

Process-informed spatial prediction for soil mapping: towards hybrid modelling frameworks for Earth science applications

Marijn van der Meij¹, Madlene Nussbaum¹

¹ Department of Physical Geography, Utrecht University, The Netherlands

Recent advances in artificial intelligence (AI) and machine learning have led to rapid progress in spatial prediction and mapping across the geosciences. Data-driven approaches, often based on Earth observation data and large environmental covariate sets, have demonstrated strong predictive performance in a wide range of applications, including geological mapping, mineral exploration, and soil property prediction.

However, these approaches also face fundamental limitations, particularly when applied in situations characterized by sparse observations, complex processes, and strong spatial heterogeneity. Especially for soil maps, predictions created by data-driven machine learning often suffer from inconsistencies. Interrelated soil properties showing typical vertical distributions are predicted one-by-one resulting in inconsistent sets of properties. Such inconsistencies introduce significant uncertainties into subsequent map applications and can lead to unrealistic outcomes, particularly when dynamic processes are wrongly specified due to unrealistic soil inputs.

This project proposes the development of a process-informed machine learning framework that integrates mechanistic knowledge with advanced machine learning techniques to generate consistent, high-resolution global soil maps. To provide global process-information to the machine learning, an application oriented numerical soil classification is developed based on new similarity metrics to compare full soil profiles. This results in a soil profile reference library, which is verified and expanded with expert knowledge. The reference library then informs two novel ML approaches for global space-time mapping: loss function-constrained and transformer-based machine learning. The newly created global soil maps are used as inputs to land surface model to simulate biomass influenced by soil nutrient dynamics. The effect of existing inconsistent and the new consistent soil maps are benchmarked.

By combining process-informed modeling with advanced uncertainty quantification, the project aims to move beyond purely data-driven GeoAI toward approaches that are both scientifically grounded and operationally useful, particularly relevant in contexts where data are limited and process understanding remains essential.