the observed waters, lowest for random forest (4.2%) and GENESIS (4.4%). Fluorite was reproduced less well (R^2 0.34–0.41). Oxic and reducing samples within the corpus showed no consistent difference, with calcite R^2 of 0.90 and 0.93 for GENESIS.

Consistency broke down in Bangladesh, where 1,984 of the 2,000 test samples met the reducing criterion. Calcite R^2 fell to between −0.34 and 0.16, gypsum to between −0.57 and −0.40, fluorite to between −0.21 and −0.07, and correct classification of calcite saturation to 56–64%. Charge-balance error rose for every method, from 7.5% in the observed waters to 10.9% for GENESIS, 13.4% for TabPFN, 15.3% for random forest and 19.7% for gradient boosting. Equilibrium consistency on familiar water is therefore shared by every accurate regressor and is not evidence that any one model has learned geochemistry, while under strong reducing shift all of them produce waters that are no longer charge-balanced, GENESIS least so.

Simulated quantum kernel

The simulated quantum kernel ranked exceedances in the held-out region with a mean AUC of 0.708, the same as the classical RBF kernel on identical rows and features (0.708, higher in 26 of 47 cells, $p = 0.66$) and the same circuit without entanglement (0.708, $p = 0.90$) (Fig. 6a, b). The three also agreed within the redox-sensitive cells (0.694–0.696) and the conservative cells (0.823–0.831, $p \geq 0.43$ in each group), so the entangling couplings added nothing measurable. The quantum kernel matched logistic regression (0.711, $p = 0.60$) and ranked below TabPFN (0.760, higher in 6 of 47 cells, $p < 10^{-4}$), fine-tuned GENESIS (0.755, $p = 0.0001$), XGBoost (0.732, $p = 0.040$) and random forest (0.731, $p = 0.026$). Its miss rate on exceedances in the held-out region,

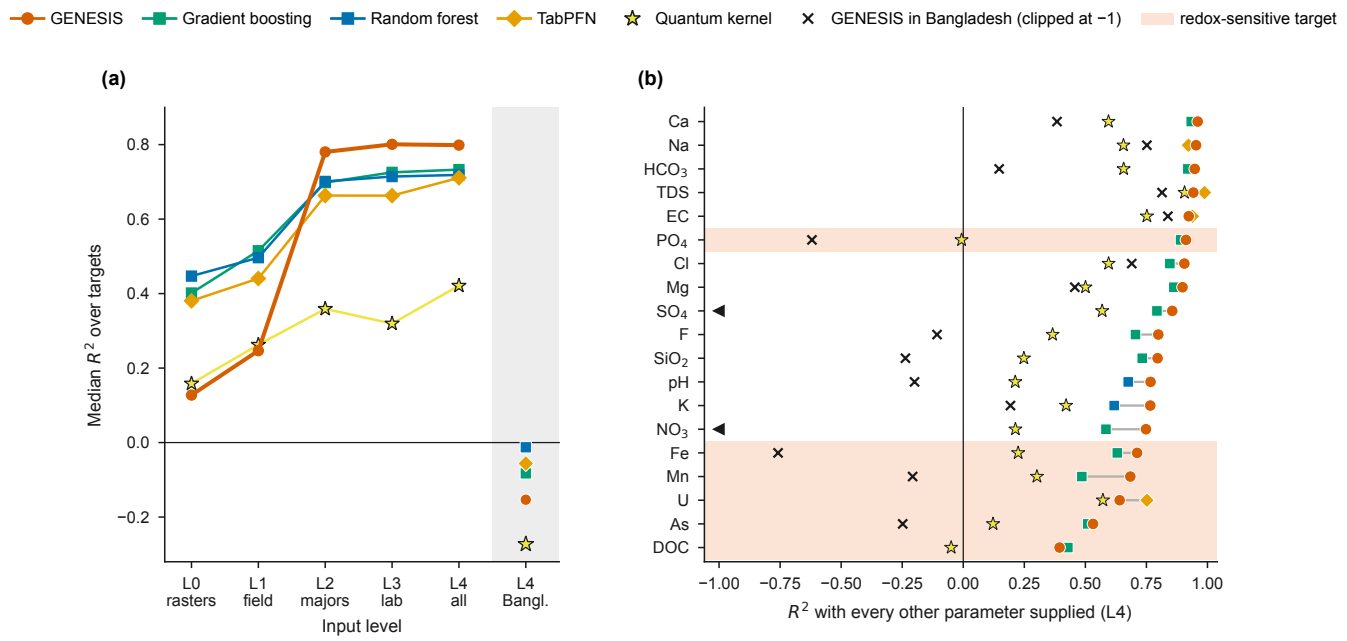


Fig. 2. Inferability from routine measurements. (a) Median R^2 over targets at site-disjoint test wells as inputs are added, from rasters only (L0) to every other parameter (L4), with the L4 values in Bangladesh at right. (b) R^2 per target at L4 for GENESIS and for the best of gradient boosting, random forest and TabPFN (marker and colour identify which), with the GENESIS value in Bangladesh marked and clipped at −1. Stars mark the simulated quantum kernel (kernel ridge regression on 12 principal components and 2,000 training samples, Methods). Shading marks redox-sensitive targets.

