International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

An Extensible Inverse Modeling Software Architecture for Parameter Distribution Estimation

Carlos Osorio[1], Matthew Over[2], Daniel P. Ames[1], Yoram Rubin[2]
[1] *Idaho State University - Geospatial Software Lab, Department of Geosciences*
Idaho Falls, Idaho, United States
[2] *University of California Berkeley, Department of Civil and Environmental*
Engineering, Berkeley, California, United States
osorcar2@isu.edu

Abstract: Inverse modeling can be a useful – though non-trivial to execute – technique for estimating unknown spatial parameter fields (e.g. hydraulic conductivity) necessary for understanding and modeling groundwater, subsurface contaminant movement, and related quantities. This paper presents the design, implementation, and a test case of an extensible software architecture intended to simplify application of inverse modeling techniques by integrating the Method of Anchored Distributions (MAD), the CUAHSI Hydrologic Information System (HIS) HydroDesktop tool, and the R statistical software package. The test case presents the inversion of model parameters related to the log-transformed hydraulic conductivity, conditional on steady-state pressure head measurements from wells in the domain. The domain is fully saturated and of unit thickness, and is bounded by no flow conditions on two sides and constant head conditions otherwise. HydroDesktop's GIS-based plugin architecture and inherent access to a large database of hydrologic data serves as the core software framework with a MAD and R integrated custom plugin. MAD is a Bayesian inversion technique for conditioning computational model parameters on relevant field observations yielding probabilistic distributions of the model parameters, related to the spatial random variable of interest, by assimilating multi-type and multi-scale data. The implementation of a desktop software tool for using the MAD technique is expected to significantly lower the barrier to using inverse modeling in education, research, and resource management. The HydroDesktop MAD plugin is being developed following a community-based, open-source approach that will help both its adoption and long term sustainability as a user tool. This presentation will briefly introduce MAD, HydroDesktop, and the MAD plugin and software development effort.

***Keywords**: Inverse modeling, hydrologic information systems, R, Bayesian inversion*
.

1 INTRODUCTION

Inversion modeling is a tool for calibration of model parameters that has applications in many of the physical sciences and engineering fields [Iden & Durner, 2007]. In groundwater applications, inversion modeling can be used to estimate parameters that describe characteristics of an aquifer. Fifteen years ago, researchers posited that the field of hydrogeology should implement inversion modeling as a standard practice [Poeter & Hill, 1997], similar to the push by the USGS in the 1970s that motivates the use of numerical forward modeling tools. But several functional obstructions have traditionally prevented widespread use of inversion models [Carrera et al., 2005]. It has been suggested that the following five issues needed to be addressed to encourage widespread use of inversion techniques: 1) incorporation of geological data; 2) improving flexibility of the code and procedures to handle any and all relevant data types; 3) accommodation of uncertainties; 4) reducing difficulty of code operation; and 5) coupling of techniques with a geographic information system (GIS) platform [Carrera et al., 2005]. The Method of Anchored Distributions (MAD) modeling technique and the associated MAD software plugin for the HydroDesktop platform addresses these shortcomings.

MAD is an inversion modeling technique [Rubin et al., 2010] that addresses the first three issues on Carrera's list; in the opinion of the authors, these issues are fundamentally accounted for by a more robust inversion approach. The incorporation of geological data is listed separately from inclusion of other data types to highlight the importance of geological processes that are often unaccounted for by zonation techniques [Carrera et al., 2005]. To this end, MAD is designed with a geostatistical framework for the Bayesian inversion of model parameters using a variety of available models, as Carrera suggests as an alternative to a zonation framework. Furthermore, MAD is a technique capable of assimilating data of any type and on any scale of the inversion problem [Rubin et al., 2010], thus making it appropriately flexible for any hydrogeological application. Finally, MAD is a Bayesian tool that can treat any parameter of interest as a random variable and hence account for uncertainty throughout the inversion process.

