# OPTIMAL MONITORING NETWORK DESIGN FOR EFFICIENT IDENTIFICATION OF UNKNOWN GROUNDWATER POLLUTION SOURCES

Om Prakash[1, 2] and Bithin Datta[1, 2]

[1] Discipline of Civil and Environmental Engineering, School of Engineering and Physical Sciences, James Cook University, Townsville QLD 4811, AUSTRALIA.

[2] CRC for Contamination Assessment and Remediation of the Environment, Mawson Lakes SA 5095, AUSTRALIA.

ABSTRACT: Application of linked simulation-optimization approach for solving groundwater identification problems is well established. Pollutant concentration measurements from different sets of monitoring locations, when used in a linked simulation-optimization approach, results in different degrees of accuracy of source identification. Moreover, the accuracy of source identification results depends on the number and spatiotemporal locations of pollutant concentrations measurements. This study aims at improving the accuracy of source identification results, by using concentration measurements from an optimally designed monitoring network. A linked simulation optimization based methodology is used for optimal source identification. Genetic programming based impact factor is used for designing the optimal monitoring network. Concentration measurement data from the designed network is then used, in the Simulated Annealing based linked simulation-optimization model for efficient source identification. The potential application of the developed methodology is demonstrated by evaluating its performance for an illustrative study area. These performance evaluation results show improvement in the efficiency in source identification when such designed monitoring networks are utilized.

*Keywords: Optimal Monitoring Network; Groundwater Pollution; Genetic Programming; Multi-Objective Optimization; Pollution Source Identification; Simulated Annealing*

## INTRODUCTION

The most common problem encountered in remediation of a polluted aquifer, is the accurate identification of pollution source locations and their release histories from sparse set of spatiotemporal concentration measurements. In scenarios where potential source locations and activity duration are known with fair degree of certainty, linked simulation-optimization based approach is often applied for solving groundwater pollution source identification problem for recreating the flux release history of the sources.

A large number of concentration measurements spread over time and space is necessary for accurate identification of pollution sources. However, long term monitoring over a large number of monitoring locations has budgetary constraints. Hence monitoring locations should be chosen such that concentration measurements from such designed monitoring network when used in a linked simulation-optimization approach, improves the accuracy of source identification results. Contrary to this, monitoring locations often consists of arbitrarily located single water supply well, or a group of arbitrarily placed wells which may not be optimally located for accurate source identification.

Design of monitoring networks may have different underlying objectives, such as optimal monitoring network for efficient identification of pollution sources [1], [2] and pollution source locations [3]. Optimization based solution for reducing the redundancy in a groundwater quality monitoring network [4], sequential monitoring network design [5], and variation of dynamic monitoring network design methodology is discussed in [6].

Most of the existing methodologies of source identification using linked simulation-optimization use concentration measurements from arbitrarily located monitoring wells. Limited amount of work has been reported for improving the accuracy of source identification using concentration measurements from an optimally designed monitoring network. A methodology is proposed for design of optimal monitoring network for improving the accuracy of source identification results using concentration measurements from a designed monitoring network. The simulated response of the aquifer is compared to the observed response of the aquifer at these optimal monitoring locations in a classical linked simulation-optimization technique.

The monitoring network designed to improve the accuracy of source identification is based on two conflicting objectives, (1) reduce the possibility of missing an actual source, and (2) decreases the degree of non-uniqueness in the set of possible aquifer responses to subjected geo-chemical stresses. The proposed methodology uses genetic programming (GP) models to calculate the impact factor of a source on a candidate monitoring location. This impact factor is utilized as the design criteria for choosing the optimal monitoring locations out of the candidate well locations. The conflicting objectives considered are, (1) maximization of the normalized impact of all potential pollution sources at selected monitoring locations and (2) to maximize the relative impact of a potential source at the chosen monitoring location. The Pareto-optimal solutions obtained from the two objective model is utilized to relate the variation in the accuracy of source identification results and the trade-off between the above stated conflicting objectives for designing an optimal monitoring network. Performance of the proposed source identification methodology is evaluated by solving an illustrative problem. The results of source identification are compared for different random monitoring networks with that of the optimal network. This method can be applied to practical scenarios where the observed concentration data is to be restricted to a few monitoring locations.

