



since 1961



Baltica

BALTICA Volume 39 Number 1 June 2026: 37–40

<https://doi.org/10.5200/baltica.2026.1.4>

Brief report

Artificial intelligence in hydrogeology: applications and an open, reproducible machine learning course

*Vytautas Samalavičius**, *Ilya Zaslavsky*

Samalavičius, V., Zaslavsky, I. 2026. Artificial intelligence in hydrogeology: applications and an open, reproducible machine learning course. *Baltica* 39 (1) 37–40. Vilnius. ISSN 0067-3064.

Manuscript submitted 4 January 2026 / Accepted 5 February 2026 / Available online 5 March 2026

© Baltica 2026

Abstract. Artificial intelligence (AI) is increasingly used in hydrogeological analysis for forecasting groundwater levels and properties, discovering hydrochemical facies and anomalies, and modelling spatio-temporal signals. Current practice emphasizes spatio-temporal validation, interpretability, and uncertainty quantification, while generative AI (GenAI) is emerging to accelerate data curation, documentation, augmentation, and code prototyping. This report summarizes representative applications of AI and machine learning (ML) in hydrogeology and describes an open, reproducible course developed for university students and professional hydrogeologists. This continuously updated online course is one of the outcomes of the GRANDE-U “Groundwater Resilience Assessment through Integrated Data Exploration for Ukraine” project. It covers Python fundamentals, environment setup, data engineering, and documented case studies in time-series and spatio-temporal modelling. Application examples are drawn from recent teaching materials and GRANDE-U and related studies and include predicting terrestrial water-storage anomalies, groundwater-level forecasting, isotope estimation from routine chemistry, risk or prospectivity mapping, and automated lineament/facies mapping.

Keywords: machine learning; groundwater; spatial data; temporal data

✉ *Vytautas Samalavičius (vytautas.samalavicius@chgf.vu.lt),  <https://orcid.org/0000-0001-5791-7483>;

Department Hydrogeology and Engineering Geology, Vilnius University, Vilnius, Lithuania;

Ilya Zaslavsky (izaslavsky@ucsd.edu),  <https://orcid.org/0000-0003-4191-8275>;

San Diego Supercomputer Center, University of California San Diego, La Jolla, CA, USA

*Corresponding author

INTRODUCTION

Hydrogeological investigations increasingly require the fusion of heterogeneous observations – borehole logs, pumping tests, hydrochemistry, geophysics, remote sensing, and climatic data – into decision-ready products.

Machine learning (ML), a subfield of artificial intelligence (AI), is particularly amenable to integration tasks through both supervised and unsupervised paradigms, with deep learning (DL) extending these capabilities to complex spatial grids and temporal sequences. In *supervised learning*, models are trained on input–output pairs to approximate a mapping from features to a specified target. This formulation encompasses *classification*, where the target is categori-

cal (e.g., assigning an observation to a discrete class), and *regression*, where the target is continuous (e.g., predicting a numerical value). Model training proceeds by minimizing a loss function that quantifies discrepancy between predictions and ground-truth labels, with performance commonly assessed using accuracy, F1, AUC, and other metrics for classification, or the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) for regression. Supervised methods are appropriate when clearly defined prediction goals and reliable labelled datasets are available.

By contrast, *unsupervised learning* seeks to characterize latent structure in data without recourse to labelled outcomes. Key tasks include *clustering*, which partitions observations into internally coherent

groups; *dimensionality reduction*, which projects data into lower-dimensional manifolds while preserving salient geometric or statistical properties; and *anomaly (or novelty) detection*, which identifies instances that deviate from learned regularities. Evaluation in this setting relies on internal validity indices, stability analyses, and downstream utility – namely, whether the discovered structure improves subsequent supervised modelling or supports expert interpretation – rather than on direct label-based metrics.

Deep learning provides a flexible function-approximation framework applicable to both supervised and unsupervised settings. Convolutional architectures are effective for signals with local spatial correlations (e.g., images, geospatial rasters), while recurrent and transformer-based architectures capture long-range dependencies in sequences and time series.

Generative AI (GenAI) refers to models that learn an approximation of the data-generating distribution and can thus synthesize new, high-fidelity samples that resemble the training data.

In teaching and practice, the recommended pipeline comprises problem definition, data collection, exploratory analysis, preprocessing, algorithm selection, cross-validated training, evaluation, hyperparameter tuning, interpretation, and deployment, with documentation at each stage.

The purpose of this brief report is to summarize current applications of artificial intelligence and machine learning in hydrogeology, with an emphasis on methods for analyzing spatial, temporal, and spatio-temporal data. It outlines key supervised, unsupervised, deep learning, and emerging generative AI approaches, highlighting representative use cases, validation practices, and interpretability considerations relevant to groundwater research and practice.

In addition, the report documents an open, reproducible machine-learning course developed for hydrogeologists. By linking methodological concepts to documented case studies and browser-based instructional materials, the report aims to support education and capacity building and to promote the rigorous, transparent, and reproducible use of AI methods in hydrogeological studies.

