A framework for accurate geospatial modeling using image ranking and machine learning

Peter Bajcsy, Yu-Feng Lin, Alex Yahja and Chulyun Kim

ABSTRACT

There is a large class of modeling problems where the complexity of the underlying phenomena is overwhelming and hence the accuracy of mathematical models is limited. Our approach to this class of problems is to design frameworks that bring together physically based and data-driven models, and incorporate the tacit knowledge of experts by providing visual exploration and feedback capabilities. This paper presents such a novel computer-assisted framework for accurate geospatial modeling applied to improve groundwater recharge and discharge (R/D) patterns. The novelty of our work is in designing a methodology for ranking and extracting relationships, as well as in developing a general framework for building accurate geospatial models. The framework combines variables derived using physically based inverse modeling with auxiliary geospatial variables directly sensed, ranks variables and extracts variable relationships using data-driven (''machine learning'') techniques, and supports partially expert-driven trial-and-error experimentation and more rigorous optimization, as well as visual explorations, to derive more accurate models for R/D pattern estimation. When the framework was tested by experts, it led to a high level of consistency between the machine-learning-based knowledge and the experts' knowledge about R/D distribution. The prototype solution of the framework is available for downloading at http://isda.ncsa.uiuc.edu/Sp2Learn/.

Key words | groundwater modeling, machine learning, optimization, pattern recognition

Peter Bajcsy (corresponding author)
Alex Yahja
National Center for Supercomputing Applications,
University of Illinois at Urbana-Champaign,
1205 W. Clark,
Urbana,
IL 61801,
USA
E-mail: *pbajcsy@ncsa.uiuc.edu*

Yu-Feng Lin
Illinois State Water Survey,
University of Illinois at Urbana-Champaign,
Urbana,
IL 61801,
USA

Chulyun Kim
Department of Software Design and Management,
Kyungwon University,
Gyeonggi-Do, 461-701,
Korea

INTRODUCTION

There is a large class of problems where the complexity of the underlying phenomena is overwhelming and therefore physics-/chemistry-/biology-based modeling is limited in its accuracy of estimations. The accuracy is typically limited due to (1) the unavoidable *uncertainty* of direct and indirect measurements feeding complex models, (2) the high *cost of measurements* collected at sufficient sampling frequencies in space, time and variable dimensions, (3) the insufficient *computational resources* to process all input measurements and execute complex models to make accurate predictions, (4) the lack of *scalability* of existing models coming from unsatisfied approximations/conditions of the models and/or implementations not utilizing computational resources, and (5) the *simplifications in modeling*, for instance, by using only a subset of variables relevant to predicting the underlying phenomena. This work is motivated by seeking approaches to overcome the inaccuracies of the modeling of complex phenomena.

The class of problems described above occurs in applications coming from the Earth Science domain, although similar problems concerned with improving modeling accuracy might be found in other domains. The applications from Earth Sciences would include the prediction of precipitation (Liu *et al.* 2008), estimation of recharge/discharge rates (Lin *et al.* 2007, 2008a) from various indirect measurements, predictions of algal biomass in rivers (Bajcsy *et al.* 2006), hypoxia along coasts (Coopersmith 2008) or simulations of optimal sewage networks (Torres 2007). For instance, prediction of

doi: 10.2166/hydro.2010.187

precipitation from the National Weather Service (NWS) Next Generation Weather Radar (NEXRAD) maps (Liu *et al.* 2008) is based upon the original prediction model that corresponds to the conversion of Level II reflectivity to rainfall rates by using the convective Z–R relationship (Fulton *et al.* 1998). Fusing accumulation gauge data to the radar data using a dynamic Bayesian method as suggested by Hill *et al.* (2009) would produce a more accurate estimate of the 20 min precipitation accumulations at the gauge locations (Liu *et al.* 2008). Thus, one can improve the accuracy of modeling of such complex phenomena by selecting new variables, assuming their relevance and by exploring their relationships to the predicted values. Similarly, estimation of groundwater recharge and discharge (R/D) rates can be based on multiple existing physically based models (Scanlon *et al.* 2002). However, the measurements of such physically based models are difficult and expensive to collect, and a physically based inverse approach provides an alternative, employing readily available information to decrease the processing time (Lin *et al.* 2008b). These models might not generate estimations at fine geospatial scales because the model assumptions have not been met. In this case, exploring the data-driven rules between additional variables, such as surface water feature proximity, topographic slope or soil type, could also lead to model improvement (Kim *et al.* 2007). For brevity, we provide only references to other studies and predictions of (a) hypoxia in the Corpus Christi Bay, TX (Coopersmith 2008) using the hydrodynamics model based on ELCIRC (Zhang *et al.* 2004) and (b) simulations of optimal redesign of a sewage network in the city of Cali, Columbia (Torres 2007) using the models for urban drainage systems based on NSGA-II (Deb *et al.* 2000) and SWMM 5.0 (Rosman 2005). All of these and many more applications share the same need for data-driven improvements of the results obtained using physically based models of complex phenomena.

Approaches to solving the above class of problems vary. In general, each specific application is decomposed according to multiple sources of inaccuracies. Each cause of accuracy limitation is tackled separately and then all partial improvements are combined. It is apparent that predicting complex phenomena accurately becomes not only a modeling effort (decoupling and linking model components into a workflow) but also an optimization and experimentation effort.

Optimization and experimentation are needed for finding the "best" combination of theoretical models, model parameters, input products and implementations of theoretical models, as well as for accessing computational resources for executing the models and obtaining predictions. It is known that optimization in high dimensional spaces and with multiple objectives can become intractable. The multi-objective optimization methods can range from traditional (Coello 1999) to more recent methods, including a fitness function (Jin *et al.* 2002) and genetic algorithms (Reed *et al.* 2001). Furthermore, all these modeling and optimization operations require human interactions with intermediate prediction results. Many times, the best predictions are decided based on considerations and tacit knowledge of a scientist or a group of scientists investigating complex phenomena. The word "tacit" is used in our paper to refer to implied but not actually expressed knowledge of scientists with various levels of certainty. Thus, especially for complex and/or incompletely defined problems, it is the cooperative problem solving approach (Jones & Jacobs 2000) that supports decision-making, planning or monitoring activities. The concept of joint human–machine cognitive systems (human-based engineering) has been introduced in Woods *et al.* (1990) and Brill *et al.* (1990). According to Brill *et al.* (1990), a human–machine decision-making system will perform better when a human is presented with a small number of different alternatives. This leads to a modeling-to-generate-alternatives (MGA) technique that combines mathematical programming models with human cognition. From our perspective, the previous work documents the need for an exploratory framework where scientists and practitioners could investigate combined solutions and experts could incorporating their tacit knowledge into modeling. Thus, the objective of our work is to design a computer-assisted optimization and exploratory framework enabling improvements of modeling accuracy.

In the rest of this paper, we present specific scientific problems of interest and an application example, the methodology, prototype solution and the experimental results obtained by applying the framework to the applications of groundwater recharge/discharge (R/D) estimations. We conclude the paper with a summary and pointers to the developed open source software called Spatial Pattern to Learn (SP2Learn).

SPECIFIC SCIENTIFIC PROBLEMS OF INTEREST

Based on the aforementioned modeling inaccuracies, the general characteristics of the applications in Earth Sciences and our objective, we focus on two problems occurring in geospatial physically based modeling: (1) variable selection and (2) variable relationships.

(1) Variable selection. *The problem of data-driven ranking of relevant input and output variables leads to a selection of a subset of variables used for modeling.* This problem arises when (a) input variables considered are not relevant to the phenomena of interest and irrelevant input variables introduce modeling inaccuracy (Bajcsy & Groves 2004) or (b) when the output variables considered are highly uncertain and inaccurate output variables have to be eliminated from the input/output relationship extraction (Bajcsy *et al.* 2007; Kim *et al.* 2007). Typically, the ranking problems cannot be approached by a brute force search of all possible subset of variables since computational requirements are prohibitive. To illustrate the computational requirements, the number of evaluations would be a summation of all subset combinations according to the following equation:

$$\sum_{i=1}^{N} \left(\frac{N}{i}\right) = 2^{N-1} - 1$$

where N is the number of variables. We approached the problem of ranking highly uncertain outputs of simulations by exploring multiple information comparison metrics that relate input and output variables. This approach assumes that (a) the relevant input variables have been selected based on the user's tacit knowledge or based on previous analyses (Bajcsy & Groves 2004) and (b) the input variables are more trustworthy and accurate than the output variables, as would be likely the case of direct versus indirect measurements. The ranking of output variables is established by searching for the highest information match between a fixed number of input variables and a subset of output variables. The details of ranking are described in the section on unsupervised ranking of variables for relationship extraction.

