AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2014WR015452

Key Points:
- CPT/SMR data used to estimate high-resolution 1-D *K* profiles
- Hydro-geophysical data integration using relevance vector machines
- *K* predictions from the learning machine agree with hydraulic tests

Supporting Information:
- Readme
- Data set

Correspondence to:
D. Paradis,
dparadis@nrcan.gc.ca

Citation:
Paradis, D., R. Lefebvre, E. Gloaguen, and A. Rivera (2015), Predicting hydrofacies and hydraulic conductivity from direct-push data using a data-driven relevance vector machine approach: Motivations, algorithms, and application, *Water Resour. Res.*, *51*, 481–505, doi:10.1002/2014WR015452.

Received 14 FEB 2014
Accepted 6 DEC 2014
Accepted article online 17 DEC 2014
Published online 22 JAN 2015

Predicting hydrofacies and hydraulic conductivity from direct-push data using a data-driven relevance vector machine approach: Motivations, algorithms, and application

Daniel Paradis[1,2], René Lefebvre[2], Erwan Gloaguen[2], and Alfonso Rivera[1]

[1]Geological Survey of Canada, Quebec City, Quebec, Canada, [2]Institut National de la Recherche Scientifique, Centre Eau Terre Environnement, Quebec City, Quebec, Canada

Abstract The spatial heterogeneity of hydraulic conductivity (*K*) exerts a major control on groundwater flow and solute transport. The heterogeneous spatial distribution of *K* can be imaged using indirect geophysical data as long as reliable relations exist to link geophysical data to *K*. This paper presents a nonparametric learning machine approach to predict aquifer *K* from cone penetrometer tests (CPT) coupled with a soil moisture and resistivity probe (SMR) using relevance vector machines (RVMs). The learning machine approach is demonstrated with an application to a heterogeneous unconsolidated littoral aquifer in a 12 km^2 subwatershed, where relations between *K* and multiparameters CPT/SMR soundings appear complex. Our approach involved fuzzy clustering to define hydrofacies (HF) on the basis of CPT/SMR and *K* data prior to the training of RVMs for HFs recognition and *K* prediction on the basis of CPT/SMR data alone. The learning machine was built from a colocated training data set representative of the study area that includes *K* data from slug tests and CPT/SMR data up-scaled at a common vertical resolution of 15 cm with *K* data. After training, the predictive capabilities of the learning machine were assessed through cross validation with data withheld from the training data set and with *K* data from flowmeter tests not used during the training process. Results show that HF and *K* predictions from the learning machine are consistent with hydraulic tests. The combined use of CPT/SMR data and RVM-based learning machine proved to be powerful and efficient for the characterization of high-resolution *K* heterogeneity for unconsolidated aquifers.

1. Introduction

Inferring the heterogeneous spatial distribution of hydraulic conductivity (*K*) in aquifers is a prerequisite to tackle groundwater flow and transport problems. Indeed, since *K* may vary over several orders of magnitude and impacts both the magnitude and direction of advective transport, the primary focus of aquifer characterization is generally on the measurement of *K* [*Koltermann and Gorelick*, 1996]. Aquifer *K* is mostly measured using hydraulic tests carried out in wells (e.g., slug tests, pumping tests). Although such tests are generally reliable sources of data about *K*, they are however costly and time consuming. Consequently, these measurements are usually available only from a few wells and at a too low spatial resolution that prevent to adequately define *K* heterogeneities at the scale needed for most practical groundwater flow and mass transport studies [*Butler*, 2005].

Due to these limitations of conventional hydraulic characterization, hydrogeophysics is increasingly recognized as an effective alternative to better image spatial distribution of hydraulic properties, which requires the translation of indirect geophysical data into hydraulic properties [e.g., *Day-Lewis et al.*, 2005; *Rubin and Hubbard*, 2005]. The value of using geophysical data for hydrogeological characterization lies in the extensive spatial coverage offered by geophysical methods, which may be helpful to provide spatial continuity in *K* heterogeneities. Reliable predictions in *K* from geophysical data should however be based on sound relations between hydraulic and geophysical data, which are usually subject to a large degree of uncertainty under field conditions [*Chen et al.*, 2001]. The major problem with the integration of hydro-geophysical data is nonuniqueness in the hydro-geophysical relations. Typical causes of nonuniqueness are the scale and the resolution disparity between hydraulic and geophysical measurements, and the uncertainty associated with field data acquisition and interpretation (e.g., noisy measurements, location errors). Another fundamental aspect with nonuniqueness is the degree of sensitivity between hydraulic and geophysical parameters,

which for a particular geological material may result in a fairly weak correlation. That problem is exacerbated under heterogeneous field conditions where sensitivities may vary for different geological materials and thus preclude reliable estimations of hydraulic properties from geophysical data. Thus, the overall motivation of this work is the need to develop efficient and robust aquifer characterization and data analysis approaches that can provide more information about *K*: higher number of control points, relatively fine vertical resolution, continuous vertical profiles, based on repeatable physical measurements. Such larger high-quality data set is needed to define *K* heterogeneity using geostatistical interpolation schemes (estimation or simulation) in order to develop more realistic numerical groundwater flow and solute transport models [*Anderson*, 1997].

This paper explores the potential of using cone penetrometer tests (CPT) coupled with a soil moisture and resistivity probe (SMR) to infer hydrofacies (HF) and *K* in unconsolidated aquifers. The paper is focused on the assessment of the usefulness of CPT/SMR soundings for HF and *K* estimation and on the integration of hydraulic and geophysical data through a learning machine approach based on relevance vector machines (RVMs). CPT/SMR is a multiparameter probe that simultaneously provides vertical profiles of mechanical (tip stress, sleeve stress, pore pressure) and electrical (dielectric constant, bulk electrical resistivity) parameters of sediments. Thus, due to the number of simultaneously measured geophysical parameters and their similar volumes of investigation, CPT/SMR soundings have the potential to reduce nonuniqueness between geophysical measurements and *K*. From a practical viewpoint, the value of using CPT/SMR data for aquifer characterization lies in the vertical decimeter-scale resolution offered by this direct-push technique [*Lunne et al*., 1997; *Schulmeister et al*., 2003], which can be hardly obtained by surface-based geophysical methods. Moreover, the number of continuous vertical profiles that can be obtain by direct-push soundings over a given period of time is significantly higher in comparison to wells or core-based hydraulic tests, allowing the definition of aquifer heterogeneities over larger investigation areas [*Lafuerza et al*., 2005; *Paradis et al*., 2014]. While there have been important improvements recently in the ability of direct-push tools to estimate *K* from hydraulic testing, such as direct-push slug testing [*Butler et al*., 2002], direct-push permeameter [*Butler et al*., 2007], direct-push injection logging [*Liu et al*., 2009; *Lessoff et al*., 2010] and hydraulic profiling [*Köber et al*., 2009], the approach followed in this paper differs significantly. Instead of relying only on direct hydraulic data, the proposed approach is based on the conversion of CPT/SMR data into indirect hydraulic data through site-specific hydro-geophysical relationships. The establishment of hydro-geophysical relationships is based on the collection of collocated data (training data set) of both geophysical and hydraulic data, where the locations of CPT/SMR sounding and well sites for hydraulic testing are carefully selected. Once those relationships are defined, CPT/SMR soundings without well installation are carried out elsewhere over the study area and direct-push data are converted into hydraulic information using previously defined relationships. With this approach, hydraulic tests are carried out with parsimony only at representative locations within the study aquifer, making the aquifer characterization process more time efficient than conventional hydraulic testing alone, as documented by *Paradis et al*. [2014]. Finally, the spatial distribution of hydraulic information over the study area can be obtained through geostatistical interpolation or simulation of direct and converted hydraulic data. Note however that the topic of geostatistical estimation is not covered in this paper.

CPT soundings for geological applications have been mostly used to deduce sediment texture from mechanical parameters [e.g., *Robertson*, 1990; *Fellenius and Eslami*, 2000]. *Farrar* [1996] also proposed a chart to evaluate *K* from sediment texture, but it only provides order-of-magnitude *K* estimates and it does not make full use of electrical parameters provided by the SMR probe. Our objective is thus to develop a general approach to define site-specific relations to reliably predict *K* from CPT/SMR data. Also, to facilitate the spatial interpretation of *K* heterogeneity over a study area and to allow a better integration with geological depositional models, we also wish to use direct-push data to define hydrofacies (HF) [e.g., *Anderson*, 1997; *Koltermann and Gorelick*, 1996; *Ouellon et al*., 2008; *Paradis et al*., 2014]. A HF is a homogeneous unit that is hydrogeologically meaningful for the purposes of flow and transport modeling [*Anderson*, 1989] and it is defined here as a distinct unit in terms of *K* distribution.

Quantitative hydro-geophysical (H-G) data integration is usually achieved by linking hydraulic and geophysical parameters through theoretical or semiempirical petrophysical relations such as Archie's law [*Archie*, 1942] [e.g., *Copty et al*., 1993; *Yamamoto et al*., 1994; *Gloaguen et al*., 2001; *Garambois et al*., 2002]. For

application in heterogeneous aquifers, this approach is however limited because petrophysical relations are often complex to model due to the strong dependence to the geological material. Thus, the site-specific applicable petrophysical model is often difficult to select and converted hydraulic data may be unreliable [*Steelman and Endres*, 2011].

A fundamentally different approach to model H-G relations involves the application of statistical techniques, either simple parametric relations [e.g., *Hyndman et al*., 2000] or more complex nonparametric methods [e.g., *Mohaghegh et al*., 1997; *Wong et al*., 1998; *Lee and Datta-Gupta*, 1999; *Chen et al*., 2001; *Chen and Rubin*, 2003; *Paasche et al*., 2006; *Shokir et al*., 2006; *Dubois et al*., 2007; *Al-Anazi et al*., 2009; *Elshafei and Hamada*, 2009; *Kharrat et al*., 2009; *Al-Anazi and Gates*, 2010a, b; *Dubreuil-Boisclair et al*., 2011; *Ruggeri et al*., 2013; *Rumpf and Tronicke*, 2014]. Unlike general theoretical or semiempirical models, statistical techniques are much more flexible and do not require prior knowledge about physical relations between various H-G parameters or geological material.

In this paper, the definition of relations to predict profiles of HF and *K* from CPT/SMR soundings is made through a nonparametric learning machine approach because of the complex relations that generally exist between H-G parameters [e.g., *Mohaghegh et al*., 1997; *Lee and Datta-Gupta*, 1999; *Dubois et al*., 2007]. Learning machines do not assume a rigid functional form, they rely on the available data to build up a model of the system, and no a priori assumptions on parameter relations are made [*Mitchell*, 1997]. Artificial neural networks (ANNs), which follow an empirical risk minimization of the training errors, are common form of learning machines that have been already considered. Different architectures of ANNs have been applied successfully for the prediction of lithofacies [*Chen and Rubin*, 2003; *Dubois et al*., 2007] or hydraulic properties in petroleum reservoirs [*Mohaghegh et al*., 1997; *Wong et al*., 1998; *Lee and Datta-Gupta*, 1999; *Shokir et al*., 2006; *Al-Anazi et al*., 2009; *Elshafei and Hamada*, 2009; *Kharrat et al*., 2009; *Iturrarán-Viveros and Parra*, 2014] from cross hole or borehole geophysics data. However, despite their potential effectiveness, ANNs present some important drawbacks [*Camps-Valls et al*., 2006]: (i) design and training often results in a complex, time-consuming task, in which many parameters must be tuned; (ii) minimization of the training errors can lead to poor generalization performance; and (iii) performance can be degraded when working with small (sparse) data sets.

To alleviate problems associated with ANNs, support vector machine (SVM) was developed to solve both classification and regression problems [*Vapnik*, 1995, 1998]. Unlike ANNs, SVM follows a structural risk minimization of generalization performance, and model complexity is controlled through a regularization term. The main idea behind SVM is to perform a linear regression in a high dimension feature space, through a kernel function, which returns a nonlinear regression in the original input space. SVM has yielded good results for the prediction of lithofacies, permeability, and porosity of petroleum reservoirs with high dimensional and sparse borehole geophysics data sets [*Al-Anazi and Gates*, 2010a, b].

The rationale for selecting RVM approaches for this study over ANNs and SVM is that many studies have shown that RVM performs better than either ANNs or SVM in many applications for accuracy and sparsity of the solution [*Khalili et al*., 2005; *Camps-Valls et al*., 2006; *Samui*, 2007; *Ghosh and Mujumdar*, 2008]. A RVM is a Bayesian extension of the SVM to solve nonlinear classification and regression models using an expectation maximization-like learning method [*Tipping*, 2001]. The most important characteristic of the RVM is to produce sparse predictive models, which are less prone to overfit the training data. Along with its ability to produce relations with good generalization capability with sparse and complex data sets, which is typical in most geosciences applications, RVM produces probabilistic outputs that can capture uncertainty in the predictions. Model selection with RVM is also easier, since it has no regularization term needing to be adjusted, and are less sensitive to model parameter setting [*Camps-Valls et al*., 2006].

The remainder of this paper is organized as follows. Section 2 describes the study area and the training data set used to develop classification and regression models. Section 3 outlines the learning machine approach and provides a description of the main algorithms, which includes fuzzy clustering and RVMs for classification and regression. Section 4 presents the development and the verification of the learning machine to recognize HF and estimate *K*. Conclusions about the key findings of this study are listed in section 5.