## METHODOLOGY

In this proposed methodology, first GP models are trained against a large set of data pattern comprising of source flux history for all the potential sources as input and corresponding aquifer response at all potential monitoring locations as the output. Based on $R^2$ value specified, a number of fittest GP models are chosen for calculating the impact factor of potential source on a monitoring location. The impact factor is calculated for all candidate monitoring locations at each monitoring time step. A multi–objective optimization formulation is applied to select the optimal set of monitoring location for source identification using linked simulation-optimization technique. In the second step, a SA based linked simulation-optimization model for source identification is solved to minimize the deviation between the simulated and measured pollutant concentrations at these optimally chosen monitoring locations.

### Genetic Programming and Impact Factor

GP is an evolutionary optimization algorithm based on the concepts of genetics and natural selection. It starts with an initial population of randomly generated computer programs and optimizes the parameter values of a given model structure within predefined parameter space to find a highly fit computer program that produces desired output for particular set of input. Each GP model is essentially a computer program that represents the mathematical relationship between the dependent variable (pollutant concentration at chosen monitoring locations and times) and the independent variables (flux values of pollutant at potential pollutant source locations).

Specified number of best individual GP models is used for computing the impact factor. The impact factor is described as a measure of how much an input variable accounts for the output result i.e., a factor by which the result would differ if the variable was removed. This essentially implies that, if by removing a variable from the mathematical function (GP model) there is a large change in the output, then the removed variable has a high impact on the output and hence the impact factor of that variable will be high. The impact factor value is then used as design criteria in a multi-objective optimization formulation for designing an optimal monitoring network.

### Multi-Objective Optimization for Monitoring Network Design Model

An SA based multi-objective optimization model is utilized for choosing monitoring locations such that the possibility of missing an actual source is reduced, and at the same time reduces the degree of non-uniqueness due to overlapping of multiple pollutant plumes. This is accomplished by an SA based multi-objective optimization model that finds monitoring well locations with the following objectives (1) finding well locations with maximum normalized impact from all the potential sources and (2) finding well locations with maximum normalized relative impact from an individual potential source over a chosen observation period. Finding well locations with maximum normalized average impact (objective 1) reduces the possibility of missing an actual source as it chooses locations where overlapping of plumes due to potential sources is maximum. This also reduces the likelihood of choosing monitoring locations where the impact of potential sources are small. However objective 1 is in conflict with objective 2 of finding well locations with maximum normalized relative impact from an individual potential source. The second objective essentially reduces the non-uniqueness due to overlapping of different pollutant plumes resulting from different sources.

A multi-objective optimization model is formulated to design an optimal monitoring network with multiple conflicting objectives. One of the objectives is traded off to improve the other objective and vice versa. The constrained method is utilized, which iteratively maximizes one of the objectives subject to the other objective achieved at a specified level. The number of monitoring wells to be selected is essentially governed by budgetary constraints. The formulation for the multi-objective optimization for monitoring network design is given in Eq. (1) through Eq. (11).

$$IF_{iob}^{S} = \sum_{t=1}^{nt} (F_{iob}^{St}) \qquad (1)$$

$IF^{S}_{iob}$ is the impact factor of source *S* on monitoring well location *iob*

$F^{St}_{iob}$ is the impact factor of source *S* on monitoring well location *iob* at stress period *t*

$$SumIF^{S}_{iob} = \sum_{k=1}^{nk} (IF^{S}_{iob})^{k} \qquad (2)$$

$SumIF^{S}_{iob}$ is the sum of the impact factors of a potential source *S* at any given monitoring location *iob* for *nk* sampling steps

$$SumIF^{norm}_{iob} = \sum_{S=1}^{nS} \frac{SumIF^{S}_{iob}}{\frac{1}{nob}\sum_{iob=1}^{nob} SumIF^{S}_{iob}} \qquad (3)$$