(2) Variable relationships. The problem of extracting relationships (e.g. rules) among input and output variables could be viewed as quantitative representations of tacit knowledge. These relationships could be viewed as an expert's knowledge about relevant variables not included in the physically based model and not having a model associated with the underlying phenomenon. The relationship extraction is a data-driven (or "machine learning") operation. It requires computer assistance in sifting through available input and output data points, and learning the model. The characteristics of data-driven models could be chosen by the modeler. For exploratory purposes, it is preferable to choose data-driven models that are human-interpretable, for instance, a model that corresponds to a set of if–then rules such as a decision tree. The use of a decision tree has been shown as human-interpretable in the applications related to environmental remediation in Farrell *et al.* (2007). We approached this problem by implementing a decision tree algorithm for extracting rules according to Shafer *et al.* (1996) and Rastogi & Shim (2000) and letting a user choose a subset of the rules on top of the existing model as described in the section on supervised variable relationship extraction using machine learning.

Solutions to both problems need *human interactions in order to optimize and explore multiple options in achieving accuracy improvements*. Both variable rankings and rules should be evaluated, selected and applied iteratively by domain experts to the results obtained from the physically based models, leading to improvements of modeling accuracy. Thus, we have designed an exploratory framework with support for trial-and-error experimentation and rigorous optimization, where the solutions to the aforementioned two problems could be embedded. The framework enables improvements of modeling accuracy by conducting interactive optimization investigations of complex phenomena. The overall schema of the framework is illustrated in Figure 1.

Related work to specific problems of interest

Our approach to solving these two problems leverages our previous work on variable ranking and selection (Bajcsy & Groves 2004; Feng 2006), as well as on data-driven extraction of information and knowledge from remote sensing imagery (Bajcsy *et al.* 2007), water quality measurements (Bajcsy *et al.*

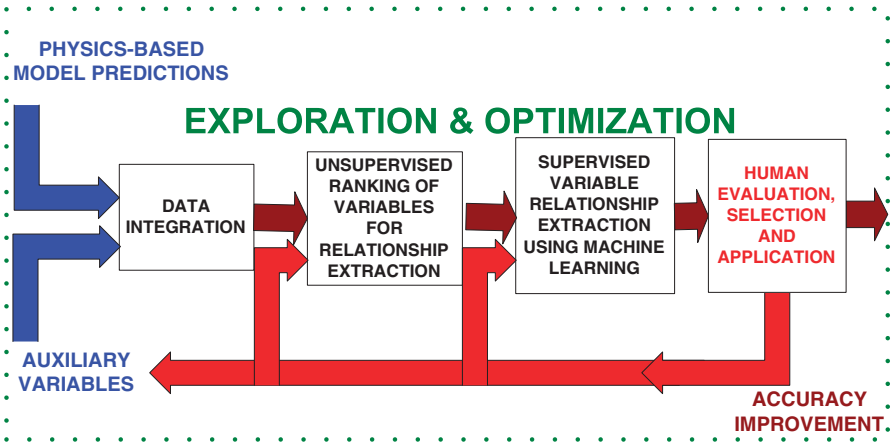

Figure 1 | An overall schema of the exploratory framework.

2006), and groundwater recharge and discharge (Kim *et al.* 2007; Lin *et al.* 2007). As introduced in the previous work on human–machine systems (human-based engineering or cognitive systems (Brill *et al.* 1990; Woods *et al.* 1990; Jones & Jacobs 2000)), building a decision support system might include interactive or non-interactive optimization and trial-and-error experimentation. Our focuses on interactive optimization and the objective of our work could be related to the work of Singh *et al.* (2008). In the work of Singh *et al.* (2008), the "Interactive Multi-Objective Genetic Algorithm" (IMOGA) was used to incorporate qualitative expert knowledge in the groundwater calibration process in contrast to more traditional "non-interactive" approaches (McLaughlin & Townley 1996; Carrera *et al.* 2005). However, while the IMOGA approach addresses a methodology for tacit knowledge extraction by mining the user feedback, it also requires the users to evaluate hundreds of solutions, causing user fatigue (Takagi 2001; Singh *et al.* 2010) and deterioration of the accuracy improvements. In comparison with the work of Singh *et al.* (2010), our work addresses the problem of "machine learning"-based modeling of variable relationships with human feedback on the quality of relationships (select if–then rules and inspect outcomes) while the previous work aims at the problem of "machine learning"-based modeling of subjective assessment of optimization and trial-and-error outcomes (label good/bad an outcome to reduce the number of outcomes to inspect). According to Babbar-Sebens & Minsker (2008), another aspect to consider is the preferences of decision-makers either related to the content of the decision support system (decision variables, constraints and objectives) or among a set of solutions presented for visual inspection. While the work of Babbar-Sebens & Minsker (2008) would concentrate on the content of the decision support system, our work aims at accommodating a decision-maker's preferences among a set of solutions represented as if–then rules of decision variables and their spatial (image) representations.

Overall, in comparison with the previous work, we introduce into the physically based modeling and calibration processes new variables of potential relevance to modeling complex phenomena. We also automate ranking variables for relationship extraction and perform the actual extraction of relationships represented by rules. Thus, the novelty of our approach is in the design of a computer-assisted framework for hybrid, physically based and data-driven, modeling with the support for interactive optimization and exploratory analyses. The data-driven part of the framework combines unsupervised and supervised "machine learning" techniques that are designed to be computationally scalable with the increasing volumes of input data.

APPLICATION EXAMPLE

Let us suppose that recharge/discharge (R/D) distribution (including rates and patterns) has been estimated using a physically based inverse model (Stoertz & Bradbury 1989) following and implemented in FORTRAN. The resulting R/D rate maps are highly uncertain due to several modeling simplifications, the lack of spatial scalability and indirect

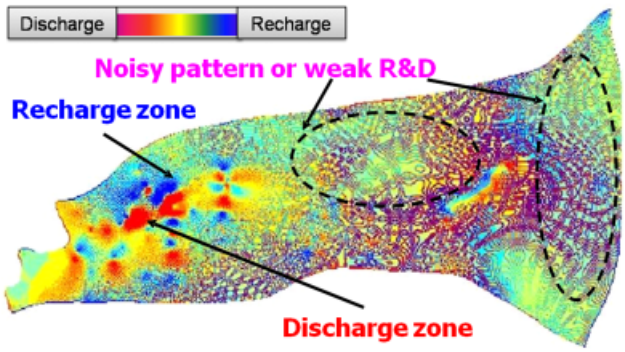

Figure 2 | The result of physically based modeling of R/D rate at 80 m × 80 m spatial grid. The regions circled with the dashed line show the lack of spatial scalability since it is highly unlikely that R/D rates would alternate rapidly over small sets of 80 m × 80 m grid cells. (Modified from Lin & Anderson (2003)).

measurements. In addition, multiple image post-processing operations with variable parameters were applied to the R/D rate estimations in order to normalize the results using the PRO-GIS plugin (Lin *et al.* 2007) and to smooth the R/D rate estimations. Thus, this is one of the problems where (a) the modeling is very complex, (b) there are several R/D rate estimation outputs with high uncertainty, (c) the cost of decreasing uncertainty by taking new measurements is very high, and (d) there is tacit knowledge among the groundwater experts that auxiliary variables such as topographic slope, soil type and proximity to surface water feature are relevant to estimating R/D patterns. Figure 2 illustrates the result of physically based modeling and its obvious lack of spatial scalability at an 80 m × 80 m spatial grid.

There are two algorithmic problems related to selection of R/D maps and to the extraction of tacit knowledge about the relationships of auxiliary variables and the R/D distribution. First, the physically based inverse modeling generates an R/D rate for each model cell (unique labels shown using a red–blue pseudo-color in Figure 2). Furthermore, the R/D rates are post-processed to remove the obvious lack of spatial scalability. This post-process also relied on subjective professional judgments; therefore multiple alternative maps might be generated as shown in Figure 3 (bottom). Before any knowledge can be extracted, the most "reliable/accurate" R/D map has to be selected and multiple R/D maps with variable numbers of labels have to be compared. This application is one example where ranking and selection of R/D maps for further learning is needed.

Second, given a set of variables including the R/D rate, the relationships among auxiliary variables and R/D distribution should be established in order to improve the accuracy of R/D maps. The form of the relationship has to be defined (e.g. a set of rules, probabilities or parametric models) and the representation of the relationships has to be learned from the data. Figure 3 shows the possible relationships among three auxiliary variables and three instances of R/D rate maps for this application, which leads to the need for computer-assisted, "machine learning" of the relationships.