2. Study Area and Hydro-Geophysical Training Data Set

Since learning machines are based on empirical data, the collection of a representative training data set for a given study area is fundamental to establish meaningful relations between hydraulic and geophysical

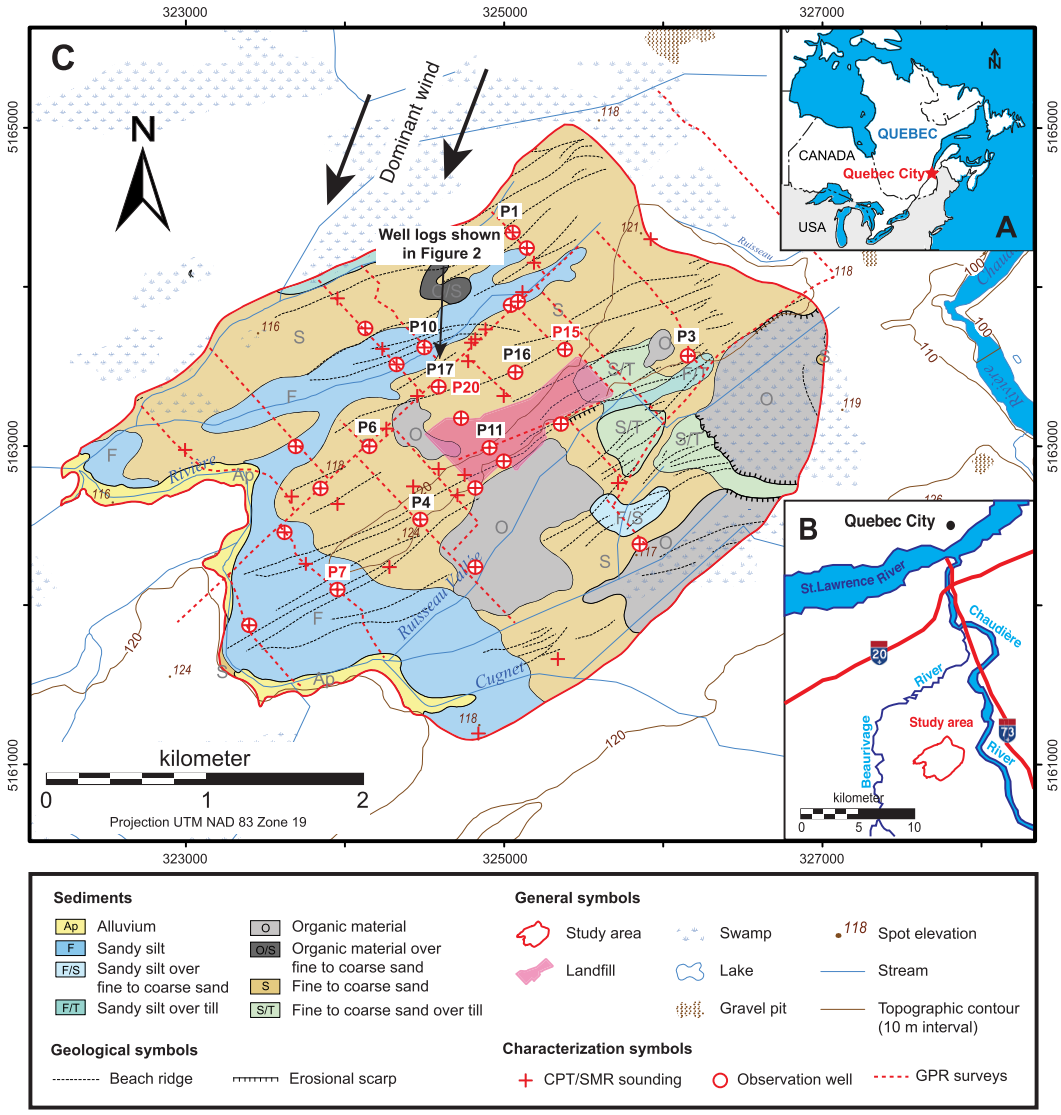

Figure 1. (a and b) General location of the St-Lambert study area, with the (c) Quaternary sediments map for the subwatershed surrounding the decommissioned sanitary landfill, showing the locations of direct-push soundings and observation wells used for aquifer characterization. The main depositional direction of sediments making up the granular aquifer is assumed to have been oblique to the orientation of the paleoshore in a littoral environment. The Quaternary map was modified from Lamarche and Tremblay (unpublished data, 2012).

parameters. *Paradis et al.* [2014] described the general data acquisition approach that was followed for the characterization of the study area, which includes the collection of the hydro-geophysical training data set used in this paper. This data acquisition approach specifically allowed, through regional geology, GPR surveys and CPT/SMR soundings analysis, the targeting of specific locations for well installations and *K* testing in order to cover the whole range of hydro-geophysical responses (*K* and CPT/SMR) observed over the study area. In this section, we thus only briefly describe the study area, provide a summary of the data acquisition process for CPT/SMR and *K* data, and describe the hydro-geophysical training data set. Field data used in this study are provided as downloadable supporting information.

2.1. Saint-Lambert Study Area

The proposed methodology was developed and applied in relation with a study carried out in St-Lambert-de-Lauzon, located 30 km south of Quebec City, Canada (Figures 1a and 1b). As illustrated in Figure 1c, the study area encompasses a 12 km^2 subwatershed surrounding a decommissioned sanitary landfill where an assessment of the migration of a leachate plume was underway [*Tremblay et al.*, 2014]. As reported by *Bolduc* [2003], the surficial sediments of the study area (Figure 1c) consist primarily of Late Quaternary sandy

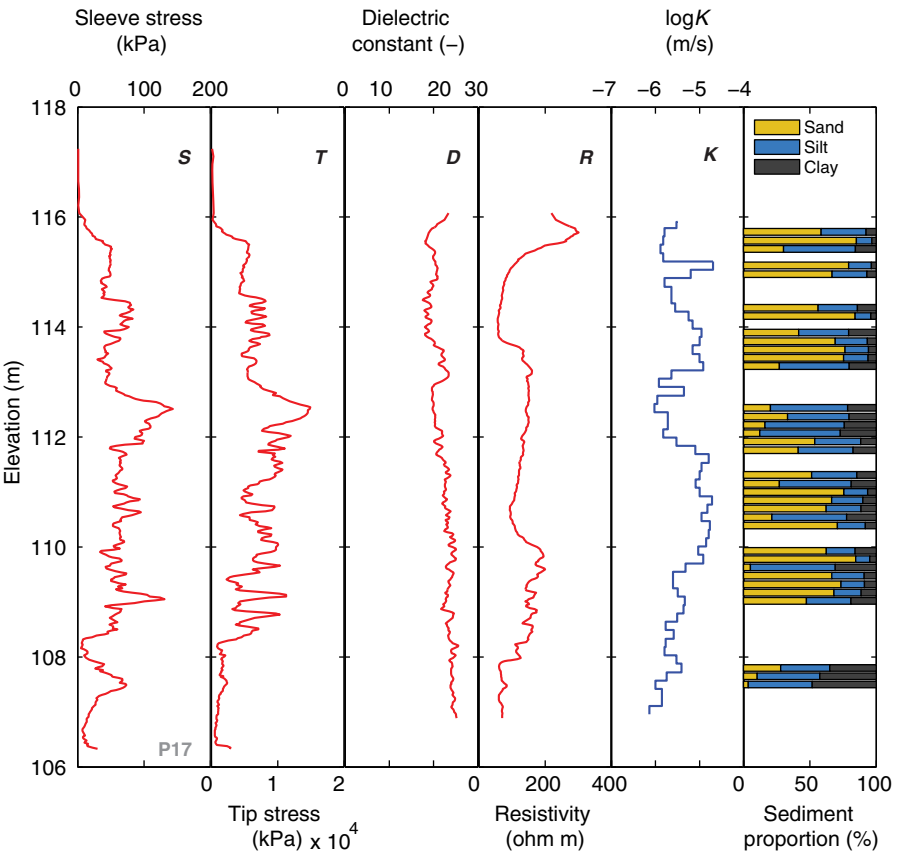

Figure 2. Example of CPT/SMR (red lines), hydraulic conductivity (blue line), and lithological profiles obtained at location P17. CPT/SMR parameters are sleeve stress (S), tip stress (T), dielectric constant (D), and electrical resistivity (R). Hydraulic conductivity (K) values were obtained from multilevel slug tests in well P17. Proportions of sand, silt, and clay are based on grain size analyses on sediment samples. P17 location is shown in Figure 1c.

and silty sediments that were deposited in the receding Champlain Sea, which was an arm of the Atlantic Ocean that had invaded the St. Lawrence Valley during the last deglaciation. More specifically, mainly longshore currents that redeposited in littoral and sublittoral settings, the sediments supplied to the Chaudière River paleodelta controlled deposition at the St-Lambert site. This is indicated in Figure 1c by the southwestward fining of the littoral sediments in conjunction with the southwest-northeast trend of the beach ridges (L. Lamarche and L. Tremblay, Géologie des formations superficielles pour le site de St-Lambert-de-Lauzon, Québec, unpublished data, 2012). These ridges and the associated nearshore bars are mostly composed of medium to fine sand while the intervening troughs are composed of finer, silty sediments with poor to very poor grain-size sorting. Thus, the littoral and sublittoral depositional environments resulted in superposition of long (>100 m) interdigitized sand and silt strata with lateral intrastratal transitions in grain size as a result of changing energy levels along Champlain Sea shorelines. A more detailed description of the aquifer heterogeneity of the St-Lambert site is provided by *Paradis et al.* [2014] and *Tremblay et al.* [2014].

2.2. Geophysical Measurements From CPT/SMR Soundings

According to the regional surficial sediments geology and GPR surveys carried out for the St-Lambert study [*Paradis et al.*, 2014], the locations of 53 CPT/SMR soundings were selected (Figure 1c). As described in the next section, the geophysical data of eight soundings are used to establish relations with hydraulic data according to the proposed methodology in this paper. The remaining 45 soundings are left as a data set of mechanical and electrical properties of sediments that eventually will be converted into *K* data and serve as a base of interpolation to image the heterogeneity in hydraulic properties of the site [e.g., *Paradis et al.*, 2014]. Direct-push soundings were carried out using a Geotech 605-D rig equipped with a CPT system including pore pressure measurement combined with a SMR probe. As illustrated in Figure 2 (red lines) for

Table 1. Original and Up-Scaled Vertical Resolutions and Vertical Support of Measurements for Direct-Push Parameters (CPT/SMR) and Hydraulic Conductivity[a]

	Original		Upscaled		
Parameter	Vertical Resolution (cm)	Vertical Support (cm)	Vertical Resolution (cm)	Data for Moving Average	Vertical Support (cm)
Direct-Push Parameters (CPT/SMR)					
Mechanical resistance: Tip stress (*T*)	2.6 ± 3.6	4	2	7	16
Mechanical friction: Sleeve stress (*S*)	2.6 ± 3.6	17	2	1	17
Dielectric constant (*D*)	2.6 ± 3.6	3	2	7	15
Bulk DC electrical resistivity (*R*)	2.6 ± 3.6	9	2	4	15
Hydraulic Parameter					
Hydraulic conductivity (*K*)	15	15	15		15

[a]Supports of measurements are approximations according to the CPT/SMR probe specifications.

location P17 (shown in Figure 1c), CPT and SMR probes allow the simultaneous measurement of two mechanical and two electrical properties of sediments, respectively, with support of measurements that range approximately from 3 to 17 cm (Table 1). A 15 cm^2 penetrometer cone with a 60° conical tip was used according to ASTM D3441 standards [*American Society for Testing and Materials (ASTM)*, 2012]. The penetrometer is advanced vertically into the soil at a constant rate of 2 cm/s, though this rate must be reduced when compact layers are met. This rate of penetration provides then vertical resolution for all CPT/SMR parameters in the range of 2 cm (Table 1). Inside the probe, two load cells independently measure the vertical stress against the conical tip and the side friction along the sleeve [*Lunne et al*., 1997]. The support of measurement of the tip stress (*T*) and sleeve stress (*S*) with the CPT probe are 4 and 17 cm, respectively, according to the probe geometry. A pressure transducer in the cone is also used to measure the pore water pressure as the probe is pushed into the ground. Despite pore pressure may be an indicator of the presence of clay, this parameter was used in this study only to correct *T* data for the overburden stress. The SMR probe is composed of four electrodes that are connected directly behind the penetrometer [*Shinn et al*., 1998]. The inner two rings are used to measure soil capacitance and the spacing between the two rings is 3 cm. The soil moisture probe operates at 100 MHz, thereby reducing the effects of the electrical conductivity of the soil on the measured dielectric constant. The instrument measures shifts in the frequency resonance of the emitted electro-magnetic signal as it passes through the soil that may be related empirically to soil moisture content or the dielectric constant (*D*). The bulk electrical resistivity (*R*) measurement employs the outer two rings of the SMR probe, that are spaced 9 cm apart, to apply the current and to measure the voltage drop (pole-pole configuration). According to the spacing between the electrodes, supports of measurement for *D* and *R* are approximately 3 and 9 cm, respectively. The probe operates at a frequency of 1000 Hz to avoid soil polarization effects.

2.3. Hydraulic Conductivity Data From Multilevel Slug Tests

Based on CPT/SMR data obtained in real time during sounding operations, 25 of the 53 direct-push soundings were converted into observation wells (Figure 1c). *K* values used to establish relations with CPT/SMR data were obtained by high-resolution multilevel slug tests in eight of the fully screened direct-push wells (wells labeled in black in Figure 1c). The remaining 17 wells were essentially used for geochemical sampling in relation to the migration of the leachate plume [*Tremblay et al*., 2014]. Each well was installed into the same hole created by the sounding to obtain colocated hydraulic and direct-push data and thus reduce uncertainty in data analysis related to disparity in interval measurements. The observation wells were installed with well screen in direct contact with sediments (without sand-pack), which is more suitable for hydraulic tests carried out over small intervals because it reduces hydraulic short-circuit and skin effects on test data. Observation wells that are fully screened across the saturated zone were also installed to provide continuous profiles of *K* and to obtain hydraulic data for all kind of sediments present over the study area.

Multilevel slug tests were made over 15 cm vertical intervals using a dual-packer assembly to isolate tested intervals [e.g., *Ross and McElwee*, 2007], as for the *K* profile shown in Figure 2. Slug tests were performed using a pneumatic method to induce an initial lowering of the water level [*Levy and Pannell*, 1991] and hydraulic responses were interpreted using the *Bouwer and Rice* [1976] method. More detailed descriptions of direct-push well installation and hydraulic testing procedure are provided by *Paradis et al*. [2014] and

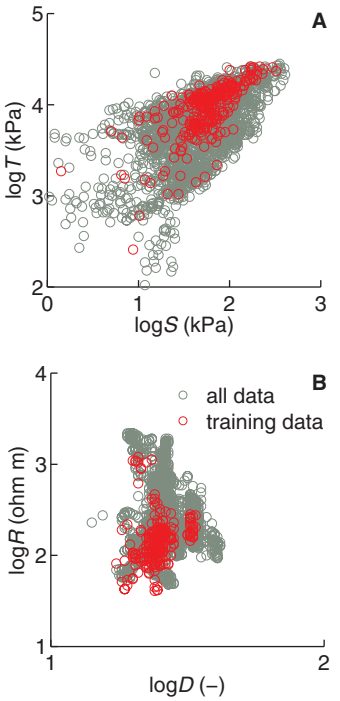

Figure 3. Distribution of the training data set relative to the range of (a) mechanical: sleeve (S) and tip (T) stresses; and (b) electrical: dielectric constant (D) and resistivity (R) responses for the 16 direct-push soundings. A direct-push interval measurement of the training data set corresponds to a colocated measurement of hydraulic conductivity (K). Direct-push measurements were transformed as summarized in Table 1 so that their vertical resolution matches the one of K measurements. The number of such colocated hydro-geophysical measurements is 280. Parameter symbols are defined in Table 1.

Paradis et al. [2011], respectively. Thus, a total of 280 intervals were tested and selected according to the range and occurrence of CPT/SMR responses, as depicted in Figure 3. Note that all available 25 wells and the eight selected wells were not systematically tested due to the time associated with slug testing (30–60 min per interval), but specific sections in selected wells deemed representative of the CPT/SMR responses were instead tested.