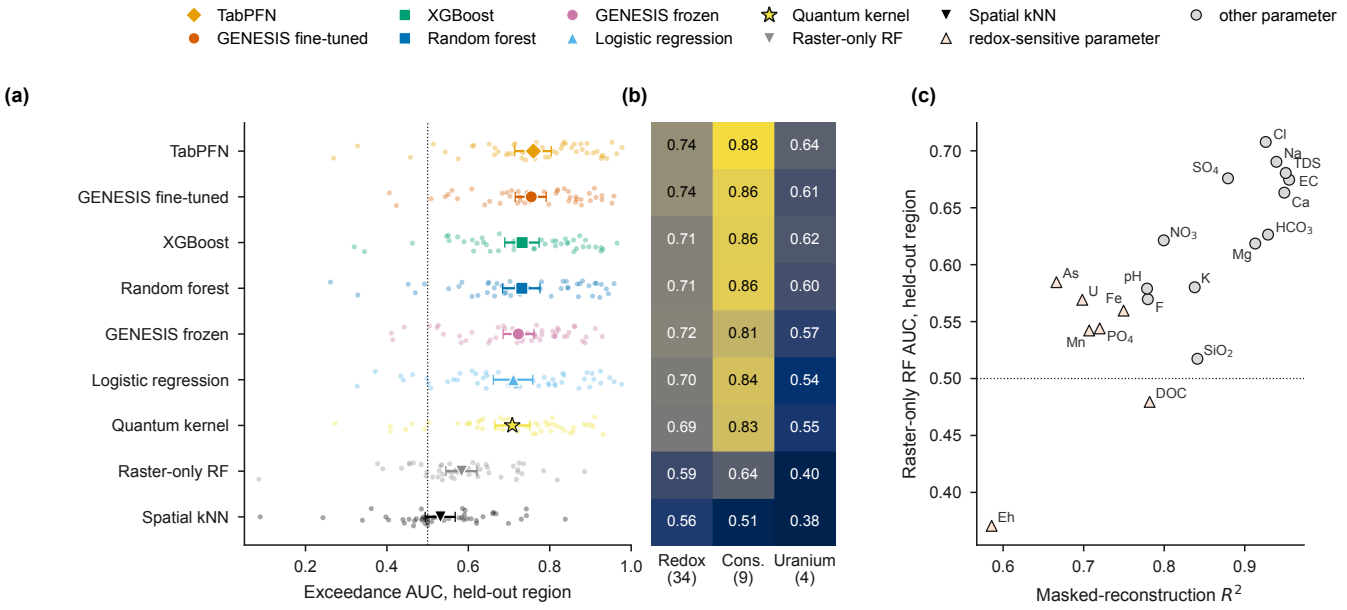


Fig. 3. Transfer to unseen regions. (a) Exceedance AUC in the held-out region for the eight methods and the simulated quantum kernel (star), with the 47 target and region cells as dots and 95% bootstrap intervals on the mean. The quantum kernel used 12 principal components of chemistry and 2,000 training pairs per cell (Methods). (b) Mean AUC by chemical group, with the number of cells in parentheses. (c) Leave-one-region-out AUC of a raster-only random forest against the masked-reconstruction R^2 of each of the 20 parameters (Spearman $\rho = 0.76$).

0.20, was lower than that of fine-tuned GENESIS (0.35, $p < 10^{-4}$), but both controls missed the same fraction, placing it among the simpler models of Fig. 6c rather than reflecting any property of the circuit. The selected bandwidth lay at the lower edge of the search range in only 5 of 141 cell and seed fits, and off-diagonal kernel values kept a median variance of 0.015, so the tie does not arise from a kernel that had collapsed to a constant.

Used for regression, the three kernels again agreed with one another and fell far below the other regressors (Figs. 2 and 5). At L4 on unseen wells the quantum kernel reached a median R^2 of 0.42, against 0.44 for the RBF kernel ($p = 0.29$) and 0.46 without entanglement ($p = 0.08$), and it trailed GENESIS, gradient boosting, random forest and TabPFN (0.71–0.80) on 18 or 19 of the 19 targets ($p < 10^{-4}$ against each). With field parameters and major ions supplied it recovered the major ions with R^2 of 0.51–0.70, against 0.83–0.95 for GENESIS, and at L4 it reached 0.12 for arsenic against 0.39–0.53. Where the entangled and unentangled circuits differed, the entangled one was worse (L3, 0.32 against 0.46, $p = 0.002$). The reconstructed waters were also less consistent, with saturation-index R^2 of 0.74–0.76 for calcite and gypsum for all three kernels against 0.87–0.94 for the other regressors, a fluorite R^2 of −1.84 for the quantum kernel and −0.64 for the RBF kernel, and a median absolute charge-balance error of 8.1–8.7% against 3.4% in the observed waters. In Bangladesh the charge-balance error of the two circuit kernels reached 31.3% with entanglement and 33.3% without it, about twice that of the RBF kernel

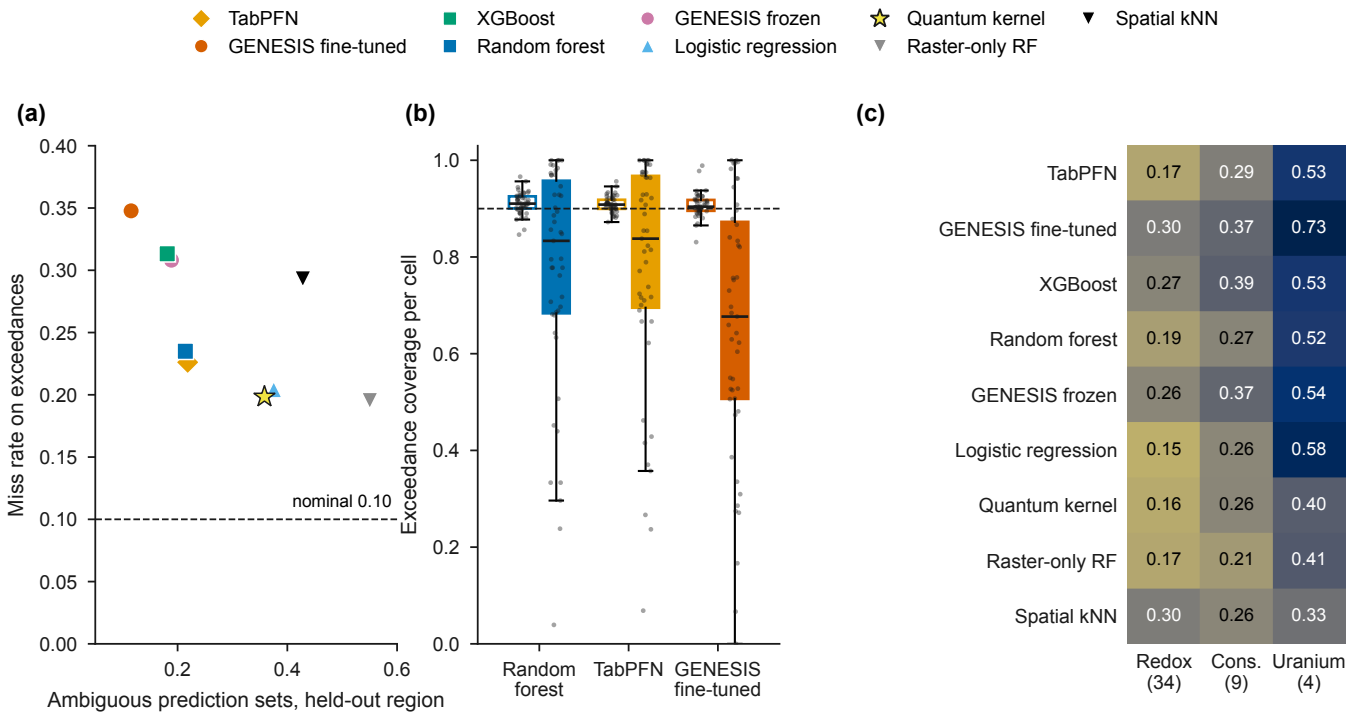


Fig. 4. Calibration under geographic shift, for split conformal prediction at a nominal miss rate of 0.10. (a) Miss rate on exceedances against the fraction of ambiguous prediction sets in the held-out region. (b) Exceedance coverage per cell in distribution (open) and in the held-out region (filled) for random forest, TabPFN and fine-tuned GENESIS. (c) Miss rate in the held-out region by chemical group. The star marks the simulated quantum kernel.

