H35M-1173 A new framework for experimental design using Bayesian Evidential Learning: the case of wellhead protection area

Robin Thibaut - Ghent University (Belgium) - PhD Student - robin.thibaut@ugent.be | Prof. Dr. Thomas Hermans - Ghent University (Belgium) | Dr. Eric Laloy - Belgian Nuclear Research Centre (SCK-CEN)

1. Create the groundwater model for the predictor and target. Sequential Gaussian Simulation is used to generate hydraulic conductivity (K) fields. A variogram model defines spatial correlation of the log10 K field. Initially, only the log10 K field's mean value is unknown.

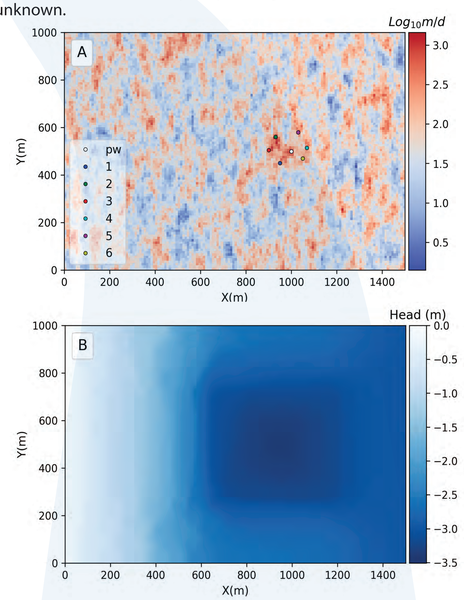

Figure1: **A**. Hydraulic conductivity field in 10-logarithmic base. The pumping well (pw) is located at (x, y) = (1000, 500)m and is surrounded by 6 injection wells. **B**. Flow solution. The direction of the natural gradient is from West to East.

2. Forward modeling: we use the model to create both the predictor and the target for our machine learning experiment. (Left: predictor; right: target). The Traveling Salesman Algorithm is used to sort the coordinates of each particle endpoint, and the resulting contour is used to compute the signed distance (SD) field of each target (Figure 3A-B).

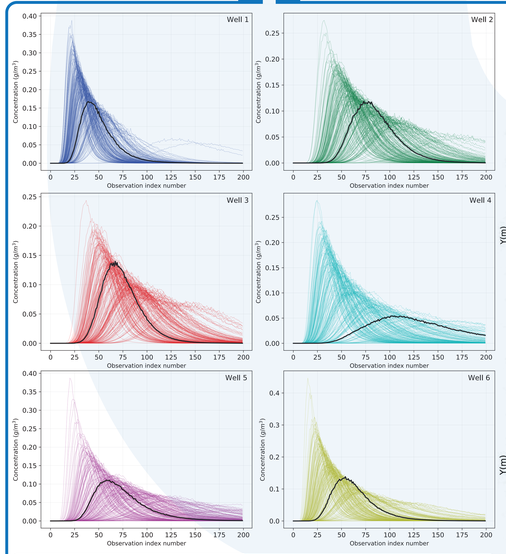

Figure 2: Predictor: breakthrough curves of tracers from each injection well, for both training and test (single sample) sets. They are all discretised and interpolated in 200 steps.

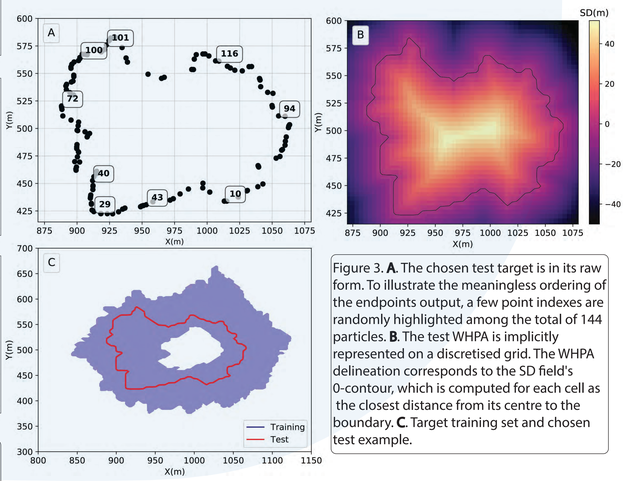

Figure 3. **A**. The chosen test target is in its raw form. To illustrate the meaningless ordering of the endpoints output, a few point indexes are randomly highlighted among the total of 144 particles. **B**. The test WHPA is implicitly represented on a discretised grid. The WHPA delineation corresponds to the SD field's 0-contour, which is computed for each cell as the closest distance from its centre to the boundary. **C**. Target training set and chosen test example.