2.4. Data Resampling and Rescaling

Statistical techniques require that the different variables be measured at the same scale and on the same support. However, this is not the case with CPT/SMR data where data are taken at a regular time interval but at a rate of penetration that is not necessarily constant. Therefore, CPT/SMR data were resampled on a regular grid of 2 cm, a vertical resolution close to the original resolution (Table 1), using trapezoidal integration [*Davis*, 1973]. In addition, the vertical support of measurement for the different hydraulic and geophysical parameters is not identical (Table 1). In order to properly compare all measurements (T, S, R, D, and K), the variations in their support need to be taken into account [*Isaaks and Srivastava*, 1989]. Hence, all the parameters with the smaller support (T, S, R, D) were upscaled to the scale of the parameter with the larger support (K). First, each CPT/SMR parameter was upscaled with a moving average to a vertical support of 15 cm that corresponds to the 15 cm intervals of the multilevel slug tests. The number of regularly spaced 2 cm data used in the moving average was varied according to the original support of each direct-push parameter (Table 1). Then, all direct-push data were resampled using linear interpolation to the 15 cm interval that corresponds to the K intervals over which hydraulic testing was carried out. Consequently, all hydro-geophysical data represent both the same vertical resolution and approximately the same vertical support of measurement (Table 1).

2.5. Descriptive Statistics of the Hydro-Geophysical Training Data Set

Since we are interested in defining relationships between K and CPT/SMR data, descriptive statistics presented here are for geophysical data available in the same intervals where K measurements are available, which together form the hydro-geophysical training data set. Statistics for the hydro-geophysical training data set are presented in Table 2, and histograms for each parameter are depicted in Figure 4. Since the range in parameter values for most parameter vary over a few orders of magnitude, a logarithmic transform was applied to make their distribution closer to a Gaussian distribution. Even though, histograms of the logarithm of geophysical parameter are all slightly asymmetric: negatively skewed for mechanical parameters ($\log S$ and $\log T$) and positively skewed for electrical parameters ($\log D$ and $\log R$). The distribution for $\log K$ is rather symmetrical and uniform (not normally distributed), which suggests weak correlations with direct-push parameters that have different distributions. Indeed, the scatterplots in Figure 4 and the Kendall rank correlation matrix in Table 3 shows no or low correlations between geophysical parameters together except for $\log S$ and $\log T$ where the correlation is relatively high ($\tau = 0.53$) and may indicate redundancy in those two parameters. The correlations between $\log K$ and geophysical parameters are generally significant but very low, except with $\log S$ where the correlation is almost null. Standard deviation for $\log K$ is also at least twice the standard deviation of geophysical parameters (even an order of magnitude greater with respect to $\log D$), which suggests that K may be more sensitive to changes in sediment types than any direct-push parameter. Thus, all the above indicates that

Table 2. Descriptive Statistics for the Logarithmic Distribution of Colocated Direct-Push and Hydraulic Conductivity Data of the Training Data Set[a]

Parameter	Number	Mean	Median	Minimum	Maximum	Range	Standard Deviation	Skewness	Kurtosis
log*S*	280	1.74	1.76	0.15	2.51	2.36	0.31	−0.96	2.94
log*T*	280	3.94	4.00	2.41	4.42	2.01	0.29	−1.44	3.63
log*D*	280	1.39	1.39	1.24	1.53	0.29	0.06	0.36	0.85
log*R*	280	2.16	2.16	1.61	3.08	1.47	0.25	1.09	3.25
log*K*	280	−5.04	−5.05	−6.24	−3.92	2.32	0.57	0.09	−1.10

[a]Parameter symbols are defined in Table 1.

the relations between *K* and direct-push parameters are not strait forward because of the weak and complex relationships among the parameters that could exist. In the next section, we describe the learning machine approach to handle this challenge.

3. Description of the Learning Machine Approach

In this section, we first present an outline of the learning machine approach proposed to establish relations between hydraulic and geophysical data. Then, we briefly review the main algorithms used by the learning machine: (i) the Gustafson-Kessel fuzzy clustering method employing an adaptive distance norm, which is an extension of the well-known fuzzy *c*-means (FCM) algorithm; and relevance vector machine (RVM) that is a Bayesian probabilistic extended linear model with a prior on the model weights to achieve sparse solutions, which can be used for (ii) regression and (iii) classification problems.

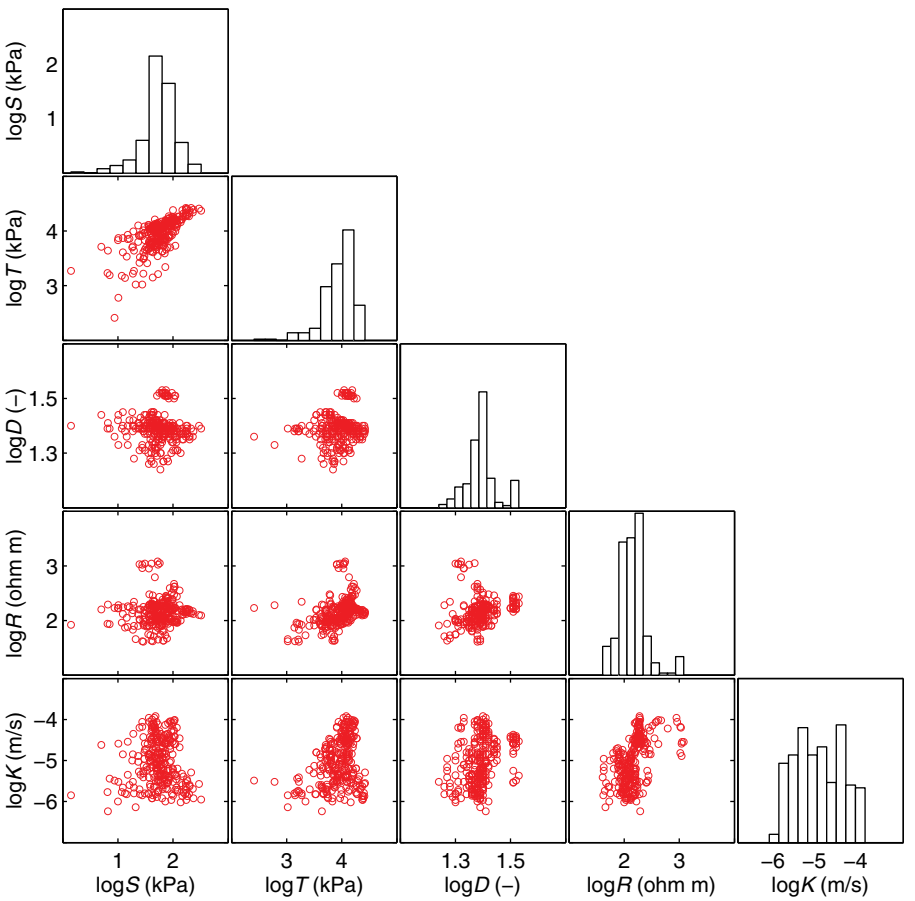

Figure 4. Matrix scatterplots with histograms for hydraulic conductivity (K) and CPT/SMR parameters (S: sleeve stress; T: tip stress; D: dielectric constant; R: electrical resistivity) for the hydro-geophysical training data set. The total number of colocated intervals is 280.

Table 3. Correlation Matrix Showing the Values of the Kendall Rank Correlation for the Logarithm of Direct-Push Data and Hydraulic Conductivity[a]

Parameter	log*S*	log*T*	log*D*	log*R*	log*K*
log*S*	1	0.53	−0.09	0.06	−0.05
log*T*		1	0.05	0.34	0.18
log*D*			1	0.30	0.26
log*R*				1	0.40
log*K*					1

[a]Parameter symbols are defined in Table 1.

3.1. Outline of the Learning Machine Approach

The proposed learning machine algorithm used to predict hydraulic information from CPT/SMR data is a two-step sequential procedure inspired by *Lee and Datta-Gupta* [1999] for permeability prediction in complex petroleum reservoirs from borehole geophysics logs, as illustrated in Figure 5. The proposed procedure requires that for each set of direct-push data, HF is first identified using HF models, and then the associated hydro-geophysical (H-G) relation is used to predict *K*. While the knowledge of HF may ease spatial interpretation of *K* and geological heterogeneity, the selection of the appropriate H-G relation through HF identification prior to *K* estimation is also important to provide accurate *K* estimate, as various geological materials represented by HF may have different hydro-geophysical behaviors.

Prior using the procedure in Figure 5 to predict HF and *K* from CPT/SMR data, HF models and H-G relations need to be developed. The development of the learning machine includes a training and verification phase. The training phase is a three-step procedure that is schematically illustrated in Figure 6 and described below:

1. Step 1—HFs definition: the first step in the training process is the definition of homogeneous groups using unsupervised fuzzy clustering (section 3.2) with *K* and direct-push data of the training data set. Clustering involves the grouping of observations in such a way that observations in the same group (cluster) are more similar to each other than to those in other groups. The main rationale behind using clustering for this study is that various geological materials may have different hydro-geophysical behaviors and clustering allows the grouping of similar hydro-geophysical characteristics without any prior geological knowledge. This grouping can thus contribute to alleviate weak correlations among hydro-geophysical parameters and thus provide more accurate regression equations between *K* and direct-push data. The integration of *K* data in the clustering process is also helpful to define HFs that are hydrogeologically meaningful. Note here that HFs are defined in the hydro-geophysical space. Moreover, to ensure optimal predictive capability of the learning machine, an exhaustive search procedure using clustering is applied to find the most relevant CPT/SMR parameters and to detect intrinsic HF structures in the hydro-geophysical training data set.

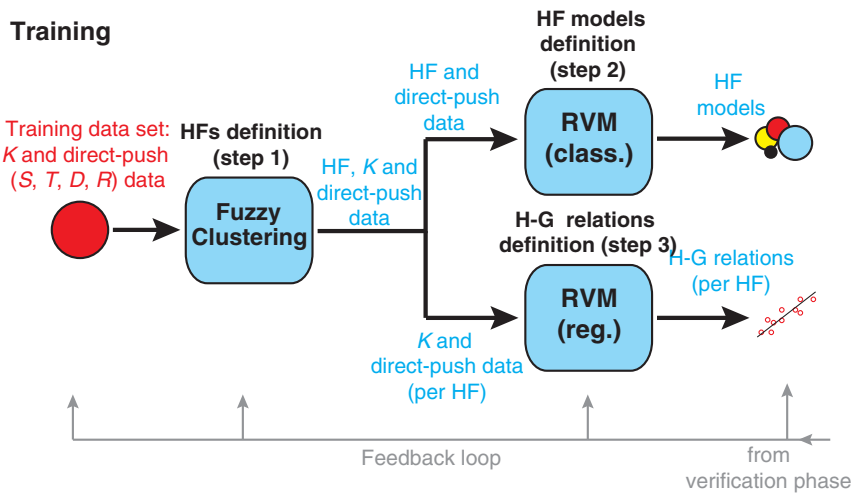

Figure 5. General stages for the training of the learning machine to define hydrofacies (HF) models and hydro-geophysical (H-G) relations from the hydro-geophysical training data set. RVM stands for relevance vector machine. Abbreviations: class., classification; reg., regression. Parameter symbols are defined in Table 1. Note that HF models and H-G relations are defined independently (in parallel).

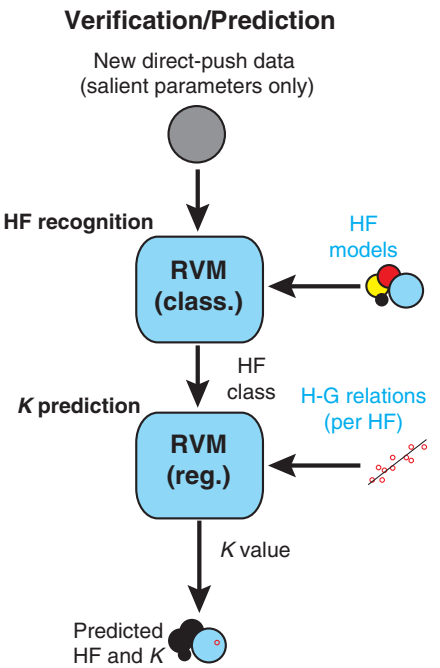

Figure 6. General stages for the prediction of hydrofacies (HF) class and hydraulic conductivity (K) value from direct-push data using the trained learning machine. This sequence is also used for the verification of the learning machine during the training phase. Parameter symbols are defined in Table 1. Note that HF recognition and K prediction are carried out sequentially.

2. Step 2—HF models definition: SINCE clustering is only a tool applicable to make associations among observations and cannot be used as a predictive tool, a multiclass RVM for classification (section 3.4) is trained to learn how to recognize each previously defined HF using CPT/SMR data alone. The training of the RVM to define site-specific HF models is made with HF labels and corresponding data of the most relevant CPT/SMR parameters, as defined by clustering. Note that HF models are defined in the geophysical space. That is, we are trying to recognize HFs based on their projection from the hydro-geophysical space to the geophysical space, as geophysical data are basic information used to predict HF and subsequently *K* (Figure 5).
3. Step 3—Hydro-geophysical relations definition: in parallel with Step 2, a site-specific hydro-geophysical (H-G) relation is developed for each HF using distinct RVMs for regression (section 3.3). Each RVM are trained using colocated *K* and direct-push data (most relevant direct-push parameters) associated to each HF, as defined by clustering.

Finally, the performance of the trained learning machine is verified using the sequential procedure illustrated in Figure 5 to predict HF and *K* from direct-push data. In this paper, the verification of the learning machine is made through an internal and an external procedure using cross validation with the training data set used for the development of the machine and with *K* data obtained from hydraulic tests not used in the development process, respectively. As depicted in Figures 5 and 6, the learning machine also includes a feedback path where the outputs obtained at any stage of the development process can be reconsidered in a previous stage. The remainder of this section presents the general algorithms used by the learning machine that involves fuzzy clustering and RVMs for both regression and classification.

3.2. Gustafson-Kessel Fuzzy Clustering

In this study, the Gustafson-Kessel (GK) algorithm [*Gustafson and Kessel*, 1979] was selected for the clustering because fuzzy *c*-means (FCM) methods are known to be stable with small and complex (e.g., outliers, overlapping clusters) data sets [*Mingoti and Lima*, 2006; *Qiu and Tamhane*, 2007; *Qiu*, 2010], as the hydro-geophysical training data set described in section 2. In particular, *Gustafson and Kessel*, [1979] extended the standard FCM algorithm [*Dunn*, 1973; *Bezdek*, 1981] by employing an adaptive distance norm, in order to detect clusters of different geometrical structures. The GK algorithm aims to find fuzzy partitioning of a given training data set, by minimizing of the basic *c*-means objective functional:

$$J(\mathbf{X};\mathbf{U},\mathbf{V},\mathbf{A})=\sum_{k=1}^{c}\sum_{i=1}^{N}\mu_{ki}^{m}D_{kiA_k}^{2} \tag{1}$$

where $\mathbf{X}=[x_{in}]$ is $N \times n$ data matrix which contains N colocated observations, each having n parameters that correspond to K and direct-push parameters. $\mathbf{U}=[\mu_{ik}]$ is $N \times c$ fuzzy partition matrix with c clusters, which represents the partial memberships of each x_i in $\mathbf{X}$. Fuzzy partition allows μ_{ik} attaining real values in $[0,1]$. $m=(1,\infty)$ is the fuzziness weighting exponent, that determines the fuzziness of the resulting clusters. $\mathbf{V}=[v_c]$ is a vector of cluster centers, which have to be determined. $D_{kiA_k}^2=\|\mathbf{x}_i-\mathbf{v}_k\|_{A_k}^2=(\mathbf{x}_i-\mathbf{v}_k)^{\mathrm{T}}A_k(\mathbf{x}_i-\mathbf{v}_k)$ is the squared inner-product distance norm for the GK algorithm. This distance norm is known as the squared Mahalanobis distance. $\mathbf{A}$ is $n \times n$ norm-inducing diagonal matrix that accounts for the variance of each parameter n. Each cluster has its own norm-inducing matrix A_k.

