Frequently Asked Questions (FAQ)

Electromagnetic & Cone Penetration Test Data Fusion on Soil Characterization

Dimitrios Madelis, Marios Karaoulis, Philippe De Smedt — EGU 2026

This supplementary FAQ accompanies the EGU 2026 poster. It addresses the questions most likely to arise after viewing the poster, with concise answers and references to the full Master Thesis where deeper detail is provided. Citations follow the poster's reference list, with additional sources where relevant.

A. Study Area & Data

Q1. How many CPT locations were used in this study, and how does that compare to total availability in the area?

A. 199 CPT logs were available in the broader Schoneveld study area, retrieved from the Dutch DINOloket database (BRO-certified .gef files). Of these, 46 fell within the 50 m proximity threshold to HEM points and were therefore usable for fusion. This is acknowledged as a borderline sample size for Machine-Learning training and is one of the main reasons the study was framed around methodology validation rather than absolute prediction performance.

Q2. Why was Schoneveld (Zeeland) selected as the study area?

A. Schoneveld combines high local CPT density in Zeeland with the heterogeneous coastal lithologies that make HEM-CPT fusion most challenging — fluvial-marine deposits, freshwater lenses over saline groundwater, and levee infrastructure (van Baaren et al., 2018; Stafleu et al., 2011). It was deliberately chosen as a stress-test for the fusion framework, not the easiest case.

Q3. What HEM system was used and what are its key technical characteristics?

A. The data come from the Fugro RESOLVE airborne system used in the FRESHEM Zeeland project (van Baaren et al., 2018). RESOLVE operates at 6 frequencies between 286 Hz and 133 kHz, with horizontal coplanar (HCP) and vertical coaxial (VCX) coil configurations suspended ~40 m below the helicopter. Survey speed of ~140-150 km/h gives a sampling distance of ~4 m along the flight line, with line spacing of 100, 200, or 300 m depending on the survey area (Siemon et al., 2009; Siemon et al., 2019).

Q4. What does the HEM "footprint" mean and why does it matter for the choice of 50 m proximity threshold?

A. The HEM footprint is the lateral area of the subsurface that effectively contributes to a single measurement. As reported in the FRESHEM Zeeland technical report, the shallow-subsurface footprint of the RESOLVE system is approximately 50 m, increasing to ~200 m at greater depth (van Baaren et al., 2018). The 50 m proximity threshold for matching CPT to HEM points is chosen to align with this shallow footprint — beyond it, the CPT and HEM measurements no longer reflect the same subsurface volume.

Q5. Why was 15 m chosen as the analysis depth?

A. CPTs in this region routinely terminate at 15-20 m depth (the depth relevant to levee stability and shallow geotechnical engineering), so 15 m is a practical common ground where most CPT profiles still have data and where HEM resolution is at its best. Below this depth, CPT coverage drops sharply and the HEM footprint expands toward 200 m, making the matching assumption progressively less valid.

Q6. What is the Soil Behaviour Type Index (I_c) and why is it the prediction target?

A. I_c (Robertson, 1990) is a continuous index derived from CPT cone resistance and friction ratio that classifies soils by their in-situ mechanical behaviour rather than grain size. It maps onto the Robertson SBT chart with characteristic zones (clay, silty clay, silty sand, sand). It was chosen as the regression target because it is

continuous (allowing standard regression metrics), already lithology-relevant (mapping directly to soil classes), and standardised (comparable across CPT datasets globally).

Q7. How many lithology classes were used, and why those four?

A. The poster shows four classes (Robertson SBT zones 3-6: clay to silty clay, clayey silt & silty clay, silty sand to sandy silt, clean sands to silty sands). The remaining Robertson zones (1, 2, 7-9) were either absent or so underrepresented in the Schoneveld dataset that including them would have produced unstable classifiers. The four retained classes cover the full range of soils encountered in the levee-relevant depth interval (Robertson, 1990).

B. Methodology

Q8. Why use Random Forest and Neural Network specifically — and not other algorithms?

