Upscaling models of solute transport in porous media through genetic programming

David J. Hill, Barbara S. Minsker, Albert J. Valocchi, Vladan Babovic and Maarten Keijzer

ABSTRACT

Due to the considerable computational demands of modeling solute transport in heterogeneous porous media, there is a need for upscaled models that do not require explicit resolution of the small-scale heterogeneity. This study investigates the development of upscaled solute transport models using genetic programming (GP), a domain-independent modeling tool that searches the space of mathematical equations for one or more equations that describe a set of training data. An upscaling methodology is developed that facilitates both the GP search and the implementation of the resulting models. A case study is performed that demonstrates this methodology by developing vertically averaged equations of solute transport in perfectly stratified aquifers. The solute flux models developed for the case study were analyzed for parsimony and physical meaning, resulting in an upscaled model of the enhanced spreading of the solute plume, due to aquifer heterogeneity, as a process that changes from predominantly advective to Fickian. This case study not only demonstrates the use and efficacy of GP as a tool for developing upscaled solute transport models, but it also provides insight into how to approach more realistic multi-dimensional problems with this methodology.

Key words | data-driven modeling, genetic programming, knowledge discovery, solute transport

David J. Hill (corresponding author)
Barbara S. Minsker
Albert J. Valocchi
Department of Civil and Environmental Engineering,
University of Illinois Urbana-Champaign,
Urbana, IL 61801,
USA
Tel: +1 (217) 714 3490
Fax: +1 (217) 333 6968
E-mail: *djhill1@uiuc.edu*

Vladan Babovic
Department of Civil Engineering,
National University Singapore,
1 Engineering Drive 2, E1A 07-03,
Singapore 117576

Maarten Keijzer
Free University Amsterdam,
Amsterdam,
The Netherlands

INTRODUCTION

Solute transport in porous media is fundamental to many significant engineering problems. Thus, modeling this process is an area of active research in many disciplines. One popular method of modeling the movement of solute through porous media involves the use of physically based mathematical equations based on conservation of momentum and mass. Darcy's Law and the advection–dispersion equation (ADE) are widely accepted as the equations governing flow and transport of groundwater and dissolved substances at the continuum scale, the length scale at which the heterogeneous aggregation of soil grains can be treated as a homogeneous spatially averaged material. It is now well recognized, however, that natural porous media exhibit significant spatial variability at the continuum scale and that this variability has a profound impact upon solute fate and transport at the larger field scale relevant to environmental and hydrological problems. The effect of this variability on solute transport is enhanced spreading, a phenomenon referred to as macrodispersion. Detailed measurements at several field sites (Sudicky 1986; Mackay *et al.* 1986; LeBlanc *et al.* 1991) have revealed that the length scale of significant conductivity variations is of the order of a few meters in the horizontal direction but only ten to twenty centimeters in the vertical direction. Therefore, computational limitations prevent the use of a transport model grid fine enough to resolve all of the spatial scales of this variability. Furthermore, many problems of environmental interest require solving the transport models many times (e.g. through the use of Monte Carlo simulations); thus, a need exists for more economical models of solute transport.

doi: 10.2166/hydro.2007.028

For this reason, much effort has been directed towards developing models that describe transport processes at a length scale larger than the continuum scale so that coarse computational grid blocks may be used. These "upscaled" models cannot explicitly resolve all of the salient features of the transport process, yet they should capture the impact of the small-scale heterogeneity in order to provide an accurate prediction of the overall plume evolution.

Traditional methods for upscaling the ADE include stochastics (Gelhar *et al.* 1979; Dagan 1984; Sposito 1997; Rubin 2003; Rubin *et al.* 2003), spatial filtering (Beckie *et al.* 1996; Beckie 1998), homogenization (Mei 1992; Wood *et al.* 2003) and statistical moments (Aris 1956; Frankel & Brenner 1989; Kitanidis 1992; Whitaker 1999). Unfortunately, although these methods are mathematically rigorous, they usually require restrictive assumptions, such as small variability, large scale separation, or ergodicity or periodicity of the medium, to achieve closure of the upscaled models.

This study develops an upscaling methodology using genetic programming (GP), a promising new tool for modeling complex phenomena whose physics are not well defined (Babovic & Abbott 1997a). For illustration, this methodology is applied to the case of developing vertically averaged models of the transport of a non-reactive solute in confined stratified aquifers. The results are compared with models developed through the method of moments (MoM), a traditional upscaling technique that is well suited for this transport configuration (Güven *et al.* 1984).

METHODS

The upscaling methodology developed in this study takes advantage of GP's ability to model complex phenomena. This section includes a description of GP, followed by the mathematical formulation of the upscaling problem addressed in this study.

Genetic programming

Genetic programming is a domain-independent method that creates a model based on input data by searching the space of possible models. This search uses operations inspired by natural evolution, which allow GP to cultivate a diverse set of approaches to solving the problem (Banzhaf *et al.* 1998). Genetic programming has shown success in many applications (e.g. Koza *et al.* 1999; Savic *et al.* 1999). Babovic & Abbott (1997b) present four applications of GP in the field of hydrology. The results of these applications illustrate the abilities of GP to: (1) model "emergent phenomena," (2) find models of data that match human derived models, (3) develop models of phenomena that are of higher quality than human derived models, and (4) find models of complex phenomena that are equally accurate, yet simpler to solve, than many human derived models.

Because this research is interested in developing mathematical models of a physical process, GP was configured to suggest models in the form of mathematical equations, a task referred to as symbolic regression. Regression is the most familiar method of determining relationships between data and known parameters. In traditional regression methods, first a model structure is selected. Then, the coefficients of that model are estimated, based on available data using a model-fitting algorithm. This method builds the user's bias into the resulting relationship through the functional form of the model chosen for regression. Symbolic regression, however, is a less biased method of determining a relationship between data and known parameters because it determines, based on the available data, not only model coefficients, but also the functional form of the model itself (Babovic & Bojkov 2001).

The process of symbolic regression begins with the establishment of a population of models that has been randomly generated from sets of independent variables and mathematical operators. Each model can be conceptualized as a hierarchy of building blocks connected via mathematical operators, each of which is a valid mathematical statement. These building blocks will hereafter be referred to as clauses. The search for models that best fit the data is directed by one or more objectives that describe the desired qualities of the model. The fitness of a candidate model is based on its fulfillment of these objectives. The search progresses as a series of iterations, known as epochs, and the population in each successive epoch is generated by selecting some of the models for propagation. Selection favors models with higher fitness. Models are propagated into the next epoch either without modification or with modification through the operations of crossover or mutation. Crossover is performed by swapping clauses between two equations, whereas mutation is performed by altering an independent variable, constant or mathematical operator in an equation.

In this research, a symbolic regression implementation known as adaptive logic programming (ALP) was used. ALP employs the concise language of logic programming to facilitate the search through the space of possible mathematical equations. This language enables convenient performance of crossover and mutation and avoidance of syntactically incorrect equations via these operations. More information regarding the ALP system can be found in Keijzer *et al.* (2001).

While other data-driven methods exist that will create black box models that map input data to outputs (e.g. artificial neural networks), symbolic regression provides the benefit of expressing the models in the language of mathematics; hence they can be analyzed for information regarding the underlying processes that created the data. This information can lead to new understandings of the physical processes being modeled.

While symbolic regression provides the advantage of constructing models without domain-specific knowledge, the field of application or desired use of the model may impose constraints. In the case of this research, three goals required the imposition of constraints on the symbolic regression task based on the desire to create: (1) physically meaningful models, (2) models that are parsimonious, and (3) models that are expressed as partial differential equations (PDEs).

Models of the physical domain must be dimensionally consistent if they are to be considered meaningful; thus, it is necessary to constrain the GP search to only dimensionally consistent equations. While this can be accomplished in many ways (e.g. Keijzer & Babovic 1999), it is most easily accomplished by converting the model parameters into dimensionless values – the strategy used in this study.

