International Journal of Earth, Energy and Environmental Sciences
ISSN: 2517-942X
Vol:7, No:2, 2013

Predicting Dispersion Coefficient in Free-Flowing Zones of Rivers by Genetic Programming

Rajeev Ranjan Sahay

***Abstract*—**Transient storage zones along the flow paths of rivers have great influence on the dispersion of pollutants that are either accidentally or otherwise led into them. The speed with which these pollution clouds get transported and dispersed downstream is, to a large extent, explained by the longitudinal dispersion coefficients in the free-flowing zones of rivers (K_f). In the present work, a new empirical expression for K_f has been derived employing genetic programming (GP) on published dispersion data. The proposed expression uses few hydraulic and geometric characteristics of a river that are readily available to field engineers. Based on various performance indices, the proposed expression is found superior to other existing expression for K_f.

***Keywords*—**Dispersion, parameter estimation, rivers, transient pollutant.

I. Introduction

POLLUTION in river catchments can originate from diffuse sources such as agriculture, point sources such as municipal and industrial effluents or incidents involving chemical or oil spillages [1]. The pollutants' concentration variation in the flowing direction of rivers is an important task for environmental engineers for devising water diversion strategies, determining self-purifying characteristics of streams, designing treatment plants, intakes and outfalls and studying environmental impact due to accidental injection of polluting effluents into streams. Near the release point the concentration field is three-dimensional but a little downstream of this point, the vertical and cross-sectional concentration variation is nearly uniform and the process of dispersion in longitudinal direction assumes importance, intensity of which is measured by the longitudinal dispersion coefficient (K). K is greatly modified by the presence of transient storage zones (TSZ) also called stagnant or dead zones. These zones are formed in rivers by quiescent backwater eddies, pool-riffle sequences in the bottom, aquatic vegetation and interstitial aquifer voids and may retain pollutants for considerable period of time before releasing them into the free-flowing river zones. Therefore, the cross-sectionally averaged longitudinal dispersion coefficient (K) is usually larger than the longitudinal dispersion coefficient in the free-flowing water zone of rivers (K_f) as K implicitly accounts for the effects of structural heterogeneity.

For practical implication, the effect of transient storage on solute exchange in rivers has been an actively researched area [2]-[12]. References [13]-[15] showed the classical advection-dispersion models inadequate in predicting dispersion in rivers while the transient storage zone model (TSM) is widely accepted in literature to be a more reliable model as the effects of dead zones are adequately represented in this model. TSM is defined by the following two mass conservation equations, one for the solute concentration dissipation in the free-flowing water zone and another in the transient storage zone of a river

$$\frac{\partial C_f}{\partial t} + U_f \frac{\partial C_f}{\partial x} - K_f \frac{\partial^2 C_f}{\partial x^2} = \varepsilon T^{-1}\left(C_s - C_f\right) \quad (1)$$

$$\frac{\partial C_s}{\partial t} = T^{-1}\left(C_f - C_s\right) \quad (2)$$

where C_f is the pollutant concentration in the free flowing water zone, C_s is the pollutant concentration in transient storage zones, t is the time elapsed since injection of the solute/pollutant, x is the longitudinal distance from place of injection of the pollutant to the investigation site, U_f is the mean flow velocity in the free flowing water zone, K_f is the longitudinal dispersion coefficient in the free-flowing water zone, ε is ratio of cross-sectional area of the transient storage to the total cross-sectional area of the channel and T is the residence time of the tracer in the transient storage zone. K_f is the most important of TSM's four key parameters as its influence on concentration variation is the greatest. The other three key parameters are T, ε and U_f. For successful application of TSM model, appropriate value of K_f is required which can best be estimated using tracer concentration profile taken from a particular reach of the stream, however, such tracer investigation is expensive and rarely done for every reach of a stream. This necessitates the usage of empirical expressions.