METHODOLOGY

Figure 1 introduced the overall schema of the framework. It aims at accuracy improvements of models of complex

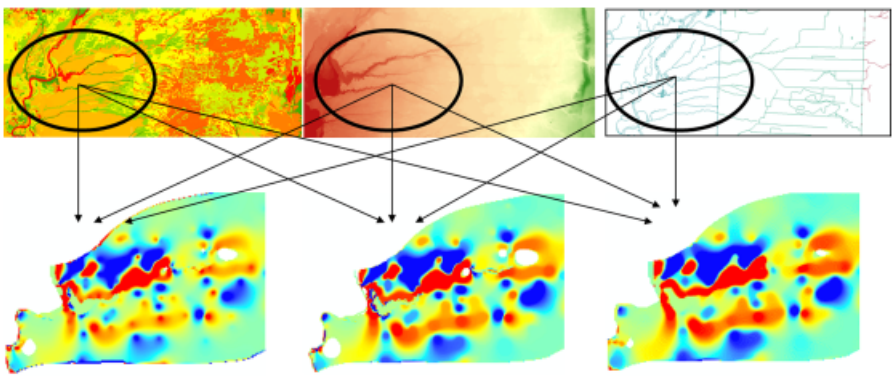

Figure 3 | Illustration of the two algorithmic problems related to selection of R/D rate maps and to extraction of tacit knowledge about the R/D relationships to other unexplored variables. Bottom row shows three instances of spatially filtered R/D rate maps using spatial average (left), normalization and total variation regularized L^1-norm function-based filter (TVL) (Lin *et al.* 2008a) (middle), and spatial average, normalization and TVL (right). Top row displays three auxiliary variables relevant to R/D rate estimation, such as soil type (left), topographic slope (middle) and proximity to river (right). The arrows refer to the two algorithmic problems of selecting which R/D rate map to select for relationship extraction, and how to extract the relationships between auxiliary variables and R/D rates.

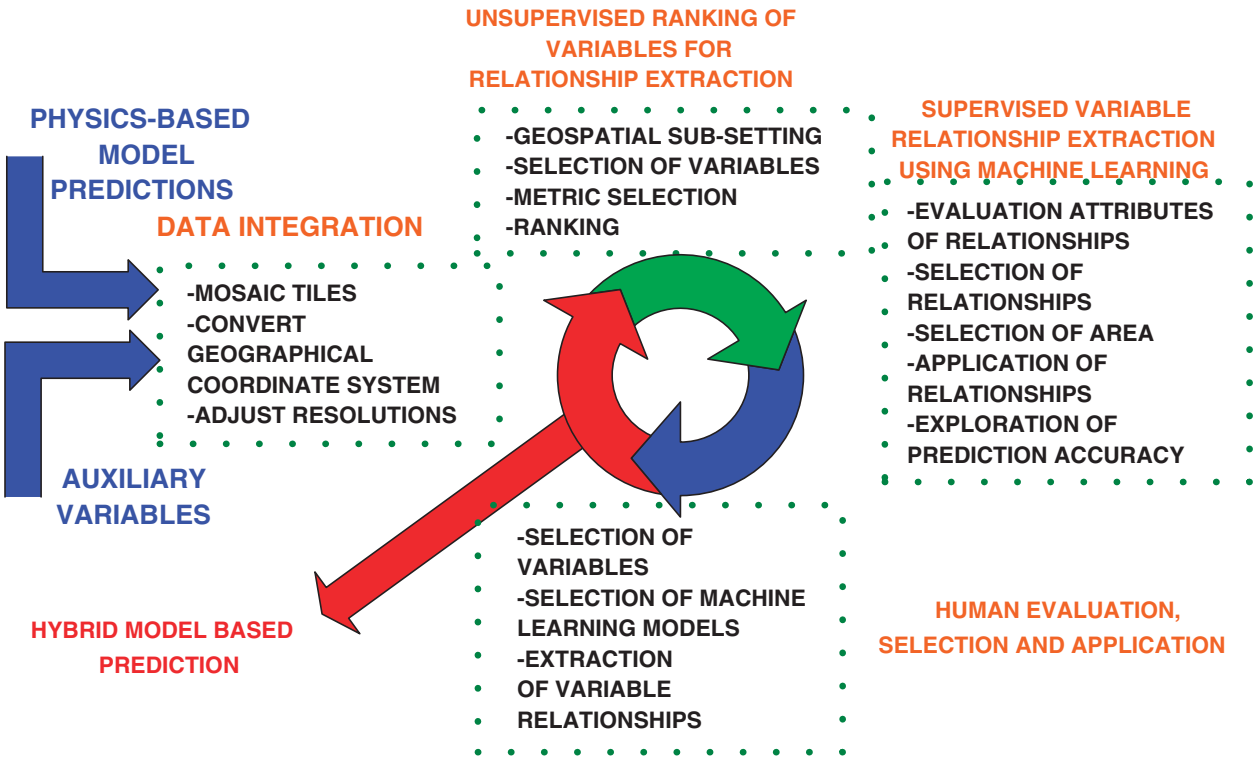

Figure 4 | A methodology for improving modeling accuracy. The methodology maps into the overall schema shown in Figure 1.

phenomena. We designed a methodology shown in Figure 4 for improving accuracy by performing data integration, unsupervised ranking of variables for relationship extraction, extracting variable relationships using "machine learning" and evaluating relationships by humans (domain experts) and applying iteratively subsets of data-driven relationships to generate more accurate estimations. The sequence of operations can form multiple closed loops and leads to a hybrid (physics and data-driven) model shown in Figure 4.

The role of the "Data Integration" component is to resolve heterogeneity of datasets in terms of their underlying spatial resolution and geographical coordinate system, as well as to combine spatial tiles of the same variables. The goal of "Unsupervised Ranking" is to establish an order of the images corresponding to the dependent variable (such as the R/D rate) before a selected image is used for extracting the relationships between independent (input) and dependent (output) variables. The role of "Relationship Extraction" is to derive statistically reliable input–output relationships that could be generalized and applied in spatial locations where the physically based models are believed to fail. Finally, the goal of "Human Evaluation" is to verify, validate and select relationships that comply with experts' knowledge. We describe next the "Unsupervised Ranking" and "Relationship Extraction" components of the methodology in detail, and then devote a section to supporting the parameter optimization needed in all components.

Unsupervised ranking of variables for relationship extraction

The goal of relating and ranking different and multi-dimensional variables can be achieved with unsupervised and supervised approaches. The unsupervised approach utilizes the fact that the more one variable relates to another the higher is the statistical correlation between the two variables. However, a statistical correlation can be applied directly to only continuous-valued variables but not to categorical variables. In order to accommodate the fact that many variables in Earth sciences are categorical, for instance, land use, land cover maps, soil maps or recharge/discharge maps, we adopted a Mutual Information measure based on Shannon's information theory (Shannon 1948). The mutual information measure can be applied to continuous and categorical variables, and it measures the

information content of a variable. The measure has been applied in hydrology in the past for studying the value of stream gauges in a network (Markus *et al.* 2003; Sarlak & Sorman 2006) or for designing optimal networks (Singh 1997). We will describe the mathematical underpinning of the unsupervised ranking of variables using mutual information measure.

Information entropy

In our case, a variable corresponds to a signal described by Shannon's theory. As information is something that reduces uncertainty, in lieu of it we can measure the uncertainty in a signal. This uncertainty measure is known as information entropy, $H(X)$, defined as

$$H(X) = -\sum_{x \varepsilon X} p(x)\log(p(x))$$

Entropy $H(X)$ is thus a measure of the amount of uncertainty associated with the value of a random variable X by considering the probability distribution $p(x)$. It is also a measure of the amount of information required on average to describe the random variable.

Figure 5 shows the entropy curve of a random variable X as a function of probability p, where p here denotes the realized probability of the random variable.

Joint entropy

As we are concerned with pairs of variables, joint entropy measures the randomness of a pair of random variables. The joint entropy $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is defined as

$$H(X, Y) = -\sum_{x \varepsilon X}\sum_{y \varepsilon Y} p(x, y)\log p(x, y)$$

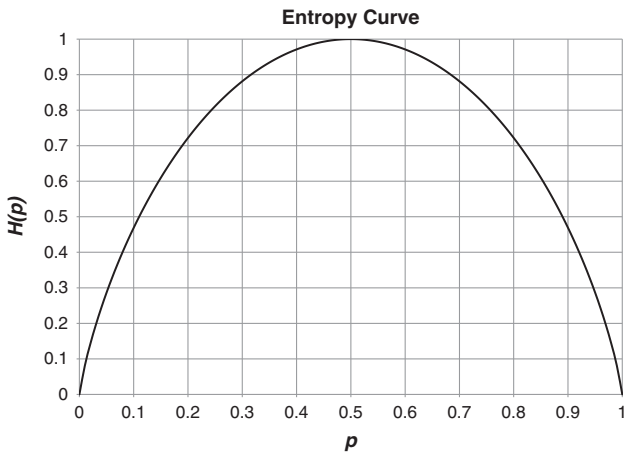

Figure 5 | Entropy curve for the realized probability p of a random variable X.