In addition to dimensional consistency, model parsimony is desired, because it removes parameters that add to model uncertainty without compromising predictive ability, and it renders models that are easier both to analyze for semantic meaning and to implement numerically. Symbolic regression will not necessarily find the most concise form of a mathematical statement. In fact, theoretical studies have shown that GP has a tendency to construct models with many extraneous clauses in an effort to protect salient clauses from the destructive effects of crossover and mutation (Banzhaf & Langdon 2002), a phenomenon commonly referred to as “bloat.” Therefore, it is often necessary for the user to simplify the resulting models into statements that are easier to implement and analyze. Useful strategies for the user to manually address model simplification include converting mathematical operators to equivalent series representations (e.g. using a Maclaurin series to represent an exponential function) and replacing clauses that approximate constant values with constant-valued parameters. Furthermore, domain knowledge can be used to modify the model to address shortcomings in its predictive ability.

Building differential equations via symbolic regression is difficult, because no general differential equation solver exists to evaluate the fitness of the candidate equations. Thus, it is important to find a method of learning differential equations without requiring integration of each candidate differential equation in the population. In this research, the upscaling problem is decomposed into a new problem that does not require the use of calculus to evaluate the objectives, as described in the next section.

MATHEMATICAL FORMULATION

Because data regarding the target phenomenon is presented to ALP as a list of examples containing several descriptive attributes and the observed response of the system, and because it is necessary for the resulting upscaled models to be easily implemented, in this study, the upscaling problem was reduced to a problem of calculating upscaled solute fluxes. The mathematical formulation starts with the ADE, as it is assumed that this model is valid for continuum-scale solute transport. Using the summation convention for repeated indices, the ADE can be expressed as

$$\frac{\partial C}{\partial t} = -\frac{\partial}{\partial x_i}(u_i C) + \frac{\partial}{\partial x_i} D_{i,j} \frac{\partial C}{\partial x_j} \tag{1}$$

where C is the continuum-scale solute concentration, t is the time, x_i is the Cartesian position vector, u_i is the pore water velocity vector, and $D_{i,j}$ is the dispersion tensor. This equation can also be expressed in terms of fluxes as

$$\frac{\partial C}{\partial t} = -\frac{\partial J_i^A}{\partial x_i} + \frac{\partial J_i^D}{\partial x_i} = -\frac{\partial J_i^T}{\partial x_i} \tag{2}$$

where J^A is the solute flux due to continuum-scale advection, J^D is the solute flux due to continuum-scale dispersion, and J^T is the total continuum-scale solute flux. By employing spatial filtering, as shown by Beckie (1998), Equation (2) can be upscaled to the block-scale equation:

$$\frac{\partial \bar{C}}{\partial t} = -\frac{\partial \bar{J}_i^T}{\partial x_i} \tag{3}$$

where the overbar indicates a spatially filtered quantity equivalent to the convolution integral of the continuum-scale quantity multiplied by a filtering function. Since volume averaging is a form of spatial filtering, the filtered terms can be thought of as block-scale averages (Nitsche & Brenner 1989; Beckie *et al.* 1996). With some algebra, the block-scale total solute flux can be divided such that

$$\bar{J}_i^T = \bar{J}_i^A + \bar{J}_i^{NA} \tag{4}$$

where $\bar{J}^A = \bar{u}_i\bar{C}$is the solute flux due to block-scale advection, and $\bar{J}^{NA}$ is the remaining non-advective solute flux. A similar decomposition of the block averaged flux was used by Efendiev *et al.* (2000). Unlike the block-scale advective flux, this latter term contains sub-grid closure quantities, and thus cannot be easily modeled at the block scale. However, by collecting data regarding the block-scale non-advective flux, ALP can be used to develop models of this flux in terms of other block-scale parameters, which will allow the solution of Equation (3). Thus, the upscaling problem is reduced to the problem of finding a model for the block-scale non-advective flux in terms of resolvable block-scale quantities, and ALP does not have to search the space of PDEs.

Data describing the block-scale non-advective flux can be generated using two numerical grids: (1) a highly resolved grid and (2) a coarse grid representing the block scale. The ADE is solved numerically on the fine grid, while the coarse grid is used for evaluating block-averaged parameters throughout the simulation. Block-scale parameters for each grid location are calculated by averaging the fine-scale parameters over the entire block, or in the case of vector quantities, such as the non-advective flux, over the appropriate block surface for each vector component.

CASE STUDY

This case study presents the development of vertically averaged models of the transport of non-reactive solutes in two-dimensional, confined, perfectly stratified aquifers (i.e. vertically varying horizontal flow parallel to the layers). This idealized transport system was selected to allow comparison of the GP-derived vertically averaged equations with those derived using the method of moments (MoM) (Taylor 1953; Aris 1956; Güven *et al.* 1984; Kitanidis 1992), a well accepted approach for perfectly stratified aquifers. This section will proceed by first describing the upscaled transport equations that can be derived using the MoM. Then, two test cases using different synthetically generated velocity fields will be defined. Next, mathematical simplifications to Equations (3) and (4), which are made possible by vertically averaging two-dimensional, confined, perfectly stratified aquifers, will be discussed. Finally, the generation of input data for ALP will be explained.

Method of moments

The MoM aims to describe the solute plume at any point in time in terms of its spatial moments. Mathematical expressions for the solute distribution's moments as functions of time can be derived from the ADE. For the case of transport in a laminar shear flow (equivalent to the case of horizontal flow in a perfectly stratified aquifer), Aris (1956) demonstrated that the MoM could be used to derive models for the temporal evolution of the spatial moments of the cross-sectionally averaged concentration. Commonly, only the zeroth, first, and second spatial moments are considered, as the models for higher-order moments are more cumbersome. The zeroth moment indicates the total solute mass in the system, the first moment indicates the mean position of the plume, and the second moment indicates the plume spread. The resulting upscaled model of transport has the same form as the ADE, except that an effective velocity vector calculated from the first moment replaces the velocity vector, and the dispersion tensor is replaced with a time-dependent macrodispersion tensor calculated via the second moment. Thus, the MoM model of the non-advective flux can be expressed as

$$\bar{J}_i^{NA} = D_i^{eff}(t)\frac{\partial \bar{C}}{\partial x_i} \tag{5}$$

where $D_i^{eff}(t)$is the macrodispersion coefficient. It can be seen that the assumption of locally Fickian macrodispersion (i.e. the assumption that plume spreading due to the variability of hydraulic conductivity can be effectively modeled as a random process) is inherent in this model. The resulting MoM model describes the solute distribution at any time as Gaussian, with the same mean and variance as the observed plume.

Because this type of model employs the assumption of locally Fickian macrodispersion, it is only valid when the plume has spread sufficiently, such that all velocities are sampled with the same frequency with which they appear in space. Furthermore, this method requires assumptions regarding the continuum-scale velocity field in order to close the equations for the spatial moments of the solute distribution. One common assumption is that of a periodic medium (Kitanidis 1992; Wood *et al.* 2003); in particular, Aris (1956) showed that for confined, perfectly stratified aquifers with flow parallel to the layers, such as those considered in this study, it is possible to rigorously derive the effective velocity and macrodispersion terms.