II. Previous Empirical Studies

Reference [16] developed the following expressions for Kf by extending the asymptotic solution of [17] for spatial variance and fitting his solution to some of the experimental data compiled by [18]

$$K_f = (0.22 + 35.6\theta)\frac{W^2 U_*}{H} \quad (3)$$

where θ is the Darcy-Weisbach friction factor, W is the stream width, H is the mean flow depth and U_* is the shear velocity.

Reference [19] developed another expression for K_f after applying the weighted one-step Huber method of nonlinear regression on published field data as

Rajeev R. Sahay is with Birla Institute of Technology, Mesra, Ranchi, 835215 India (phone: 91-9431382737; fax: 0651-227-5401; e-mail: rajeev_sahay@yahoo.com).

$$\frac{K_f}{HU_*} = 0.583 \left(\frac{W}{H}\right)^{1.287} \left(\frac{U}{U_*}\right)^{0.562} \quad (4)$$

where U is mean flow velocity.

Reference [20] emphasized the importance of channel sinuosity and Peclet number besides other hydraulic and geometric characteristics of rivers in solute exchange mechanisms associated with transient storage zones and applied the robust minimum covariance determinant method on published field data to derive the following formula in non-dimensional format and showed its superiority over other reported expressions of the time

$$\frac{K_f}{HU_*} = 43.928 \left(\frac{W}{H}\right)^{0.2176} \left(\frac{U}{U_*}\right)^{8453} P^{-0.042} S_i^{-1.6981} \quad (5)$$

where S_i is the channel sinuosity and P is the Peclet number.

III. Derivation of New Expression

Existing expressions for the parameter K_f were evaluated and found to give large deviation between the measured and the predicted values. The present study aims at deriving a more reliable expression for K_f in a dimensionless format by applying GP on the published field data. For deriving and verifying the new expression, dispersion data consisting of 55 sets of observations from USA Rivers was utilized (Table I). Comparable datasets were chosen for deriving and verifying the new expression to avoid any bias in modeling. There were two reasons for selecting this data: (i) it represents a wide range of geometric and flow characteristics of streams and (ii) it was earlier used by other investigators [19] and [20]. Thus, results from the proposed and other reported expressions for K_f can be compared well.

A. Genetic Programming (GP)

Genetic Programming is an artificial intelligence algorithm based on the Darwinian evolutionary processes of reproduction, mutation, recombination and selection for finding functional relationships among the constituents in a system. GP, though a branch of genetic algorithm (GA), has distinctive differences. The main difference lies in the representation of the solution. GP creates computer programs as the solution of a problem, whereas, GA creates a string of numbers that represents the solution.

The GP approach starts with the user's defining the task and identifying mathematical functions (e.g. divide, log, sin, tanh, etc.), the automated computer programs will be using during computation. Thereafter, an initial population of programs is generated randomly which GP translates, compiles, executes and based on fitness function evaluates. Some of the best performing programs are then selected for reproduction which are combined or mutated into offspring to make up for the next generation of programs. The process is repeated until a termination criterion is met. The criterion may be either the number of generations or the change in the fitness value of programs between two consecutive generations. In general, the fitness values of later generations should improve, though we can't expect the best solution to be found in the final generation. A good illustration of the working of GP can be found in the books by [21] or [22]. The working structure of a GP system is shown in Fig. 1.

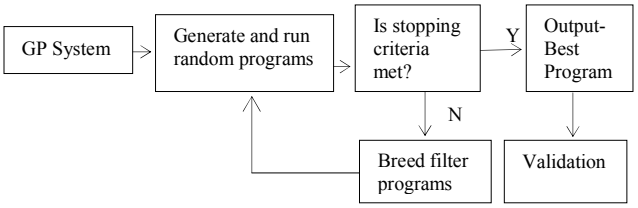

Fig. 1 Working structure of Genetic Programming

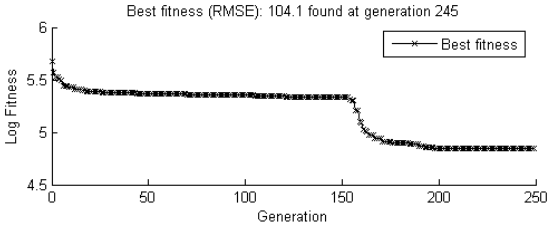

Fig. 2 Fitness vs. generation number (derivation dataset)