For a predefined numbers of c clusters, the minimization of J that is carried out with respect to the partition matrix and the prototypes gives rise to the structure in $\mathbf{X}$. The generic optimization scheme involves a sequence of iterations, in which we successively update the values of the partition matrix:

$$\mu_{ki}=\frac{1}{\sum_{j=1}^{c}\left(D_{kiA_k}/D_{jiA_k}\right)^{2/(m-1)}},\ 1\le i\le c,\quad 1\le k\le N \tag{2}$$

and the centers:

$$\mathbf{v}_i=\frac{\sum_{i=1}^{N}\mu_{ki}^{m}\mathbf{x}_i}{\sum_{i=1}^{N}\mu_{ki}^{m}},\ 1\le k\le c \tag{3}$$

This iterative process terminates when the difference between the fuzzy partition matrices in the following iterations is lower than a maximum termination tolerance value. The matrices $A_k=[\rho_k\det(\mathbf{F}_k)]^{1/n}\mathbf{F}_k^{-1}$ are also used as optimization variables in the c-means functional, thus allowing each cluster to adapt the distance norm to the local topological structure of the data. F_k is the fuzzy covariance matrix of the kth cluster defined by:

$$\mathbf{F}_k=\frac{\sum_{i=1}^{N}\mu_{ki}^{m}(\mathbf{x}_i-\mathbf{v}_k)(\mathbf{x}_i-\mathbf{v}_k)^{\mathrm{T}}}{\sum_{i=1}^{N}\mu_{ki}^{m}},\ 1\le k\le c \tag{4}$$

and it is updated in addition to the partition matrix and the centers in the iterative process leading to the minimization of J.

3.3. RVM for Regression

In a regression problem, a predictor model $y(\mathbf{x})$ is inferred from a set of input data $\{\mathbf{x}_i\}_{i=1}^{N}\in R^d$ along with corresponding responses (targets) $\{t_i\}_{i=1}^{N}\in R$. The objective is to make accurate predictions of the targets t_i (e.g., K) from new values of $\mathbf{x}_i$ (e.g., direct-push parameters). A common approach to express $y(\mathbf{x})$ is as an extended linear model with a set of M kernel functions $\{\phi_j(\mathbf{x})\}_{j=1}^{M}$, of the following form:

$$y_i=y(\mathbf{x}_i;\mathbf{w})=\sum_{j=1}^{M}w_j\phi_j(\mathbf{x}_i)+w_0=\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_i) \tag{5}$$

where y_i are the model targets, $\mathbf{w}=[w_0,w_1,\ldots,w_M]^{\mathrm{T}}$ are the weights of the model, w_0 represents the bias in the regression model, $\phi_j(\mathbf{x}_i)$ is the response of the jth kernel function to input data, $\mathbf{x}_i$, and $\boldsymbol{\phi}(\mathbf{x}_i)=[1,\phi_1(\mathbf{x}_1),\ldots,\phi_M(\mathbf{x}_i)]^{\mathrm{T}}$. Note that equation (5) follows the standard probabilistic formulation, where observed targets, t_i differ from the corresponding model targets, y_i by a Gaussian noise of zero mean and variance σ^2, i.e., $e_i=t_i-y_i\sim N(0,\sigma^2)$. The extended linear model described in equation (5) is thus a linearly weighted sum of M kernel functions, where the weights of the model can be inferred using standard procedures for linear models, and nonlinear kernel functions can be employed to model complex training data set. Then to obtain a predictor model from equation (5), the kernel function has to be chosen and the weights of the model need to be estimated.

Once the basis functions of the extended linear model described in equation (5) are defined, a likelihood function is first used for estimating the model weights, $\mathbf{w}$. By assuming an independent zero mean Gaussian noise model of variance σ^2, the likelihood of the complete data set can be written as:

$$\begin{aligned}p(\mathbf{t}|\mathbf{w},\sigma^2)&=\prod_{i=1}^{N}N(t_i|w_i,\sigma^2)\\&=(2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\|^2\right\}\end{aligned} \tag{6}$$

where $\boldsymbol{\Phi}$ is a $N\times(M+1)$ design matrix with that contains the responses of all kernel functions $\boldsymbol{\phi}(\mathbf{x}_i)$ to the input data $\mathbf{x}_i$.

Then, within a probabilistic Bayesian framework, the likelihood defined in equation (6) is regularized with an a priori model of weight distribution to alleviate overfitting problems. For regression problems, a

noninformative Gaussian prior distribution of zero mean and variance $\alpha_j \equiv 1/\sigma_{w_j}^2$ is indeed imposed over each weight:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{j=1}^{M} N(w_j|0, \alpha_j^{-1}) \tag{7}$$

where sparsity is obtained by the use of M independent hyperparameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_M)^{\mathsf{T}}$, one per weight to moderate the strength of the prior. To complete the specification of this hierarchical prior, hyperpriors are also defined over α and σ^2 with Gamma distributions as proposed by *Tipping*, [2001].

Finally, having defined the prior and the likelihood, Bayesian inference proceeds by computing, from Bayes' rule, the posterior over all unknowns given the data:

$$p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2|\mathbf{t}) = p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2)p(\boldsymbol{\alpha}, \sigma^2|\mathbf{t}) \tag{8}$$

Now, the posterior distribution over the weights $p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2)$ is be computed analytically using:

$$\begin{aligned} p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2) &= \frac{p(\mathbf{t}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})}{p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2)} \\ &= (2\pi)^{-(N+1)/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{w}-\boldsymbol{\mu})\right\} \end{aligned} \tag{9}$$

where the posterior covariance and mean are respectively:

$$\Sigma = (\sigma^{-2}\Phi^{\mathsf{T}}\Phi + \mathbf{A})^{-1} \tag{10}$$

$$\boldsymbol{\mu} = \sigma^{-2}\Sigma\Phi^{\mathsf{T}}\mathbf{t} \tag{11}$$

with $\mathbf{A} = \mathrm{diag}(\alpha_0, \alpha_1, ..., \alpha_N)$.

And, the hyperparameter posterior $p(\boldsymbol{\alpha}, \sigma^2|\mathbf{t})$ is approximated as a delta-function at its most probable values $\boldsymbol{\alpha}_{\mathrm{MP}}$, σ_{MP}^2, which for the case of uniform hyperpriors leads to the maximization of the following marginal likelihood:

$$\begin{aligned} p(\boldsymbol{\alpha}, \sigma^2|\mathbf{t}) \propto p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2) &= \int p(\mathbf{t}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \\ &= (2\pi)^{-N/2}|\sigma^2\mathbf{I} + \Phi\mathbf{A}^{-1}\Phi^{\mathsf{T}}|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{t}^{\mathsf{T}}(\sigma^2\mathbf{I} + \Phi\mathbf{A}^{-1}\Phi^{\mathsf{T}})^{-1}\mathbf{t}\right\} \end{aligned} \tag{12}$$

where $\mathbf{I} = \Sigma^{-1}\Sigma$. Values of $\boldsymbol{\alpha}$ and σ^2 that maximize the marginal likelihood are obtained using an iterative approximate Expectation-Maximization (EM) procedure, as described by *Tipping* [2001]. Essentially, this procedure proceeds by iterative computation of the updating rules for $\boldsymbol{\alpha}$ and σ^2, defined as:

$$\alpha_i^{\mathrm{new}} = \frac{(1-\alpha_i\Sigma_{ii})}{\mu_i^2} \tag{13}$$

$$(\sigma^2)^{\mathrm{new}} = \frac{\|\mathbf{t}-\Phi\boldsymbol{\mu}\|^2}{N-\Sigma_i(1-\alpha_i\Sigma_{ii})} \tag{14}$$

concurrent with updating the posterior statistics Σ and $\boldsymbol{\mu}$ from equations (10) and (11), until some suitable convergence criteria is satisfied. In the iterative maximization of equation (12), many of the hyperparameters α_j tend to infinity and the corresponding weights w_j are thus deleted from the model, as well as their associated kernel functions $\phi_j(\mathbf{x})$, leading to a sparse solution. The remaining observations that have nonzero weights are the relevance vectors.

At the convergence of the hyperparameter estimation procedure, predictions are made on the basis of the posterior distribution over the weights, conditioned on maximizing values $\boldsymbol{\alpha}_{\mathrm{MP}}$ and σ_{MP}^2. Thus, given new input data $\mathbf{x}_*$, the probability distribution of the corresponding target y_* is given by the Gaussian predictive distribution:

$$p(t_*|\mathbf{t}, \boldsymbol{\alpha}_{\mathrm{MP}}, \sigma_{\mathrm{MP}}^2) = \int p(t_*|\mathbf{w}, \sigma_{\mathrm{MP}}^2)p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}_{\mathrm{MP}}, \sigma_{\mathrm{MP}}^2)d\mathbf{w} \sim N(t_*|y_*, \sigma_*^2) \tag{15}$$

where the mean and the variance of the prediction are, respectively:

$$y_* = \boldsymbol{\mu}^{\mathrm{T}} \phi(\mathbf{x}_*) \tag{16}$$

$$\sigma_*^2 = \sigma_{\mathrm{MP}}^2 + \phi(\mathbf{x}_*)^{\mathrm{T}} \Sigma \phi(\mathbf{x}_*) \tag{17}$$

The predictive mean is thus the basis functions evaluated for the new input data $\mathbf{x}_*$ weighted by the posterior mean weights (relevance vectors), whereas the predictive variance comprises the sum of the estimated noise on the data (first term) and due to uncertainty in the prediction of the weights (second term).

3.4. RVM for Classification

RVM for classification follows an essentially identical framework as previously detailed for regression, except that the likelihood function is adapted to account for the target quantities (discrete data). In this section, we consider a two-class classification problem with a set of input data $\{\mathbf{x}_i\}_{i=1}^N \in R^d$ along with corresponding targets $\{t_i\}_{i=1}^N$ (e.g., HF) that may take discrete values of 0 or 1 (class labels). Thus, applying the logistic sigmoid link function $\sigma(y) = 1/(1+e^{-y})$ to $y(\mathbf{x};\mathbf{w})$ and adopting a Bernoulli distribution to account for the discrete probability distribution of the target data, the likelihood function is expressed as:

$$P(\mathbf{t}|\mathbf{w}) = \prod_{i=1}^{N} \sigma\{y(\mathbf{x}_n;\mathbf{w})\}^{t_i} \left[1-\sigma\{y(\mathbf{x}_n;\mathbf{w})\}\right]^{1-t_i} \tag{18}$$

A Bernoulli distribution is a discrete distribution having two possible outcomes n that takes value of 1 with success probability p and value of 0 with failure probability $q = (1-p)$ with probability density function of $P(n) = p^n(1-p)^{1-n}$. And, the logistic sigmoid function is a S-shaped curve between 0 and 1 that is used to model the Bernoulli probability distribution as a continuous variable. Note that there is no noise variance σ^2 expressed in the likelihood function.

The likelihood defined in equation (18) is then regularized with an a priori model of Gaussian distribution of zero mean and variance α_j imposed over each weight, as defined in equation (7). Hyperpriors are also defined with Gamma distributions, but only over α because there is no noise variance σ^2 considered for classification problems. Finally, from Bayes' rule, and considering uniform hyperpriors, the posterior over all unknowns is given as:

$$p(\mathbf{w},\boldsymbol{\alpha}|\mathbf{t}) = p(\mathbf{w}|\mathbf{t},\boldsymbol{\alpha}) P(\mathbf{t}|\boldsymbol{\alpha}) \tag{19}$$

For the classification case, the posterior distribution over the weights $p(\mathbf{w}|\mathbf{t},\boldsymbol{\alpha})$ cannot be evaluated analytically, and posterior statistics are evaluated using the Laplace approximation [*MacKay*, 1992], as proposed by *Tipping* [2001]. With this approach, since $p(\mathbf{w}|\mathbf{t},\boldsymbol{\alpha}) \propto P(\mathbf{t}|\mathbf{w}) p(\mathbf{w}|\boldsymbol{\alpha})$, the most probable weights $\mathbf{w}_{\mathrm{MP}}$ are found by iteratively maximizing the following logistic log likelihood function over the weights $\mathbf{w}$:

$$\log\{P(\mathbf{t}|\mathbf{w}) p(\mathbf{w}|\boldsymbol{\alpha})\} = \sum_{i=1}^{N} [t_i \log y_i + (1-t_i)\log(1-y_i)] - \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} \tag{20}$$

with $y_i = \sigma\{y(\mathbf{x}_i;\mathbf{w})\}$. And, the posterior covariance Σ is obtained by a quadratic approximation to the log-posterior around its mode that is estimated using the previous logistic log likelihood. Thus, at the mode of the posterior distribution, the covariance and most probable weights are, respectively:

$$\Sigma = (\Phi^{\mathrm{T}}\mathbf{B}\Phi + \mathbf{A})^{-1} \tag{21}$$

$$\mathbf{w}_{\mathrm{MP}} = \Sigma \Phi^{\mathrm{T}} \mathbf{B} \mathbf{t} \tag{22}$$

where $\mathbf{B} = \mathrm{diag}(\beta_1, \beta_2, \ldots, \beta_N)$ is a diagonal matrix with $\beta_i = \sigma\{y(\mathbf{x}_i)\}[1-\sigma\{y(\mathbf{x}_i)\}]$.

Identical to the regression case, the hyperparameter posterior $P(\boldsymbol{\alpha}|\mathbf{t})$ is approximated as a delta-function at its most probable value $\boldsymbol{\alpha}_{\mathrm{MP}}$, and values of $\boldsymbol{\alpha}$ that maximize the marginal likelihood are obtained iteratively using the posterior statistics Σ and $\mathbf{w}_{\mathrm{MP}}$ from equations (21) and (22), and updating hyperparameters $\boldsymbol{\alpha}$ until convergence:

$$\alpha_i^{\text{new}} = \frac{(1-\alpha_i \Sigma_{ii})}{\mathbf{w}_{\text{MP}}^2} \tag{23}$$

At the end of the maximization procedure of the marginal likelihood, predictions are made on the basis of the posterior distribution over the weights, conditioned on maximizing values $\boldsymbol{\alpha}_{\text{MP}}$. Thus, given new input data $\mathbf{x}_*$, the probability distribution of the corresponding target y_* is given by the logistic sigmoid predictive distribution:

$$y_* = \sigma\left\{\mathbf{w}_{\text{MP}}^{\text{T}} \phi(\mathbf{x}_*)\right\} \tag{24}$$

which vary from 0 to 1. Thus, the probability of membership to one of the class (class 0 or 1) can be evaluated using equation (24).