$SumIF^{norm}_{iob}$ is the normalised sum of impact factor at any monitoring location *iob* due to all the potential sources *nS* for all *nk* monitoring time steps

*nt* is the total number of stress periods

*nk* is the total number of monitoring time steps

$\frac{1}{nob}\sum_{iob=1}^{nob} SumIF^{S}_{iob}$ is the average impact factor due to a source *S* at all monitoring well locations *nob*

$$^{\mathrm{Re}\,l}SumIF_{iob} = Max\{SumIF^{S}_{iob}\} - ((\sum_{S=1}^{nS}(SumIF^{S}_{iob})) - Max\{SumIF^{S}_{iob}\}) \qquad (4)$$

$^{\mathrm{Re}\,l}SumIF_{iob}$ is the relative impact factor due to all the sources *nS* at a given monitoring well location *iob*

$$^{\mathrm{Re}\,l}SumIF^{norm}_{iob} = \frac{^{\mathrm{Re}\,l}SumIF_{iob}}{\frac{1}{nS}\sum_{S=1}^{nS} SumIF^{S}_{iob}} \qquad (5)$$

$^{\mathrm{Re}\,l}SumIF^{norm}_{iob}$ is the normalized relative impact factor at monitoring well location *iob* for all potential sources

$\frac{1}{nS}\sum_{S=1}^{nS} SumIF^{S}_{iob}$ is the average impact factor at monitoring well location *iob* for all potential sources

The two objectives *F1* and *F2* of the multi-objective optimization model for optimal monitoring network design for accurate identification of unknown pollution sources is defined by Eq. (6) and Eq. (8) respectively.

$$MaximizeF1 = \sum_{iob=1}^{nob} SumIF^{norm}_{iob} f_{iob} \qquad (6)$$

$$MaximizeF2 = \sum_{iob=1}^{nob} {}^{\mathrm{Re}\,l}SumIF^{norm}_{iob} f_{iob} \qquad (7)$$

$$\sum_{iob=1}^{nob} f_{iob} \le \alpha \qquad (8)$$

$\alpha$ is integer constant representing the maximum number of wells that can be chosen

$f_{iob}$ represent the binary decision variable to select a monitoring well location $f_{iob} \equiv \{0,1\}$ such that when $f_{iob}$ value equal to 1 representing monitoring well to be selected at location *iob,* and zero otherwise

$$\sum_{iob=1}^{nob} {}^{\mathrm{Re}\,l}SumIF^{norm}_{iob} f_{iob} - \lambda \ge 0 \qquad (9)$$

$$MaxF2 \ge \lambda \qquad (10)$$

$$F2_{MaxF1} \le \lambda \qquad (11)$$

The two objective multi-objective optimization model is solved using the constrained method where one of the objective function (F1) is maximized keeping minimum level of satisfaction (λ also termed as the trade-off constant) of the second objective function (F2) as shown in Eq. (9). All solutions lying on the Pareto-optimal front corresponds to a different monitoring network.

## Linked Simulation-Optimization Model for Source Identification

Groundwater flow (MODFLOW) and solute transport (MT3DMS) simulation model are used to simulate the physical process within the optimization model.Simulated Annealing (SA) is used as an optimization algorithm to solve the optimization problem. Candidate values of unknown variables (source fluxes) are generated in optimizations algorithm for simulations of flow and transport models. The difference between simulated and observed pollutant concentrations are computed, and finally obtain an optimal solution that minimizes the difference between observed and simulated values.