Synthetic aquifers

The process of defining the properties of the synthetic aquifers for this study was guided by the desire to facilitate comparison with the MoM and maximize the generalizability of the aquifers. Assuming a two-dimensional system where flow is parallel to the x axis, and the z axis represents the aquifer depth, any arbitrary velocity profile can be discretized into small sub-layers of constant velocity. Because of this, and because this study is confined to vertically averaged blocks, a two-dimensional computational grid is necessary to fully resolve the fine-scale concentration distribution, whereas the aquifer's vertically averaged counterpart can be resolved with a one-dimensional computational grid. The two-dimensional representation will hereafter be referred to as the fine-scale representation. The fine- and block-scale (coarse-scale) representations are illustrated in Figure 1. In order to facilitate comparison with the MoM, the following two velocity distributions were considered:

$$u(z) = h^2\left(1 - \frac{z^2}{h^2}\right) \tag{6a}$$

$$u(z) = 0.5\cos\left(2\pi\frac{z}{h}\right) + 0.5\cos\left(4\pi\frac{z}{h}\right) + 1 \tag{6b}$$

where h is the aquifer depth. The flow distributions described by Equations (6a,b) will hereafter be referred to as parabolic and cos–cos, respectively. Two distributions were chosen in order to demonstrate the generality of the derived upscaled models to solute transported by different velocity distributions. The parabolic distribution was selected because many MoM studies address transport by this distribution (e.g. Aris 1956; Güven *et al.* 1984), while the cos–cos distribution was selected because it varies more sharply than the parabolic distribution. These velocity distributions were applied by creating 100-layer synthetic aquifers and defining the flow rate in each layer, such that the total mass of water passing through the layer was equivalent to the total mass of water that would pass through the same discretized region using Equations (6a,b). The number of layers for the aquifers was selected in order to minimize the differences between the discretized distribution of the aquifers and the continuous distribution functions, thus facilitating comparisons with the MoM upscaled equations, since these equations use continuous velocity distributions. The depth of the aquifers was chosen

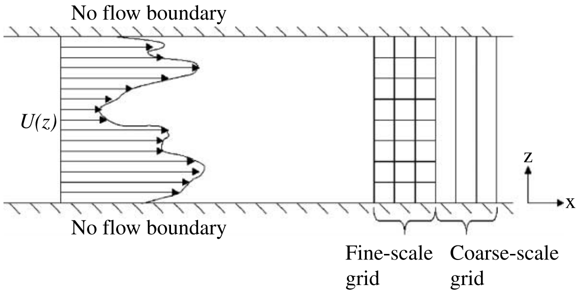

Figure 1 | Schematic of a two-dimensional perfectly stratified aquifer indicating both the fine- and block-scale computational grids as well as the fine-scale velocity distribution.

to be 1 m. The transverse dispersion coefficient within the aquifers was specified to be 0.01 m/s^2. Since it has been shown that the longitudinal spreading of the plume due to aquifer heterogeneity is significantly larger than that due to continuum-scale dispersion (Gelhar *et al.* 1979), the latter was ignored. The exact values of these parameters, however, are unimportant, because, as will be discussed shortly, dimensionless parameters are used to describe the aquifers. Thus, the numerical results of the transport simulation can be scaled to represent a large number of aquifer geometries and transverse dispersion conditions.

Simplifying the numerical formulation

The use of vertically averaged representations of two-dimensional aquifers allows two simplifications to be made to Equations (3) and (4). First, the subscripts (i) can be dropped from the vector quantities (e.g. $\bar{J}_i^{NA}$) because the vectors have only one component (i.e. x directional). Thus, Equations (3) and (4) can be combined and simplified to

$$\frac{\partial \bar{C}}{\partial t} = -\frac{\partial}{\partial x}\bar{u}\bar{C} - \frac{\partial \bar{J}^{NA}}{\partial x}. \tag{7}$$

The second simplification involves converting from a Cartesian to a Lagrangian coordinate system that moves at the vertically averaged pore water velocity. In this coordinate system, the position vector is $x_R = x - \bar{u}t$, and Equation (7) becomes

$$\frac{\partial \bar{C}}{\partial t} = -\frac{\partial \bar{J}^{NA}}{\partial x_R} \tag{8}$$

This simplification allows for convenient implementation of the upscaled solute transport model, because the block-scale advective flux is implicitly accounted for by the Lagrangian coordinate transformation. Therefore, the block-scale solute concentration can be calculated using Equation (8), where $\bar{J}^{NA}$is modeled by ALP.

Since dimensionless training data are used to implicitly constrain ALP to dimensionally consistent equations, the following transformations were used to convert dimensional data into dimensionless data:

$$\phi = \frac{C}{C_0} \tag{9a}$$

$$\tau = \frac{tD_T}{h^2} \tag{9b}$$

$$\xi = \frac{x - \bar{u}t}{h} \tag{9c}$$

$$\psi^{NA} = J^{NA}\frac{h}{D_T C_0} \tag{9d}$$

where C_0 is the input concentration and D_T is the transverse dispersion coefficient.

Generation of training data

In order for ALP to develop models of the block-scale non-advective flux, it is necessary to provide a set of training examples that contain block-scale descriptive attributes and the resulting block-scale non-advective flux observed in the aquifer being studied. Training examples were collected from the aquifer with the parabolic flow distribution for a pulse input of solute. The time evolution of the continuum-scale solute distribution was solved using a numerical finite difference solution to the ADE, which employed operator splitting to separate the modeling of the advective and dispersive processes. A third-order explicit total variation diminishing (TVD) method (Leonard 1988) was used to solve the advection term, while an implicit method was used for the dispersion term. This method for solving the ADE was selected to minimize prediction errors of the numerical solution to the ADE and, thus, to minimize errors in the training data. The fine-scale solution used a highly refined computational grid to minimize numerical errors. The block-scale computational grid was defined such that each block spanned the entire depth of the aquifer and had the same length as the fine-scale grid.

The descriptive attributes selected to describe the block-scale non-advective flux at the block interface consisted of the block-scale concentration ($\bar{\phi}$) at the upstream and downstream block centers; the position of observed block interface (ξ); the time of observation (τ); the block-scale concentration gradient ($d\bar{\phi}/d\xi$) at the block interface, as well as at three upstream locations; the second spatial derivative of the block-scale concentration ($d^2\bar{\phi}/d\xi^2$) at the block interface and at three upstream locations; and the mixed space/time derivative of the block-scale concentration ($d^2\bar{\phi}/d\xi d\tau$) at the block interface and at three upstream locations. The values of the block-scale derivatives were estimated using discrete approximations. Unless otherwise indicated, these attributes were recorded at the block interfaces. These attributes were selected because they appeared in upscaled transport equations derived using traditional methods (Gelhar & Axness 1983; Beckie 1998; Efendiev *et al.* 2000). Because there are many block interfaces and time steps in the numerical simulation of the fine-scale transport model, it was necessary to select only a subset of the salient examples for training. Examples from a particular block interface were considered salient only if the solute concentration at some point within a five-block neighborhood surrounding the interface was non-zero. The examples used to train ALP were selected at random from the set of salient examples that were recorded during every tenth time step of the simulation.

Parameterization of ALP

ALP requires many user-selected parameters to define the search for good models of the training examples, including objective, functional set, population size, and number of training epochs. Unfortunately, there is little theoretical work to direct the selection of these parameters, and thus, a large number of experiments were performed with different values in order to find good solutions.

The objective parameter defines the criteria by which to evaluate the quality of the derived equations for predicting the target attribute. ALP implements several types of objectives, including both goodness-of-fit (e.g. sum of squared errors) and parsimony (e.g. equation length) objectives; furthermore, ALP permits multi-objective searches. In this study, the objectives were selected to be the correlation coefficient (r^2) between the candidate equation and the training data and the equation length. The r^2 statistic indicates the degree to which the relationship between two variables is linear; thus, it is insensitive to relational constants, such as scale or shift (Devore 1995). The r^2 statistic was selected for this latter property because it does not require ALP to find the correct value of the relational constants, a task that is generally difficult for GP (Koza 1992). Scale constants refer to constants that are multiplied to or divided from a function, while shift constants refer to constants that are added to or subtracted from a function. When using the r^2 objective, it is necessary to determine both the scale and shift constants *a posteriori*. In this research, the scale and shift parameters were determined by performing linear regression of the data pairs composed from the observed non-advective flux in the training data and the model prediction of this flux. The slope and y intercept of this line are the scale and shift parameters, respectively. The equation length was selected to encourage ALP to explore equations of varying complexity and to control bloat. Because ALP uses a multi-objective search, controlling bloat in this manner will not eliminate clauses that improve the model's r^2 value.