Mutual information

Mutual information is a measure of the amount of information that one random variable contains about another random variable. It is also a reduction of the uncertainty of one random variable due to knowledge of the other variable. In other words, how well one variable X describes another variable Y is represented by mutual information $I(X, Y)$ as defined below:

$$I(X, Y) = \sum_{x \varepsilon X}\sum_{y \varepsilon Y} p(x, y)\log\frac{p(x, y)}{p(x)p(y)}$$

$$I(X, Y) = H(X) + H(Y) - H(X, Y)$$

The mutual information measure can be applied for ranking random variables if they have the same dimensionality. That is, all instances of X have the same dimensionality and all instances of Y have the same dimensionality, but X and Y might not have the same dimensionality. Figure 6 shows the Venn diagram relating information entropy $H(X)$ and $H(Y)$, joint entropy $H(X, Y)$ and mutual information $I(X, Y)$. As shown, mutual information $I(X, Y)$ is the intersection between $H(X)$ and $H(Y)$.

Normalized information measures

It is apparent that the mutual information metric depends on the number of labels in the evaluated map (signal). Thus, mutual information could be used for comparing predicted maps only if the number of predicted labels remained constant. However, this is not always the case, as illustrated in the example scenario. The instances of the predicted variable – representing the predicted map – do not have the same dimensionality in this case.

One of the approaches to the unequal number of labels is to normalize mutual information. According to Figure 6, we could normalize (a) mutual information by the entropy of the variable representing predicted outputs, (b) mutual informa-

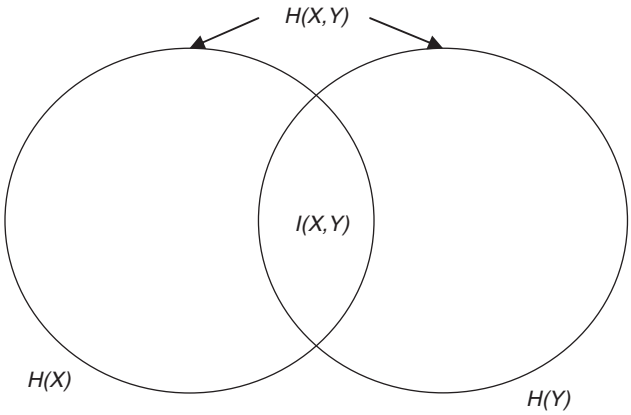

Figure 6 | Venn diagram relating information entropy, joint entropy and mutual information.

tion by the joint information entropy of input and output variables or (c) joint information by the sum of input and output entropies (denoted as the entropy correlation coefficient). Let us assume that X are the auxiliary variables (inputs) and Y are the predicted variables (outputs). Then, these three normalized metrics can be defined as follows:

Mutual information normalized by output entropy (MI/OE):

$$\frac{MI}{OE} = \frac{H(X) + H(Y) - H(X,Y)}{H(Y)}$$

Mutual information normalized by joint entropy (MI/JE):

$$\frac{MI}{JE} = \frac{H(X) + H(Y) - H(X,Y)}{H(X,Y)}$$

Entropy correlation coefficient (ECC):

$$ECC = 2 - 2/NMI(X,Y)$$

where NMI is the normalized mutual information metric, defined as

$$NMI(X,Y) = \frac{H(X) + H(Y)}{H(X,Y)}$$

The MI/OE metric accounts for variability of the number of labels in the output variable. The MI/JE and ECC metrics take into account the variability in both input and output variables. We will use all three metrics MI/OE, MI/JE and ECC for unsupervised ranking.

Supervised variable relationship extraction using machine learning

Once the predicted (output) variables were ranked and selected for further processing, data-driven ("machine learning") techniques have to be selected and optimized for extracting relationships between input and output variables. Among the multitude of supervised data-driven techniques (Han & Kamber 2001), we selected a decision tree algorithm due to its intuitive interpretation. Based on the interactive nature of computer-assisted accuracy evaluations, it is critical for users (domain experts) to understand the relationships suggested derived by "machine learning". The decision tree is one of the few data-driven models that could be easily represented by a set of ifthen statements (rules), visualized as graphs, efficient in expressing complex rules, and viewed as a self-learning expert's system (Feigenbaum *et al.* 1994).

Although there are many implementations of the decision tree technique, very few of them would provide computational scalability. The most frequently used decision tree implementations are CART (classification and regression trees) and C4.5 algorithms (Quinlan 1993). CART algorithms would create nodes by optimizing the Gini impurity (squared probabilities of membership for each target category in the node) while C4.5 algorithms would optimize entropies in the node. Besides the node splitting criteria, the computational scalability of these algorithms becomes very important to Earth scientists as they work with large volumes of data represented by maps, remote sensing images and ground measurements over larger spatial areas and temporal windows. We followed and extended the scalable decision tree algorithm called SPRINT according to Shafer *et al.* (1996) that can be parallelized easily on networked computers with private memory and disks. In addition, we have focused on computational efficiency of building and pruning a decision tree classifier which improves performance, especially when dealing with large volumes of data. We have incorporated into our design several approaches of the PUBLIC implementation documented in Rastogi & Shim (2000). The computational savings come from the fact that (a) decision tree building and pruning operations are integrated, and (b) the building operation is parallelized by data placement and workload balancing in a shared-nothing computational environment with message passing interfaces. Thus, the developed

implementation of the decision tree technique is specially designed for processing large volumes of data as occurring in Earth Sciences.

The best decision tree is selected by the minimum description length (MDL) criterion (Krichevsky & Trofimov 1981; Grunwald 2005). Since given supervised data can be explained by different decision trees, we need a method to select the best alternative among possible decision trees. According to Occam's Razor, "All other things being equal, the simplest solution is the best." Thus we take the smallest one among all decision trees which can explain the given data. The MDL criterion is a formalized theory of Occam's Razor and, in order to measure the size of each decision tree, we use the description length cost following the MDL criterion (Shafer *et al.* 1996; Rastogi & Shim 2000). The cost of the description length consists of three contributions: (1) the cost of encoding tree structure, (2) the cost of encoding a split condition in each inner node and (3) the cost of encoding a data record in leaf nodes. The first cost contribution is measured by the number of nodes in the tree. An inner node is represented by 1 and a leaf node is represented by 0. The following tree is encoded as "10,100" with the depth-first ordering – see Figure 7.

The second cost contribution is the number of possible split conditions. For n number of categories in an inner node, there are 2^n numbers of possible conditions. Among them, the empty set and the whole set are not meaningful and thus the encoding cost contribution is calculated as $\log(2^n - 2)$. For the third cost contribution, we refer to the equation below from Krichevsky & Trofimov (1981):

$$\sum_i n_i \log \frac{n}{n_i} + \frac{k-1}{2} \log \frac{n}{2} + \log \frac{\partial k/2}{A(k/2)}$$

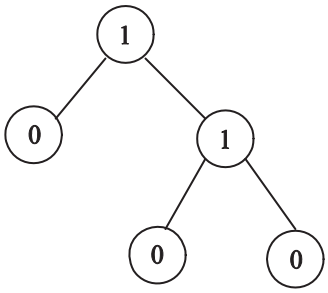

Figure 7 | Simple tree structure.

where k is the number of classes, n is the number of records and n_i is the number of records in a class i. The sum of these three cost contributions is the description length cost of a decision tree.

We use the MDL criteria two times. First, we select the best decision tree for any given data. This selection is performed during the pruning step. We decide the best size of the decision tree using MDL criteria. Second, we evaluate all variables represented by the data with MDL criteria. As we discussed in the previous sections, the input data consist of multiple variables chosen by the unsupervised method. Thus, there are other possible combinations of variables. If the best description length cost from a variable set is smaller than that from another variable set, we can say the former is a better selection of variables than the latter. This information can be fed back to the variable selection stage as shown in Figure 1.

The result of a decision tree algorithm is not only a set of rules but also the confidences of those rules. The confidence of a rule "if X = x then Y = y" is the ratio of the number of data points following the rule to that of all data points with Y = y. This confidence is typically reported as a percentage.

Parameter optimization and exploratory views

In a typical data-driven study, modeling assumptions and parametric set-ups of models raise the questions about optimality. Figure 4 illustrates a sequence of operations that require optimization and perhaps trial-and-error experimentation. First, the data integration operation needs an "optimal" target spatial resolution, projection and datum. Due to the fact that no projection or datum would be superior to others, we approached the problem by letting a user select one of the data files. The projection and datum of the selected file become the target parameters. The spatial resolution would also be set to the resolution of the file selected.