The last two items on Carrera's list are accounted for in the development of easy to use software that lowers the bar for stochastic inversion modeling and is constructed on the HydroDesktop GIS platform. The software is being developed with a GUI that guides users through formulating the inversion problem and manages appropriate data throughout the process. The need for a hydrologist to fully understand the statistics of an approach and compile large amounts of simulated data traditionally is requirements of an inversion model application. But as a default, the GUI automates the statistics and manages the data, allowing users to focus on specifying a better forward model or the implications of the predictions of the calibrated model, instead of the minutia of the inversion methodology. Lastly, HydroDesktop is an extensively used platform, familiar to many hydrologists as a search and discovery tool for hydrologic information [Piasecki et al., 2010], with a catalog of available GIS tools. The GIS characteristics of HydroDesktop allow multiple problem formulations and efficient association of geographic data with inverse models.

The objective of this paper is to present the software architecture of a plug-in to the existing HydroDesktop application, further reducing limitations to the widespread use of inversion modeling in hydrogeologic investigations. To demonstrate the usability of the plug-in, introduce the GUI, and briefly summarize the MAD methodology, a simple inversion modeling case study is performed using steady state pressure head data determined for a two-dimensional synthetic flow problem on an isotropic, heterogeneous hydraulic conductivity field.

2. METHODS

MAD is a Bayesian approach to inversion modeling that relies on the following proportionality,

$$f(\boldsymbol{\vartheta}|\boldsymbol{z}_a, \boldsymbol{z}_b) \propto f(\boldsymbol{z}_b|\boldsymbol{\vartheta}, \boldsymbol{z}_a) f(\boldsymbol{\vartheta}|\boldsymbol{z}_a) \qquad [1]$$

Where $f(\boldsymbol{\vartheta}|\boldsymbol{z}_a)$ is the joint prior distribution of the anchors conditional on Type A data vector $\boldsymbol{z}_a$ only; $f(\boldsymbol{z}_b|\boldsymbol{\vartheta}, \boldsymbol{z}_a)$ is the likelihood of observing the Type B data vector $\boldsymbol{z}_b$ given the anchors and Type A data; and $f(\boldsymbol{\vartheta}|\boldsymbol{z}_a, \boldsymbol{z}_b)$ is the joint posterior distribution of the anchors conditional on both Type A and Type B data. Two key elements of MAD are presented in Equation (1): the first is anchors $\boldsymbol{\vartheta}$, which are statistical devices placed in the model domain that capture information about the target variable of inversion, and the second is the categorical approach to data classification as either Type A or as Type B. Type A data is directly related to the target variable of inversion, and Type B data is related to the target variable through some function, either numeric or analytic [Rubin et al., 2010]. Note that Equation (1) is a reduced version of Bayesian methodology core to MAD [Rubin et al., 2010], but is appropriate for determining the posterior model parameters of the simple case study presented in Section 3.

2.1 MAD Architecture

The MAD technique is applied through three modular structures called blocks [Rubin et al., 2010]. Block I is the preprocessing module, which works primarily with Type A data and defines Type B data. Block I functionality is focused on obtaining and classifying types of data and defining parameters for the conditional field realizations. Block II is an external module that calculates the forward model simulations of the physical process being studied on the conditional field realizations of the spatial variable of interest generated in Block I. Block III evaluates the likelihood function by comparing the Type B data measurements to their simulated counterparts obtained in Block II. This is a prediction block that requires the selection of the probability density estimation method, the number of Type B measurements, and the variable output to be visualized.

The MAD technique described here is well-suited to implementation within hydrological software with GIS and statistical capabilities. For the purposes of this paper, HydroDesktop serves as such an application, offering the required capabilities in an open source software development environment with support for extensions and plugins.