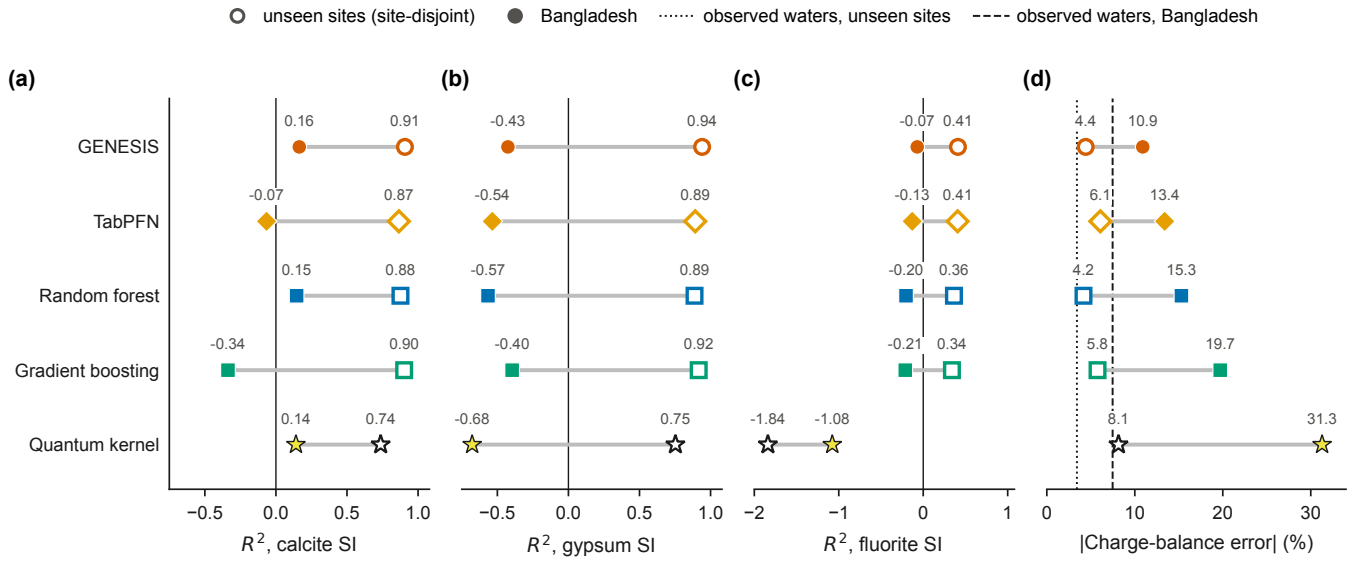


Fig. 5. Physical consistency of reconstructed waters at site-disjoint wells (open markers) and in Bangladesh (filled markers), for the four regressors at level L4. (a, b, c) R^2 of calcite, gypsum and fluorite saturation indices computed from reconstructed against observed waters. (d) Median absolute charge-balance error of reconstructed waters, with the observed waters' error as vertical lines. The fifth row is the simulated quantum kernel.

(15.2%).

Discussion

What routine monitoring can and cannot replace

The inferability results draw a line through the parameter list that monitoring programmes can use directly. Field parameters and major ions fix Ca, Mg, Na, Cl, HCO_3, SO_4, EC and TDS at unseen wells with R^2 above 0.83, which is expected because charge balance and carbonate equilibrium tie these quantities together. A model only has to recover constraints that the water already obeys. Where budgets force a choice, these are the analyses that can be thinned first. The redox-sensitive species on the other side of the line gained accuracy only when every other parameter was supplied. Arsenic never exceeded R^2 0.53 for any method, so the shortfall lies in what routine chemistry carries about arsenic and not in the choice of model. Arsenic mobilization depends on the reductive dissolution of iron oxides and on sediment age and organic carbon supply[26–28], and these controls leave only a partial imprint on the major-ion composition of the water. The practical consequence is that arsenic must continue to be measured wherever it is a health concern, and that a reconstructed arsenic value should not stand in for a missing analysis.

In the greedy measurement sets, Fe and Mn were chosen first for each other and Fe first for arsenic, mirroring the coupled reduction of iron and manganese oxides[29,30]. HCO_3 was chosen first for uranium and fluoride, consistent with carbonate complexation of uranium and the association of fluoride with sodium

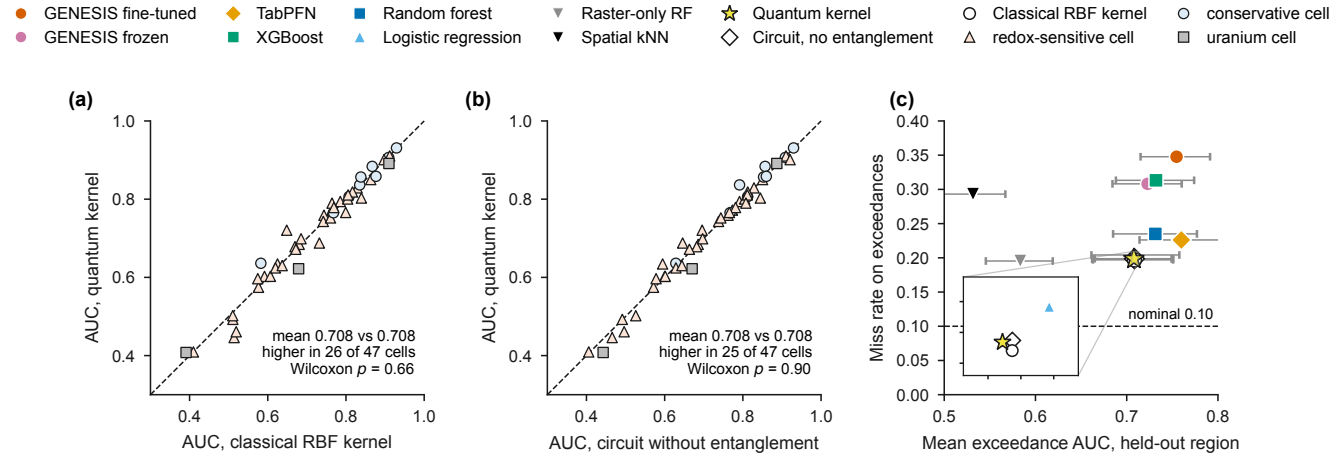


Fig. 6. Simulated quantum kernel against its matched controls. (a) Exceedance AUC in the held-out region for each of the 47 cells, quantum kernel against a classical RBF kernel on the same rows and features. (b) Quantum kernel against the same circuit without entangling couplings. Dashed lines mark equality, and p values are from two-sided Wilcoxon tests. (c) Mean AUC against miss rate on exceedances in the held-out region for all eleven classifiers, with 95% bootstrap intervals over cells. The inset enlarges the cluster at AUC 0.708–0.711 and miss rate 0.197–0.204, where the quantum kernel, both controls and logistic regression coincide.