Second, the spatial grid cells of interest have to be selected from the set of overlaid maps. The selection criteria vary, depending on the user's knowledge and preferences. Our approach was to develop multiple masking options for selecting data points (vectors of variables for co-located measurements). The masking options include (1) a hand-drawn rectangular- or oval-shaped region, (2) a set of geo-point locations of variable radii loaded from either CSV or

Excel files, (3) a set of irregular regions delineated by several geo-spatial boundaries loaded from ESRI Shapefiles, (4) a set of geo-points matching categories in a loaded map with a categorical variable, (5) a set of geo-points satisfying a range criterion in a loaded map with continuous variable and (6) any Boolean combinations of the masks obtained by one of these options.

Third, the operation of unsupervised ranking requires choosing the metric for ranking (MI/OE, MI/JE and ECC) and selecting input and output variables to be evaluated. Although we provide user interfaces only to three metrics for ranking, we have conducted research with many other metrics as well. The metrics provided here are the most general in the sense that they work with continuous and categorical variables because the metrics measure entropy, regardless of variable ranges. A user can choose a specific metric according to the variability of inputs and outputs.

Fourth, the operation of supervised "machine learning" of variable relationships lets users experiment with different subsets of inputs and outputs, and evaluate the decision trees with or without pruning using the minimum description length (MDL) criterion (Grunwald 2005). The MDL criterion can be viewed as learning about the best data compression by finding regularities in the data (or finding goodness-of-fit tradeoffs). Thus, if unsupervised ranking would lead to several predicted maps with similar ranking scores and the same number of labels then we could prefer relationships derived from the decision tree characterized by the smallest MDL. In this case, MDL provides an extra indicator for evaluating multiple decision trees and is one component of computer-assisted explorations.

Fifth, the result of "machine learning" is a set of rules "if $X = x1$ then $Y = y1$" with their confidence $= P\%$. The rules are inspected by a user and a subset of rules is selected. A set of selected rules is applied to the predicted output map Y by forcing $Y = y1$ at every grid cell where $X = x1$. Thus, after forcing the rules, inaccurate grid cells ($X = x1$, $Y = y2$) will be updated to ($X = x1$, $Y = y1$). We have approached the problem of rule selection by building an image color based and text label based rule representation that would be easy to read and comprehend. The problem of rule applications is currently supported by showing updated values, the difference and the locations of updates.

While the current framework has plenty of room for automating selections by designing sophisticated optimization strategies, we have left several optimization operations in the hands of a user for trial-and-error experimentation. The reason behind our design leaning toward computer-assisted optimization and trial-and-error experimentation rather than automated optimization lies in the very many unknowns about how experts select optimal parameters and variable configurations. We believe that there is still a lot of intuition and experience involved when parameter optimization is performed. It is true especially for the class of problems where the complexity of the underlying phenomena is overwhelming. Therefore, providing exploratory frameworks with (1) metrics, such as the three unsupervised ranking metrics (MI/OE, MI/JE and ECC) and the MDL criterion, (2) "machine learning" tools like decision trees and (3) visualizations of data and rules are essential for improving accuracy. Exploratory views lead ultimately to a better choice of physically based models based on visual inspection of spatial coherence of new estimations, and distribution and magnitude of grid cells that have been updated.

PROTOTYPE SOLUTION

We have prototyped open source software called Spatial Pattern To Learn (SP2Learn) according to the methodology described in the previous section. SP2Learn can be viewed as an encapsulated workflow for (1) loading multiple raster files (images), (2) integrating and mosaicking all raster datasets to form a stack with consistent spatial resolution as well as geographic projection, (3) loading other files (boundaries, points or images) to create a mask for pixel selection purposes, (4) integrating the existing stack of raster images with other masking information, (5) selecting boundaries or image regions of interest and extracting variables from the stack of images, (6) performing information-content-based ranking of selected input and output variables, (7) extracting relationships using data-driven, decision-tree-based, modeling for ranked input and output variables, (8) analyzing data-driven models to assign a relevance coefficient to input variables, (9) evaluating and selecting rules (relationships) to be applied and (10) applying the rules at the pixel level to predict output variables. All aforementioned steps are supported by visualizations (color, grayscale or pseudo-color) of input, intermediate and output data sets, as well as the data models.

According to Figure 1, the workflow steps (1)–(5) belong to the data integration component, step (6) is the ranking component, steps (7) and (8) support extraction of relationships and steps (9) and (10) are designed for exploratory evaluations, selections and applications of rules to improve estimation accuracy.

In order to bring together so much functionality, the architecture of SP2Learn leverages several technologies. The majority of the SP2Learn code is based on Image to Learn (Im2Learn) developed at NCSA (Kooper *et al.* 2008). The Im2Learn makes additional calls to the Hierarchical Definition Format (HDF) library to open many raster files from NASA and to the MODIS Reprojection Tool (MRT) to perform geographic re-projections. The code for the workflow framework is adopted from the GeoLearn software (Bajcsy *et al.* 2007) developed by NCSA. The sequence of steps is divided into "Load raster", "Registration", "Create Mask", "Map Ranking", "Attribute Selection", "Rules" and "Apply Rules" operations in the prototype workflow.

The methodology for going from raw data to knowledge is generic, and can be customized by adding or removing steps. The software solution is content-agnostic, regardless of the fact that our main application driver came from the groundwater domain. The implementation of SP2Learn is following a "linear workflow" paradigm where any step can be inserted into the workflow of steps by extending a template Java class called WizardStep. Thus, the functionality of a step can be extended or overridden by plugging in and out the extensions of the WizardStep class and the flow of steps would change accordingly. The plugging mechanism for the extensions of the WizardStep class is to add an additional step in the main Java class (i.e. addStep(spDecisionTree); and to maintain input and output data values of the step in the function of switching steps (i.e. switchStep(oldstep, newstep)).

Regarding the format of the final results (rules relating variables), there are two types of conditions among the reported rules, positive and negative. Conditions with "in" are positive conditions and those with "not in" are negative conditions. For positive conditions with multiple values, the relationships are formed using "or" while for negative conditions, the relationships are formed using "nor". For example, the rule "if condition = (proximity to water is far) AND (topographic slope is between 0 and 0.3) AND (soil type is labeled as Mh with pseudo-color representation (red = 60, green = 255, blue = 0)) then conclusion = (R/D rate for estimation labeled as zbard_mod1 is in the interval [74,255,234])" would be described in the file as follows:

```
<RULE ID = 162 NUM_OF_CASES = 5512 CONFIDENCE = 20.46%>
  <CONDITION>Attr. Ex_WI_water in {not_near_water} AND Attr.
Ex_WI_slope in {0-0.3} AND Attr. Ex_WI_soiltype in
{Mh,60,255,0}</CONDITION>
  <CONCLUSION>Ex_WI_zbard_mod1 is 74,255,234 </CONCLUSION>
</RULE>
```

EXPERIMENTAL RESULTS

In order to evaluate the framework in the context of the application example, we first established reference data points for the optimal number of R/D zones and the commonly accepted relationships among R/D rates and other variables. These relationships are currently considered as tacit knowledge of experts and would be supported (or not) by the data-driven analyses. Next, we compared the optimal number of R/D zones and the relationships obtained using our framework with the reference data points.

Reference data points

The field application is to estimate groundwater recharge and discharge of the Buena Vista Groundwater Basin in Wisconsin. The basin is well understood and hence the accuracy of the estimation can be validated by domain experts. The previous studies concluded that the accuracy of R/D estimation is highly sensitive to the number and the shape of R/D zones at this site. The past studies used a physically based model based on a finite difference mass balance approach (Stoertz 1989; Stoertz & Bradbury 1989; Lin & Anderson 2003) and concluded that the optimal number of R/D zones is approximately 40 based on the analysis between 13 and 50 zones which was identified as the feasible range at this particular site. However, the delineation of R/D zone boundaries generated by previous studies heavily relied on subjective professional judgment. The subjectivity comes from the fact that the boundaries are formed depending on the users' skills and knowledge needed to cross-reference the pattern in

multiple R/D rate outcomes obtained from various combinations of image processing methods. An analyst needs to generate several alternative R/D conceptual models and compare them with field information, such as land coverage, soil type maps, surface water distribution and topographic slope. Then the analyst makes subjective delineations based on their professional judgment without any quantifiable reference. Thus, the delineation of the R/D zones could be considered only as rough guidelines and not as exact reference data points.

As for the expected relationships, we consulted with the experts any anticipated relationships. Some of the relationships are listed below:

- If proximity to nature water is near then R/D rate might not be small.
- If soil type is clay then R/D rate is likely to be small.
- If soil type is loam then R/D rate is likely to be medium.
- If soil type is sand then R/D rate is likely to be high.
- If topographic slope is high then the R/D rate is likely to be small.
- If topographic slope changed from high to low the R/D rate is high unless the soil type is clay.

The terms "small", "medium" and "high" serve only as a qualitative index to reflect the tacit knowledge of experts based on the individual field site. These terms might not have a consistent value range for general use. which also motivates the use of our framework.