For a model with more than two classes, a multiclass classification approach for which a series of binary classifications is performed should be adopted. Two of the common methods for multiclass classification include the one-against-all (1AA) and the one-against-one (1A1) techniques. The 1A1 approach is adopted here because it generally produces better classification performance over the 1AA approach [*Allwein et al*., 2000; *Hsu and Lin*, 2002]. In the 1A1 approach, each class is compared to each other class [*Hastie and Tibshirani*, 1998] and a binary model is built to discriminate between each pair of classes, while discarding the rest of the classes. This requires building $c(c-1)/2$ binary models. When testing new input data $\mathbf{x}_*$, a voting is performed among the various binary models and the class with the maximum number of vote wins, and this class label is assigned to the new input data $\mathbf{x}_*$.

4. Development of the Learning Machine

In this section, we present results of the training of the learning machine, as proposed in section 3, using the training data set described in section 2. Predictions for HF and *K* from the trained learning machine are also verified through a cross-validation procedure using the training data set, and with *K* data from hydraulic tests not used for the development of the learning machine. In addition, robustness as a function of reduced training data set and computational cost of the learning machine are also discussed.

4.1. Fuzzy Clustering for HF Definition

The first step of the training phase is the definition of HFs through fuzzy clustering (Figure 5). The clustering was carried out with the Matlab Fuzzy Clustering and Data Analysis Toolbox [*Balasko et al*., 2005] with all the 280 colocated *K* and CPT/SMR measurements of the training data set. To avoid the largest-valued parameters to bias the clustering, the logarithm distribution of the original data for each parameter was normalized to a common scale using their respective distribution range. Every clustering simulation was also initialized with random seeds. Experiments were done with different seeds not reported here and those results were very similar to the ones given in this paper.

In this study, a systematic procedure was applied to search for the combination of CPT/SMR parameters and the number of HFs to include in the clustering to obtain the learning machine with the best predictive capability. As suggested in Table 3, each direct-push parameter contains information about *K* to various degrees, which used alone or in combination with other parameters may affect in different ways the predictive capability. Also, according to the geological materials present over the study area, hydro-geophysical responses may show different behaviors than what we wish to detect in order to have meaningful relations for each material. Thereby, the systematic search procedure indicated the most salient CPT/SMR parameters and the distinct structures in the training data set.

The predictive capability of the learning machine during clustering was estimated as the overlapping between HFs defined in the geophysical space, which results from the lost of information due to the non-uniqueness between hydraulic and geophysical parameters. It should be remembered that HFs are defined here in the hydro-geophysical space with *K* and direct-push data, but the prediction of HF and *K* is made in the geophysical space only using direct-push data. As illustrated in Figure 6, an improper HF recognition during the prediction process will indeed results in the selection of the wrong hydro-geophysical relation to predict *K*. To assess the degree of HF overlapping for each examined combination of geophysical

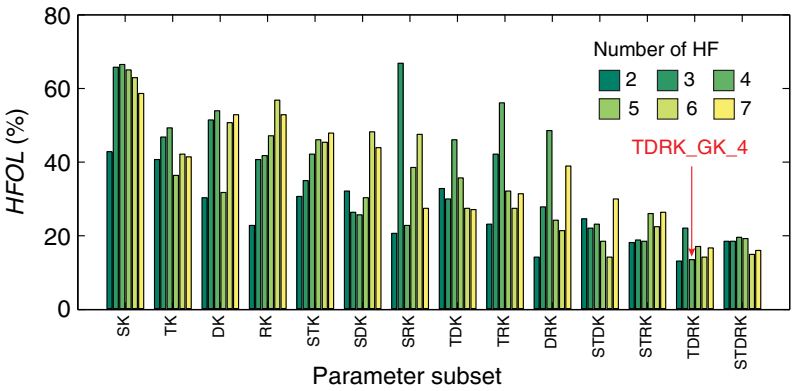

Figure 7. Hydrofacies (HF) overlapping index (HFOL) values resulting from the projection of the HFs defined in the hydro-geophysical space with Gustafson-Kessel (GK) clustering to the geophysical space for different number of HFs and subset of direct-push parameters. Note that hydraulic conductivity (K) is included in all clustering subsets. The arrow indicates the parameter subset (parameters T, D, R, and K with 4 HFs: TDRK_GK_4 subset) with the best predictive capability based on HFOL values, which was used to illustrate the development of the learning machine. Parameter symbols are defined in Table 1.

parameters and HF number, we define the HF overlapping index (*HFOL*) as the percentage of misclassified observations resulting from the projection of HFs from the hydro-geophysical to the geophysical space:

$$HFOL = \frac{|HF_{hg} - HF_g|}{|HF_{hg} \cup HF_g|} \text{x}100 \tag{25}$$

where HF_{hg} and HF_g are HF labels in the hydro-geophysical and geophysical spaces, respectively. This is simply the number of HF labels different to both spaces divided by the total number of HF labels in both spaces. A lower value of *HFOL* indicates less overlapping between HFs and a better potential of recognition from direct-push data. *HFOL* is estimated through clustering by imposing HF coordinate centers obtained in the hydro-geophysical space to new HFs using only geophysical data. Note that only the coordinate centers of the direct-push parameters are used for the new clustering to simulate dimensionality reduction from the hydro-geophysical space to the geophysical space. *HFOL* is thus obtained by comparing labels for each observation in the two spaces after fuzzy memberships are transformed to integer numbers by using the HF label with the maximum fuzzy membership.

The results of the exhaustive search procedure for direct-push parameters and HF number are presented in Figure 7. During this procedure, all combinations of geophysical parameters with various numbers of HFs were individually clustered using the GK algorithm to find HF structures with the lowest *HFOL* value (less HF overlapping). A total of 84 combinations were thus examined with the number of HFs varied between 2 and 7, and *K* data included in all clustering experiments. The fuzziness weighting exponent m was fixed to a value of 2 for all experiments. Note that varying *m* between 1 and 4 did not provide significant differences in *HFOL* values for our data set.

Several observations can be made from Figure 7. First, HF overlapping generally decreases with the number of geophysical parameters used to define HFs, as expressed by decreasing *HFOL* values. For instance, *HFOL* for a subset using only one geophysical parameter is up to 66% (e.g., SK with 4 HFs), while HF overlapping is as low as 13% for subsets using three or four direct-push parameters (e.g., TDRK with 4 HFs). Our interpretation is that for perfectly correlated geophysical parameters with *K*, only one geophysical parameter is necessary to predict *K* from geophysical parameters, and dimensionality reduction is not leading to HF overlapping (assuming noise-free data). However, with weakly correlated geophysical parameters with *K*, and possibly heterogeneous correlations varying according to the geological materials, more geophysical parameters are needed. For that case, any degree of correlation is stretching the cloud of observations in the direction of the correlation. The larger the correlation, or anticorrelation, the stronger is the stretching. Thus, the larger is number of geophysical parameters, even with weak correlation with *K*, the larger is the distortion of the cloud and the more distinct are the HFs projected in the geophysical space.

Another important observation from Figure 7 is that some CPT/SMR parameters are better suited to define HFs with minimal overlapping. According to Figure 7, the clustering experiment that provides lower *HFOL* value is the subset with direct-push parameters *T*, *D*, and *R* with 4 HFs (thereafter referred as the TDRK_GK_4

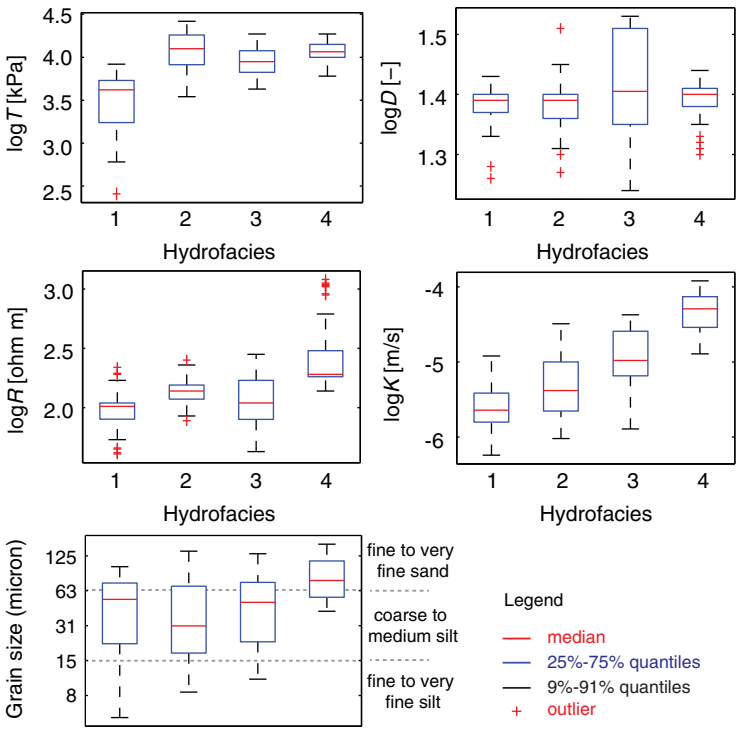

Figure 8. Data distribution for each of the four hydrofacies (HF) resulting from the clustering experiment with the best predictive capability (TDRK_GK_4 subset; see Figure 7) along with the distribution of the mean grain size for sediment associated to each HF. Mean grain-size values are based on sieve analyze of 62 sediment samples colocated with direct-push and hydraulic measurement intervals [see *Paradis et al.*, 2014]. Parameter symbols are defined in Table 1.

subset), which do not includes parameter *S*. Examination of the correlation matrix in Table 3 suggests that the null correlation of S with K and the redundancy with T (high correlation) may explains why this parameter was not retained in the search procedure with geophysical parameters. From Figure 7, we also observe that for the same number of parameters, subsets with *S* generally present the higher *HFOL* values.

Data distribution for each HF and parameter for the TDRK_GK_4 subset are illustrated in Figure 8, where we assigned observations to the highest HF membership. The median and range of values for *K* and each retained direct-push parameter are fairly distinct between HFs with only a few outliers. Particularly, the median values for *K* gradually increase from HF1 to HF4, with slight overlaps between HFs that may be attributed to the complexity of the hydro-geophysical responses and to the transitional nature of the littoral depositional environment. Moreover, each HF presents distinct profiles of *K* and direct-push parameters, as expected from the various sediments composing the aquifer that may present different hydro-geophysical responses. As depicted in Table 4, clustering also results in distinct rank correlation between parameters for each HF and higher rank correlations between geophysical parameters and *K*.

Table 4. Correlations Matrix Showing the Values of the Kendall Rank Correlation for the Logarithm of Direct-Push Parameters and Hydraulic Conductivity for Each Hydrofacies of the TDRK_GK_4 Subset[a]

Parameter	log*T*	log*D*	log*R*	log*K*	Parameter	log*T*	log*D*	log*R*	log*K*
Hydrofacies 1					Hydrofacies 2				
log*T*	1	0.17	0.34	0.18	log*T*	1	−0.22	0.06	−0.39
log*D*		1	−0.18	0.13	log*D*		1	0.50	0.36
log*R*			1	0.02	log*R*			1	0.12
log*K*				1	log*K*				1
Hydrofacies 3					Hydrofacies 4				
log*T*	1	0.51	0.58	0.38	log*T*	1	−0.26	0.07	0.26
log*D*		1	0.70	0.44	log*D*		1	−0.40	−0.08
log*R*			1	0.27	log*R*			1	−0.04
log*K*				1	log*K*				1

[a]Parameter symbols are defined in Table 1.

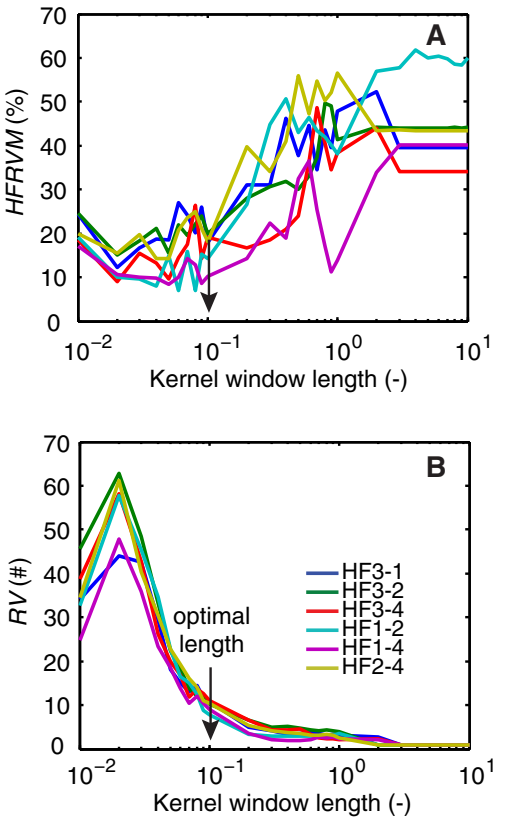

Figure 9. Graph of: (a) the hydrofacies (HF) misclassification error (HFRVM) associated with the classifier performance of the relevance vector machine (RVM), and (b) the number of relevance vector (RV), versus the kernel window length. The arrow on each figure indicates the optimal kernel window length for the training of the RVM for classification. Each curve represents a pair of HF used for the one-against-one (1A1) classification. Results are for a Gaussian kernel function.

We note that few other parameter subsets could provide alternative solution to the TDRK_GK_4 subset (e.g., TDRK_GK_2, TDRK_GK_6, STDK_GK_6, DRK_GK_2). But, we selected the TDRK_GK_4 subset because it separates the data set into 4 HFs that can be correlated to available lithological information to allow a better integration with the littoral depositional model [*Paradis et al*., 2014]. These alternative-clustering solutions could however be used within a geostatistical framework to account for model selection uncertainties. In this paper, only the TDRK_GK_4 subset is explored. Note that the reexpression of the principal component analysis of the data set and features construction as product of original parameters were also tested [e.g., *Guyon and Elisseeff*, 2003]. Those experiments did not however provide lower *HFOL* values than TDRK_GK_4, likely due to the loss of information caused by the filtering process and dimensionality reduction.