bicarbonate waters. Because these selections were made without geochemical input, their agreement with known couplings is a check on the method, although the selections do not rescue arsenic and they name the measurements that help most, not the ones that make reconstruction reliable.

Why raster-based hazard maps lose skill in new regions

Hazard maps built from environmental rasters performed well under random cross-validation for every parameter (AUC 0.74–0.89) and lost much of that skill when a whole region was held out (0.37–0.71). Random cross-validation therefore measures how well a raster model interpolates between neighbouring wells, not how well it predicts in a new place, the gap that spatial validation exposed for large ecological maps [17,20]. The loss was ordered by how well chemistry determines each parameter, so that conservative ions, whose concentrations follow climate, evaporation and lithology at regional scale, kept the most skill. Mn, DOC and Eh, which depend on local redox conditions in the subsurface, lost it entirely. Surface layers describe climate, soil and land cover but not the sediment age, buried organic matter and flow paths that set redox state at depth [10,27]. A map built from them would be expected to transfer worst for the species these subsurface conditions control, as observed.

Two results temper this reading, the first being that within the US network, where one protocol applies throughout, the ordering was weak and not significant ($\rho = 0.31$), while within the French network it was strong (0.77). Part of the cross-region gradient may therefore come from differences in sampling, filtration or analytical practice between data sources, which would leave the ranking of parameters intact but would be confounded with region. Second, detection limits do not explain it, since every labelling threshold lay above every source's reporting limit. The practical message does not depend on resolving this, because whatever its origin the loss is invisible under random cross-validation, and it is largest for the contaminants of most concern after arsenic. A raster model given no chemistry was also the weakest transfer method of the eight (AUC 0.58), so any measured chemistry from a target region carries more information about exceedance than the global layers do.

What pretraining contributes

The comparison separates two claims that are often made together for foundation models in environmental science. The first, that pretraining on a large domain corpus yields a representation useful in new regions, was not supported, because the frozen GENESIS probe ranked exceedances no better than random forest or logistic regression. The second, that a pretrained network adapted to the task ranks better than conventional models, was supported. Generic pretraining achieved the same result, since TabPFN, trained only on synthetic tables and given a quarter of the training rows, matched the fine-tuned GENESIS on transfer. The advantage of the pretrained classifiers over random forest and XGBoost (about 0.02–0.03 AUC) therefore reflects the class of model and its adaptation, not groundwater-specific knowledge. This is consistent with evidence that tree and linear models are strong but beatable on tabular data [16,22,23].

Domain pretraining did leave a measurable trace in reconstruction, since given full chemistry at unseen sites GENESIS was the most accurate regressor (median R^2 0.80 against 0.71–0.73), with its largest margins on Mn, NO_3 and Fe. Its reconstructions stayed the most charge-balanced in Bangladesh. With rasters alone it was the weakest, and in Bangladesh its reconstruction advantage disappeared. The defensible claim is that masked pretraining on 2.1 million samples produces a strong reconstructor within the distribution it was trained on, which is useful for gap-filling in well-monitored networks, but not a model that carries groundwater knowledge into new hydrogeological settings. A groundwater foundation model should therefore be judged against a generic tabular model and a tuned tree ensemble on held-out regions before any claim of transfer is made.

The quantum kernel leads to the same conclusion from another direction. An entangling circuit, an unentangled version of it and a classical Gaussian kernel produced indistinguishable rankings and miss rates on every chemical group, consistent with large benchmarks in which classical baselines match or beat quantum classifiers on classical data and removing entanglement costs nothing [31]. Out-of-region skill in this comparison was therefore limited by the shift in chemistry between regions and by the number of training rows a method can use, not by the expressiveness of the similarity measure. The regression axes make the second limit visible, since all three kernels, restricted to 2,000 training samples, reconstructed chemistry far less accurately and less consistently than models trained on up to 150,000, the classical kernel by a similar margin to the quantum one. Quantum models for water quality have so far been tested within single sites or basins, and the result here suggests they should be reported against a classical kernel on the same features and under regional hold-out before any advantage is claimed.

Accuracy and trustworthy uncertainty under shift

The calibration results are the most consequential for practice, because a monitoring agency acts on stated risk levels and not on a ranking score. Conformal calibration delivered its nominal coverage in distribution for every method, and every method fell below it in unseen regions, so a guarantee calibrated on one set of networks does not transfer to another. The loss was not confined to redox species, which indicates a general change in the chemistry of exceeding samples between regions, of the kind that conformal guarantees, which require exchangeable calibration and test data [24,25], cannot cover.

The more troubling finding is that ranking skill and reliability moved in opposite directions. Fine-tuned GENESIS ranked as well as any method yet missed 35% of exceedances, compared with 23–24% for TabPFN and random forest, because it returned the sharpest predictions, with ambiguous sets for only 12% of samples. Fine-tuned deep networks are known to produce overconfident probabilities [32], and calibration of deep models degrades further under distribution shift [33]. The present results show the same behaviour in a groundwater setting, where its consequence is concrete, since an overconfident model directs field teams away from wells that do exceed. At the other extreme, logistic regression and the raster-only forest reached low miss rates only by declining to decide for 38–55% of samples. No method was both sharp and safe out of region, and a miss rate reported without the fraction of ambiguous predictions overstates how useful a method is.

Recalibrating with up to 200 labelled samples from the new region restored nominal coverage only by making most prediction sets ambiguous, so a small local campaign cannot repair the calibration of a transferred model. Model-based screening in a new region should therefore prioritize sampling, not replace it. The miss rate under transfer, not the in-distribution coverage, is the number to use when deciding how many wells must still be tested.