2.2 HydroDesktop and Extensibility Support

HydroDesktop is an open source installable desktop software application developed using the C# programming language as a client tool for the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) Hydrologic Information System (HIS). Its primary purpose is to support discovery, download, display, editing, and analysis of hydrologic and climate data. It uses the WaterOneFlow Web service application programming interface (API) for communicating with the HIS servers to retrieve data and metadata, and to search the HIS Central metadata catalog which provides the ontology keywords/concepts about the WaterOneFlow web services. The modular structure of HydroDesktop allows for integrating different functions through a plugin interface that creates a barrier between core functionality and extended functionality and allows for multiple developers in a global open source software ecosystem to create tools and plugins independently. HydroDesktop builds on the open source DotSpatial (www.dotspatial.org) GIS programming library and supports external connections to web services as well as locally installed analytical tools (R, MatLab, Excel, etc.) (Figure 1). Retrieved HIS data is stored in a SQLite database to be processed and visualized. This application allows visualizing, summarizing, and converting data units and formats.

The DotSpatial open source GIS library provides all of the geographic and spatial data analysis and display functionalities within HydroDesktop. The library is divided in several individual packages that support various GIS functions including: data, control, projection, analysis, symbology and extension package.

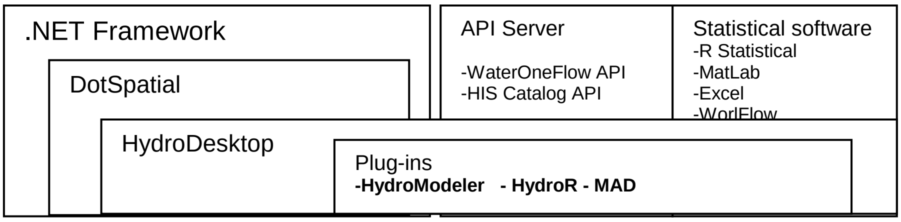


Figure 1 HydroDesktop Architecture

Its architecture allows the creation of new GIS applications and the implementation of new GIS functions using an extensible plug-in system. In brief, a new plug-in extends an *Extension* class which imports the features of the main program through of *AppManager* class. This structure provides all mechanisms for controlling the entire environment of HydroDesktop (Figure 2).

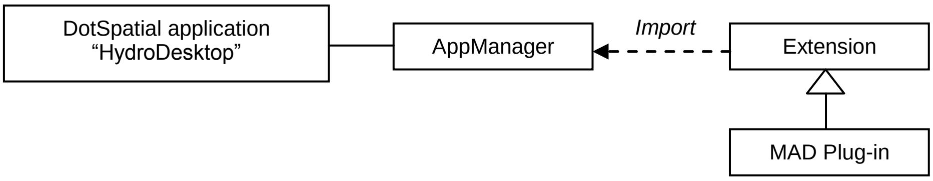


Figure 2 DotSpatial extension profile.

2.3 Application Design

The MAD plugin user interface allows one to specify each of the requisite parameters of MAD in a logical "wizard-style" manner. In Block I, the domain area definition, which is essential in the MAD procedure, adapts to spatial distribution of measurements and allows configuring the number of columns, rows, orientation, and cell size. Another screen allows for the management of measurement data (i.e. Type A) as well as the definition of Type B data. Another screen allows for the management of any measurement data. Here the user can specify variable names and types (e.g. hydraulic conductivity as Type A and pressure head as Type B) as well as their values at specified locations in the domain. So called anchor points are placed in the domain area using the HydroDesktop GIS capabilities either manually or automatically, using a grid or random option. Next, the user must specify the geostatistical mode and its parameters (mean, variance, length scale, etc.) from which realizations of field data will be generated. The application supports both isotropic and anisotropic models. Finally, the number of realizations and samples are be defined for executing the simulation.

Block II calculates the forward model simulated output dependent on the conditional realizations of the spatial random field. The MAD technique does not establish a specific forward model. The application accepts the result of other hydrologic modeling software for this process. Block III requires joining the results of the forward models and the selection of Type B variable(s) that will be used in the parametric or non-parametric likelihood calculation. The output of the MAD plug-in application is the determination of the posterior conditional distributions of the geostatistical parameters, anchors, and the target variable in the domain area.