4.2. Multiclass RVM Training for HF Models Definition

In order to build predictive HF models for our study site, a multiclass RVM is trained to recognize HFs of the TDRK_GK_4 subset using data for parameters *T*, *D* and *R* (Step 2 in Figure 5). Note that HF data used for the training are integer numbers obtained from the transformation of fuzzy memberships resulting from clustering. Figures 9a and 9b show graphs used for the selection of the optimal kernel window length of the multiclass RVM using a Gaussian kernel function. To find the optimal kernel window length, the classifier performance (Figure 9a) and complexity (Figure 9b) are assessed for various kernel functions and kernel window lengths using training and testing data sets. The classifier performance is defined here by the HF misclassification error associated to the RVM (*HFRVM*) that is evaluated using an equation similar to equation (25):

$$HFRVM = \frac{|HF_c - HF_{RVM}|}{|HF_c \cup HF_{RVM}|} \text{x}100 \tag{26}$$

where HF_c and HF_{RVM} are HF labels from clustering and RVM classification, respectively. Note that HF_c are integers obtained from the transformation of fuzzy memberships, whereas HF_{RVM} are integers resulting from the multiclass voting process. A lower value of *HFRVM* indicates a better HF predictive capability of the RVM. The model structural complexity (sparsity) of the classifier is expressed by the total number of relevance vectors used by the RVM classifier as follow:

$$RV = \sum_{k=1}^{c} RV_k \tag{27}$$

where RV_k is the number of the relevance vectors per HF. A lower RV value produces a smoother solution.

Thus, for a given kernel function and a kernel window length, relevance vectors are first determined with the procedure in section 3.4 using the training data set. Then, HFs are predicted using the testing data set with equation (24) and previous relevance vectors. Finally, performance and sparsity of the classifier are assessed using equations (26) and (27), respectively. This procedure was repeated for different kernel window lengths to produce Figures 9a and 9b with a Gaussian kernel function. Note that we are using

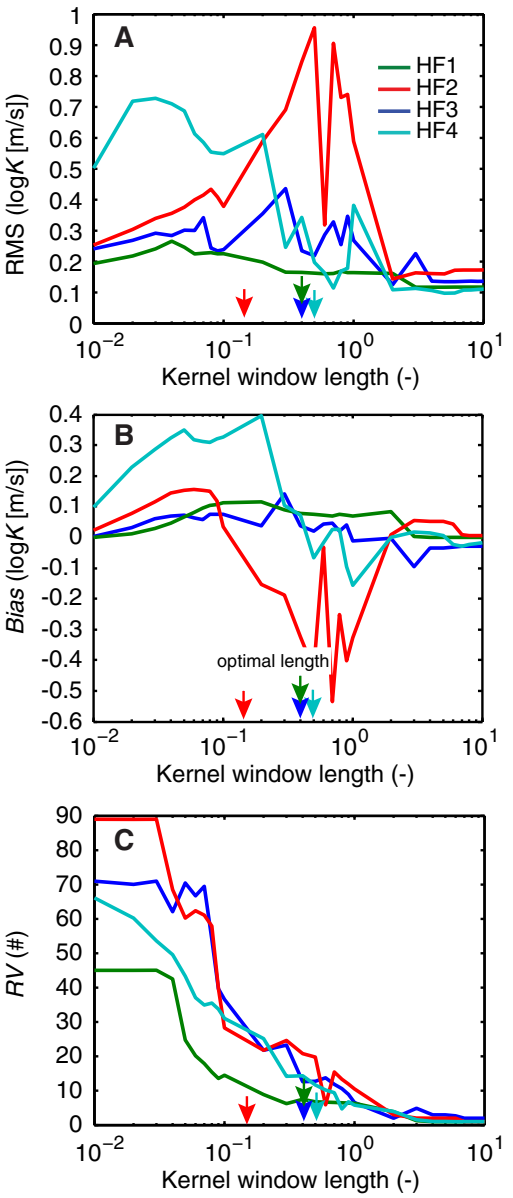

Figure 10. Graph of: (a) the root-mean-square (RMS) error, (b) the mean error (Bias), and (c) the number of relevance vector (RV), versus the kernel window length. Those statistics were used to assess the regressor performance of the relevance vector machine (RVM) for each regression model associated to the four hydrofacies (HF) of the TDRK_GK_4 subset. Arrows indicate the optimal kernel window lengths for each of the four RVMs. Results are for a Laplace kernel function.

normalized parameters and kernel window widths; so different parameters can thus be plotted on the same axis in Figures 9a and 9b. Different kernel functions can also be tested. The supervised classification with RVM was carried out with the SPARSEBAYES Matlab Toolbox [*Tipping and Faul*, 2003; *Tipping*, 2009] using Bernoulli likelihood.

To avoid bias in the selection of training and testing data sets, a 10-fold cross-validation procedure [*Geisser*, 1975] was followed to produce Figures 9a and 9b. Cross validation with partial data splitting is reported to be a robust procedure for model selection of classification problems [*Arlot and Celisse*, 2010]. With this procedure the entire training data set is split randomly into 10 groups of similar size and each group is used in turn as a testing set, while the other nine groups are used together to form a single training set. The average value of the 10 experiments for each kernel window length of a given kernel function is then used to plot *HFRVM* and *RV* curves. We note in Figures 9a and 9b that statistics for *HFRVM* and *RV* are also provided for each of the six binary models used by the 1A1 approach for the training of the multiclass RVM.

According to Figures 9a and 9b, we selected a unique kernel window length of 0.1 as the optimal value for best predictive capability of the RVM classifier with a Gaussian kernel function. Note that other kernel functions were also tested, but those experiments did not provide better predictive capabilities as verified by the procedure presented in section 4.4. The selection of the kernel window length followed the elbow-criterion, to find the optimal number of relevance vectors that strike a balance between overfitting and oversmoothing the testing data, while obtaining a RVM classifier with low *HFRVM* value. On one hand, using a large number of relevance vectors generally leads to a good classifier performance with training data, but to poor generalization capability when used with testing data because the overfitted model describes the random error instead of the underlying relationship [*Tetko et al.*, 1995]. On the other hand, using very few relevance vectors leads to poor classifier performance with both training and testing data due to the oversmoothing of the underlying relationship.

4.3. RVM Regression Training for H-G Relations Definition

Similar to the definition of HF models with RVM classification, kernel functions and kernel window lengths are tested to define H-G relations with the best predictive capabilities using *K*, *T*, *D*, and *R* data of the TDRK_GK_4 subset (Step 3 in Figure 5). As illustrated in Figures 10a–10c for a given Laplace kernel function, a H-G relation is independently defined for each of the four HFs. Note that for developing H-G relation for a given HF, we assign each observation (*K*, *T*, *D*, *R* data) to the HF with the maximum membership after the

fuzzy memberships of each observation obtained from clustering are transformed to integer numbers. Reliability of the H-G relations is defined in terms of goodness-of-fit statistics that also reflect the adequacy and significance of the predicted model. These key statistics are mean error (Bias) and root-mean-square error (RMS):

$$Bias = N^{-1}\sum_{i=1}^{N}(t_{*i} - y_{*i}) \tag{28}$$

$$RMS = \sqrt{N^{-1}\sum_{i=1}^{N}(t_{*i} - y_{*i})^2} \tag{29}$$

Additionally, we used the number of relevance vector defined in equation (27) as an index of structural complexity of the RVM regressor.

For the training of RVM regressors, a LOO cross-validation procedure [*Stone*, 1974] is followed, where each observation is successively left out from the entire training data set and used for testing. This procedure is generally well suited for model selection of regression problems and produces almost unbiased assessment of performances [*Arlot and Celisse*, 2010]. Thus, for a given kernel function and kernel window length, relevance vectors are determined with the procedure described in section 3.3 using all available data except one observation, and *K* prediction is made using equation (16) with the observation left out. This process is repeated until all observations have been used as testing data and the average for all experiments is used to assess performances (Figures 10a and 10b) and sparsity (Figure 10c) of the RVM regression models. The supervised regression was carried out with the SPARSEBAYES Matlab Toolbox [*Tipping and Faul*, 2003; *Tipping*, 2009] using Gaussian likelihood. According to Figures 10a–10c, the optimal kernel window lengths for HF1 to HF4 are 0.4, 0.15, 0.4, and 0.5, respectively. Note that the Laplace kernel function used in Figures 10a–10c provided the best predictive capabilities as verified by the procedure presented in section 4.4.

4.4. Cross Validation of HF and *K* Predictions

In this section, we assess the error associated with the application of the previously trained learning machine to identify HF and predict *K* from CPT/SMR data for our study site. The verification process follows the same sequential steps for prediction illustrated in Figure 6 using training and testing data sets. First, the training data with parameters of the RVMs (kernel function and window length) are used to build HF models and H-G relations. Then, for a given CPT/SMR observation (*T*, *D*, *R* for this example) of the testing set, HF is predicted using HF models, and the H-G relation corresponding to this HF is then applied to estimate *K* using the same CPT/SMR data. Predicted HF and *K* are then compared to known values of the testing set to assess performances. Note that predicted HFs here are the result of the multiclass voting process. To assess classification and regression errors, we randomly selected 80% of the available colocated H-G data and used it as a training set while the remaining 20% was used as a testing set. This procedure was repeated 100 times to provide error distributions associated with the selection of training and testing sets. The same RVM parameters (kernel function and window length) found in sections 4.2 and 4.3 were used for all simulations.

In order to illustrate the performance of the learning machine, we selected the cross-validation simulation with median HFRVM value, along with corresponding *K* values predicted for this simulation, as depicted in Figures 11a and 11b. Figure 11a presents a confusion matrix comparing the HF classification obtained by the RVM classifier to the original classification made by clustering for the simulation with the median *HFRMS* value. The classification error is fairly well distributed over all HFs with *HFRMS* for each HF ranging from 10% to 21%, with a median *HFRMS* value for all HFs of 14%. The median value obtained from the cross-validation procedure is similar to the *HFOL* value of 13% obtained from clustering to evaluate the degree of HF overlapping associated with nonuniqueness between *K* and CPT/SMR parameters. This means that the classification with the RVM is almost perfect, as expressed by close *HFRMS* and *HFOL* values, and the obtained *HFRMS* value is associated with nonuniqueness as discussed in section 4.2.

Figure 11b also presents a scatter plot comparing log*K* estimates obtained by the RVM regressors to the value obtained from slug tests for the same testing data set used in Figure 11a. The correlation coefficient between predicted and field log*K* estimates is 84% and there is no bias in the estimate as the regression

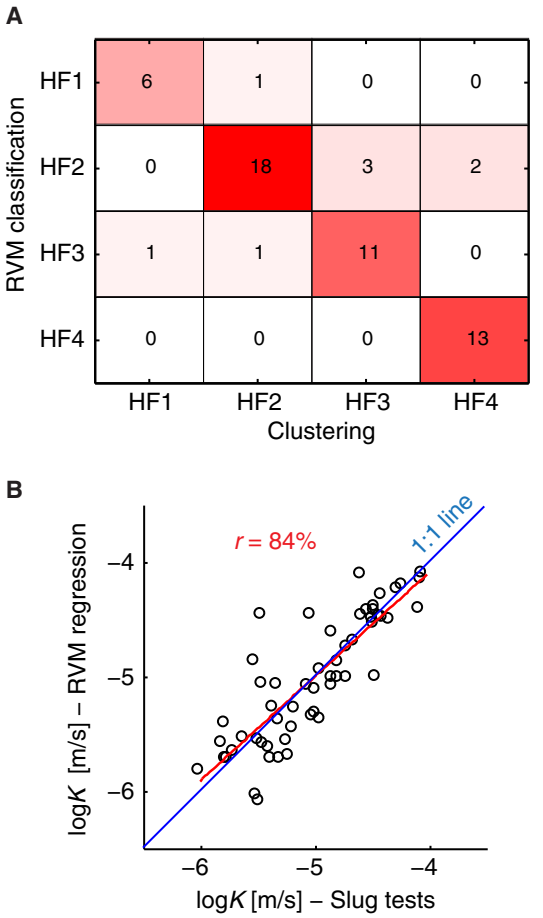

Figure 11. (a) With a testing data set (n = 56 observations), the confusion matrix compares the hydrofacies (HF) classification obtained by the relevance vector machine (RVM) classifier to the original classification made by fuzzy clustering (TDRK_GK_4 subset). This classification corresponds to the simulation using the median HF misclassification error (HFRVM) associated with the RVM classifier. The diagonal indicates the observations for which both classifications are identical (HF1 6/7, 86%; HF2 18/20, 90%; HF3 11/14, 79%; HF4 13/15, 87%; overall 48/56, 86%). Off-diagonal observations were misclassified by RVM classification (HF1 1/7, 14%; HF2 2/20, 10%; HF3 3/14, 21%; HF4 2/15, 13%; overall 8/56, 14%). (b) Comparison of the logarithm of the hydraulic conductivity (K) measured with multilevel slug tests with the estimation made using RVM regressors with the same verification data set shown in Figure 11a.

line overlaps the 1:1 perfect fit line. This is in agreement with results of the cross validation for all the 100 simulations that indicate that there is no significant bias in *K* estimates, as expressed by a median *Bias* value of 0.016, and median *RMS* error of 0.327, which represents approximately 14% of the total range in log*K* values. According to the sequential procedure to estimate *K* in Figure 6, we note that reported error for predicted *K* is cumulative and depends on both the capability to recognize HFs, which depends on nonuniqueness and accuracy of the HF models, and the accuracy of the H-G relations used to make *K* predictions.

4.5. External Verification of *K* Predictions

To further assess the prediction capabilities of the developed learning machine to estimate *K*, we predict *K* values for 3 wells (wells labeled in red in Figure 1c) where colocated CPT/SMR and *K* data were available. A total of 64 *K* data were obtained from flowmeter tests according to the field data acquisition procedure provided by *Paradis et al.* [2011], while direct-push data were rescaled to the same 15 cm intervals of the flowmeter tests following the procedure in section 2.4. Direct-push values used for external verification are within the range of geophysical responses of the training data set (Figures 3 and 4 and Table 2), except for few *D* measurements at well P7 that are slightly above the training range. Predictions of *K* values followed the predictive procedure in Figure 6 with the same kernel functions and kernel window lengths previously used for cross-validation (internal verification) purposes.