Physical consistency as a check

Reconstructed waters reproduced calcite and gypsum saturation and stayed close to observed charge balance at unseen sites for every accurate regressor, so equilibrium consistency on familiar water is a property of accuracy and not evidence that a particular model has learned geochemistry. The informative case was Bangladesh, where almost every test sample was reducing and consistency failed for every method, with charge-balance error rising from 7.5% in the observed waters to 10.9–19.7% for the four main regressors. Charge balance and saturation indices need no labels, so they could serve as an inexpensive warning that reconstructions have left the conditions a model was trained on. This use has not been tested here as a shift detector and is offered as a direction for future work.

Limitations

The corpus is dominated by European and North American networks, and Bangladesh is the only strongly reducing region held out, so the redox results rest on one setting in which reducing conditions and regional shift coincide. The weak gradient within the US network leaves open how much of the cross-region pattern reflects protocol and how much reflects geochemistry. TabPFN was limited to a 10,000-row context, so its results are a lower bound, and GENESIS was evaluated at one encoder size. The quantum kernel was simulated exactly at 12 qubits on 2,000 training rows per cell or fit, so neither larger circuits nor hardware noise were tested. In the transfer comparison the GENESIS classifiers also saw the environmental rasters, which the other chemistry models did not, although rasters alone transferred poorly and the frozen probe, which also saw them, did no better than random forest. The thresholds for Fe, Mn and PO_4 are operational rather than health-based, the uranium results rest on four cells, and Eh lacked site-disjoint test rows. The thermodynamic check used Davies activities and free-ion speciation without ion pairs, which is adequate for comparing reconstructed with observed waters by identical equations but not for absolute saturation states in saline or strongly complexing waters.

Conclusions

This study compared eight machine-learning methods and a simulated quantum kernel for groundwater chemistry on identical splits drawn from 2.1 million samples, along four axes that a water manager would ask about. The central finding is that ranking accuracy and trustworthy uncertainty separate under geographic shift. Out of region, every method fell below its nominal 90% exceedance coverage, and the best-ranking fine-tuned Transformer missed the most exceedances (35%), while the methods with low miss rates achieved them through ambiguous predictions. Measured chemistry transferred to new regions where environmental rasters and location did not (AUC 0.71–0.76 against 0.58 and 0.53), and the loss of raster-only skill followed the redox control of each parameter, a loss that random cross-validation conceals. Routine measurements recovered the major ions, EC and TDS of unseen wells (R^2 0.83–0.95) but not arsenic (at most 0.53), which must therefore still be measured. Groundwater-specific pretraining added nothing frozen and, once fine-tuned, matched a generic tabular foundation model, and a simulated quantum kernel matched its classical counterpart (AUC 0.708 for both), while GENESIS yielded the most accurate and most charge-balanced reconstructions within the training distribution. Equilibrium consistency was shared by every accurate regressor and failed for all of them in the reducing aquifers of Bangladesh. Machine learning is therefore best used in groundwater monitoring to prioritize where to sample rather than to replace sampling, and models intended for new regions should be reported with their out-of-region miss rate and ambiguity alongside their ranking skill. The harmonized corpus, fixed splits and per-sample predictions are released so that later methods can be scored on the same footing.

Methods

Data and harmonization

The corpus combines 2,087,970 multi-parameter water chemistry samples from four public monitoring databases, namely GEMStat v3 [9] with 943,058 samples, French ADES through the Hub'Eau interface [34] with 693,720, EEA Waterbase [35] with 288,319, and USGS NWIS through the Water Quality Portal with 162,873. All records were mapped to a vocabulary of 20 parameters, comprising the contaminants As, F, U, NO_3 and Mn, the redox indicators Fe, Eh, DOC and PO_4, the major ions Ca, Mg, Na, K, Cl, HCO_3 and SO_4, the bulk properties pH, EC and TDS, and SiO_2. Dissolved fractions were preferred over total ones, and plausibility filters removed physically impossible values. Each sampling site was described by 18 environmental rasters extracted through Google Earth Engine [36]. The rasters cover terrain (elevation and slope), soil (sand, clay, organic carbon, pH and bulk density), land surface (land cover, population density, and the mean and standard deviation of NDVI), climate (precipitation, temperature, actual and potential evapotranspiration, and aridity) and GRACE water storage (anomaly mean and trend). These rasters are the inputs that a global hazard map relies on, and they are supplied to every method that is allowed them.

Single samples serve the inferability and physical-consistency axes, while temporal pairs, two samples from the same site at least five years apart with at least three parameters in common, serve transfer and calibration, where the task is to predict exceedance at the later sample from chemistry at the earlier one,

and the corpus yields 58,722 such pairs. A Bangladesh dataset of 2,788 groundwater samples from 1,507 sites, of which 1,807 come from the authors' 2020–2021 field campaign and 981 from the 2012–2013 Bangladesh Water Development Board campaign [37], was withheld from every stage of training, and wells sampled in both campaigns give 224 Bangladesh pairs. Bangladesh is used as a stress test because its Holocene aquifers are strongly reducing [26,38], so it lies outside the redox conditions that dominate the European and North American records.

Methods compared

Of the eight methods evaluated (Table 1), logistic regression on raw chemistry sets a linear floor. Spatial k-nearest-neighbour interpolation predicts the mean label of the 25 nearest training sites by great-circle distance and so represents what interpolation from location alone can recover. A raster-only random forest uses the 18 environmental rasters and no chemistry, following the design of global hazard maps [2,3]. Random forest and gradient boosting on raw chemistry represent strong conventional machine learning. TabPFN v2 [16], a Transformer pretrained on synthetic tabular tasks and applied without further training, represents generic pretraining with no groundwater data. GENESIS represents groundwater-specific pretraining and enters twice, once as a logistic probe on its frozen representation and once fine-tuned end to end, so that the value of pretraining can be separated from the value of task tuning.

GENESIS is a Transformer encoder [39] with pre-layer normalization [40], in which each measured parameter and each raster becomes a token whose embedding sums a parameter-type vector, a learned projection of the value and a water-body-type vector. Skewed parameters (18 of 20) were transformed as $\log(1 + x)$ before z-scoring, while pH and Eh were z-scored directly. The Base configuration used throughout has 8 layers, 8 attention heads and a width of 256, giving 8.99 million parameters. It was pretrained by masked reconstruction, in which 20% of the chemistry tokens of each sample are hidden and reconstructed from the remaining chemistry and the rasters, which are never masked, in two stages (mask ratio 0.20 for 100 epochs, then 0.25 for 50 epochs) with AdamW [41]. The pretraining validation set was 50,000 samples drawn at random (seed 42) at the sample level, and architecture and pretraining details are given in the Supplementary Material.