$$MinimizeF = \sum_{k=1}^{nk}\sum_{iob=1}^{nob} ABS(cest^{k}_{iob} - cobs^{k}_{iob}) \qquad (12)$$

$$cest^{k}_{iob} = f(q_s, C_s) \quad f(q_s, C_s, cest^{k}_{iob}) \; v \; iob, k \qquad (13)$$

$q_s C_s$ is the pollutant source fluxes

$q_s$ is the volumetric flux

$C_s$ is the concentration of the sources or sinks

$ABS$ is the absolute difference

$cest^{k}_{iob}$ is the simulated concentration

$nk$ is the total number of monitoring time steps

$nob$ is the total number of observation wells

$cobs^{k}_{iob}$ is the observed concentration

## PERFORMANCE EVALUATION

The performance of the developed methodology was evaluated for the study area shown in Fig. 1, with hydro-geological parameters as given in Table 1. Three sources with three stress periods of 500days each were considered. The pollutant flux from each of the sources is assumed to be constant over a stress period. Four temporal concentration measurements at each potential location (starting t = 1600 days) are taken after every 200 days.

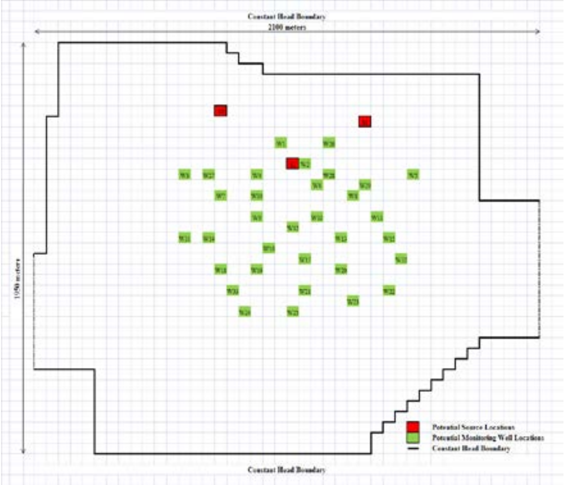

Fig. 1 Plan view of the Illustrative Study Area

Table 1 Hydro-geological Parameters

| Parameter | Unit | Value |
|---|---|---|
| Maximum length of study area | m | 2100 |
| Maximum width of study area | m | 1950 |
| Saturated thickness, *b* | m | 30 |
| Grid spacing in *x*-direction, Δ*x* | m | 50 |
| Grid spacing in *y*-direction, Δ*y* | m | 50 |
| Grid spacing in *z*-direction, Δ*z* | m | 30 |
| Hydraulic conductivity, *K* | m/d | 20 |
| Effective porosity, θ | | 0.3 |
| Longitudinal Dispersivity, α*L* | m/d | 20 |
| Transverse Dispersivity, α*T* | m/d | 10 |
| Horizontal Anisotropy | | 1 |
| Initial contaminant | g/l | 0 |
| concentration | | 0 |
| Diffusion Coefficient | g/s | 0-100 |
| Contaminant Flux | | |

## Genetic Programming Impact Factor

The impact factor is calculated for 3 sources and 25 (W1 to W25 as shown in Fig. 1) potential monitoring well locations. The input data consists of flux values for each source at every stress period. The corresponding output data consists of the resulting pollutant concentration measurement due to these source fluxes at all the 25 potential monitoring well at t = 1600, t = 1800, t = 2000 and t = 2200 days. 3000 data patterns are used of which 50% is used for training, 40% for validation, and 10% for testing the GP models. DiscipulusTM 5.1 (RML Technologies, Inc.) is used for training, validation and testing. Based on $R^2$ fitness value, top 30 GP models are used for computing the impact factor. The impact factor for all the potential monitoring locations at every sampling time step is calculated likewise which is then used to calculate the normalized relative impact factor ${}^{\mathrm{Re}\,l}SumIF_{iob}^{norm}$ and normalised sum of impact factor due to all the potential sources $SumIF_{iob}^{norm}$.