AI applications in hydrogeology: current practice and emerging roles for generative AI

In supervised classification, typical targets include groundwater potential, vulnerability, and risk classes. Predictors commonly include terrain attributes, lithology, land cover, climatic indices, and distances to hydrologic features. Various ML models (e.g., random forests) serve as strong baselines for tabular predictors and can be calibrated to produce probability maps. In regression, common targets include

hydraulic properties (e.g., hydraulic conductivity and storage), recharge, groundwater heads, and isotope geochemistry.

The models trained with spatially blocked cross-validation often outperform single-equation formulas and quantify the marginal value of key predictors such as grain-size percentiles (Araya, Ghezzehei 2019; Rehman *et al.* 2022; Samalavičius *et al.* 2024). Isotope content from routine chemistry augmented with coordinates and depth has demonstrated high predictive skill when evaluated on held-out regions or even entire basins, providing the reconnaissance of recharge patterns at moderate cost and supporting spatial gap-filling by predicting values in unsampled areas (Cemek *et al.* 2022; Samalavičius *et al.* 2025).

Time-series and spatio-temporal modelling are key use cases in hydrogeology. The feature-based pipelines that use lagged meteorological drivers (e.g., precipitation and evapotranspiration), seasonal encodings, groundwater abstraction records, and river stage data provide reproducible baselines for groundwater-level forecasting. With sufficiently long and information-rich records, ML and DL algorithms often outperform traditional approaches under standard error metrics (Feng *et al.* 2024). Remote sensing-based fusion (Solovey *et al.* 2025a, b) extends these ideas to GRACE (Gravity Recovery and Climate Experiment)-derived terrestrial and groundwater waters to range anomalies, where monthly precipitation, evapotranspiration, and soil-moisture proxies act as predictors. To avoid overly optimistic performance estimates, evaluation should use rigorous spatial and temporal block splits that keep training and testing separated in both space and time. This setup also makes it easier to interpret which predictors matter most (for example, through feature importance analysis) and to identify dominant hydro-climate controls. These approaches show strong potential for estimating groundwater storage from GRACE-type gravity observations, including AI-based downscaling that translates coarse monthly signals into higher-resolution maps by combining satellite products with hydro-climate and land surface predictors.

Generative AI (GenAI) constitutes the next pragmatic step in this toolkit. Large language models can accelerate code authoring, documentation, and unit-test generation; model-card drafts and plain-language summaries can improve reproducibility and communication. In data engineering, GenAI can assist with harmonizing variable names and units, writing parsers for heterogeneous spreadsheets or PDFs, and annotating coordinate reference systems and provenance. Generative models (for example, diffusion and transformer-based approaches) can support gap-filling and super-resolution for gridded datasets and controlled

scenario augmentation for rare events, provided synthetic outputs are clearly labelled, traceable, and excluded from validation sets to prevent leakage.

ML course for hydrogeologists

To demonstrate these practices and encourage their adoption, an open, continually updated course has been assembled for senior undergraduates, graduate students, and professionals. Materials are hosted as Google Colab notebooks to minimize setup burden and ensure reproducibility through pinned environments and seeded runs. Each module includes runnable examples designed to be adapted to local datasets. The syllabus comprises: (a) introduction to Python programming with scientific Python idioms; (b) environment setup, including Colab configuration, package pinning, and lightweight version control; (c) data engineering, covering ingestion and validation of tabular, raster and time-series sources, reconciling coordinate reference systems and units, performing raster-vector joins, feature engineering (lags, encodings, distances), and quality assurance/quality control (QA/QC); and (d) documented case studies. Case studies include multivariate groundwater-level forecasting with baseline and deep sequence models, spatio-temporal prediction of GRACE-derived water-storage anomalies from climate and land-surface drivers, isotope modelling using routine ions plus spatial context, and sinkhole formation risk assessment.

The course is designed as a reproducible, browser-based laboratory where students can modify parameters, rerun models, and explore interpretability methods without requiring local installation. The pedagogical design of this course draws on recent developments in spatial data science and GeoAI education, emphasizing project-based, reproducible workflows for geospatial analysis.

Spatial Data Science integrates geographic information systems, statistics, and machine learning to extract knowledge from location-referenced data. It explicitly accounts for spatial relationships, autocorrelation, coordinate systems, and geovisualization, which are essential elements in environmental and hydrogeological modelling. Spatial data are inherently interdependent, requiring models and validation strategies that respect geographic structure and scale (Vierø, Szell 2024). Robust experimental design – such as spatially blocked or temporally stratified cross-validation, or leave-location-out sampling – reduces spatial and temporal leakage and provides more reliable performance estimates for environmental prediction.