Table 1 lists the inputs, pretraining and evaluation axes of each method. Logistic regression, spatial interpolation and the GENESIS classifiers produce exceedance probabilities and enter the transfer and calibration axes only. The inferability and physical-consistency axes need regressors, for which GENESIS reconstructs the target by masking it and gradient boosting, random forest and TabPFN regress it directly.

A simulated quantum kernel enters as a ninth method on all four axes, as a support vector classifier for transfer and calibration and as kernel ridge regression for inferability and physical consistency. A quantum kernel encodes each sample as the state of a quantum circuit and measures the similarity of two samples as the overlap of their states, and a support vector machine then classifies with that similarity [42]. The 20 chemistry parameters were log-transformed as for GENESIS, median-imputed, standardized and reduced by principal component analysis to 12 components, one per qubit, which retained a median 83% of the variance. Each sample was encoded by two layers of Hadamard gates, each followed by single-qubit Z rotations and ZZ couplings between neighbouring qubits on a ring, an entangling circuit of the type introduced by Havlíček et al. [42]. All rotation angles were multiplied by a bandwidth factor, which counters the concentration of kernel values towards a constant as qubits are added [43,44]. The kernel was computed exactly by statevector simulation on a laptop CPU, which is tractable at 12 qubits and free of the sampling and gate noise of quantum hardware, so it is an optimistic estimate of what the same circuit would achieve on a device. Kernel support vector machines scale poorly with sample size, therefore each cell trained on a stratified subsample of 2,000 pairs from its training partition. The bandwidth (0.01–0.8) and the regularization constant (0.1, 1 or 10) were chosen by AUC on the validation partition, and decision values were mapped to probabilities by Platt scaling [45] on the same rows, a monotone mapping that leaves the conformal sets unchanged. Two controls received identical rows and features, namely the same circuit without the ZZ couplings, whose states are unentangled and whose kernel is classically trivial, and a classical support vector machine with a Gaussian (RBF) kernel. A quantum advantage would require the entangled kernel to outperform both. For regression the same three kernels were used in kernel ridge regression on the test rows of the other regressors, each fit drawing 2,000 training samples from the same training pool and 500 further samples on which the bandwidth and the ridge penalty (0.001–10) were chosen by R^2, with the inputs of each level reduced to 12 principal components as above. The other regressors trained on up to 150,000 samples, so a gap between the kernels and those regressors measures the training-size limit of kernel methods as well as the kernel, and only the comparison among the three kernels isolates the circuit.

Axis A, inferability from routine measurements

Inferability was measured as the accuracy with which each parameter could be reconstructed from progressively richer sets of other measurements, ordered to mirror how monitoring programmes add analyses. Level L0 supplies the rasters alone, L1 adds the field parameters pH and EC, L2 adds the major ions Ca, Mg, Na, K, Cl, HCO_3, SO_4 and TDS, L3 adds NO_3, F, SiO_2 and DOC, and L4 supplies every parameter except the target, with the rasters available at every level. Parameters outside a level were removed from the input rather than masked, which matches how the corpus is naturally incomplete.

Test rows had to be unseen by every model, and because pretraining split the corpus at the sample level, a site can contribute to both training and validation, so the test set was restricted to groundwater validation samples from sites with no training sample, giving 1,173 site-disjoint samples. Within each target, every level was scored on the same rows, those where the target and the major ions Ca, Mg, Na, Cl, HCO_3 and SO_4 were all measured. This left 85–333 rows per target for 19 of the 20 parameters, with Eh lacking test rows. The Bangladesh samples formed a second test set of up to 2,710 rows per target for 18 parameters, U and DOC being too sparse there. GENESIS reconstructed the masked target with its pretrained weights and no further training. Gradient boosting (300 iterations, learning rate 0.1) and random forest (200 trees, minimum leaf size 5, 30% row subsampling) were trained on up to 150,000 randomly drawn groundwater training samples with the same inputs. TabPFN received a random context of 10,000 of those samples, its design limit, so its results are a lower bound. Accuracy is the coefficient of determination R^2 in the normalized space of each parameter, with 95% intervals from 500 paired bootstrap resamples.

A minimum measurement set was derived for the seven contaminants and redox indicators by greedy forward selection. Starting from pH and EC, the parameter that most raised the reconstruction R^2 of the target was added at each step, up to four additions. To keep the search from inflating the reported accuracy, selection used 800 validation samples from sites that also appear in training, while scoring used the site-disjoint rows above.

Axis B, transfer to unseen regions

Transfer was evaluated by leave-one-region-out validation of exceedance prediction, with the temporal pairs assigned to 15 regions, namely the USA (NWIS), France (ADES), the European

Table 1. Methods compared, their inputs and pretraining, and the axes on which each was scored (A inferability, B transfer, C calibration, D physical consistency).

Method	Inputs	Pretraining	Axes
Logistic regression	chemistry	none	B, C
Spatial kNN ($k = 25$)	coordinates	none	B, C
Raster-only random forest	18 environmental rasters	none	B, C
Random forest	chemistry (+ rasters in A, D)	none	A–D
XGBoost / gradient boosting	chemistry (+ rasters in A, D)	none	A–D
TabPFN v2	chemistry (+ rasters in A, D)	synthetic tabular tasks	A–D
GENESIS, masked reconstruction	chemistry + rasters	2.1 M water samples	A, D
GENESIS, frozen probe	chemistry + rasters	2.1 M water samples	B, C
GENESIS, fine-tuned	chemistry + rasters	2.1 M water samples	B, C
Quantum kernel (simulated)[a]	12 principal components of chemistry (+ rasters in A, D)	none	A–D

[a] Support vector classifier in B and C and kernel ridge regression in A and D, each with two controls on identical rows and features, the same circuit without entanglement and a classical RBF kernel.

Union (EEA), eleven GEMStat country sets and Bangladesh. Seven targets were scored against the WHO guidelines for As (10 µg/L), F (1.5 mg/L), NO_3 (50 mg/L) and U (30 µg/L)[6], and 0.3 mg/L for Fe, 0.4 mg/L for Mn and 0.5 mg/L for PO_4. A target and region formed a cell when the held-out region held at least 10 exceedances and 10 non-exceedances and the remaining regions supplied at least 200 training pairs with at least 10 exceedances, giving 47 of the 105 possible cells. Of these, 34 are redox-sensitive (As, Fe, Mn, PO_4), 9 conservative (NO_3, F) and 4 uranium cells.