The functional set defines the mathematical operators that can be used to relate the attributes to the target value. Several functional sets were evaluated, including the set of all arithmetic operators (e.g. $x+y$, where x and y are attributes), the set of all arithmetic and geometric operators (e.g. $\sin(x)$, where x is an attribute), and the set of all arithmetic operators and the exponential function (i.e. e^x, where x is an attribute). It was observed that the latter two functional sets found many large equations that fit the data well and consisted of long chains of sine/cosine terms or exponential functions, respectively. It is well known that any continuous function can be represented by an infinite series of sine/cosine or exponential terms through the formation of Fourier or Taylor series. Because the exponential operator is easier to simplify than the set of geometric operators, the functional set including both arithmetic operators and the exponential function was used in order to retain the expressivity facilitated by exponential functions without overwhelming the GP results with difficult-to-analyze solutions.

The population size specifies how many candidate equations participate in the search for good equations. Larger populations contain a greater variety of clause building blocks from which to derive new candidate equations; however, larger populations also increase the time it takes to evaluate one epoch of GP. Therefore, it is necessary to have a population that is large enough to represent an adequate number of discrete clauses, yet small enough to allow a reasonable computation time. According to the guideline presented by Sastry *et al.* (2003) the population would have to contain over 11 million individuals in order to guarantee a good supply of building blocks in this research. However, preliminary results indicated that a population size of 1000 individuals was sufficient to produce good results. For this reason, a population size of 1000 candidate equations was chosen.

The number of epochs specifies how many iterations of the genetic operations the population of candidate equations is subjected to. Langdon & Poli (2002) showed that more epochs result in a larger number of extraneous clauses in each candidate equation in the final epoch. Therefore, it is common to use only a few training epochs but perform GP many times (Koza 1992). In this paper, each run of ALP will be referred to as an experiment. After all the experiments have been completed, the results from each experiment are merged, eliminating all the results that are dominated by results from different experiments, resulting in a "front" of non-dominated (and thus Pareto optimal) equations for modeling the training data. The candidate equations are then evaluated for semantic meaning, as well as for goodness of fit. In this study, hundreds of experiments were performed, during which the candidates were evolved for 50 epochs.

In addition to these parameters, a crossover rate of 0.8, a mutation rate of 0.1, and binary tournament selection were used. These values reflect those recommended by the developers of the ALP system from extensive trials on many different functions (e.g. Keijer *et al.* 2001; Babovic *et al.* 2001; Keijzer 2002; Keijzer & Cattolico 2002).

RESULTS

The results from many experiments of ALP compose a Pareto front of non-dominated solutions to the GP task. In this case, the front consisted of many equations with nearly equivalent r^2 values but widely varying lengths. In general, longer equations tended to have slightly higher r^2 values. Analysis of these models, however, showed that the majority of the equations along the front contained the same clause, along with many irrelevant or nearly-irrelevant clauses that could be removed without significantly reducing the ability of the equations to fit the training data. These extraneous clauses were considered to be the result of GP bloat and, thus, were removed from the equations, resulting in a consensus on a final model for the data.

Three characteristic results from along the Pareto front were

$$\bar{\psi}_i^{NA} = \bar{\phi}_{i-\Delta\xi/2}\frac{\xi}{\tau} \tag{10a}$$

$$\bar{\psi}_i^{NA} = \exp\left(\exp\left(\bar{\phi}_{i-\Delta\xi/2}\frac{\xi}{\tau}\right)\right) \tag{10b}$$

$$\bar{\psi}_i^{NA} = \exp\left(\exp\left(0.85\frac{\partial\bar{\phi}}{\partial\xi}\right)\right) - \frac{\left(\left(\frac{\bar{\phi}_{i-\Delta\xi/2}+\bar{\phi}_{i+\Delta\xi/2}}{2}\right)\frac{\xi}{\tau} - \exp\left(0.85\frac{\partial\bar{\phi}}{\partial\xi}\right)\right)}{\exp(\exp(\bar{\phi}_{i+\Delta\xi/2}))} \tag{10c}$$

where the subscript i indicates the ith block interface. These results fit the training data with r^2 values of 0.95, 0.97, and 0.95, respectively; thus, the models are of similar quality, but they differ greatly with regard to semantics and complexity. Equation (10a) contains only three parameters, all of which contribute significantly to its quality. However, it is interesting that the model only includes the dimensionless concentration upstream of the block face. If this parameter is replaced with the dimensionless concentration at the block face (calculated as the average of the concentration upstream and downstream of the block face), the resulting equation becomes

$$\bar{\psi}_i^{NA} = \frac{\bar{\phi}_{i-\Delta\xi/2}+\bar{\phi}_{i+\Delta\xi/2}}{2}\frac{\xi}{\tau} \tag{11}$$

which also has an r^2 value of 0.95. Thus, the replacement of the upstream concentration with the block face centered concentration does not improve (or reduce) the quality of the model's fit with the non-advective flux data, but it does improve the performance of this model for prediction of the block-scale solute distribution. The improvement in the solute distribution prediction occurs because, at the downstream edge of a solute plume, the model shown in Equation (10a) will predict zero solute flux, whereas the model shown in Equation (11) will predict a finite non-advective solute flux, the latter case being the physically plausible model response. This discrepancy in flux prediction between the two models occurs because, at the downstream edge of the plume (i.e. at position i), the concentration a bit further downstream (i.e. at position $i + \Delta\xi/2$) is zero, whereas the concentration just upstream (i.e. at position $i - \Delta\xi/2$) is non-zero.

Equation (10b) can also be reduced to Equation (11). Recall that the exponential function is equivalent to the Maclaurin series:

$$\exp(a) = \sum_{n=0}^{\infty} \frac{a^n}{n!}. \tag{12}$$

Since the magnitude of the product within the exponentials in Equation (10b) is always less than 1, the terms in the series get smaller as $n \to \infty$; therefore, all but the first two terms of the series can be ignored. If this procedure is followed for both exponential functions, the resulting model is a linear function of Equation (10a). Since the r^2 statistic is insensitive to scale and shift parameters, the r^2 value of the approximation to Equation (10b) is equal to that of Equation (10a), namely, 0.95. In return for the reduction in performance caused by this approximation, there is a substantial increase in both semantic meaning and ease of implementation of this model.

Equation (10c) can also be reduced to Equation (11) through evaluation of its clauses. Equation (10c) can be divided into four clauses such that

$$\bar{\psi}_i^{NA} = clause_1 - (clause_2 - clause_3)/clause_4. \tag{13}$$

Using the training data, the minimum, mean, and maximum values of *clause₁* can be calculated to be 2.705, 2.718, and 2.720, respectively. Because the clause has a small range, it can be replaced by its mean with little loss of generality. The same is true for $clause_3$ and $clause_4$. Since $clause_2$ is Equation (11), it can be seen that the approximation of Equation (10c) is a linear function of Equation (11); thus, the r^2 value is 0.95, which is equivalent to the r^2 value of Equation (10c). Therefore, there is no measurable predictive ability lost by using Equation (11) to approximate Equation (10c).

In the preceding discussion, it was demonstrated that different length models from the set of Pareto optimal solutions could be simplified to the same model without a significant loss of predictive ability. However, if the longer models have a higher r^2 value, why were they not preferred? The answer is twofold. First, there is a precedent in learning theory to prefer simpler models to more complex models with similar predictive abilities (i.e. Occam's razor) (Duda *et al.* 2001). Second, the longer models are often too complex to be implemented numerically. Furthermore, a t test with a 95% significance level showed that the difference in r^2 values between the longer models and Equation (11) is insignificant. Note that, due to space constraints, only a few of the shorter equations were discussed; the same techniques can be applied to the longer equations along the Pareto front, often resulting in Equation (11). This consensus between models strengthens the claim that Equation (11) best models the non-advective flux.