Pedagogical approaches to spatial data science and GeoAI emphasize project-based, practice-driven integration of geospatial and computational methods

(Gao *et al.* 2023; Zaslavsky, Kudinov 2021). Many university courses now combine Python, GIS, and ML through open digital environments such as Colab or Jupyter, interactive maps, and reproducible workflows. Incorporating GeoAI pedagogy into hydrogeological education ensures that students not only learn model architecture and evaluation metrics but also acquire spatial thinking, validation discipline, and the ability to implement robust spatial analysis workflows, which are critical competencies for responsible AI use in Earth and environmental sciences.

The course is continuously updated: notebooks are versioned and periodic release snapshots are archived for stable citation and teaching, and new modules are added as additional studies are completed, with consolidated links and webinar resources available at <https://www.grande-u.org/tasks/task-3-ai-modeling>. The course was developed as one of the GRANDE-U (“Groundwater Resilience Assessment through Integrated Data Exploration for Ukraine”) project tasks and is maintained by the Department of Hydrogeology and Engineering Geology, Vilnius University, Lithuania, and the San Diego Supercomputer Center at the University of California San Diego, USA.

CONCLUSIONS

AI supports many hydrogeological tasks: mapping, forecasting, risk assessment, hydrochemical interpretation, and image-based delineation. Deep models address complex spatio-temporal structure; generative AI is poised to act as a productivity and data-quality multiplier when deployed under rigorous validation and governance. The open, dynamic course outlined here lowers the entry barrier by providing documented, browser-based notebooks that convert these concepts into reproducible practice.

ACKNOWLEDGEMENTS

The GRANDE-U “Groundwater Resilience Assessment through Integrated Data Exploration for Ukraine” project (NSF Award 2409395) unites researchers from six countries – the U.S., Ukraine, Poland, Lithuania, Latvia, and Estonia. Vilnius University has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-IMPRESU-24-3.

REFERENCES

- Araya, S.N., Ghezzehei, T.A. 2019. Using Machine Learning for Prediction of Saturated Hydraulic Conductivity and Its Sensitivity to Soil Structural Perturbations. *Water Resources Research* 55, 5715–5737. <https://doi.org/10.1029/2018WR024357>

- Cemek, B., Arslan, H., Küçüktopcu, E., Simsek, H. 2022. Comparative analysis of machine learning techniques for estimating groundwater deuterium and oxygen-18 isotopes. *Stochastic Environmental Research and Risk Assessment* 36, 4271–4285. <https://doi.org/10.1007/s00477-022-02262-7>
- Feng, F., Ghorbani, H., Radwan, A.E. 2024. Predicting groundwater level using traditional and deep machine learning algorithms. *Frontiers in Environmental Science* 12. <https://doi.org/10.3389/fenvs.2024.1291327>
- Gao, S., Hu, Y., Li, W. (Eds) 2023. *Handbook of Geospatial Artificial Intelligence*, 1st ed. CRC Press. <https://doi.org/10.1201/9781003308423>
- Rehman, Z. ur, Khalid, U., Ijaz, N., Mujtaba, H., Haider, A., Farooq, K., Ijaz, Z. 2022. Machine learning-based intelligent modeling of hydraulic conductivity of sandy soils considering a wide range of grain sizes. *Engineering Geology* 311. <https://doi.org/10.1016/j.enggeo.2022.106899>
- Samalavičius, V., Vanhala, E.K.-M., Lekstutytė, I., Gadeikienė, S., Gadeikis, S., Žaržojus, G. 2024. Hydraulic conductivity determination of Lithuanian soils using machine learning. *Baltica* 137–150. <https://doi.org/10.5200/baltica.2024.2.5>
- Samalavičius, V., Gadeikienė, S., Žaržojus, G., Gadeikis, S., Lekstutytė, I. 2025. Oxygen-18 prediction using machine learning in the Baltic Artesian Basin groundwater. *Stochastic Environmental Research and Risk Assessment* 39, 765–787. <https://doi.org/10.1007/s00477-024-02896-9>
- Solovey, T., Śliwińska-Bronowicz, J., Janica, R., Brzezińska, A. 2025a. Terrestrial water storage changes in the Bug river transboundary catchment observed by GRACE and water balance analysis. *Meteorology, Hydrology and Environmental Monitoring* 2024, 4–16. <https://doi.org/10.15407/Meteorology2024.06.004>
- Solovey, T., Śliwińska-Bronowicz, J., Janica, R., Brzezińska, A. 2025b. Assessment of the Effectiveness of GRACE Observations in Monitoring Groundwater Storage in Poland. *Water Resources Research* 61. <https://doi.org/10.1029/2024WR038888>
- Vierø, A.R., Szell, M. 2024. Defining spatial data science as an interdisciplinary field. *Geo: Geography and Environment* 11. <https://doi.org/10.1002/geo2.125>
- Zaslavsky, I., Kudinov, D. 2021. Teaching Spatial Data Science and Deep Learning. In: *GIS for Science: Maps for Saving the Planet*, Vol. 3. Redlands, CA: Esri Press, 172–181 pp.