Figure 12 presents a composite well plot comparing log*K* estimates obtained by the learning machine to the value obtained from flowmeter tests for the same 15 cm intervals. Several observations can be made from Figure 12. First, the correlation coefficient and *RMS* error between predicted and flowmeter log*K* estimates are 84% and 14%, respectively, which is similar to results of the previous cross-validation procedure. Indeed, because the actual predictive process for external verification uses 100% of the available training data set instead of 80% for the cross validation, equal or slightly better *K* estimates could be expected with new data representative of the data set used for the training of the learning machine. We should note however that the learning machine was trained with *K* data from slug tests that may differ from *K* estimates from flowmeter tests. Indeed, the study of *Paradis et al.* [2011] that compared *K* estimates from flowmeter and multilevel slug tests, for 123 of the 280 intervals of the training data set used to develop the learning machine, showed correlation coefficient and *RMS* error of 88% and 10%, respectively. While there are certainly differences between the various methods to estimate *K*, it appears that the proposed indirect method to estimate *K* from CPT/SMR data compares fairly well with direct methods based on hydraulic tests (slug and flowmeter tests). Finally, although the general trend in log*K* values is similar for both predicted and flowmeter distributions, the range in predicted log*K* is slightly narrower than the

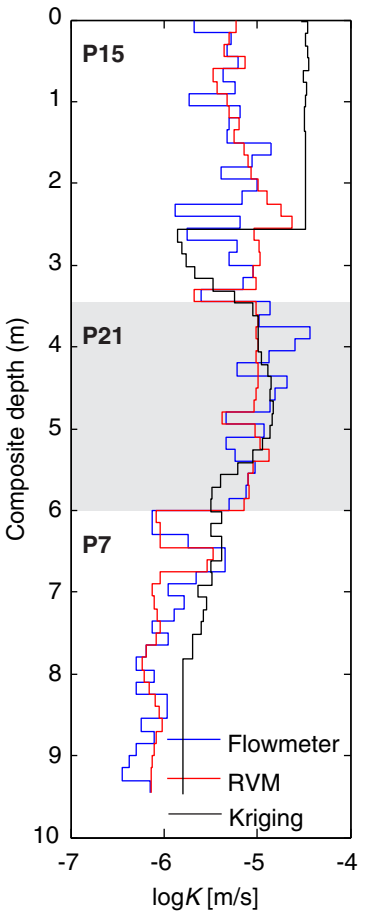

Figure 12. Composite well plot comparing estimated hydraulic conductivity (K) values from flowmeter tests, the RVM-based learning machine, and kriging interpolation for wells P15, P21, and P7. Locations of the well are indicated in Figure 1c (wells labeled in red). The 64 K values from flowmeter tests and the corresponding direct-push data were not used for the training of the learning machine.

observed range. This indicates that the learning machine smooths log*K* estimations, which is inherent to any estimation process (classification and regression).

4.6. *K* Estimates at New Locations: Geophysical or Spatially Informed?

In previous sections, we demonstrated that CPT/SMR data could provide accurate information about *K* through a learning machine process. However, it may be interesting to see if such an approach is worth the effort in term of prediction accuracy with respect to a kriging approach that make uses only of available direct *K* data from hydraulic testing. This would help answer a fundamental question for the study site, which is to assess whether accuracy of *K* predicted at new locations is the result of spatial correlation between *K* estimates at sampled and new locations, as spatial correlation could be implicitly inscribed into the hydro-geophysical relations that we previously defined, or instead if the geophysical data really contributing to information about *K* at new locations.

Figure 12 presents *K* estimates at P15, P21, and P7 using kriging of the 280 *K* values of the training data set. The modeling of the spatial structure for kriging interpolation consists in distinct three nested structures in the horizontal and vertical directions. The modeled variogram defined using *K* data estimated from slug tests along P4, P6, P11, P17, P10, P1, and P3 (Figure 1) has the following characteristics: nugget value of 0.023 $(m/s)^2$ and two spherical models (horizontal ranges = 500/4000 m and vertical ranges = $3.1/1 \times 10^6$ m) with sills = 0.129/0.5 $(m/s)^2$. Those semivariogram parameters were needed to match the nonstationary (quasi-linear) experimental semivariogram, which is the result of the transitional littoral environment of the study area that shows spatially varying grain-size sediments, and then hydraulic properties, according to the distance from the paleoshoreline [*Paradis et al.*, 2014]. Due to the strong difference in scale between horizontal and vertical semivariograms, a search radius procedure for interpolation using the six closest observations was defined to avoid numerical instability and to compensate for the nonstationary of the geostatistical model.

Examination of Figure 12 reveals that kriging results do not match very well flowmeter estimates, especially for wells P7 and P15 that are far (more than few hundreds meters) from well with available direct *K* data (Figure 1). Except for well P21, that is very close to well P17 (<10 m), information about *K* at new locations is hardly provided by spatial correlation and converted geophysical data better predict *K* information over the study area. This is obviously expected from the complex geology and the size of the study area with respect to the number of direct *K* data available, which both preclude developing a meaningful geostatistical model and conditioning of the interpolation.

4.7. Robustness of the Learning Machine to Reduce Training Data Set

An important unknown in aquifer characterization is the number of data needed to appropriately characterize a site, which are often limited due to the cost associated to data collection. In the context of this study, it is worth asking when a sufficient number of data have been acquired to train the learning machine. Thus, we tested the robustness of the learning machine as a function of its accuracy when reducing the size of

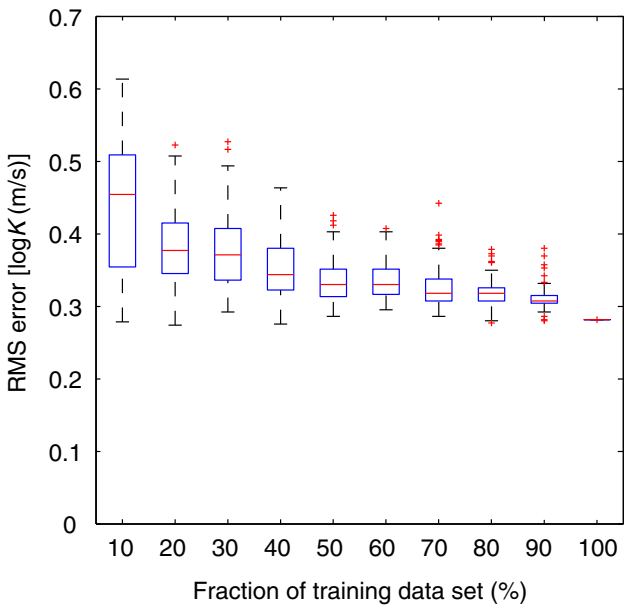

Figure 13. Whisker plot showing the evolution of the RMS error evaluated using the flowmeter testing set (see Figure 12) versus the size of the training data set used by the learning machine. For each whisker box, 100 random training data set were simulated, except for the data set using 100% of all the 280 observations of the training data set.

the training data sets. Figure 13 shows the evolution of the *RMS* error evaluated using the flowmeter testing set versus the size of the training data set used by the learning machine. For each decimated data set, we randomly selected 100 training data sets and plotted the *RMS* error distribution as illustrated in Figure 13. Note that the full data set with 100% use all of the 280 observations of the training data set, and its statistics corresponds to the external verification discussed in section 4.5. Predictions for all data sets followed the procedure in Figure 6 using previous optimal RVMs parameters.

Several observations can be made from Figure 13. First, while the median *RMS* error increases with reducing data set size, the minimum *RMS* error is approximately constant for all reduced data sets and close to the *RMS* error with 100% of the training data set. This suggests that as long as the reduced training data set has the same statistical characteristics as the full training data set, the accuracy of the predictions will remain the same even with a reduced data set. RVMs are thus robust predictors with sparse data set, as theoretically claimed. However, accuracy of K predictions may decrease drastically with a reduced data set, as expressed by the spreading between minimal and maximal *RMS* errors. This raises the question of representativeness of a training data set to train a learning machine, and consequently to properly characterize hydraulic properties of a site. While analysis of Figure 13 cannot tell us whether the actual training data set used to develop the learning machine for our site is adequate, the use of a larger data set certainly increases the chance to get a representative training data set, as expressed by the narrowing of *RMS* error spreading for larger data sets. Finally, the spreading of the *RMS* error for different reduced data sets may be an indication of the complexity of the training data set. Statistically, more homogenous data sets would present narrower *RMS* error spreading, and better precision for a reduced data set.

4.8. Computational Cost

Typically, the computational cost of RVMs increases with the number of observations used since the complexity and memory storage of the computational operations scale with the square and the cubic of the number of basis functions M, respectively. Thus, for problems with a high number of training data ($N > 1000$), using RVM could be prohibitively expensive [e.g., *Khader and McKee*, 2014]. In our case study, the computational burden was not a problem because training a single model requires only a few seconds using 280 observations (QuadCore i7@2.2Ghz on OSX platform). The computational cost for our application is thus mostly related to the number of simulations used for cross validation (e.g., kernel window length selection, performances assessment). For instance, the computation of Figure 9 for classification model selection was carried out in less than 15 min. Note that in this study, we used a fast implementation of RVMs that optimizes the marginal likelihood function through sequential addition and deletion of candidate basis functions [*Tipping and Faul*, 2003].

5. Summary and Conclusions

This paper presented a learning machine approach to define site-specific relationships in order to estimate aquifer hydraulic properties based solely on geophysical measurements. Specifically, we explored the use of

CPT/SMR soundings data for *K* estimation using a statistical framework combining fuzzy clustering and RVMs. HFs reflecting geological materials present within the studied aquifer were first extracted from a training data set composed of *K* data measured in wells using 15 cm vertical resolution packer slug tests and CPT/SMR data that include resistance to penetration (tip stress, *T*), mechanical friction (sleeve stress, *S*), dielectric constant of bulk sediments (*D*) and DC electrical resistivity (*R*). All colocated *K* and CPT/SMR data were up-scaled to a common vertical resolution of 15 cm for the purpose of establishing H-G relations. RVMs for classification and regression were then trained independently using previous clustering data to define predictive HF models and H-G relations, respectively. Accuracy of HF and *K* estimates using the developed learning machine was assessed, through a cross-validation procedure with the training data set and by external verification with *K* data not used during the training process, to evaluate the potential of CPT/SMR data for *K* estimation. Important conclusions and observations resulting from this study, which can be generalized to successfully employing geophysical data for hydraulic property characterization, include the following:

The combined use of CPT/SMR soundings and RVM-based learning machine hold the potential to estimate *K* at a high-vertical resolution under real field conditions, as indicated by results of the cross-validation and external verification. Factors that contributed to these promising results are twofold:

1. First, in addition to the vertical decimeter-scale resolution offered by direct-push soundings, the multiparameter CPT/SMR probe contributes to reduce nonuniqueness between geophysical and hydraulic parameters by providing a series of complementary geophysical parameters to correlate with *K*. As seen in Figure 7, a better potential of accurate estimation of hydraulic data is generally achieved when using more geophysical parameters. However, despite up to four geophysical parameters from the CPT/SMR probe were available to correlate with *K*, it was not possible to completely resolve nonuniqueness, as expressed by the nonzero degree of HFs overlapping. This remaining uncertainty is inherent to measurement and interpretation errors, but also to the fundamental nature of the relationships between geological materials and corresponding hydro-geophysical responses. While the degree of limitation to fully represent hydraulic properties through indirect geophysical approaches may vary from site to site, the choice and number of geophysical parameters to use are crucial to ensure meaningful hydrogeological interpretation of geophysical data.
2. Second, the learning machine composed of fuzzy clustering and RVMs offers a robust and flexible approach to build meaningful relations between hydraulic and geophysical data. Indeed, the division of the training data set in HFs with distinct hydro-geophysical responses, which can be associated to different geological materials, contributes to alleviate the complexity in the established relations. In addition, RVMs for classification and regression are effective to establish HF models and H-G relations with good generalization capabilities, which is critical for successfully employing geophysical data for hydraulic properties characterization.

As nonparametric learning machines are based on empirical data, the selection of a representative training data set is fundamental to establish meaningful relationships for a particular study area. Three aspects here have to be pointed out:

1. First, as suggested in Figure 13, no single statistical algorithm can compensate for an unrepresentative training data set. Whether a training data set is "truly" representative of a specific study area is an unanswered question because the "reality" would be always hidden. In this context, a comprehensive data acquisition approach should be developed and adopted [e.g., *Bradford and Babcock*, 2013; *Paradis et al.*, 2014] to ensure a more representative training data set. Such approaches are needed to ensure the coverage of the entire range of hydraulic and geophysical responses present over a study area.
2. Second, the acquisition of *K* data for aquifer characterization based on H-G relations leads to new ways to target hydraulic tests, which are carried out in the perspective of providing *K* values over the observed range of geophysical responses in a given study area. Such a perspective and approach can contribute to a more efficient aquifer characterization process because less time-consuming hydraulic tests are needed. This is especially true for study area with complex geology where difficult to obtain direct *K* data from hydraulic tests in few wells are not sufficient to provide a meaningful geostatistical model.
3. Third, high-resolution hydraulic testing has to meet more specific criteria than conventional testing over long screens. Notably, in the design of suitable observation wells for reliable *K* estimates (e.g., sand-pack free and fully screened well), and more efficient hydraulic testing approaches (e.g., direct-push hydraulic testing, flowmeter tests) to acquire larger high-resolution *K* training data sets.

Acknowledgments
The authors would like to acknowledge the important technical support provided by J.-M. Ballard and D. Martin as well as Y. Michaud and D. Kirkwood for their support. The Geological Survey of Canada (Groundwater Geoscience Program), the Régie intermunicipale de gestion des déchets des Chutes-de-la-Chaudière, and NSERC Discovery Grants held by E.G. and R.L supported this study. This is an Earth Science Sector contribution 20130456. This paper has benefited from the generous comments of G. Bohling, L. Bentley, R. Martel, and two anonymous reviewers.

References

Al-Anazi, A., and I. D. Gates (2010a), A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs, *Eng. Geol.*, *114*, 267–277.

Al-Anazi, A., and I. D. Gates (2010b), On the capability of support vector machines to classify lithology from well logs, *Nat. Resour. Res.*, *19*, 125–139.

Al-Anazi, A., I. D. Gates, and J. Azaiez (2009), Fuzzy logic data-driven permeability prediction for heterogeneous reservoirs, paper SPE 121159 presented at the Society of Petroleum Engineers of the EUROPEC/EAGE Annual Conference and Exhibition, Soc. of Pet. Eng., Amsterdam.

Allwein, E., R. Shapire, and Y. Singer (2000), Reducing multiclass to binary: A unifying approach for margin classifiers, *J. Mach. Learn. Res.*, *1*, 113–141.

American Society for Testing and Materials (ASTM) (2012), D5778-12: Standard test method for electronic friction cone and piezocone penetration testing of soils, ASTM International, West Conshohocken, Pa.

Anderson, M. P. (1989), Hydrogeologic facies models to delineate large-scale spatial trends in glacial and glaciofluvial sediments, *Geol. Soc. Am. Bull.*, *101*, 501–511.

Anderson, M. P. (1997), Characterization of geological heterogeneity, in *Stochastic Subsurface Hydrology*, edited by G. Dagan and S. P. Neuman, 23–43, Cambridge Univ. Press, Cambridge, U. K.

Archie, G. E. (1942), The electrical resistivity log as an aid in determining some reservoir characteristics, *Pet. Trans. AIME*, *146*, 54–62.

Arlot, S., and A. Celisse (2010), A survey of cross-validation procedures for model selection, *Stat. Surv.*, *4*, 40–79.

Balasko, B., J. Abonyi, and B. Feil (2005), *Fuzzy Clustering and Data Analysis Toolbox for Use With Matlab*, Univ. of Veszprem, Veszprem, Hungary. [Available at http://www.mathworks.com/matlabcentral/fileexchange/7473, last accessed 29 Mar. 2011.]