2.4 Implementation Process

The MAD plug-in software was implemented in three parts, following the steps of the MAD analytical procedure. A Windows form containing parameter options and relevant GIS tools was created for each MAD block. Block I creates the domain area through a *DomainArea* class object which contains all geographic information and the description of the study area. The vector grid generated in Block I uses the information from the *DomainArea* class. The measurements and anchors, which are a type of *DataEvent* Class, depend directly on definition of the domain area. When the domain changes, the *DataEvent* instances update the relative position in the grid and synchronize the information with the R statistical software. The simulation is executed within R. Each step in Block I generates its correspondent R instruction codes using the *Rcode* class. The *Rcode* class connects the MAD plug-in with R, calling the functions from the R executable file.

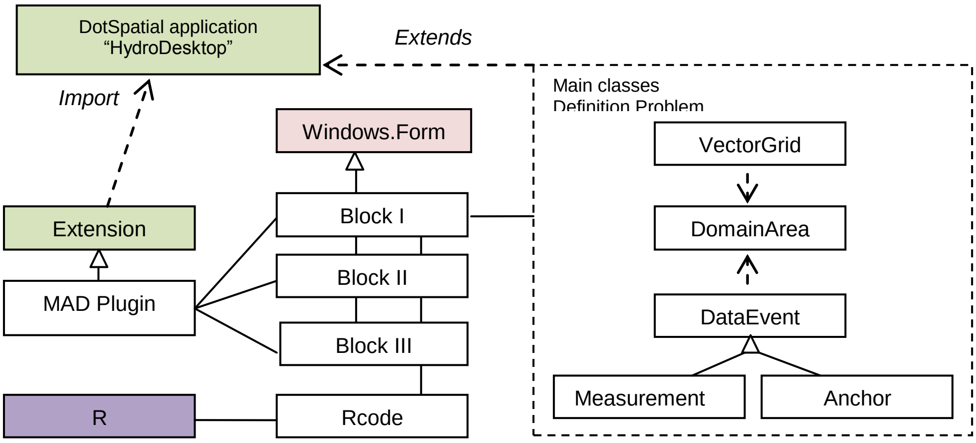
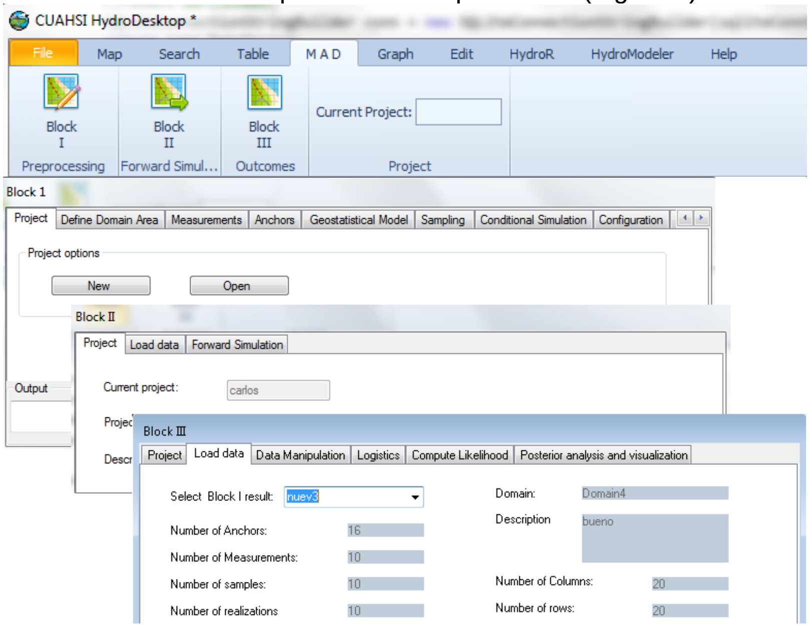


Figure 3 Main UML classes diagram of MAD Plug-in.

Block II retrieves information from Block I and executes the forward model. For the purposes of this case study, the algorithm for the forward process is stored in the *Rcode* class. However, this procedure can be done with other hydrologic modelling programs. Block III generates the R instructions for estimating likelihoods and visualizes the final distributions of all parameters. Additionally, the plug-in creates a workspace with all information produced by a project. The MAD plug-in adds a menu with the three steps of the MAD procedure (Figure 4).