Experimental data

In order to evaluate the framework against the reference data points, we have processed several maps representing auxiliary variables (or inputs) occurring in the anticipated relationships or believed by experts to be relevant to predicting R/D rates. The images included maps of surface water feature, soil type and topographic slope for the Buena Vista Groundwater Basin in Wisconsin. The auxiliary variables are shown in Figure 8.

The outputs representing R/D rates for the chosen region in Wisconsin were obtained by running a physically based mass balance model, GRADE-GIS (Lin *et al.* 2008b), and post-processed by several image filtering techniques with multiple algorithmic parameters found in PRO-GIS (Lin *et al.* 2008b). The outcomes of modeling and post-processing were R/D rate maps with the variable number of zones ranging from 6 to 74. Figure 10 shows examples of R/D rates with 6, 11, 21, 56 and 74 zones (labels). The 11, 21 and 56 zone maps were improved from the previous studies (Lin & Anderson 2003; Lin *et al.* 2008b). In order to test the efficiency of the learning process, two more maps were generated (6 and 74 zones) with minimum effort and no professional judgment involved. Some zones in the 11-zone map were consolidated to generate a 6-zone map. Some zones in the 56-zone map were divided to generate a 74-zone map. Therefore, we will have two test maps which should not have the best score to test our algorithm. These two maps with 6 and 74 zones will also verify if the algorithm is biased to the low or high zone numbers.

Evaluations of consistency of results

Optimal number of zones. Before we executed the experiments with real data, we explored the characteristics of the three metrics on a set of synthetic images. Table 1 summarizes one of many combinations with the four-region input map and a variable number of regions and variable layouts of output maps. One could observe that, as the number of output regions (zones) deviates from the number of input regions, the scores decrease. Similarly, as the layout of output regions

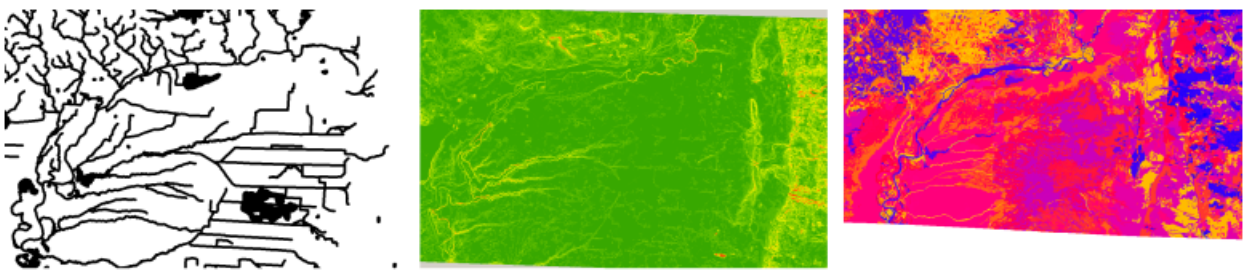

Figure 8 | Auxiliary variables related to predicting R/D rate according to expert's tacit knowledge. Left to right – surface water (black refers to proximity to water), topographic slope (the brightness refers to the elevation gradient) and soil type (each color refers to one type of soil).

Table 1 | Synthetic input and output images and their corresponding scores obtained based on the three metrics (ECC, MI/OE, MI/JE)

Input (4 zones, fixed layout)										
Output layout										
Number of zones in output	1	2	4	4	4	4	4	8	16	16
ECC=2-2/NMI	n/a	0.67	1	0.96	0.81	0.69	0.50	0.80	0.67	0.67
MI/OE	n/a	1	1	0.96	0.82	0.72	0.50	0.67	0.50	0.51
MI/JE	0	0.5	1	0.93	0.68	0.52	0.33	0.67	0.50	0.51

deviates from the layout of input regions the scores decrease. Each metric has a different rate in which it decreases but the overall characteristics follow both of the above observations.

Given the real data (the five R/D pattern outcomes shown in Figure 9) based on the processes and guidelines provide by (Lin *et al.* 2008a), the unsupervised ranking approach leads to the optimal number of zones to be 56. The evaluation of the optimal number of zones is based on inspecting the results obtained using the three metrics (ECC, MI/OE, MI/JE) and the three input maps (slope, water proximity and soil-type maps). Table 2 summarizes all scores for all combinations of inputs, outputs and metrics. Figure 10 shows the total score equal to the sum of scores per input as a function of the metric used. The larger the score per metric the larger the mutual information between inputs and the output R/D map. The maximum score in Figure 10 was detected for the number of R/D zones equal to 56, which implies the highest mutual information between the three inputs and the output R/D map with 56 zones.

While the evaluations led to 56 zones as the optimal number, one could study the influence of each individual input map on the optimal number of zones. For example, if we had considered water proximity only then 21 zones would be the optimal number. The choice of 74 zones would be the

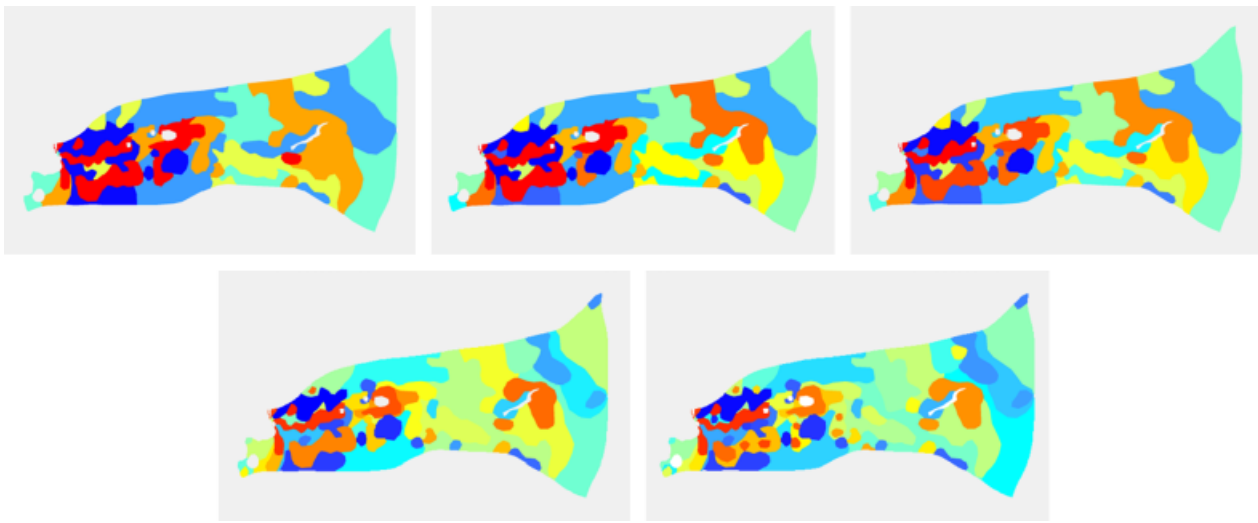

Figure 9 | Water recharge/discharge maps with variable number of zones. Top row (left to right) – 6, 11, and 21 zones. Bottom row (left to right) – 56 and 74 zones.

Table 2 | Summary of all ranking scores of R/D maps with different numbers of zones based on three metrics (ECC, MI/OE, MI/JE) and all inputs (slope, water proximity and soil-type maps)

Input	Output R/D map with N zones	ECC = 2-2/NMI	MI/Oe	MI/JE
Water proximity	6	0.010726449	0.006722343	0.005392
Water proximity	11	0.01892856	0.011398513	0.009555
Water proximity	21	0.024750596	0.014331343	0.01253
Water proximity	56	0.021710729	0.012275529	0.010974
Water proximity	74	0.0219443	0.012329061	0.011094
Soil type	6	0.164027408	0.208722068	0.089341
Soil type	11	0.199772593	0.224341959	0.110971
Soil type	21	0.290938692	0.285648085	0.170233
Soil type	56	0.344388722	0.309535545	0.208013
Soil type	74	0.335197348	0.293961019	0.201344
Slope	6	0.02318693	0.019126788	0.011729
Slope	11	0.026830879	0.020445647	0.013598
Slope	21	0.062206744	0.043709236	0.032102
Slope	56	0.067464672	0.045048071	0.03491
Slope	74	0.066610942	0.043866766	0.034453
Sum of all input scores	6	0.197940787	0.234571199	0.106462
Sum of all input scores	11	0.245532032	0.25618612	0.134123
Sum of all input scores	21	0.377896033	0.343688664	0.214865
Sum of all input scores	56	0.433564123	0.366859145	0.253897
Sum of all input scores	74	0.42375259	0.350156846	0.24689

second-best overall. This choice is always outperformed by the choice of 56 zones, except for the case when only water proximity would be considered which leads to very similar scores. Overall, the lower scores on 6-zone and 74-zone maps indicated that the algorithm would not select maps obtained by maximizing or minimizing the number of zones.