A. Both methods are well-established in geophysical-geotechnical fusion (Coelho & Karaoulis, 2022). Random Forest gives interpretable feature importances out-of-the-box and is robust to noise without extensive tuning. The Neural Network was included to evaluate whether non-linear, multi-layer representation learning offered an advantage on this dataset. Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel is acknowledged as a promising alternative for future work, particularly for classification of overlapping classes, but was outside the scope of this comparison.

Q9. How were hyperparameters selected for the Random Forest?

A. RandomizedSearchCV with 150 iterations across a wide grid: n_estimators (100-1000 in steps of 100), max_depth (5 to 50, plus unlimited), min_samples_split (2-20), min_samples_leaf (1-10), max_features ('sqrt', 'log2', 0.2, 0.5, 0.8), and bootstrap on/off (Madelis, 2025, §4). The inner cross-validation used GroupKFold to prevent leakage between training and validation folds.

Q10. How was the Neural Network architecture chosen?

A. A custom random search over fully connected architectures with Batch Normalization, ReLU activations, and Dropout layers. Hyperparameters explored: hidden layer sizes (32/64/128 in layer 1; 16/32/64 in layer 2), dropout rates (0.1, 0.2, 0.3), learning rates (1e-3, 5e-4, 1e-5), and batch sizes (32, 64, 128). The Adam optimizer was used, with early stopping triggered when validation loss did not improve over a 20-epoch patience window.

Q11. What is GroupKFold cross-validation and why was it used here?

A. GroupKFold ensures that all samples sharing the same "group" (in this case, all data points belonging to the same CPT profile) appear in either the training fold or the validation fold — never both. This is essential for spatial data: a standard random split would put data points from the same vertical CPT profile on both sides of the train/test boundary, producing artificially optimistic performance through subtle spatial leakage. GroupKFold (here with 5 folds, grouped by CPT profile name) gives a realistic estimate of how the model generalises to genuinely unseen locations.

Q12. What features were used as ML inputs?

A. Five features per HEM-CPT point: X coordinate, Y coordinate, depth (m), electrical resistivity (Ohm·m), and distance to coastline (m). Coordinates and depth provide spatial context; resistivity is the primary geophysical signal; coastline distance encodes the salinity/lithology gradient that strongly conditions resistivity in coastal aquifers (van Baaren et al., 2018). Coastline distance is computed via an alpha-shape coastline polygon and a per-point Euclidean distance calculation.

Q13. What does the CPT-to-HEM interpolation actually do?

A. The interpolation is the operation that makes the entire fusion possible. It exists to solve a fundamental geometric mismatch between the two datasets: CPT data are vertically continuous and dense (I_c measured

every ~1 cm down a single profile, typically to 15-20 m depth), while HEM data are horizontally dense but vertically coarse (a 1D inversion produces only 20 layers, with thicknesses ranging from 0.5 m at the surface to several metres at depth). A given CPT profile and the nearest HEM sounding therefore describe the same column of subsurface but at completely different sampling rates — the CPT might have 1500 I_c values over 15 m, while the HEM produces 20 resistivity values over the same interval.

This mismatch makes direct fusion impossible. A machine learning model that takes resistivity as input and predicts I_c needs paired (resistivity, I_c) examples at the *same depth*; without that, there is nothing to learn from. The interpolation creates these pairs: for each HEM layer in the matched region, the smoothed CPT I_c profile is linearly interpolated onto the layer midpoint, producing exactly one I_c value per HEM layer per matched location. The result is a labelled training dataset with consistent dimensionality — every HEM point now has a feature vector (X, Y, depth, resistivity, coastline distance) and a target (I_c, lithology class) at the same depth.

The choice of linear interpolation (rather than nearest-neighbour or higher-order schemes) is deliberate: linear preserves the smooth lithological transitions that the Savitzky-Golay filter is meant to retain, while avoiding the over-smoothing that splines would introduce at sharp clay-sand boundaries. The Savitzky-Golay step before interpolation is also essential — it removes the centimetre-scale CPT spikes that would otherwise create noisy interpolated values that misrepresent true soil behaviour.