For each cell, the pairs outside the held-out region were partitioned into training (70%), validation (10%), calibration (10%) and in-distribution test (10%) sets, deterministically per target, region and seed, and the held-out region formed the out-of-region test set, with every method seeing the identical partition under three seeds (42, 43, 44). Random forest (400 trees, balanced class weights) and logistic regression (balanced class weights) used median imputation and standardization, XGBoost (400 trees, depth 6, learning rate 0.05) and TabPFN handled missing values natively, and TabPFN received a stratified context of 10,000 training pairs. Both GENESIS variants received the 18 rasters as context tokens, as in pretraining, whereas the other chemistry models used chemistry alone. The frozen GENESIS probe fitted a logistic regression to the pooled encoder representation. Fine-tuned GENESIS trained the full encoder and a classification head with AdamW (learning rate 2×10^{-5}, weight decay 0.01, batch size 128) on a class-weighted binary cross-entropy for up to 20 epochs, stopping when the validation partition failed to improve for four epochs. Skill is the area under the ROC curve (AUC) on the out-of-region test set.

A separate audit asked whether raster-only hazard models transfer in proportion to how well a parameter can be reconstructed from chemistry. For each of the 20 parameters a random forest (300 trees, minimum leaf size 5) was fitted on the 18 rasters to a site-level label marking values above that parameter's 80th percentile of site medians, which fixes prevalence at 20% for every parameter. Each parameter was subsampled to 3,000 sites in each of five repeats, so sample size was also matched. Leave-one-region-out AUC was then compared with the masked-reconstruction R^2 of the same parameter by Spearman correlation. The audit was repeated within single regions and single data sources, and against reporting limits, to separate a geochemical gradient from differences in sampling protocol.

Axis C, calibration under geographic shift

Calibration was tested with class-conditional (Mondrian) split conformal prediction[24,25] at a miscoverage level $\alpha = 0.10$. For each class c the nonconformity score is one minus the predicted probability of that class, and the threshold q_c is the $\lceil (n_c + 1)(1 - \alpha) \rceil$-th smallest calibration score among the n_c calibration pairs of class c. A class enters the prediction set when its score does not exceed q_c, so a set may contain "exceeds", "below", both or neither. Calibration used out-of-region pairs only, and coverage was checked twice, on the in-distribution test set where the guarantee should hold and on the held-out region where a model would be deployed. Mondrian calibration remains valid under a pure change in prevalence, so any loss of coverage in the held-out region indicates that the chemistry of exceedances itself differs between regions.

The quantity that matters for health is coverage of the exceedance class, because an exceeding sample whose set omits "exceeds" is a missed exceedance that the method claimed to cover. Results are reported as the miss rate, one minus that coverage, together with the fraction of ambiguous sets containing both classes, since a method can always lower its miss rate by declaring more samples ambiguous. As a remedy check, each method was recalibrated on k = 25, 50, 100 or 200 labelled pairs drawn from the held-out region (50 draws each) and scored on the remaining pairs of that region.

Axis D, physical consistency of reconstructed waters

Physical consistency was judged by whether reconstructed waters obey two constraints that no model was given, electroneutrality and mineral equilibrium. Each of Ca, Mg, Na, K, Cl, HCO_3, SO_4, F and pH was reconstructed in turn at level L4, and the reconstructed composition was evaluated by the same equations as the observed one, so approximation error in the equations cancels in the comparison. The charge-balance error is

$$\mathrm{CBE} = 100 \times \frac{\sum \text{cations} - \sum \text{anions}}{\sum \text{cations} + \sum \text{anions}}, \qquad (1)$$

with both sums in meq/L. Saturation indices of calcite, gypsum and fluorite used Davies activity coefficients at 25 °C and free-ion concentrations without ion pairs, with carbonate derived from HCO_3 and pH ($pK_2 = 10.33$) and log K_{sp} of −8.48 for calcite, −4.58 for gypsum and −10.6 for fluorite. Agreement is the R^2 between saturation indices computed from reconstructed and observed waters, and the median absolute CBE of reconstructed waters is compared with that of the observed ones. Samples were classed as reducing when Fe $\geq$ 0.1 mg/L or Mn $\geq$ 0.05 mg/L, a reduced form of the McMahon and Chapelle scheme[46] since dissolved oxygen is outside the vocabulary. The test rows were the 142 site-disjoint groundwater samples with complete major ions (48 of them reducing) and 2,000 Bangladesh samples drawn at random.

Statistical comparison

Cell metrics were averaged over the three seeds before any test, so each of the 47 cells contributes one value per method. Methods were compared by the two-sided Wilcoxon signed-rank test on paired cell values, coverage in and out of distribution by the

same test, and mechanism groups by a one-sided Mann–Whitney test. For axis A, methods were compared by the Wilcoxon test on paired per-target R^2. Confidence intervals on mean AUC and coverage are percentile intervals from 5,000 bootstrap resamples of cells.

Data availability

The fixed site-disjoint split and leave-one-region-out partitions, the per-sample predictions of every method, harmonized data tensors and the Bangladesh field data are archived at Zenodo (`https://doi.org/10.5281/zenodo.20631197`). Training data derive from public databases, namely GEMStat v3 (Zenodo record 18459694), USGS NWIS through the Water Quality Portal, French ADES through Hub'Eau and EEA Waterbase. Raw GEMStat records are subject to the UN GEMS/Water data agreement and are not redistributed.

Code availability

The analysis code, including the simulated quantum kernel, is available at `https://github.com/rakibhhridoy/genesis-arsenic-bangladesh`. The partitions are seeded from the target, region and seed, so they can be regenerated from the code alone.

Acknowledgements

The UNEP GEMS/Water Programme (GEMStat), the U.S. Geological Survey (NWIS, through the Water Quality Portal), the Bureau de Recherches Géologiques et Minières and the Hub'Eau platform (ADES), and the European Environment Agency (Waterbase) are gratefully acknowledged for openly available monitoring records. Environmental rasters were derived from public satellite products accessed through Google Earth Engine. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions

M.R.H. conceived the study, curated the data, developed the methods and software, performed the analysis, prepared the figures and wrote the original draft. S.Z. reviewed and edited the manuscript. A.Z. supervised the work, provided resources, and reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

References

[1] T. Gleeson, L. Cuthbert, G. Ferguson, and D. Perrone. Global groundwater sustainability, resources, and systems in the Anthropocene. *Annual Review of Earth and Planetary Sciences*, 48:431–463, 2020.

[2] J. Podgorski and M. Berg. Global threat of arsenic in groundwater. *Science*, 368(6493):845–850, 2020.

[3] J. Podgorski and M. Berg. Global analysis and prediction of fluoride in groundwater. *Nature Communications*, 13:4232, 2022.