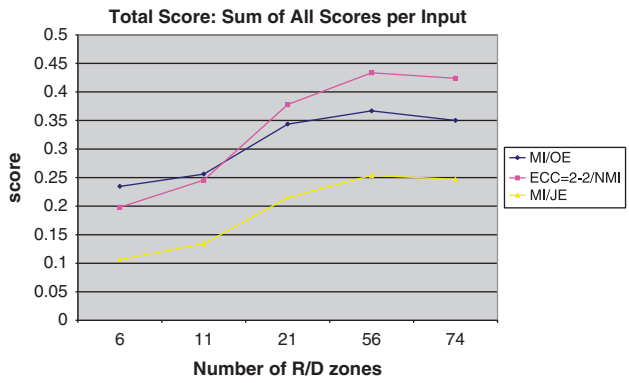

Figure 10 | The SP2Learn ranking scores of R/D maps with different number of zones based on three metrics (ECC, MI/OE, MI/JE) and all inputs (slope, water proximity and soil-type maps).

The results are consistent with "approximately 40 zones" as previous studies stated by Lin & Anderson (2003) and Lin *et al.* (2008a) (see also the section on Reference Data Points). If the preparation of R/D rate maps could be controlled in terms of the number of labels and the number of zones could vary continuously then we would be able to explore all possible outcomes as a function of the number of zones. However, we could not generate outcomes with an *a priori* defined number of zones and therefore the optimal number of zones is computed only over the set of available values.

Extracted relationships

We have extracted relationships from the R/D map with 56 zones since that particular R/D map was chosen based on the previous analysis. Next, we selected the most reliable rules recommended by the SP2Learn decision-tree-based data-driven model in order to explore the consistency between the rules derived by using our framework and the relationships among variables anticipated by experts.

The criteria for selecting the most "statistically" reliable rules were: (1) more than 1000 cases (image pixels) have the R/D label and (2) more than 60% of all cases satisfying (1) also satisfy the rule conditions (also denoted as the confidence of a rule).

We have identified seven rules (Rule ID 2329, 2333, 2444, 2446, 2449, 2664, 2960) meeting the criteria above. The rules are summarized in Table 3. Out of these seven rules, all of them have one part of the condition that relates to soil type. This might indicate that the soil type has the highest pattern correlation with R/D rate pattern.

We divided these soil-type conditions into those that referred to the presence or absence of a certain soil type. Among the conditions referring to the presence of a certain soil type, Kr and Rf soil type families occurred in rules predicting medium recharge rates (R/D rate RGB color value=112–255–210; Kr and Rf presence was one part of four rules: Rule 2333, 2444, 2446 and 2960). According to the soil database at the USDA Natural Resources Conservation Service, the main texture of the Kr soil family is Kranski loamy sand which has moderate to high permeability. Note that the KrC and KrD soils are obtained by interactively selecting Portage County, Wisconsin from the Soil Data Mart database webpage http://soildatamart.nrcs.usda.gov/Default.aspx. The main texture of the Rf soil family is Richford loamy sand which has also moderate to high permeability like the Kr soil family. However, the KrC (6–12% slopes) and KrD (12–20% slopes) soils are located at places with high topographic slopes (6–20%) which would significantly increase surface runoffs and decrease the R/D rates. A reader can find detailed description of KrC and KrD at the URL: http://mmas-mapping.soils.wisc.edu/soil_descriptions/portage_soil_descriptions.html#KrC. For the same reason, the R/D rate at the locations with RfB and RfC soil types would also decrease due to the geographic collocation with moderate to high topographic slopes (2–12%).

Rule 2329 with a medium discharge rate pattern corresponded to the WyB soil family, which is Wyocena sandy loam with low to moderate permeability. The second condition of Rule 2329 shows low topographic slope which would be favorable to maintain R/D rates in a medium range. Rule 2449 with low recharge pattern corresponded to the Ov soil family, which is Oesterle loam with silty subsoil variant. The Ov soil family has low to moderate permeability. However, the second condition of Rule 2449 shows very low topographic slope (0–0.3%) which would increase recharge rates in a medium range. Rule 2664 with low recharge or discharge pattern corresponded to the MgC soil type. The main texture of MgC is Mecan sandy loam which has low to moderate permeability. However, MgC is located at places with high topographic slopes (6–12%) which will significantly increase surface runoffs and decrease the R/D rate.

The seven most "statistically" reliable rules have suggested relationships between R/D patterns and auxiliary variables, such as soil types and topographic slope. These suggested relationships are consistent with the tacit knowledge of experts. Based on the experts' knowledge, the Buena Vista Groundwater Basin is located at the Central Sand Plains area in Wisconsin, also well known for its high recharge and discharge condition due to the prevalence of sandy soil. Furthermore, all seven most reliable rules predict a slow to

Table 3 | Top seven rules selected based on the analyses

Rule-ID	Num of Cases	Confidence	R/D Rate	Condition
2329	2290	90.48%	Mid-Discharge	Soil type is MyB and topographic slope is less than 1.7
2333	1284	88.94%	Mid-Recharge	Soil type is KrD
2444	2914	78.35%	Mid-Recharge	Soil type is KrC and topographic slope is less than 4.9 or greater than 9.5
2446	1137	88.74%	Mid-Recharge	Soil type is RfC and there is no water
2449	1451	60.79%	Mid-Recharge	Soil type is Ov and topographic slope is less than 0.3
2664	1055	64.36%	Low-Recharge or Low-Discharge	Soil type is MgC
2960	1550	64.39%	Mid-Recharge	Soil type is RfB and topographic slope is greater than 4.9

medium R/D rate pattern which would be difficult to identify and quantify in the presence of all other relevant contributors to R/D rate patterns. Therefore, the extraction and data-driven quantification of these most reliable rules provided valuable information to the domain scientists trying to understand the complexity of R/D rate patterns at this field site.

SUMMARY

This paper presented a framework for accurate geospatial modeling using image ranking and "machine learning". We have demonstrated how to improve our understanding of complex underlying physical phenomena and increase the accuracy of geospatial models by (a) incorporating auxiliary variables, (b) ranking variables for relationship extraction and (c) extracting variable relationships. The framework supports modeling optimizations, trial-and-error experimentation and visual explorations, and aims at extracting tacit knowledge of domain experts. It could also be viewed as an informatics workflow with a hybrid modeling approach, where physically based modeling and data-driven modeling approaches are combined. The novelty of our work is in designing a methodology for ranking and extracting relationships, as well as in developing a general framework for building accurate geospatial models. This methodology provides a quantitative and systematic approach that could bridge the gap between the traditional subjective approaches for initiating conceptual models and advanced stochastic and uncertainty analyses.

The framework was applied to the problem of modeling groundwater recharge and discharge (R/D) rates that has been known to be very complex. We used the physically based R/D rate model that is initialized by only a small set of field measurements, such as hydrologic conductivity, water table level and bedrock elevation. In our framework, we explored the accuracy improvements of the model by using data-driven analyses. The analyses employed several image de-noising techniques with a decision tree "machine learning" technique, and used several remote sensing and terrestrial raster measurements, for example, topographic slope, soil type and proximity to water bodies. For the specific application, the ranking approach in SP2Learn provided quantifiable reliability indices of R/D maps which enabled hydro-geologists to recognize R/D patterns and compare R/D estimations with a more objective approach.

In addition, the "machine learning" process also summarized the rules and provided the capability for recommending a new optimal R/D pattern. The summary of the rules is an example how the expert knowledge was mapped by the framework into a set of concrete rules so that the knowledge could be disseminated and re-used in a more efficient way. The framework also confirmed and refined the expert knowledge which leads to better understanding of some of the complexities of the field site. The experts now have the capability to map the knowledge into a set of rules, and improve the maps in a more systematic and reliable approach using the prototype solution called Spatial Pattern To Learn. The software solution is available for downloading at http://isda.ncsa.uiuc.edu/Sp2Learn/. In the near future, we plan demonstrating the scalability of the software as was outlined in the text.

ACKNOWLEDGEMENTS

The work was funded by the NCSA Faculty fellowship program and by the Illinois Department of Natural Resources (IDNR). The authors would like to acknowledge NCSA and IDNR for the funding provided.

REFERENCES

Babbar-Sebens, M. & Minsker, B.S. 2008 Standard Interactive Genetic Algorithm (SIGA): a comprehensive optimization framework for long-term ground water monitoring design. *J. Wat. Res. Plann. Mngmnt.* **134**(6), 538–547.

Bajcsy, P. & Groves, P. 2004 Methodology for hyperspectral band selection. *Photogramm. Engng. Remote Sensing J.* **70**(7), 793–802.