## Optimal Monitoring Network and Arbitrary Monitoring Network

The optimal monitoring design model is solved using normalized impact factor values as calculated above. The value of the minimum satisfaction level of objective function F2 is varied from a minimum -1.7 to a maximum of 8.06 to obtain 12 different Pareto-optimal solutions representing different Pareto-optimal monitoring networks: MN1to MN12. To show the comparison, 10 arbitrary monitoring networks ARMN1 to ARMN10 are chosen. A total of 6 monitoring wells are selected in each monitoring network.

## Source Identification using Data from Pareto-Optimal Monitoring Networks and Arbitrary Monitoring Network

To evaluate the performance, observed concentration measurements are generated synthetically. These simulated measurements are then perturbed with random error term (maximum deviation of 10 percent of the measured concentration data) to incorporate realistic measurement errors (Eq. 17 to Eq. 18). A linked simulation-optimization model is solved using measurements from 12 Pareto-optimal monitoring networks (MN1to MN12) and 10 arbitrary monitoring networks (ARMN1to ARMN10). The source identification model is solved with error free data and perturbed erroneous data.

$$ {}^{Pert}cobs_{iob}^{k} = cobs_{iob}^{k} + err \tag{14}$$

$$ err = \mu per \times rand \tag{15}$$

${}^{Pert}cobs_{iob}^{k}$ is the perturbed concentration value

$cobs_{iob}^{k}$ is the numerically simulated concentration value

$err$ is the error term

$\mu per$ is the specified maximum deviation expressed as a fraction < 1

$rand$ is a random fraction between -1 and +1

## RESULTS AND DISCUSSION

The Pareto-optimal solution for the two-objective optimal monitoring network design model is shown in Fig. 2. The first objective function values F1 is plotted against the minimum satisfaction level of the second objective function value F2.

The non-inferior solutions show the conflicting nature of the two objective functions and their trade-off.

The results of source flux identification solution results obtained by using linked simulation-optimization model is compared for all the 12 Pareto-optimal monitoring locations (MN1 to MN12), obtained as solutions, using error free and perturbed error data (Fig 3 and Fig 4).