Q14. Why a Savitzky-Golay filter on the CPT data and not a simpler moving average?

A. CPT measurements often contain centimetre-scale spikes from gravel encounters or instrument noise that do not reflect true soil behaviour. Savitzky-Golay smooths these spikes by fitting low-degree polynomials in moving windows, preserving the shape of genuine lithological transitions far better than a moving average — which would systematically blunt the boundaries between clay and sand units (Savitzky & Golay, 1964). This matters because the regression target is precisely those I_c transitions.

Q15. How was class imbalance in the lithology classification handled?

A. Class 4 (clayey silt & silty clay) dominates the dataset, while classes 3 and 6 are much less represented. Manual class weights were applied during training to compensate, in proportion to inverse class frequency. This is a partial mitigation — the more substantial limitation, addressed in the conclusions, is that adjacent Robertson SBT classes overlap in their I_c values, so even a perfect classifier would have a non-zero confusion rate at boundaries.

Q16. How was data leakage between training and test prevented?

A. Two safeguards: (1) the train/test split is profile-level — all rows belonging to one CPT profile are assigned to either train or test, never split between them; and (2) the same fixed split (saved as train_names.npy / val_names.npy) is reused for both regression and classification, so a CPT profile cannot leak from one task into the other. This is more conservative than typical row-level splits in Machine Learning benchmarks.

C. Results & Interpretation

Q17. What is the actual Neural Network performance on regression, and how does it compare to Random Forest?

A. On the held-out validation set: Random Forest achieved RMSE = 0.29, the Neural Network RMSE = 0.33. The poster shows RF results because they are stronger, but the NN is competitive and would likely close the gap with a larger training set. The same is not true in classification, where the NN clearly outperforms RF.

Q18. What is the Random Forest classification accuracy, and why does the Neural Network win this task?

A. Random Forest reached overall classification accuracy of ~43%, the Neural Network ~56%. The NN's advantage is consistent with its ability to learn smoother non-linear decision boundaries between

overlapping classes — important here because Robertson SBT classes are not crisply separated in Ic space, especially at the silty clay / clayey silt and silty sand / sandy silt boundaries.

Q19. RMSE = 0.29 corresponds to $R^2 \approx 0.57$ — is this a strong result?

A. In context, yes. The I_c range in the dataset is approximately 1.5-3.5, so RMSE = 0.29 represents ~14.5% of the data range and is below the typical width of Robertson SBT classification zones (~0.35-0.55 in I_c). This means most regression predictions fall within the correct lithological class, which is the operationally relevant criterion. The modest R^2 is largely attributable to the noise in the underlying CPT measurements rather than model limitations.

Q20. Where does the model perform worst, and why?

A. Two main failure modes are documented: (i) sharp lithological transitions where resistivity changes abruptly between adjacent HEM layers, since linear interpolation across the boundary smears the true I_c step; and (ii) intervals with sparse CPT coverage (fewer than ~10 CPT points contributing to that depth range), where the model has insufficient examples to learn stable resistivity-Ic relationships. Both are honest data limitations rather than algorithmic failures.

Q21. Why are misclassifications concentrated at the boundaries between adjacent Robertson SBT classes?

A. The Robertson chart partitions a continuous physical quantity (I_c) into discrete classes at fixed thresholds. Soils whose I_c happens to sit near a class boundary are physically intermediate and behave intermediately — the model is correct in placing them near the boundary, but the discrete metric counts them as errors. This pattern reflects the underlying continuity of soil behaviour rather than a model weakness (Robertson, 1990; Robertson & Cabal, 2022).

D. Method Validity, Limitations & Future Work

Q22. How does this study compare to Coelho & Karaoulis (2022), the closest published work?