Bajcsy, P., Kim, C.-Y., Li, Q., Kooper, R., Mehra, V., Robertson, R. & Kumar, P. 2007 GeoLearn: an exploratory framework for extracting information and knowledge from remote sensing imagery. In: *Proc. 32nd International Symposium on Remote Sensing of Environment Sustainable Development Through Global Earth Observations, San Jose, Costa Rica.*

Bajcsy, P., Kooper, R., Marini, L., Clutter, D. & Markus, M. 2006 Visualization and data mining tools applied to algal biomass prediction in Illinois streams. In: *Proc. 7th International Conference on Hydroinformatics, Nice, France.* Research Publishing, pp. 926–933.

Brill, E.D., Jr., Flach, J.M., Hopkins, L.D. & Ranjithan, S. 1990 MGA: a decision support system for complex, incompletely defined problems. *IEEE Trans. Syst. Man Cybern.* **20**(4), 745–757.

Carrera, J., Alcolea, A., Medina, A., Hidalgo, J. & Slooten, L.J. 2005 Inverse problem in hydrogeology. *Hydrogeol. J.* **13**(1), 206–222.

Coello, C. A. C. 1999 An updated survey of evolutionary multiobjective optimization techniques: State of art and future trends. *Evolutionary Computation, 1999. CEC 99.* **1**, 13 (doi: 10.1109/CEC.1999.781901).

Coopersmith, E. J. 2008 *Understanding and Forecasting Hypoxia using Machine Learning Algorithms*. Masters thesis, Univeristy of Illinois at Urbana-Champaign, Urbana, IL.

Deb, K., Pratap, A., Agrawal, S. & Meyarivan, T. 2000 *A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II*. Indian Institute of Technology, Kanpur.

Farrell, D. M., Minsker, B., Tcheng, D., Searsmith, D., Bohn, J. & Beckman, D. 2007 Data mining to improve management and reduce costs of environmental remediation. *J. Hydroinf.* **9**(2), 107–121.

Feigenbaum, E. A., Friedland, P. E., Johnson, B. B., Nii, H. P., Schorr, H., Shrobe, H. & Engelmore, R.S. 1994 Knowledge-based systems research and applications in Japan. *AI Mag.* **15**(2), 15.

Feng, W.-W. 2006 *Relevance Assignment and Fusion of Multiple Learning Methods on Geo-Science Data Analysis*. MS thesis, Department of Computer Science, University of Illinois at Urbana-Champaign.

Fulton, R. A., Breidenbach, J. P., Seo, D.-J., Miller, D. A. & O'Bannon, T. 1998 The WSR-88D rainfall algorithm. *Weather Forecasting* **13**, 377–395.

Grunwald, P. 2005 Introducing the minimum description length principle. In: *Introducing the Minimum Description Length Principle. Theory and Applications* Grünwald, P. D., Myung, I. J. & Pitt, M. A. MIT Press, Cambridge, MA, pp 3–22.

Han, J. & Kamber, M. 2001 *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, CA.

Hill, D. J., Minsker, B. S. & Amir, E. 2009 Real-time Bayesian anomaly detection in streaming environmental data. *Water Resources Research* **45**(W00D28), 16..

Jin, Y., Olhofer, M. & Sendhoff, B. 2002 A framework for evolutionary optimization with approximate fitness functions. *IEEE Trans. Evol. Comput.* **6**(5), 481–494.

Jones, P. M. & Jacobs, J. L. 2000 Cooperative problem solving in human-machine systems: theory, models, and intelligent associate systems. *IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev.* **30**(4), 397–407.

Kim, C.-Y., Lin, Y.-F. & Bajcsy, P. 2007 *A Framework For Geospatial Models From Sparse Field Measurements Using Image Processing And Machine Learning*. Microsoft eScience Workshop at RENCI, University of North Carolina at Chapel Hill, Chapel Hill, NC.

Kooper, R., Clutter, D., Lee, S.-C. & Bajcsy, P. 2008 *Im2Learn (Image to Learn) Manual*. NCSA, UIUC, Urbana, IL.

Krichevsky, R. & Trofimov, V. 1981 The performance of universal encoding. *IEEE Trans. Inf. Theory.* **27**(2), 8.

Lin, Y.-F. & Anderson, M. P. 2003 A digital procedure for ground water recharge and discharge pattern recognition and rate estimation. *Ground Wat.* **41**(3), 306–315.

Lin, Y.-F., Bajcsy, P. Valocchi, A. J., Kim, C. & Wang, J. 2007 Evaluation of alternative conceptual models using interdisciplinary information: an application in shallow groundwater recharge and discharge. In: *Proc. American Geophysical Union Fall Meeting*. American Geophysical Union, San Francisco, CA.

Lin, Y.-F., Wang, J. & Valocchi, A. J. 2008a A new GIS approach for estimating shallow groundwater recharge and discharge. *Trans. GIS* **12**(4), 459–474.

Lin, Y.-F., Wang, J. & Valocchi, A. J. 2008b PRO-GRADE: GIS toolkits for ground water recharge and discharge estimation. *Ground Wat.* **47**(1), 122–128.

Liu, Y., Hill, D., Rodriguez, A., Marini, L., Kooper, R., Futrelle, J., Minsker, B. & Myers, J. 2008 Near-real-time spatiotemporal precipitation virtual sensor creation based on NEXRAD level. In: *Proc. 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Irvine*, CA. ACM, New York, USA.

Markus, M., Knapp, H. V. & Tasker, G. D. 2003 Entropy and generalized least square methods in assessment of the regional value of streamgages. *J. Hydrol.* **283**(1–4), 14.

McLaughlin, D. & Townley, L. R. 1996 A reassessment of the groundwater inverse problem. *Wat. Res. Res.* **32**(5), 1131–1161.

Quinlan, J. R. 1993 C4.5: *Programs for Machine Learning*. Morgan Kaufmann, San Fransisco.

Rastogi, R. & Shim, K. 2000 PUBLIC: a decision tree classifier that integrates building and pruning. *Data Mining Knowledge Discovery* **4**, 315–344.

Reed, P., Minsker, B. & Goldberg, D. E. 2001 A multiobjective approach to cost effective long-term groundwater monitoring using an elitist nondominated sorted genetic algorithm with historical data. *J. Hydroinf.* **3**(2), 71–89.

Rosman, L. A. 2005 *Storm Water Management Model User's Manual*. US Environmental Protection Agency, Cincinnati, OH.

Sarlak, N. & Sorman, A. U. 2006 Evaluation and selection of streamflow network stations using entropy methods. *Turkish J. Engng. Environ. Sci.* **30**, 9.

Scanlon, B. R., Healy, R. W. & Cook, P. G. 2002 Choosing appropriate techniques for quantifying groundwater recharge. *Hydrogeology Journal* **10**, 18–39.

Shafer, J., Agrawal, R. & Mehta, M. 1996 SPRINT: a scalable parallel classifier for data mining. In: *Proc. 22nd International Conference on Very Large Data Bases (VLDB), Mumbai (Bombay)*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, pp. 544–555.

Shannon, C. E. 1948 A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423, 623–656.

Singh, A., Minsker, B. S. & Valocchi, A. J. 2008 An interactive multi-objective optimization framework for groundwater inverse modeling. *Adv. Wat. Res.* **31**, 1269–1283.

Singh, A., Minsker, B. S. & Bajcsy, P. 2010 Image-based machine learning for reduction of user-fatigue in an interactive model calibration system. *J. Comput. Civil Engng.* **24**(3), 241–251.

Singh, V. P. 1997 The use of entropy in hydrology and water resources. *Hydrol. Process.* **11**, 39.

Stoertz, M. W. 1989 *A New Method for Mapping Groundwater Recharge Areas and for Zoning Recharge for an Inverse Model*. PhD thesis,

Department of Geology and Geophysics, University of Wisconsin–Madison, Madison, WI.

Stoertz, M. W. & Bradbury, K. R. 1989 Mapping recharge areas using a groundwater-flow model - a case-study. *Ground Wat.* **27**(2), 220–228.

Takagi, H. 2001 Interactive evolutionary computation: fusion of the capabilities of EC optimization and human evaluation. *Proc. IEEE* **89**(9), 1275–1296.

Torres, S. A. 2007 *Towards a Demonstrator of an Urban Drainage Decision Support System*. MSc thesis, UNESCO-IHE Institute for Water Education, Delft.

Woods, D. D., Roth, E. M. & Benett, K. 1990 Explorations in joint human-machine cognitive systems. In: *Cognition, Computing and Cooperation* (Robertson, S., Zachary, W. & Black, J.B. (Eds.). Ablex, pp 123–158.

Zhang, Y.-L., Baptista, A. M. & Myers, E. P. 2004 A cross-scale model for 3D baroclinic circulation in estuary-plume-shelf systems: I. Formulation and skill assessment. *Continental Shelf Res.* **24**, 2187–2214.

First received 9 October 2009; accepted in revised form 22 January 2010. Available online 4 October 2010