It should now be clear that many of the Pareto-optimal results of the GP task can be reduced to one common equation shown in Equation (11). This model will hereafter be referred to as the sub-grid advective (SGA) model for reasons that will become clear shortly. The SGA model is strongly correlated with the non-advective solute flux from a pulse input of solute in both the aquifer with the parabolic velocity distribution and the aquifer with the cos–cos velocity distribution, with r^2 values of 0.95 and 0.93, respectively, whereas the MoM model (Equation (5)) is only weakly correlated with the observed non-advective flux, indicated by r^2 values for the parabolic and cos–cos velocity distributions of 0.1 and 0.24, respectively. This result enables two conclusions. First, since flux data from the cos–cos velocity distribution was not used for training, this result indicates that the SGA model generalizes to different flow conditions than those used for training. This generality suggests that the SGA model describes the mechanism of macrodispersion, rather than merely being a concise representation of the training data. Second, because a strong correlation exists between the SGA model and the observed non-advective flux, but not between the MoM model and the observed non-advective flux, this result indicates that the SGA model is a better predictor of the observed non-advective flux than the MoM model. This result may appear surprising because the MoM model should be correct at late times, when the assumption of Fickian macrodispersion is valid. However, the r^2 metric considers the model residuals holistically with respect to time, and the model residuals are more likely to be large in magnitude at early times than at later times because the magnitude of the non-advective flux is larger at early times. Therefore, the r^2 metric is biased towards early time behavior. Since r^2 was used as the goodness-of-fit metric, this latter conclusion suggests why ALP did not create any models similar to the MoM model.

In order to determine the scale and shift constants, linear regression between the SGA model (converted back into dimensional form) and the observed non-advective flux was performed. This regression indicated approximate scale and shift parameters of 1 and 0, respectively, resulting in the equation:

$$\bar{J}^{NA} = \bar{C}\frac{x_R}{t} \tag{14}$$

where x_R is the position in the Lagrangian coordinate system. This model describes the non-advective flux of solute in the aquifer with the parabolic flow profile. The zero value of the shift parameter is expected, because a non-zero shift parameter would indicate that a significant component of the non-advective flux could not be modeled by the SGA model. Equations (8) and (14) can be solved numerically to predict the time evolution of a pulse input of solute in the aquifer with the parabolic flow profile at times greater than zero (since time appears in the denominator).

Figure 2 compares the performance of the SGA upscaled model with the MoM upscaled model for predicting the evolution of the solute plume. Because the SGA model is not valid at very early times, the fine-scale model was used to predict the plume evolution for the first 30 time steps of the simulation (until $\tau = 0.03$), before the SGA and MoM models took over the prediction. It can be seen that, at early times, the SGA model more closely approximates the plume shape, whereas at later times, the MoM model produces a more accurate result. However, it is important to note the y axis scale when comparing the two predictions, because the maximum absolute error between the MoM model and the fine-scale model is much larger than the maximum absolute error between the SGA model and the fine-scale model, as depicted in case 0 of Figure 3. This latter result may be misleading because the MoM model was derived such that the error between the predicted and observed values of the first two spatial moments of the plume is minimized, while the SGA model (Equation (11)) was developed with the goal of minimizing the total error. Thus, a comparison that invokes absolute errors will be biased towards the SGA model, while a comparison based on moments will be biased towards the MoM model. However, a comparison of the time evolution of the first two moments, calculated by the SGA and MoM models, to the first two moments of the plume, calculated using the ADE, indicates that both the SGA and MoM capture the time evolution of the zeroth, first and second spatial moments well with average errors of less than one-half percent. Therefore, in our numerical experiment, the two models perform equally well when compared via spatial moments.

Analyzing the second term in the SGA illustrates that the block-scale non-advective flux can be attributed to solute advection that occurs below the block scale:

$$\frac{x_R}{t} = \frac{x - \bar{u}t}{t} = u - \bar{u} = u' \tag{15}$$

Equation (15) relates the position of the solute to u', the deviation (from the mean velocity) of the unresolved velocity. Thus, the SGA model approximates the solute transported by unresolved velocity variations using block-scale resolvable parameters. In fact, it can be shown that the SGA model is quite capable of reproducing the solute plume evolution of a pulse input in a perfectly stratified system with no transverse mixing.

Figure 2 indicates that, at early times, the time evolution of the solute distribution in the aquifer appears purely advective, and that the time evolution of the plume at late times is well described by the MoM model. However, at intermediate times, the solute plume behaves in a manner consistent with a combination of the pure advection and Fickian macrodispersion cases, where the influence of the SGA model decreases with time, and the influence of the MoM model increases with time. Thus, a new model, which is a hybrid of the SGA and MoM models, is suggested:

$$\bar{J}^{NA} = F(t)^* \underbrace{\left(D_\infty^{eff}\frac{\partial\bar{C}}{\partial x}\right)}_{MoM\,\text{model}} + (1 - F(t))^* \underbrace{\left(\bar{C}\frac{x_R}{t}\right)}_{SGA\,\text{model}} \tag{16}$$

where $F(t)$ is a continuous function over all values of t and has a minimum value of zero that occurs at $t = 0$ and a

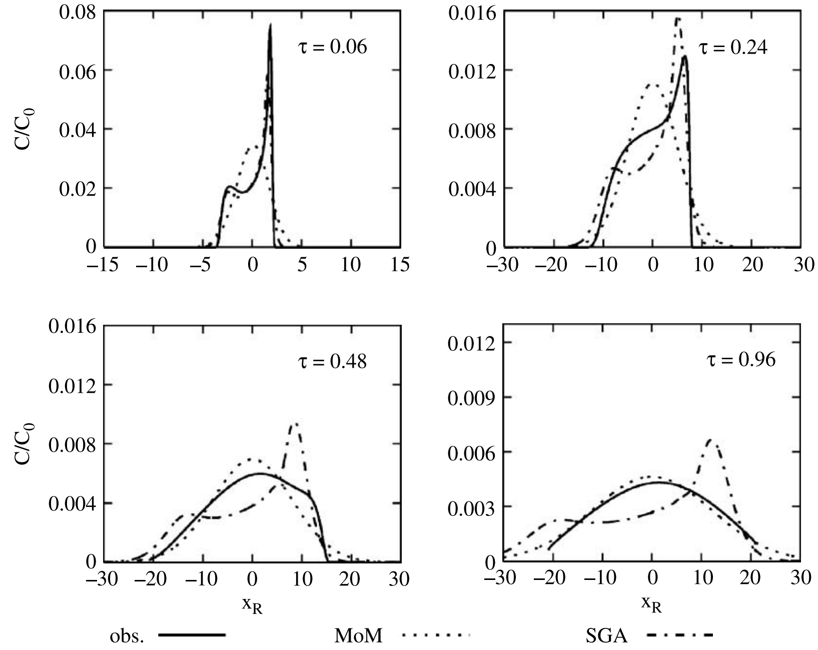

Figure 2 | Comparison of the MoM and SGA upscaled models for predicting the vertically averaged time evolution of a pulse input in the synthetic aquifer with parabolic flow.

maximum value of 1 that occurs at very late time, and D_∞^{eff} is the asymptotic coefficient of macrodispersion suggested by the MoM. The function $F(t)$, which will hereafter be referred to as the mixing function, controls the influence of both the SGA and MoM models over time, allowing the SGA model to dominate the behavior of the solute distribution at early times and allowing the MoM model to dominate its behavior at later times. This model is consistent with a conceptual model of the transport process in which, at early times, insufficient solute has been exchanged between the layers, such that the process is similar to the pure advection process described by the SGA model. As a larger quantity of solute samples more of the flow paths in the individual layers, a larger fraction of the transport process behaves in a manner consistent with Fickian macrodispersion; once sufficient time has passed for the average solute behavior to be consistent with having sampled all the flow paths, the process is well described by Fickian macrodispersion. Equations (7) and (16) can be solved to predict the time evolution of a pulse input of solute in an aquifer at times greater than zero.