Current Project:

Current project: carlos

Select Block I result: nuev3

Number of Anchors: 16

Number of Measurements: 10

Number of samples: 10

Number of realizations 10

Domain: Domain4

Description bueno

Number of Columns: 20

Number of rows: 20

Figure 4 MAD plug-in.

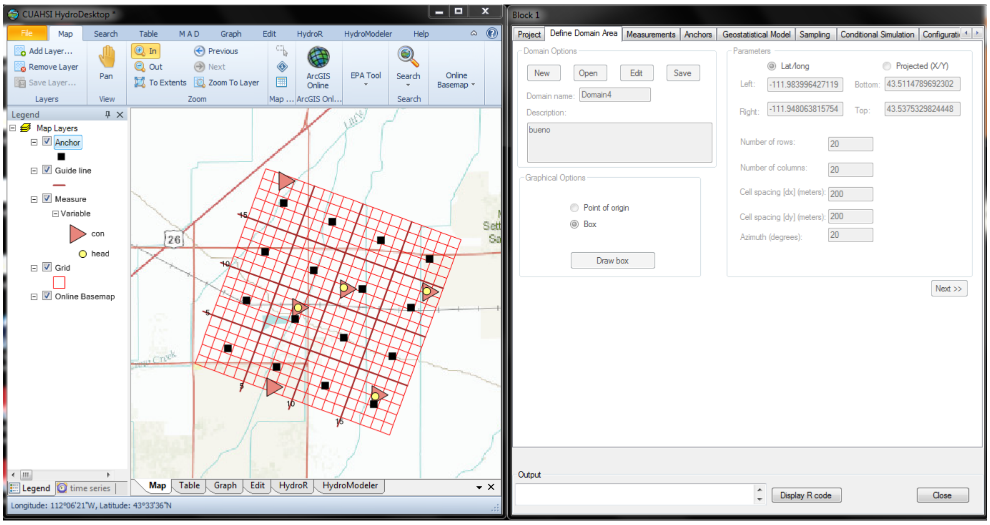
3 RESULTS

3.1 Demonstration of the MAD plugin

The following simple case study presents an application of Equation (1) used to determine the joint posterior anchor distribution $f(\vartheta|z_a, z_b)$ conditional on log-transformed conductivity measurements (Type A) and steady-state pressure head measurements (Type B). Estimating the joint posterior anchor distribution is a two step procedure. First, samples are drawn from the joint prior distribution of the anchors$f(\boldsymbol{\vartheta}|\boldsymbol{z_a})$, which are used to generate ensembles of conditional realizations of the target variable field. Secondly, the likelihood function $f(\boldsymbol{z_b}|\boldsymbol{\vartheta},\boldsymbol{z_a})$ is estimated parametrically or non-parametrically by solving the numeric or analytic equations describing the Type B variables over the ensemble of conditional realizations, and comparing those solutions with measurements of the same Type B variables collected on site. The goal of this analysis – performed entirely in the R statistics software via the MAD plug-in – is to derive probabilistic distributions of the log transform of the hydraulic conductivity at various locations in the domain where measurement data is unavailable.

To begin the MAD analysis, the MAD plug-in user interface guides a user through several data input and definition stages that are necessary to formulate the problem. The information required to proceed with drawing samples from the prior distribution includes: a modeling coordinate system; the spatial locations and values of the Type A data to be conditioned on measured *in situ*, if any; the spatial locations of the anchors; and a model linking the Type A variable to the target variable.

In the example problem, the modeling domain is a 20 by 20 grid with a cell size of 200 by 200 meters and is specified by a point of origin, which can be located through a variety of available base maps in HydroDesktop. The Type A measurements are collected at 6 wells in the domain and are assumed to be error-free, but generally MAD can account for measurement error. The GUI adopts a general approach to variable definition that allows users to name and classify variables that are used in the inversion; in this case, log conductivity was defined as a Type A variable and saved as a double. Then 16 anchors are placed in the domain to capture local, sub-domain information. Figure 4 shows a screenshot after the data input process is complete, with 16 squares depicting anchor locations and 6 triangles depicting measurement locations.