3. Principal Component Analysis (PCA) is used to aggregate the correlated variables into a few independent Principal Components (PC's) for both the target and the predictor.

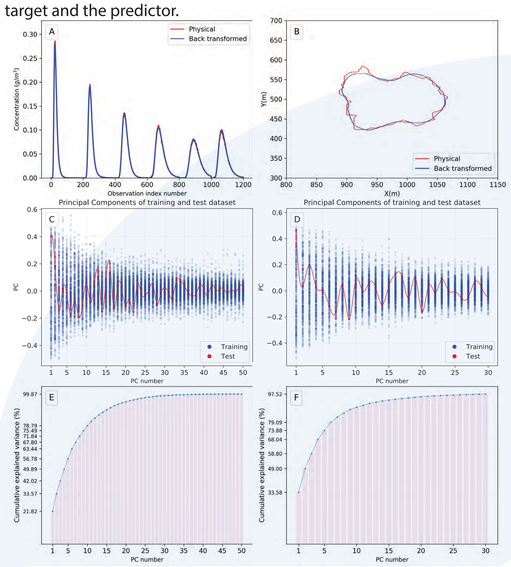

Figure 4. **A**. The full predictor of the test set is the concatenation of all breakthrough curves (black curves in Figure 2). PCA with 50 PCs allows to recover the original curves while smoothing out noise present in the original dataset. **B**. Zero-contour of the test target's original SD compared to the zero-contour of the test target's projected SD, which was then back-transformed with its 30 PCs. **C**. Predictor training set and projected test set PCs. **D**. PCs of the target training set's SD and the projected test set's SD. **E**. Cumulative explained variance for the PCs of the predictor training set. **F**. Cumulative explained variance for the PCs of the SD of the target training set.

4. Canonical Correlation Analysis (CCA) transforms the two sets into pairs of Canonical Variates (CVs). This is the learning step: to establish a multivariate relationship between predictor and target.

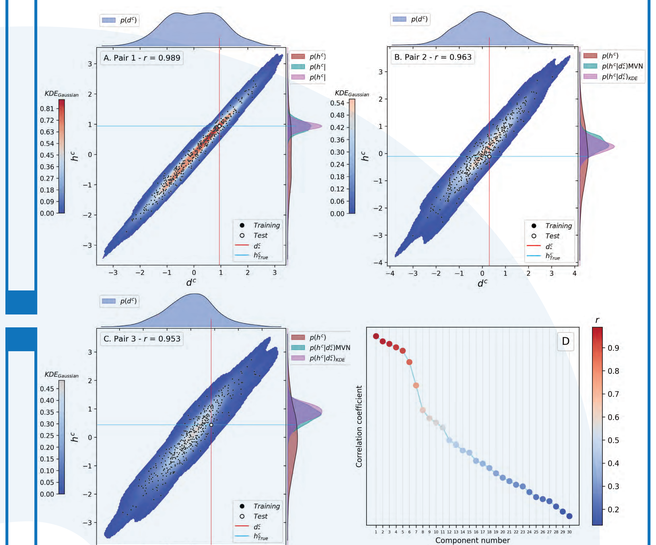

Figure 5. **A**, **B**, **C**. Canonical variates bivariate distribution plots for the 4 firsts pairs of the training set, and the canonical space projection of the selected test predictor and associated test target (see notches). The posterior distribution of h computed according to BEL and KDE can be compared on the y marginal plot. **D**. Decrease of the canonical correlation coefficient r with the number of CV pairs for the training set.

5. To predict the distribution of the unknown WHPA given the trained regression model and the observed value, samples are drawn from the posterior distribution of the target in canonical space and then back-transformed to the original space.

