Geophysical Research Abstracts
Vol. 20, EGU2018-4357, 2018
EGU General Assembly 2018

Determining Hydrogeological Site Similarity using Hierarchical Agglomerative Clustering

Nura Kawa (2), Karina Cucchi (3), Yoram Rubin (3), and Falk Heße (1)

(1) Department of Computational Hydrosystems, Helmholtz Center for Environmental Research - UFZ, (2) University of California Berkeley, Department of Statistics, (3) University of California Berkeley, Department of Civil and Environmental Engineering

Hydrogeological site characterization of an aquifer is both difficult and costly due to large variability of many subsurface properties. However, to effectively manage groundwater resources, a good representation of such properties, i.e. the hydraulic conductivity, is necessary. In such a sparse data context, it is necessary to use all available data to reduce the uncertainty in groundwater flow and transport predictions. Bayesian methods optimally suited to provide a framework wherein heterogeneous data sources can be joined to represent the available knowledge base on a given situation. This is achieved by distiguishing between two different sources of data. First, case specific data which are represented in the likelihood and, second, background data which are represented in the prior distribution. Using Bayes' theore, these two distributions can be joined into a full knowledge representation.

Such background knowledge should come from ex-situ sites to avoid correlations with in-situ measurements that are available. A prior distribution should therefore be based on data transferred from other, already sampled, sites. To reduce the uncertainty in the inferred prior distirbution as much as possible, the sites used for the information transfer should be as similar as possible to the site under investigation. To that end, a notion of site similarity is necessary.

In this study, we introduce a data-driven notion of site similarity in order to provide guidelines for the selection of relevant sites, replacing the traditional literature review with tools from machine learning. Our approach uses hierarchical agglomerative clustering to categorize sites into groups based on observable characteristics, such as environment type or rock type. We illustrate the value of the methodology by applying it to the construction of informative priors for the distribution of hydraulic conductivity, using data from a large open-source data base; the World-Wide HYdrogeological Parameters DAtabase (WWHYPDA).

We find that the use of such a data-driven notion of site similarity improves the predictive ability of our prior distribution in most of the investigated cases. While some cases show no improvement, we also saw an increase in predictive uncertainty in some other cases, which points towards necessary improvements in both the used database as well as our hierachcial Bayesian model used for deriving the prior distribution. We conclude that there is now a unique opportunity to combine hydrogeological experience with large-scale databases and data-driven methods to improve the predictive capability of geostatistical modeling. Additionally, encouraging the increased adoption of open-source hydrogeological databases would increase the size, quality and availability of hydrogeological data for future studies.