Bezdek, J. C. (1981), *Pattern Recognition With Fuzzy Objective Function Algoritms*, Plenum, N. Y.

Bolduc, A. (2003), Géologie des formations superficielles: Charny (Québec), Dossier public 1776, Commission géologique du Canada, échelle 1/50000.

Bouwer, H., and R. C. Rice (1976), A slug test method for determining hydraulic conductivity of unconfined aquifers with completely or partially penetrating wells, *Water Resour. Res.*, *12*, 423–428.

Bradford, J. H., and E. Babcock (2013), The need to adapt the exploration model from the oil patch to contaminated-site characterization: A case from Hill AFB, Utah, USA, *Leading Edge*, *32*(7), 750–756.

Butler, J. J., Jr. (2005), Hydrogeological methods for estimation of hydraulic conductivity, in *Hydrogeophysics*, edited by Y. Rubin and S. Hubbard, pp. 23–58, Springer, N. Y.

Butler, J. J., Jr., J. M. Healey, G. W. McCall, E. J. Garnett, and S. P. Loheide II (2002), Hydraulic tests with direct-push equipment, *Ground Water*, *40*(1), 25–36.

Butler, J. J., Jr., P. Dietrich, V. Wittig, and T. Christy (2007), Characterizing hydraulic conductivity with the direct-push permeameter, *Ground Water*, *45*(4), 409–419.

Camps-Valls, G., L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, J. Amorós-López, and J. Calpe-Maravilla (2006), Retrieval of oceanic chlorophyll concentration with relevance vector machines, *Remote Sens. Environ.*, *105*, 23–33.

Chen, J., and Y. Rubin (2003), An effective Bayesian model for lithofacies estimation using geophysical data, *Water Resour. Res.*, *39*(5), 1118, doi:10.1029/2002WR001666.

Chen, J., S. Hubbard, and Y. Rubin (2001), Estimating the hydraulic conductivity at the south oyster site from geophysical tomographic data using Bayesian techniques based on the normal linear regression model, *Water Resour. Res.*, *37*, 1603–1613, doi:10.1029/2000WR900392.

Copty, N., Y. Rubin, and G. Mavko (1993), Geophysical-hydrological identification of field permeabilities through Bayesian updating, *Water Resour. Res.*, *29*, 2813–2825.

Davis, J. C. (1973), *Statistics and Data Analysis in Geology*, John Wiley, N. Y.

Day-Lewis, F. D., K. Singha, and A. Binley (2005), Applying petrophysical models to radar traveltime and electrical resistivity tomograms: Resolution-dependent limitations, *J. Geophys. Res.*, *110*, B08206, doi:10.1029/2004JB003569.

Dubois, M. K., G. C. Bohling, and S. Chakrabarti (2007), Comparison of four approaches to a rock facies classification problem, *Comput. Geosci.*, *33*, 599–617.

Dubreuil-Boisclair, C., E. Gloaguen, D. Marcotte, and B. Giroux (2011), Heterogeneous aquifer characterization from ground-penetrating radar tomography and borehole hydrogeophysical data using nonlinear Bayesian simulations, *Geophysics*, *76*(4), J13–J25.

Dunn, J. C. (1973), A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *J. Cybern.*, *3*, 32–57.

Elshafei, M., and G. Hamada (2009), Neural network identification of hydrocarbon potential of shaly sand reservoirs, *J. Pet. Sci. Technol.*, *27*, 72–82.

Farrar, J. A. (1996), Research and standardization needs for direct push technology applied to environmental site characterization, in *Sampling Environmental Medias, ASTM Spec. Tech. Publ. 1282*, edited by J. H. Morgan, pp. 93–107, Am. Soc. for Test. and Mater., Philadelphia, Pa.

Fellenius, B. H., and A. Eslami (2000), Soil profile interpreted from CPTu data, paper presented at Year 2000 Geotechnics, Geotechnical Engineering Conference, Asian Inst. of Technol., Bangkok, 27–30 November.

Garambois, S., P. Sénéchal, and H. Perroud (2002), On the use of combined geophysical methods to assess water content and water conductivity of near-surface formations, *J. Hydrol.*, *259*(1-4), 32–48.

Geisser, S. (1975), The predictive sample reuse method with applications, *J. Am. Stat. Assoc.*, *70*, 320–328.

Ghosh, S., and P. Mujumdar (2008), Statistical downscaling of GCM simulations to streamflow using relevance vector machine, *Adv. Water. Resour.*, *31*, 132–146.

Gloaguen, E., M. Chouteau, D. Marcotte, and R. Chapuis (2001), Estimation of hydraulic conductivity of an unconfined aquifer using cokriging of GPR and hydrostratigraphic data, *J. Appl. Geophysics*, *47*(2), 135–152, doi:10.1016/S0926-9851(01)00057-X.

Gustafson, D. E., and W. C. Kessel (1979), Fuzzy clustering with a fuzzy covariance matrix, in *Proceeding of IEEE Conference on Decision and Control Including the 17th Symposium on Adaptive Processes*, pp. 761–766, IEEE, San Diego, Calif.

Guyon, I., and A. Elisseeff (2003), An introduction to variable and feature selection, *J. Mach. Learn. Res.*, *3*, 1157–1182.

Hastie, T., and R. Tibshirani (1998), Classification by pairwise coupling, *Ann. Stat.*, *26*, 451–471.

Hsu, C. W., and C. J. Lin (2002), A comparison of methods for multi-class support vector machines, *IEEE Trans. Neural Networks*, *13*, 415–425.

Hyndman, D., J. Harris, and S. Gorelick (2000), Inferring the relation between seismic slowness and hydraulic conductivity in heterogeneous aquifers, *Water Resour. Res.*, *36*, 2121–2132.

Isaaks, E. H., and R. M. Srivastava (1989), *An Introduction to Applied Geostatistics*, Oxford Univ. Press, N. Y.

Iturrarán-Viveros, U., and J. O. Parra (2014), Artificial Neural Networks applied to estimate permeability, porosity and intrinsic attenuation using seismic attributes and well-log data, *J. Appl. Geophys.*, *107*, 45–54, doi:10.1016/j.jappgeo.2014.05.010.

Khader, A. I., and M. McKee (2014), Use of a relevance vector machine for groundwater quality monitoring network design under uncertainty, *Environ. Model. Software*, *57*, 115–126, doi:10.1016/j.envsoft.2014.02.015.
Khalili, A., M. N. Almasri, M. McKee, and J. J. Kaluarachchi (2005), Applicability of statistical learning algorithms in groundwater quality modeling, *Water. Resour. Res.*, *41*, W05010, doi:10.1029/2004WR003608.
Kharrat, R., R. Mahdavi, H. Bagherpour, and S. Hejri (2009), Rock type and permeability prediction of a heterogeneous carbonate reservoir using artificial neural networks based on flow zone index approach, paper SPE 120166 presented at SPE Middle East Oil and Gas Show and Conference, Soc. of Pet. Eng., Bahrain, 15–18 March.
Köber, R., G. Hornbruch, C. Leven, L. Tischer, J. Grossmann, P. Dietrich, H. Weiss, and A. Dahmke (2009), Evaluation of combined direct-push methods used for aquifer model generation, *Ground Water*, *47*(4), 536–546, doi:10.1111/j.1745-6584.2009.00554.x.
Koltermann, C. E., and S. M. Gorelick (1996), Heterogeneity in sedimentary deposits: A review of structure-imitating, process-imitating, and descriptive approaches, *Water Resour. Res.*, *32*, 2617–2658.
Lafuerza, S., M. Canals, J. L. Casamor, and J. M. Devincenzi (2005), Characterization of deltaic sediment bodies based on in situ CPT/CPTU profiles: A case study on the Llobregat delta plain, Barcelona, Spain, *Mar. Geol.*, *222-223*, 497–510.
Lee, S. H., and A. Datta-Gupta (1999), Electrofacies characterization and permeability predictions in carbonate reservoirs: Role of multivariate analysis and nonparametric regression, paper presented at SPE Annual Technical Conference and Exhibition, Soc. of Pet. Eng., Houston, Tex., 3–6 October.
Lessoff, S. C., U. Schneidewind, C. Leven, P. Blum, P. Dietrich, and G. Dagan (2010), Spatial characterization of the hydraulic conductivity using direct-push injection logging, *Water Resour. Res.*, *46*, W12502, doi:10.1029/2009WR008949.
Levy, B. S., and L. Pannell (1991), Evaluation of a pressure system for estimating in-situ hydraulic conductivity, in *Proceedings of 5th National Outdoor Action Conference*, pp. 131–146, NWWA, Dublin Ohio.
Liu, G., J. J. Butler Jr., G. C. Bohling, E. Reboulet, S. Knobbe, and D. W. Hyndman (2009), A new method for high-resolution characterization of hydraulic conductivity, *Water Resour. Res.*, *45*, W08202, doi:10.1029/2009WR008319.
Lunne, T., P. K. Robertson, and J. J. M. Powell (1997), *Cone Penetration Testing in Geotechnical Practice*, Spon Press, N. Y.
MacKay, D. J. C. (1992), The evidence framework applied to classification networks, *Neural Comput.*, *4*, 720–736.
Mingoti, S. A., and J. O. Lima (2006), Comparing SOM neural network with fuzzy *c*-means, *k*-means and traditional hierarchical clustering algorithms, *Eur. J. Oper. Res.*, *174*, 1742–1759.
Mitchell, T. (1997), *Machine Learning*, McGraw-Hill, N. Y.
Mohaghegh, S., B. Balan, and S. Ameri (1997), Permeability determination from well log data, *SPE Form. Eval.*, *12*, 170–174.
Ouellon, T., R. Lefebvre, D. Marcotte, A. Boutin, V. Blais, and M. Parent (2008), Hydraulic conductivity heterogeneity of a local deltaic aquifer system from the kriged 3D distribution of hydrofacies from borehole logs, Valcatier, Canada, *J. Hydrol.*, *351*, 71–86.
Paasche, H., J. Tronicke, K. Holliger, A. G. Green, and H. R. Maurer (2006), Integration of diverse physical-property models: Subsurface zonation and petrophysical parameter estimation based on fuzzy c-means cluster analyses, *Geophysics*, *71*(3), H33–H44, doi:10.1190/1.2192927.
Paradis, D., R. Lefebvre, R. H. Morin, and E. Gloaguen (2011), Permeability profiles in granular aquifers using flowmeters in direct-push wells, *Ground Water*, *49*, 534–547.
Paradis, D., L. Tremblay, R. Lefebvre, E. Gloaguen, A. Rivera, M. Parent, J.-M. Ballard, Y. Michaud, and P. Brunet (2014), Field characterization and data integration to define the hydraulic heterogeneity of a shallow granular aquifer at a sub-watershed scale, *Environ. Earth Sci.*, *72*, 1325–1348, doi:10.1007/s12665-014-3318-2.
Qiu, D. (2010), A comparative study of the k-means algorithm and the normal mixture model for clustering: Bivariate case, *J. Stat. Plann. Inference*, *140*, 1701–1711.
Qiu, D., and A. C. Tamhane (2007), A comparative study of the k-means algorithm and the normal mixture model for clustering: Univariate case, *J. Stat. Plann. Inference*, *137*, 3722–3740.
Robertson, P. K. (1990), Soil classification using the cone penetration test, *Can. Geotech. J.*, *27*, 151–158.
Ross, H. C., and C. D. McElwee (2007), Multi-level slug tests to measure 3-D hydraulic conductivity distributions, *Nat. Resour. Res.*, *16*, 67–79.
Rubin, Y., and S. Hubbard (2005), *Hydrogeophysics*, Springer, Dordrecht, Netherlands.
Ruggeri, P., J. Irving, E. Gloaguen, and K. Holliger (2013), Regional scale integration of multiresolution hydrological and geophysical data using a two-step Bayesian sequential simulation approach, *Geophys. J. Int.*, *194*, 289–303.
Rumpf, M., and J. Tronicke (2014), Predicting 2D geotechnical parameter fields in near surface sedimentary environments, *J. Appl. Geophys.*, *101*, 95–107.
Samui, P. (2007), Seismic liquefaction potential assessment by using relevance vector machine. *Earthquake Eng. Eng. Vibration*, *6*, 331–336.
Schulmeister, M., J. J. Butler Jr., J. Healey, L. Zheng, D. Wysocki, and G. McCall (2003), Direct-push electrical conductivity logging for high-resolution hydrostratigraphic characterization, *Ground Water Monit. Rem.*, *23*, 52–62.
Shinn, J. D., D. A. Timian, R. M. Morey, G. Mitchell, C. L. Antle, and R. Hull (1998), Development of a CPT deployed probe for in situ measurement of volumetric soil moisture content and electrical resistivity, *Field Anal. Chem. Tech.*, *2*, 103–110.
Shokir, E. M. El-M., A. Ateeq, and A. Al-Sughayer (2006), Permeability estimation from well log responses, *J. Can. Pet. Technol.*, *45*, 41–46.
Steelman, C. S., and A. L. Endres (2011), Comparison of petrophysical relationships for soil moisture estimation using GPR ground waves, *Vadose Zone J.*, *10*(1), 270–285.
Stone, M. (1974), Cross-validatory choice and assessment of statistical predictions, *J. R. Stat. Soc., Ser. B*, *36*, 111–147.
Tetko, I. V., D. J. Livingstone, and A. I. Luik (1995), Neural network studies: 1. Comparison of overfitting and overtraining, *J. Chem. Inf. Comput. Sci.*, *35*, 826–833.
Tipping, M. E. (2001), Sparse Bayesian learning and the relevance vector machine, *J. Mach. Learn. Res.*, *1*, 211–244.
Tipping, M. E. (2009), SPARSEBAYES: An efficient Matlab implementation of the sparse Bayesian modelling algorithm (Version 2.0). [Available at http://www.miketipping.com.]
Tipping, M. E., and A. C. Faul (2003), Fast marginal likelihood maximisation for sparse Bayesian models, in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, edited by C. M. Bishop and B. J. Frey, Key West, Fla.
Tremblay, L., R. Lefebvre, D. Paradis, and E. Gloaguen, (2014), Conceptual model of leachate migration in a granular aquifer derived from the integration of multi-source characterization data (St-Lambert, Canada), *Hydrogeol. J.*, *22*(3), 587–608, doi:10.1007/s10040-013-1065-1.
Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer, N. Y.
Vapnik, V. (1998), *Statistical Learning Theory*, John Wiley, N. Y.
Wong, P. M., D. J. Henderson, and L. J. Brooks (1998), Permeability determination using neural networks in the Ravva Field, Offshore India, *SPE Form. Eval.*, *1*, 99–104.
Yamamoto, T., T. Ney, and M. Kuru (1994), Porosity, permeability, shear strength: Crosswell tomography below an iron foundry, *Geophysics*, *59*, 1530–1541, doi:10.1190/1.1443542.