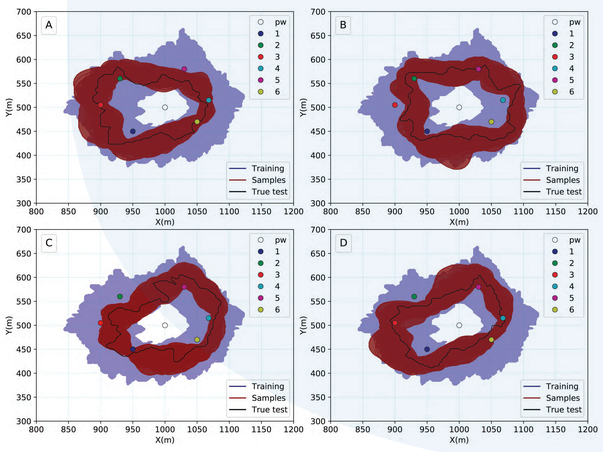

Figure 6. BEL-derived posterior predictions for 4 different tests WHPAs. Subfigure **A** displays the chosen test example associated with Figures 3.

6. Experimental design: BEL can be used to quantify the amount of information delivered by each possible data source, allowing us to make the best injection well location choice. The Modified Hausdorff distance (MHD) is our metric.

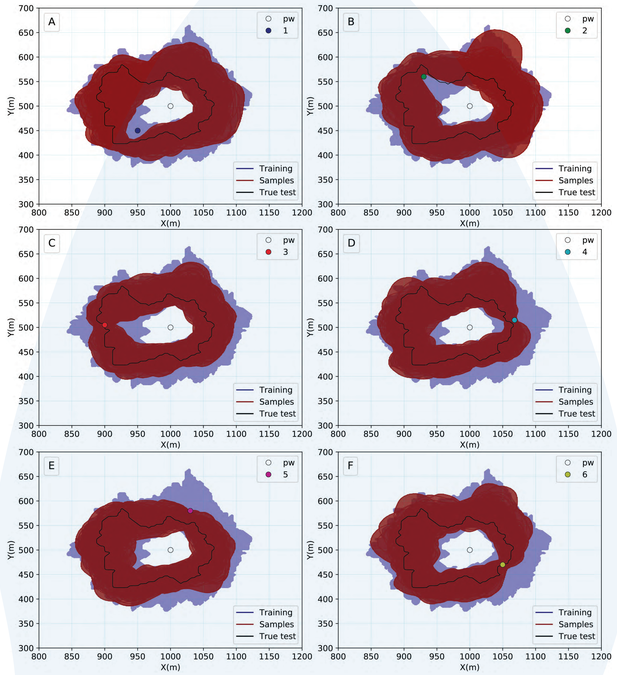

Figure 7. WHPA predictions for each well 1, 2, 3, 4, 5, 6 for the chosen test example.

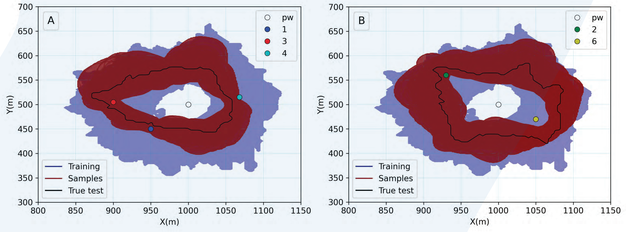

Figure 8. WHPA predictions for multiple-wells combinations, both performed with a training set of 1000 samples and test set of 250 samples. **A**. Prediction using wells 1, 3, 4. **B**. Prediction using wells 2, 6.

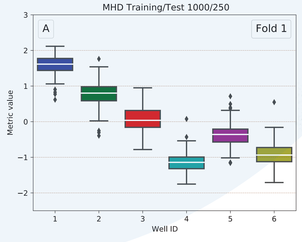

Figure 9. Boxplots of the standardised MHD distance for each well and one fold for a 1000-sample training dataset and 250-sample test dataset.

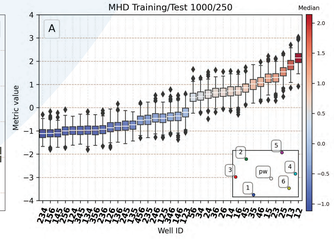

Figure 10. Boxplots of the standardised MHD distance for all 2 and 3 well combinations on a 1000-sample training dataset and a 250-sample test dataset.

GHENT UNIVERSITY
FACULTY OF SCIENCES

python powered

SKBEL on GitHub
Python module for implementing the Bayesian Evidential Learning framework built on top of scikit-learn

AGU FALL MEETING
New Orleans, LA & Online Everywhere
13–17 December 2021

Publication
Thibaut, R., E. Laloy, and T. Hermans, 2021
A new framework for experimental design using Bayesian Evidential Learning
The case of wellhead protection area
Journal of Hydrology, 603, 126903.

scikit learn
machine learning in Python