Longitude: 112°06'21"W, Latitude: 43°33'36"N

Domain name: Domain4

Description: bueno

○ Point of origin
● Box

● Lat/long ○ Projected (X/Y)

Left: -111.983996427119 Bottom: 43.5114789692302

Right: -111.948063815754 Top: 43.5375329824448

Number of rows: 20

Number of columns: 20

Cell spacing [dx] (meters): 200

Cell spacing [dy] (meters): 200

Azimuth (degrees): 20

Figure 4 Screenshot of definition study area

Determination of the model linking Type A data with the anchors is finalized through the following assumption: the target variable field is an isotropic Gaussian random field with an exponential covariance model with error-free mean, variance, and integral scale; note that MAD can generally account for uncertainty in any of the covariance model parameters, can handle anisotropy, and can work with non-Gaussian fields. The assumption of Gaussianity and covariance model means the joint prior distribution of the anchors conditional on Type A data $f(\boldsymbol{\vartheta}|\boldsymbol{z_a})$, is multivariate normal and determined via simple kriging.

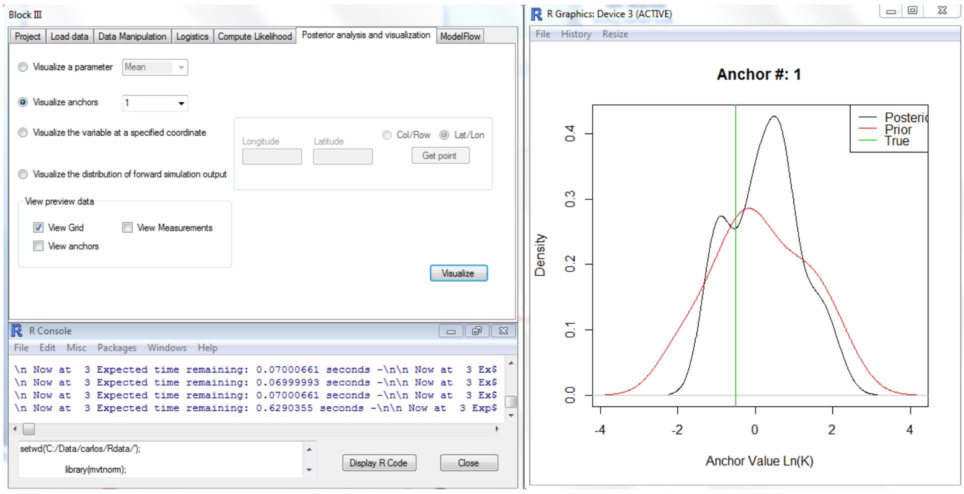
After the joint prior distribution of the anchors is defined, there is an intermediate data generation stage. In the example problem, samples are drawn from the joint prior distribution of the anchors; the 16 samples at the anchors, in conjunction with the 6 Type A measurements, are all used as conditioning points in the target variable field. Owing to the inherent nature of random fields, there are an infinite number of conditional fields that obey a given covariance model and its parameters, so each sample from the anchor distribution is used to create an ensemble of conditional target variable field realizations. In this example, 6 samples were drawn from the prior distribution of the anchors, conditional on Type A data only, and 100 realizations of the target variable were generated for each sample.

Before the likelihood function can be computed, the analytical or numerical equations for the Type B variables on the target variable field must be solved for every realization for every sample, which is often the most time-consuming step of an inversion. In the example problem, the numerical equation that relates the target variable field to the Type B data is the 2-D steady state head equation, with the elevation head subtracted from the solution. Equation 2 details the 2-D steady state head equation and the necessary boundary conditions in the local coordinate system,

$$0 = \frac{\partial}{\partial x}\left(K(x,y)\frac{\partial h(x,y)}{\partial x}\right) + \frac{\partial}{\partial y}\left(K(x,y)\frac{\partial h(x,y)}{\partial y}\right)$$

$$\frac{\partial h(x,0)}{\partial y} = 0, \qquad \frac{\partial h(x,4000)}{\partial y} = 0, \qquad h(0,y) = 1, \qquad h(4000,y) = 0 \qquad [2]$$