[4] J. Podgorski and M. Berg. Manganese, a hidden threat to global water quality and health. *Nature Water*, 4:954–964, 2026.

[5] L. Rodriguez-Lado, G. Sun, M. Berg, et al. Groundwater arsenic contamination throughout China. *Science*, 341(6148):866–868, 2013.

[6] World Health Organization. *Guidelines for Drinking-water Quality*, 4th edition incorporating the 1st and 2nd addenda. WHO, Geneva, 2022.

[7] A. H. Smith, E. O. Lingas, and M. Rahman. Contamination of drinking-water by arsenic in Bangladesh: a public health emergency. *Bulletin of the World Health Organization*, 78(9):1093–1103, 2000.

[8] P. Ravenscroft, H. Brammer, and K. Richards. *Arsenic Pollution: A Global Synthesis*. Wiley-Blackwell, 2009.

[9] UNEP GEMS/Water. GEMStat – Global Water Quality Database, version 3. Zenodo, record 18459694, 2024.

[10] A. J. Tesoriero, S. A. Wherry, D. I. Dupuy, and T. D. Johnson. Predicting redox conditions in groundwater at a national scale using random forest classification. *Environmental Science & Technology*, 58(11):5079–5092, 2024.

[11] H. Cao, X. Xie, J. Shi, G. Jiang, and Y. Wang. Siamese network-based transfer learning model to predict geogenic contaminated groundwaters. *Environmental Science & Technology*, 56(15):11071–11079, 2022.

[12] H. Cao, X. Xie, Z. Xiao, W. Liu, and K. Feng. Cross-regional transfer learning for interpretable prediction of groundwater uranium contamination. *Environmental Science & Technology*, 60:16088–16098, 2026.

[13] Y. Zheng, X. Zhang, Y. Zhou, Y. Zhang, T. Zhang, and R. Farmani. Deep representation learning enables cross-basin water quality prediction under data-scarce conditions. *npj Clean Water*, 8:33, 2025.

[14] N. Mehrabi, D. A. Polya, and D. Han. Machine learning models of the geospatial distribution of groundwater quality: a systematic review. *Water*, 17:2861, 2025.

[15] R. Bommasani, D. A. Hudson, E. Adeli, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[16] N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister, F. Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025.

[17] P. Ploton, F. Mortier, M. Réjou-Méchain, N. Barbier, N. Picard, V. Rossi, et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11:4540, 2020.

[18] H. Meyer, E. Pebesma. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9):1620–1633, 2021.

[19] A. M. J.-C. Wadoux, G. B. M. Heuvelink, S. de Bruin, and D. J. Brus. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, 457:109692, 2021.

[20] H. Meyer, E. Pebesma. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13:2208, 2022.

[21] H. Cao, X. Xie, Z. Xiao, W. Liu. Transferability of machine learning models for geogenic contaminated groundwaters. *Environmental Science & Technology*, 58(20):8783–8791, 2024. doi:10.1021/acs.est.4c01327.

[22] L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

[23] D. McElfresh, S. Khandagale, J. Valverde, V. Prasad C, G. Ramakrishnan, M. Goldblum, and C. White. When do neural nets outperform boosted trees on tabular data? In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

[24] V. Vovk, A. Gammerman, G. Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[25] A. N. Angelopoulos, S. Bates. Conformal prediction: a gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

[26] R. T. Nickson, J. M. McArthur, P. Ravenscroft, W. G. Burgess, and K. M. Ahmed. Mechanism of arsenic release to groundwater, Bangladesh and West Bengal. *Applied Geochemistry*, 15(4):403–413, 2000.

[27] S. Fendorf, H. A. Michael, and A. van Geen. Spatial and temporal variations of groundwater arsenic in South and Southeast Asia. *Science*, 328(5982):1123–1127, 2010.

[28] P. L. Smedley and D. G. Kinniburgh. A review of the source, behaviour and distribution of arsenic in natural waters. *Applied Geochemistry*, 17(5):517–568, 2002.

[29] P. N. Froelich, G. P. Klinkhammer, M. L. Bender, N. A. Luedtke, G. R. Heath, D. Cullen, P. Dauphin, D. Hammond, B. Hartman, and V. Maynard. Early oxidation of organic matter in pelagic sediments of the eastern equatorial Atlantic: suboxic diagenesis. *Geochimica et Cosmochimica Acta*, 43(7):1075–1090, 1979.

[30] D. R. Lovley. Dissimilatory Fe(III) and Mn(IV) reduction. *Microbiological Reviews*, 55(2):259–287, 1991.

[31] J. Bowles, S. Ahmed, M. Schuld. Better than classical? The subtle art of benchmarking quantum machine learning models. arXiv:2403.07059, 2024.

[32] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70:1321–1330, 2017.

[33] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, J. Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems 32*, 2019.

[34] BRGM. ADES, Accès aux Données sur les Eaux Souterraines. Hub'Eau API, `https://hubeau.eaufrance.fr`, 2024.

[35] European Environment Agency. Waterbase, Groundwater quantitative and chemical status. `https://www.eea.europa.eu/data-and-maps/data/waterbase-water-quality`, 2024.

[36] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017.

[37] M. Shamsudduha, R. G. Taylor, M. I. Haq, S. Nowreen, A. Zahid, and K. M. U. Ahmed. The Bengal Water Machine: Quantified freshwater capture in Bangladesh. *Science*, 377(6612):1315–1319, 2022.

[38] P. Bhattacharya, A. H. Welch, K. G. Stollenwerk, M. J. McLaughlin, J. Bundschuh, and G. Panaullah. Arsenic in the environment: Biology and Chemistry. *Science of the Total Environment*, 379(2-3):109–120, 2007.

[39] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[40] R. Xiong, Y. Yang, J. He, et al. On layer normalization in the Transformer architecture. In *Proceedings of ICML*, pages 10524–10533, 2020.

[41] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *Proceedings of ICLR*, 2019.

[42] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, J. M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567:209–212, 2019.

[43] R. Shaydulin, S. M. Wild. Importance of kernel bandwidth in quantum machine learning. *Physical Review A*, 106:042407, 2022.

[44] S. Thanasilp, S. Wang, M. Cerezo, Z. Holmes. Exponential concentration in quantum kernel methods. *Nature Communications*, 15:5200, 2024.

[45] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, Cambridge, MA, 1999.

[46] P. B. McMahon, F. H. Chapelle. Redox processes and water quality of selected

principal aquifer systems. *Ground Water*, 46(2):259–271, 2008.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarypreprint.pdf