A. Coelho & Karaoulis (2022) established the foundational framework of fusing CPT and resistivity data with ML (RF and NN) to produce 3D subsoil schematisations — and already used a CPT-level 80/20 split to avoid data leakage. The present study extends that framework in three concrete ways: (i) GroupKFold cross-validation (5 folds, grouped by CPT profile) replaces the single 80/20 split, providing a more robust performance estimate less dependent on a single random draw; (ii) an explicit classification task (Robertson SBT lithology classes) is added alongside regression — Coelho & Karaoulis (2022) predict only continuous I_c, not discrete lithological classes; and (iii) coastline distance replaces the geomorphological map as a spatial feature, which is more physically motivated for coastal salinity gradients and eliminates the need for an external geological dataset. Additionally, this study applies the framework to airborne HEM data (not ground-based ERT), which covers a far larger area but requires a more complex vertical interpolation step to align the coarse HEM layer structure with the fine-resolution CPT profiles (Coelho & Karaoulis, 2022).

Q23. How far from a CPT can predictions be trusted? What is the spatial decorrelation distance?

A. This is the open question highlighted in the poster's take-home message and remains unanswered quantitatively. The conservative interpretation is that prediction reliability degrades beyond the HEM shallow footprint (~50 m), since beyond that distance the resistivity at the prediction location and the resistivity that contributed to the matched CPT-HEM training pair come from different subsurface volumes. Quantifying this rigorously — for instance via leave-one-CPT-out experiments parametrised by distance — is the planned next step.

Q24. Could the framework transfer to other coastal areas with different geology?

A. The framework is designed to be transferable: the inputs (HEM resistivity + spatial features) and outputs (I_c, lithology) are universal, and Robertson's SBT classification was developed precisely to be site-independent. However, the trained model weights are not transferable — every new region requires its own training data because the local relationship between resistivity and Ic depends on porosity, mineralogy, and porewater chemistry. Validating the framework (not the trained models) on a second coastal area is identified as the priority for future work.

Q25. What are the practical implications for levee safety and coastal water management?

A. The framework allows continuous 3D maps of geotechnical properties to be produced from existing HEM coverage, with much higher spatial density than CPT alone permits. For levee assessment this means: (i) identifying potentially weak silty/clayey horizons at depth across entire embankments rather than at sparse CPT locations; (ii) targeting future CPT campaigns to areas of highest predictive uncertainty rather than uniform spacing; and (iii) integrating the resulting models into existing engineering workflows for stability and seepage analysis (van Baaren et al., 2018; Stafleu et al., 2011).

References

Coelho, Z. B., & Karaoulis, M. (2022). Data fusion of geotechnical and geophysical data for three-dimensional subsoil schematisations. *Advanced Engineering Informatics*, 53. https://doi.org/10.1016/j.aei.2022.101671

Robertson, P. K. (1990). Soil classification using the cone penetration test. *Canadian Geotechnical Journal*, 27(1), 151-158.

Robertson, P. K., & Cabal, K. L. (2022). *Guide to Cone Penetration Testing for Geotechnical Engineering* (7th ed.). Gregg Drilling.

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8), 1627-1639.

Siemon, B., Christiansen, A. V., & Auken, E. (2009). A review of helicopter-borne electromagnetic methods for groundwater exploration. *Near Surface Geophysics*, 7(5-6), 629-646.

Siemon, B., van Baaren, E., Dabekaussen, W., Delsman, J., Dubelaar, W., Karaoulis, M., & Steuer, A. (2019). Automatic identification of fresh-saline groundwater interfaces from airborne electromagnetic data in Zeeland, the Netherlands. *Near Surface Geophysics*, 17(1), 3-25. https://doi.org/10.1002/nsg.12028

Stafleu, J., Maljers, D., & Gunnink, J. L. (2011). 3D modelling of the shallow subsurface of Zeeland, the Netherlands. *Netherlands Journal of Geosciences*, 90(4), 293-310.

van Baaren, E., Delsman, J., & Karaoulis, M. (2018). *FRESHEM Zeeland Technical Report*. Deltares, 1209220-000-BGS-0030.

Contact

Dimitrios.Madelis@UGent.be | mkaraoulis@geo.auth.gr | Philippe.DeSmedt@UGent.be