After the ensembles of pressure head simulations are compiled, the likelihood function of observing the Type B data given the anchors and Type A data $f(\boldsymbol{z}_b|\boldsymbol{\vartheta},\boldsymbol{z}_a)$ can be estimated by fitting the joint distribution and evaluation of the Type B field measurements. In the example problem, the likelihood function is computed non-parametrically using a Gaussian kernel density approach. Figure 5 demonstrates the difference between the prior distributions of the anchors conditional on Type A data only and the posterior distributions of the anchors conditional on Type A and Type B data.



○ Visualize a parameter: Mean

● Visualize anchors: 1

○ Visualize the variable at a specified coordinate ○ Col/Row ● Lat/Lon

○ Visualize the distribution of forward simulation output

☑ View Grid ☐ View Measurements

☐ View anchors

```
\n Now at  3 Expected time remaining: 0.07000661 seconds -\n\n Now at  3 Ex$
\n Now at  3 Expected time remaining: 0.06999993 seconds -\n\n Now at  3 Ex$
\n Now at  3 Expected time remaining: 0.07000661 seconds -\n\n Now at  3 Ex$
\n Now at  3 Expected time remaining: 0.6290355 seconds -\n\n Now at  3 Exp$
```

```
setwd('C:/Data/carlos/Rdata/');
        library(mvtnorm);
```

Figure 5 Output of posterior versus prior distribution in an anchor

Finally, the plug-in contains an interface for estimating the distribution of the parameters involved in the inversion, and the distribution of the target variable in any place in the study area.

4 CONCLUSIONS AND DISCUSSION

HydroDesktop offers an extensible architecture for developing modelling applications in a GIS environment. Its integration with the R statistics software creates a robust analysis platform for hydrological projects. These characteristics supported the implementation of the MAD technique, which requires several inputs and the transformation of geographical features to model parameters. The MAD plug-in presented here captures all MAD requirements in a generic interface that can be applied in different hydrologic problems. Some achievements of this extensible application are as follows:

- The MAD plug-in combines the flexibility of R scripting with a friendly GUI interface for estimating the parameter distribution of inverse modelling problems.
- The parameters and inputs in the MAD plug-in are managed in a workspace that allows users to compare results.
- The plug-in accepts and returns data in standard GIS format such as shapefile, Grid ASCII.
- Uncertainty estimation can be obtained anywhere in study area using graphical interface.

Prospective improvements:

- The generation of fields in Block I and the execution of forward models in Block II are parts of the MAD technique that consume significant CPU time and could be managed by a parallelization technique. The reduction of time-consumption will be studied for the next version of MAD plug-in using a high performance computing (HPC) approach.
- The forward model process in the current version is executed within the R statistics software. In the future, the MAD plug-in will support external forward modelling applications in Block II using, for example, MODFLOW.

References

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

Iden, S. C., and W. Durner (2007), Free-form estimation of the unsaturated soil hydraulic properties by inverse modeling using global optimization, Water Resources Research., 43, W07451, doi:10.1029/2006WR005845.

Piasecki, M., D. P. Ames, J. Goodall, J. Horsburgh, D. Maidement, D. Tarboton, I. Zalavsky (2010) Development of an information system for hydrologic community, 9th International Conference on Hydroinformatics, Tianjin, China.

Poeter, EP and M.C. Hill, (1997), Inverse Methods: A Necessary Next Step in Groundwater Modeling, Ground Water, v. 35, no. 2, pp. 250-260.

Rubin, Y., X. Chen, H. Murakami, M. Hahn (2010) A Bayesian approach for inverse modeling, data assimilation, and conditional simulation of spatial random fields. WATER RESOURCES RESEARCH, VOL. 46, W10523, doi:10.1029/2009WR008799.