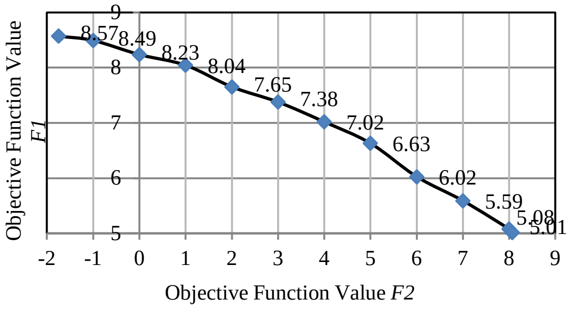

Fig. 2 Pareto Optimal Solution Front

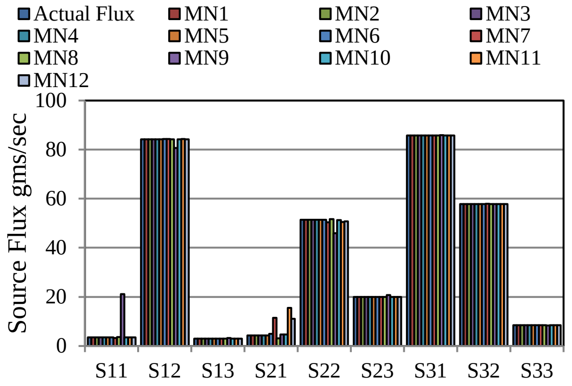

Fig. 3 Identification Results using Error free data

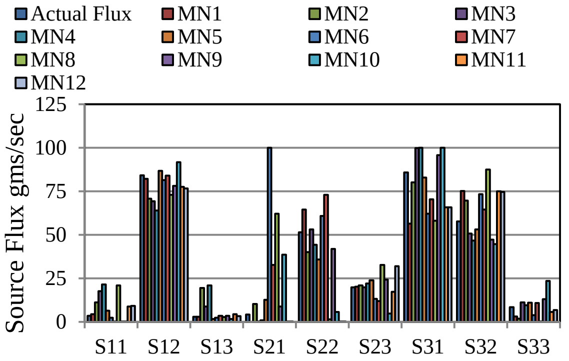

Fig. 4 Identification Results using Erroneous data

To choose the best monitoring network out of the 12 Pareto-optimal monitoring networks (MN1 to MN12), absolute difference of actual source flux and the estimated source flux for all the 12 Pareto-optimal monitoring networks is calculated using error free data and erroneous data (Fig. 5 and Fig. 6).

Absolute difference of actual source flux and the estimated source flux for all the 12 Pareto-optimal monitoring networks (MN1 to MN12) with error free data and erroneous data show similar trend. The average of the absolute difference, between the actual source flux and the estimated source flux is minimum for monitoring network 5 (MN5), hence can be designated as the optimal monitoring network. The source identification model is solved for 10 arbitrary networks (ARMN1 to ARMN10) with both error free data and erroneous data. The estimated flux values using the arbitrary networks is averaged and compared with the actual flux values and estimated flux value from monitoring network 5 (MN5), both for error free data and erroneous data (Fig. 7 and Fig. 8). It is seen that the estimated flux using monitoring network 5 (MN5), are closer to the actual flux values as compared to the flux estimated using the arbitrary networks (AVG-AR).

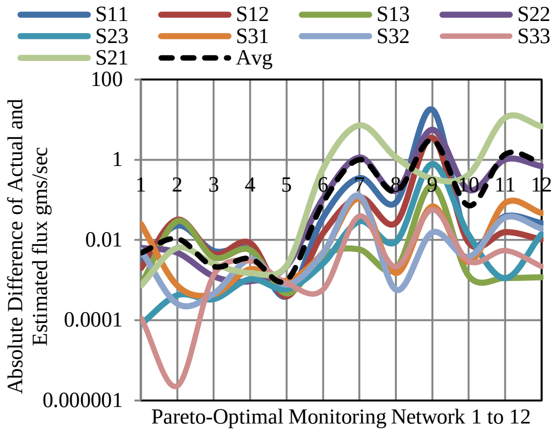

Fig. 5 Difference of Actual Flux and Estimated Flux-Error free Data

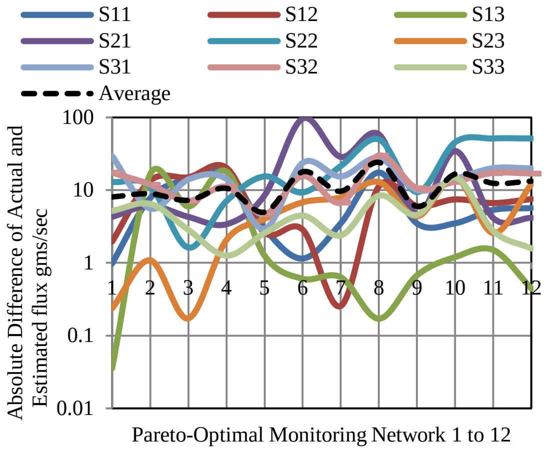

Fig. 6 Difference of Actual Flux and Estimated Flux-Erroneous Data

## CONCLUSIONS

Not all monitoring locations are ideally located for accurate identification of source flux using linked simulation-optimization. The solution results in the illustrative example problem show that the accuracy of source flux identification varies when using pollutant concentration measurement data from different monitoring locations. A methodology has been developed for designing an optimal monitoring network for accurate source flux identification. Concentrations measurements from such a designed monitoring network when used in source flux identification, can improve the accuracy of the source identification results.