In the discussion above, the mixing function $F(t)$ was intentionally vaguely defined because the optimal function may vary depending on transport conditions. In this research, a sigmoid function:

$$F(t) = \left[1 + \exp\left(-\left(t\frac{D_t}{h^2} - a\right)/b\right)\right]^{-1} \tag{17}$$

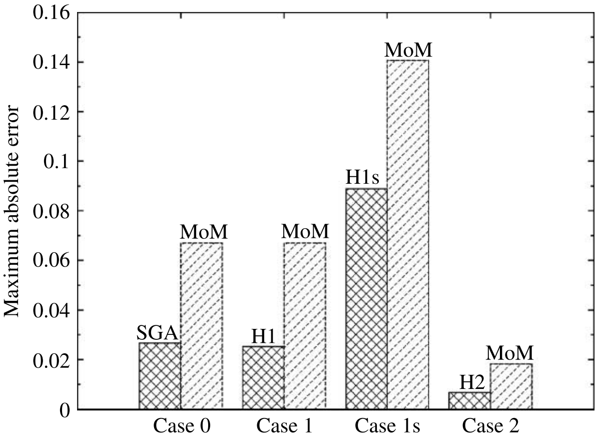

Figure 3 | Comparison of the maximum absolute errors between ALP derived upscaled models (SGA, H1, H1s, and H2) and the MoM upscaled model. Cases 0 and 1 refer to the transport of a pulse input in the synthetic aquifer with parabolic flow, case 1s refers to the transport of a finite width input in the synthetic aquifer with parabolic flow, and case 2 refers to the transport of a pulse input in the synthetic aquifer with cos–cos flow.

was chosen to demonstrate how the model described by Equation (16) performs on the aquifers considered in this study. Note that, in this equation, time is normalized by the transverse dispersivity and aquifer depth, resulting in an equation that generalizes to other aquifers with similar flow distributions but different depths and transverse dispersion coefficients. The parameters for the sigmoid were chosen via a manual trial-and-error approach using visual inspection of the solute distribution shapes to guide the search. Nevertheless, the upscaled models created performed well on a variety of transport conditions, including different initial conditions of the solute input into the aquifer with parabolic flow and different velocity distributions. Since, as described previously, the SGA model is not valid at very early times, for the experiments described below, the fine-scale model was used to predict the plume evolution for the first 30 time steps ($\tau \leq 0.03$) of the simulation, after which the hybrid and MoM models took over the prediction.

For the case of a pulse input into the aquifer with parabolic flow (which is equivalent to ALP's training conditions), values of 0.3 and 0.01 were chosen for the sigmoid function parameters, *a* and *b*, respectively. It can be seen in Figure 4 that the hybrid model with these parameter values, hereafter referred to as H1, performs better when compared to the MoM model. At early times, model H1 preserves the favorable behavior of the SGA model, predicting the bimodal solute distribution with high accuracy. At late times, model H1 retains the benefits of the MoM model, predicting a unimodal, nearly Gaussian plume. At intermediate times, model H1 performs quite well at capturing the peak concentration and shape of the plume's leading edge. It also does quite well at capturing the overall shape when compared to the MoM model. Furthermore, the maximum absolute concentration deviation between model H1's prediction and the observed plume shape is less than that of either the SGA or MoM models alone, as illustrated in Figure 3.

The predictive abilities of the upscaled model H1, however, are not restricted to the conditions on which it was developed. Using superposition, this model can be extended to the cases of an instantaneous finite width input, a finite duration input, or even a continuous input, by summing the effects of multiple pulse inputs each calculated with model H1. For example, Figure 5 shows the superior

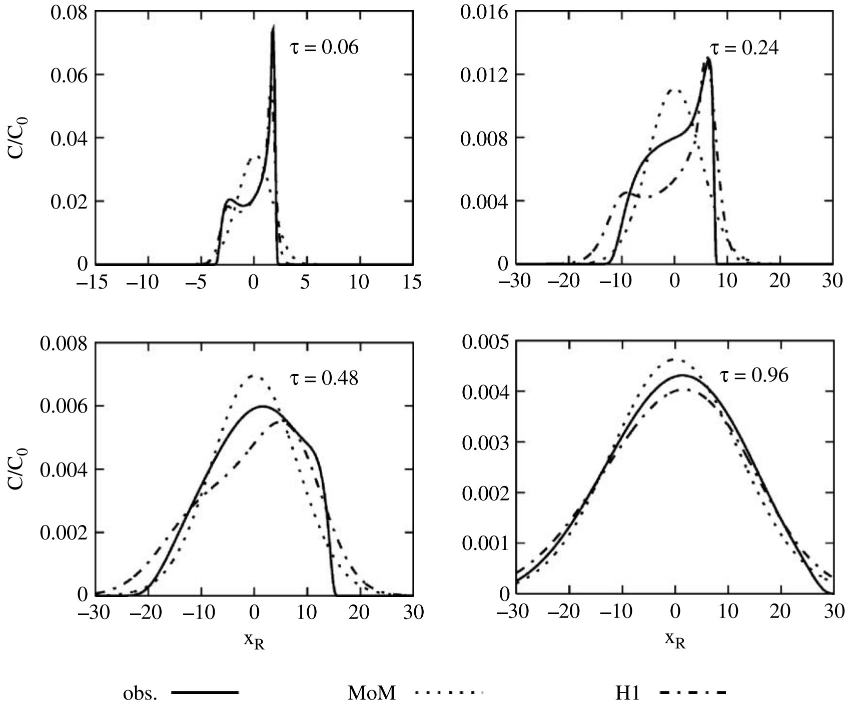

Figure 4 | Comparison of the MoM and H1 upscaled models for predicting the vertically averaged time evolution of a pulse input in the synthetic aquifer with parabolic flow.

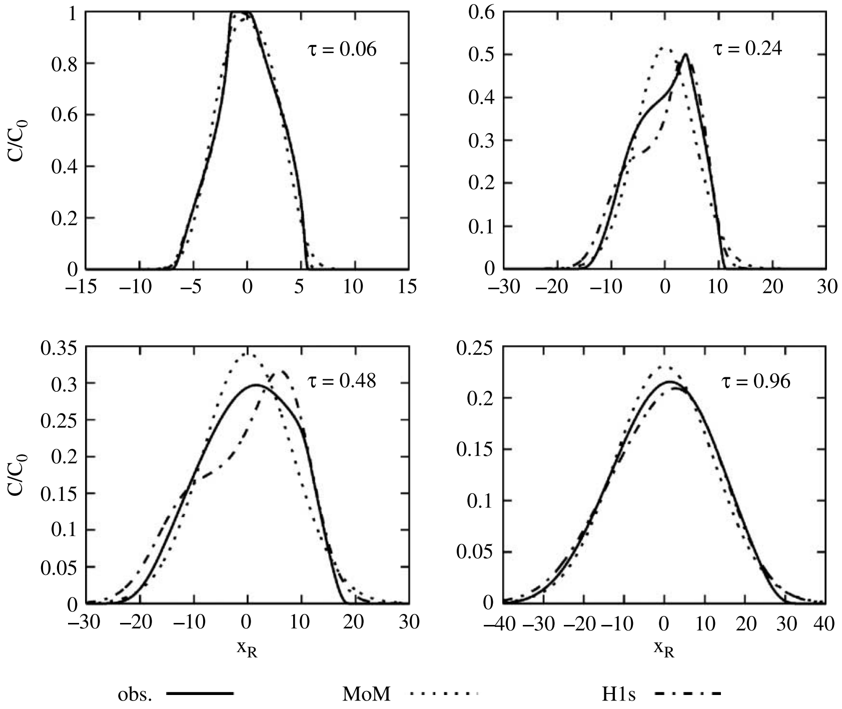

Figure 5 | Comparison of the MoM and H1s upscaled models in predicting the vertically averaged time evolution of an instantaneous finite width input of solute in the synthetic aquifer with parabolic flow.

performance of the superposed H1 model, hereafter referred to as H1s, for the case of an instantaneous input over a finite width of the aquifer with parabolic flow. In particular, model H1s more accurately predicts the shape of the plume's leading edge, as well as the peak concentration location, than the MoM model at all times. Additionally, as shown in Figure 3, the maximum absolute error between model H1s and the observed plume shape is smaller than that of the MoM model alone. Thus, though model H1 was developed for particular initial conditions, it generalizes well to other input conditions.

