EGU26-20968
EGU General Assembly 2026

Conditional Subsampling of Legacy Boreholes for Subsurface Model Validation

Pablo De Weerdt[1], Stijn Luca[2], and Ellen Van De Vijver[1]
[1]Department of Environment, Ghent University, Ghent, Belgium
[2]Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium

Practical constraints often force modellers to rely on legacy data rather than targeted new data collection relying tailored sampling design for subsurface modelling. While these pre-existing datasets enable model development and gap identification, their spatial density and distribution may not always meet the desired resolution or precision. Consequently, strategic subsampling for calibration and validation is essential to ensure a robust and accurate performance assessment of the resulting models. While cross-validation techniques are commonly applied to maximize data utility, their application in spatial modelling yields overoptimistic performance estimates with high variance, particularly when data are clustered. Probabilistic-based sampling is known to tackle bias, but its effectiveness remains poorly understood for spatially sparse and clustered legacy data.
This research evaluates the impact of subsampling methods on the validation of spatial interpolation techniques. Conditional versus random subsampling is compared for different subsample sizes in terms of actual model performance with particular attention to geostatistical concepts that additionally take into account spatial autocorrelation within subsurface data. Legacy boreholes spanning over a century with sparse and clustered spatial distribution were queried to model peat content in 3D. Conditioning relied on 2D legacy attributes such as age, spatial coordinates, and target feature statistics. We also investigated how the complexity of spatial variation (represented in different models with varying anisotropic autocorrelation) influenced performance by populating the existing borehole configuration with three 3D target features: two more spatially continuous synthetic and one heterogeneous, real field dataset. First results suggest that variance of validation results reduced exclusively in the heterogeneous case, provided the validation subset was large enough (35%) to incorporate the cumulative peat content within a borehole as a 2D attribute. These results underscore the resilience of conditioned probabilistic subsampling over alternative validation methods for legacy-based modelling.