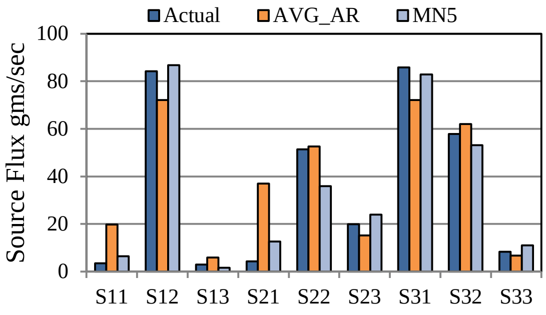

Fig. 7 Comparison of Identification Results using Optimal Network and Arbitrary Networks: Error free Data

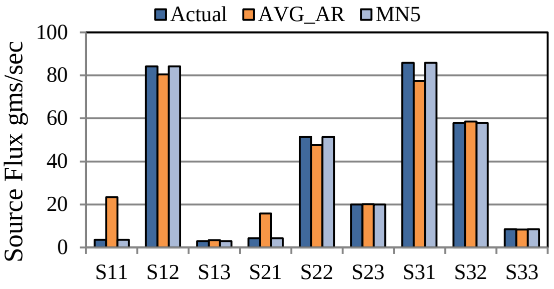

Fig. 8 Comparison of Identification Results using Optimal Network and Arbitrary Networks: Erroneous Data

An optimal monitoring network design for source flux identification is a multi-objective problem. It requires the right balance between well locations where the impact of all the potential sources are significantly large, reducing the possibility of missing an actual source and, well locations where non-uniqueness due to overlapping of source plumes is less. In all real world problems of source identification, the degree of uncertainty in terms of source locations and aquifer response to subjected geo-chemical stress are large. Moreover, the number of monitoring wells to be implemented for concentration measurement data is governed by budgetary constraints. The proposed methodology can be applied to polluted aquifer sites. This method can decrease such uncertainties using limited number of monitoring wells which otherwise will have to be reduced by implementing large number of monitoring wells, resulting in increased capital cost. This method can increase the accuracy of source identification with concentration measurement data from limited number of monitoring wells in a designed monitoring network.

## ACKNOWLEDGEMENTS

We authors would like to thank CRC-CARE and James Cook University, Australia for partial funding and support for this research.

## REFERENCE

[1] Mahar, P.S., and B. Datta (1997), Optimal monitoring network and ground-water-pollution source identification. Journal of Water Resource Planning and Management, 123(4):199–207.

[2] Datta, B., O. Prakash, S. Campbell, and G. Escalada (2013), Efficient Identification of Unknown Groundwater Pollution Sources using Linked Simulation-optimization incorporating Monitoring Location Impact Factor and Frequency Factor, Water Resources Management., DOI 10.1007/s11269-013-0451-8.

[3] Prakash, O., and B. Datta (2012), Sequential optimal monitoring network design and iterative spatial estimation of pollutant concentration for identification of unknown groundwater pollution source locations, Environment Monitoring Assess., DOI 10.1007/s10661-012-2971-8.

[4] Dhar, A., and B. Datta (2010), Logic-Based Design of Groundwater Monitoring Network for redundancy Reduction, Journal of Water Resource Planning and Management, 136,88(2010).

[5] Dhar, A., and B. Datta (2007), Multi-objective design of dynamic monitoring networks for detection of groundwater pollution, Journal of Water Resource Planning and Management, 133(4):329–338.

[6] Sreenivasulu, C., and B. Datta (2008), Dynamic optimal monitoring network design for transient transport of pollutants in groundwater aquifer, Water Resources Management, 22(6):651–670.

*Int. J. of GEOMATE, March, 2014, Vol. 6, No. 1 (Sl. No. 11), pp. 785-790.*
MS No. 3248 received on June 16, 2013 and reviewed under GEOMATE publication policies.

Pertinent discussion including authors' closure, if any, will be published in the March. 2015 if the discussion is received by Sept, 2014.
**Corresponding Author: Om Prakash**