Furthermore, the hybrid model can be adapted for use in the aquifer with cos–cos flow simply by changing the parameters a and b of the sigmoidal mixing function to 0.5 and 0.0875, respectively. These parameters were found using another trial-and-error fit. A comparison of the performance of this adapted hybrid model, hereafter referred to as H2, with the MoM model, for predicting the time evolution of an instantaneous pulse input into the aquifer with cos–cos flow is shown in Figure 6. Even for velocity distributions on which the SGA model was not developed, the hybrid model outperforms the MoM model in predicting the plume shapes, especially in capturing the shape of the plume's leading edge, as well as in predicting the magnitude and location of the peak concentration, and in minimizing the maximum absolute error, as shown in Figure 3.

The applicability of the general hybrid model to a range of initial conditions and velocity distributions suggests that this model does not serve simply as a surrogate for the observed data used to train it, but actually describes the processes that drive the solute's macrodispersion. For that reason, it can be suggested that the behavior of the macrodispersion changes from a process that manifests itself as advective at the block scale to a process that manifests itself as Fickian at the block scale. Furthermore, since $F(t)$ had to be re-parameterized for model H2, but not model H1s, this experiment suggests that the behavior of this change is a function of the velocity field, rather than of the initial solute distribution. This knowledge could be used to develop a relationship between the mixing function parameters and the velocity distribution, so that a trial-and-error fitting procedure would no longer be necessary.

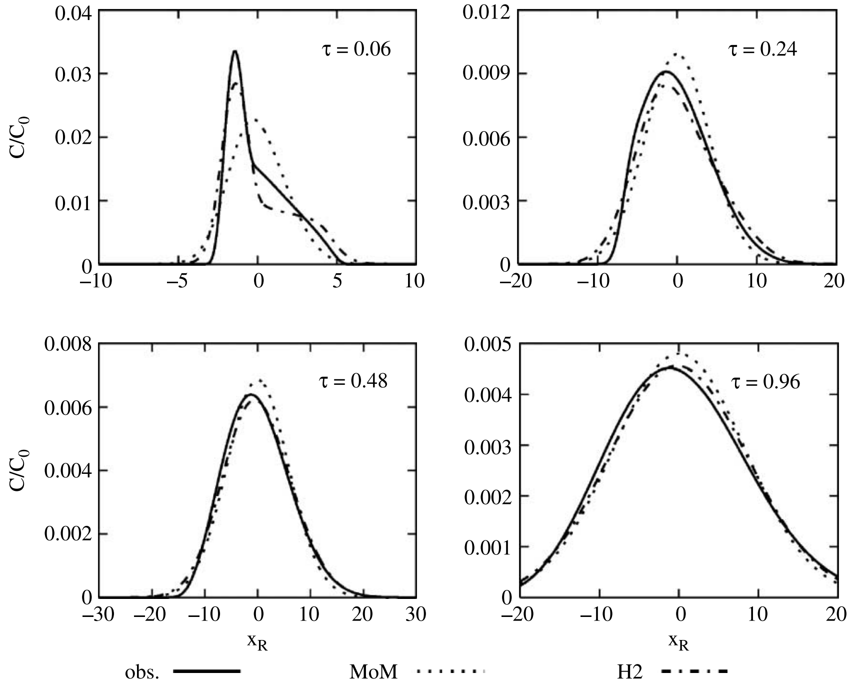

Figure 6 | Comparison of the MoM and H2 upscaled models for predicting the vertically averaged time evolution of a pulse input in the synthetic aquifer with cos–cos flow.

DISCUSSION

The case study presented in this paper illustrates several benefits of using GP as a research tool. First, GP cannot only be used to accurately model training data, but also to produce mathematical models that researchers can understand, unlike other data-driven approaches to modeling (e.g. neural networks). This representation facilitates the interpretation of model semantics, as was illustrated in the case study, when the ALP derived model was related to subgrid advection. Furthermore, the representation of models as equations renders them capable of being modified to incorporate domain knowledge to improve applicability. This is especially beneficial in the case of ill-defined modeling tasks. For example, in the case study, ALP's objective was to find a model that fitted the non-advective flux data well using the r^2 statistic. However, groundwater researchers evaluate the quality of upscaled solute transport models based on their ability to predict the time evolution of plumes. This latter objective is difficult to express mathematically and would require the solution of a PDE for each population member during each epoch. This would require both an automated numerical implementation scheme and a long time to evaluate each training epoch. Therefore, significant economy is realized by decomposing the problem and later reconstructing it from its constituents. The reconstruction process is facilitated by the mathematical form of the GP models, which allowed the combination of the SGA model with a model of Fickian macrodispersion to improve late-time performance. The mathematical representation of the GP derived models also facilitates the integration of these models into more complex modeling tasks. For example, ALP helped create a model of the block-scale non-advective flux, which could then be integrated into a PDE describing the solute plume evolution. Finally, because researchers can interpret the mathematical equations produced by GP, these equations can be used to gain insight into the predominant processes that create the training data. For example, the case study shows how the GP search encouraged the development of a conceptual model of macrodispersion that changes from a predominantly advective to a predominantly Fickian process.

The results of the case study also provide some insight into how to approach upscaling solute transport models to multidimensional blocks using GP. This task requires learning a model for a multi-dimensional vector quantity. Therefore, ensuring that mass continuity is conserved will be more difficult than in the one-dimensional case presented here, and new objectives may be necessary to guide the GP search towards methods that conserve mass. Furthermore, a method should be sought to reduce the observed bias of the r^2 metric for capturing early time behavior.

CONCLUSION

This study presents promising initial results from a novel data-driven approach to upscaling solute transport models. A methodology was developed such that the problem of upscaling models of solute transport from the fine scale to the block scale was reduced to finding a model of the block-scale non-advective flux. To demonstrate this method, a case study was performed, in which vertically averaged models were developed for the transport of solute in perfectly stratified aquifers by flow parallel to the layers. The many Pareto optimal equations found by ALP were analyzed to discover a consensus equation that described the advection of solute by fine-scale velocity variations from the vertically averaged velocity that could be expressed entirely in terms of block-scale parameters.

When this model was used to predict the time evolution of the solute distribution, the short-term predictions were of high quality, but this was not the case with the long-term predictions. This result may be due to a bias in the ALP fitness function toward capturing early time behavior. This model, however, was determined by the consensus of many searches to best capture the behavior of the non-advective flux, thus compelling the development of a new hybrid model of the non-advective flux that changed from an advective to a Fickian process. This new hybrid model was shown to be applicable to a variety of initial conditions and flow distributions, rather than merely the conditions used to train GP, suggesting that the new hybrid model describes the mechanism of macrodispersion, rather than simply being a surrogate for the training data.

Though the case study develops vertically averaged models of solute transport under relatively simple flow conditions (i.e. two-dimensional, steady state flow in a confined, perfectly stratified aquifer of infinite extent), the results presented in this study are promising. Data-driven modeling using GP is a novel approach to the upscaling problem, and to our knowledge, no previous studies exist in which data-driven modeling techniques have been used to develop semantically meaningful upscaled solute transport models. As demonstrated here, GP can be used as a tool to inspire researchers to develop novel solutions that may not be immediately obvious. The success of the hybrid model for predicting the evolution of the solute plume indicates that the GP upscaling methodology may also be successful for modeling more complex systems. Furthermore, the results of the case study provide insight into how to approach more complex transport conditions, as well as multi-dimensional blocks.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under grant no. BES 97-34076 CAR. Any opinions, findings and conclusions or recommendations expressed in this material are those of the writers and do not necessarily reflect the views of the National Science Foundation. This work was also partially supported by the University of Illinois Research Board as well as by the University of Illinois International Programs in Engineering's Office through an International Graduate Travel Fellowship. The authors would also like to thank Delft Hydraulics for hosting the first author during part of this study.

REFERENCES

Aris, R. 1956 On the dispersion of a solute in a fluid flowing through a tube. *Proc. R. Soc. Ser. A., Math. Phys. Sci.* **235** (1200), 67–77.

Babovic, V. & Abbott, M. B. 1997a The evolution of equations from hydraulic data part I: theory. *J. Hydr. Res.* **35** (3), 397–410.

Babovic, V. & Abbott, M. B. 1997b The evolution of equations from hydraulic data part II: applications. *J. Hydr. Res.* **35** (3), 411–430.

Babovic, V. & Bojkov, V. H. 2001 *Runoff Modelling with Genetic Programming and Artificial Neural Networks*. D2K Technical Report D2K TR 0401-1, Danish Hydraulic Institute – Water and Environment, Hørsholm, Denmark.

Babovic, V., Keijzer, M., Aquilera, D. & Harrington, J. 2001 An evolutionary approach to knowledge induction: genetic programming in hydraulic engineering. In *Proceedings of the World Water and Environmental Resources Congress*, doi: 10.1061/40569.

Banzhaf, W. & Langdon, W. B. 2002 Some considerations on the reason for bloat. *Genetic Programme. Evolv. Mach.* **3**, 81–91.

Banzhaf, W., Nordin, P., Keller, R. E. & Francone, F. D. 1998 *Genetic Programming – An Introduction; On the Automatic Evolution of Computer Programs and Its Applications*. Morgan Kaufmann. Heidelberg.

Beckie, R. 1998 Analysis of scale effects in large-scale solute-transport models. In *Scale Dependence and Scale Invariance in Hydrology*, (ed. G. Sposito) Cambridge University Press, Cambridge. pp. 314–334.

Beckie, R., Aldama, A. A. & Wood, E. F. 1996 Modeling the large scale dynamics of saturated groundwater flow using spatial-filtering theory: 1. Theoretical development. *Water Resour. Res.* **32** (5), 1269–1280.

Dagan, G. 1984 Solute transport in heterogeneous porous formations. *J. Fluid Mech.* **145**, 151–177.

Devore, J. L. 1995 *Probability and Statistics for Engineering and the Sciences*, 4th edn. International Thompson Publishing Company, Pacific Grove.

Duda, R. O., Hart, P. E. & Stork, D. G. 2001 *Pattern Classification*. Wiley-Interscience, New York.

Efendiev, Y., Durlofsky, L. J. & Lee, S. H. 2000 Modeling of subgrid effects in coarse-scale simulations of transport in heterogeneous porous media. *Water Resour. Res.* **36** (8), 2031–2041.

Frankel, I. & Brenner, H. 1989 On the foundations of generalized Taylor dispersion theory. *J. Fluid Mech.* **20**, 97–119.

Gelhar, L. W. & Axness, C. L. 1983 Three-dimensional stochastic analysis of macrodispersion in aquifers. *Water Resour. Res.* **19** (1), 161–180.

Gelhar, L. W., Gutjahr, A. L. & Naff, R. L. 1979 Stochastic analysis of macrodispersion in a stratified aquifer. *Water Resour. Res.* **15** (6), 1387–1397.

Güven, O., Molz, F. J. & Melville, J. G. 1984 An analysis of dispersion in a stratified aquifer. *Water Resour. Res.* **20** (10), 1337–1354.

Keijzer, M. 2002 *Scientific Discovery using Genetic Programming*. PhD thesis, Danish Technical University, Lyngby, Denmark, March.

Keijzer, M. & Babovic, V. 1999 Dimensionally aware genetic programming. In *Proceedings of the Genetic and Evolutionary Computation Conference, Orlando* (ed. W. Banzhaf, J. Daida, A. E. Eiben, M. H. Garzon, V. Honavar, M. Jakiela & R. E. Smith), Morgan Kaufmann, San Mateo, CA. pp. 42–49.

Keijzer, M., Babovic, V., Ryan, C., O'Neill, M. & Cattolico, M. 2001 Adaptive logic programming. In *Proceedings of the Genetic and Evolutionary Computation Conference, San Francisco, July* (ed. L. Spector, E. D. Goodman, A. Wu, W. B. Langdon, H. -M. Voigt, M. Gen, S. Sen, M. Dorigo, S. Pezeshk, M. H. Garzon & E. Burke), Morgan Kaufmann, San Mateo, CA. pp. 42–49.

Keijzer, M. & Cattolico, M. 2002 An example of the use of context-sensitive constraints in the ALP system. In *GECCO 2002: Proceedings of the Bird of a Feather Workshops, Genetic and Evolutionary Computation Conference, New York, July* (ed. A. M. Barry), pp. 128–132.

Kitanidis, P. K. 1992 Analysis of macrodispersion through volume-averaging moment equations. *Stochast. Hydrol. Hydraul.* **6**, 5–25.

Koza, J. R. 1992 *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press. Cambridge, MA.

Koza, J. R., Bennett, F. H., Andre, D. & Keane, M. A. 1999 *Genetic Programming III: Darwinian Invention and Problem Solving*. Morgan Kaufmann. San Mateo, CA.

Langdon, W. B. & Poli, R. 2002 *Foundations of Genetic Programming*. Springer-Verlag. Berlin.

LeBlanc, D. R., Garabedial, S. P., Hess, K. M., Gelhar, L. W., Wuadri, R. D., Stollenwerk, K. G. & Wood, W. W. 1991 Large-scale natural gradient tracer test in sand and gravel, Cape Cod, Massachusetts 1: experimental design and observed tracer movement. *Water Resour. Res.* **27** (5), 895–910.

Leonard, B. P. 1988 The ULTIMATE conservative difference scheme applied to unsteady one-dimensional advection. *Comput. Methods Appl. Mech. Eng.* **88**, 17–74.

Mackay, D. M., Freyberg, D. L., Robberts, P. V. & Cherry, J. A. 1986 Natural gradient experiment on solute transport in a sand aquifer I: approach and overview of plume movement. *Water Resour. Res.* **22** (13), 2017–2029.

Mei, C. C. 1992 Method of homogenization applied to dispersion in porous media. *Transport Porous Med.* **9**, 261–274.

Nitsche, L. C. & Brenner, H. 1989 Eulerian kinematics of flow through spatially periodic models of porous media. *Arch. Rational Mech. Anal.* **107** (3), 225–292.

Rubin, Y. 2003 *Applied Stochastic Hydrogeology*. Oxford University Press. Oxford.

Rubin, Y., Bellin, A. & Lawrence, A. E. 2003 On the use of block-effective macrodispersion for numerical simulations of transport in heterogeneous formations. *Water Resour Res.* **39** (9), 1242–1252.

Sastry, K., O'Reilly, U.-M., Goldberg, D. E. & Hill, D. 2003 Building-block supply in genetic programming. In *Genetic Programming in Theory and Practice* (ed. R. Riolo & B. Worzel), Kluwer, Dordrecht. pp. 137–152.

Savic, D., Walters, G. A. & Davidson, J. W. 1999 A genetic programming approach to rainfall-runoff modelling. *Water Resour. Manag.* **13**, 219–231.

Sposito, G. 1997 Ergodicity and the 'scale effect'. *Adv. Water Resour.* **20** (5–6), 309–316.

Sudicky, E. A. 1986 Natural gradient tracer experiment on solute transport in a sand aquifer: spatial variability of hydraulic conductivity and its role in the dispersion process. *Water Resour. Res.* **22** (13), 2069–2082.

Taylor, G. 1953 Dispersion of soluble matter in solvent flowing slowly through a tube. *Proc. R. Soc. Ser. A: Math. Phys. Sci.* **219** (1137), 186–203.

Whitaker, S. 1999 *The Method of Volume Averaging*. Kluwer, Dordrecht.

Wood, B. D., Cherblanc, F., Quintard, M. & Whitaker, S. 2003 Volume averaging for determining the effective dispersion tensor: closure using periodic unit cells and comparison with ensemble averaging. *Water Resour. Res.* **39** (8), 1210.