In the present work, GPTIPS ([23]), an open source GP program, was applied on published dispersion data (Table I) with the following settings: population size = 150, number of generation = 200, tournament size = 4, elitism = 0.02% of population, maximum depth of tree = 3, maximum number of genes allowed in an individual tree = 4 (W/H, U/U*, Pe and Si) and function node set = {plus, minus, times, power, sine and cosine}. The above setting was used to minimize the root mean square prediction error between measured and predicted values on the derivation datasets. The model that performed the best on the verification datasets was chosen. After several runs, the best fitness was achieved at generation number 245 (Fig. 2) and, accordingly, the expression for dimensionless parameter was obtained as

$$\frac{K_f}{HU_*} = -12.59 + \frac{0.0001164 \left(\frac{U}{U_*}\right)^3 \frac{W}{H}}{PCos\left(\frac{W}{H}\right)} + 0.0082 \frac{\frac{W}{H}\left(\frac{U}{H_*}-1\right)-7.49}{Sin\left(\frac{W}{H}-7.24\right)} + 3.84\, Sin\left(\frac{W}{H}\right) S_i \frac{W}{H} + 3.84 \frac{\frac{W}{H}+\frac{U}{U_*}}{Sin(S_i)} - 0.766\, Sin\left(\frac{W}{H}\right)\frac{W}{H}\frac{U}{U_*} - 1.532\, Sin\left(\frac{W}{H}\right)\frac{W}{H} \quad (6)$$

TABLE I
DISPERSION DATA AT 55 SITES ON USA RIVERS [20]

River	*W/H*	*U/ U**	P_e	S_i	$K_x/HU*$	
Green & Duwamish, WA*	13.778	5.681	0.08	1.39	7.9	[24]
Green & Duwamish, WA	27.417	7.724	2.22	1.39	23.2	
Copper creek, VA	32.600	2.923	0.05	1.23	164.2	[25]
Copper creek, VA*	45.053	1.560	0.09	1.23	104.2	
Copper creek, VA	21.964	1.233	0.02	1.23	54.3	
Powel river, TN	41.024	1.669	0.03	1.38	88.8	
Clinch river, VA*	46.233	2.730	0.09	1.73	81.9	
Copper creek, VA	24.159	5.088	0.07	1.23	292.1	
Clinch river, VA	25.625	7.073	0.10	1.73	74.0	
Coachella canal, CA*	15.321	17.104	0.58	1.04	131.4	
Coachella canal, CA*	15.761	17.391	0.58	1.04	120.6	
Clinch river, VA	25.625	6.422	0.10	1.73	101.4	
Copper creek, VA	32.000	2.543	0.03	1.23	164.3	
Missouri river	55.122	15.862	0.14	2.00	1874.1	
Antiem creek, Md	32.973	3.002	0.32	1.91	220.2	[18]
Antiem creek, Md*	25.890	5.469	0.33	1.91	259.2	
Antiem creek, Md	18.015	4.710	0.56	1.91	80.2	
Antiem creek, Md*	51.311	5.517	0.39	1.91	200.9	
Monocacy river, MD	88.673	5.062	0.17	1.39	483.9	
Monocacy river, MD	130.930	3.176	0.10	1.39	258.9	
Monocacy river, MD	57.663	5.196	0.26	1.39	158.9	
Monocacy river, MD*	68.903	0.873	0.01	1.39	240.9	
Conococheague creek, MD	89.000	4.319	0.16	2.27	703.4	
Conococheague creek,MD*	95.070	1.930	0.05	2.27	1130.0	
Conococheague creek, MD	53.895	9.038	0.38	2.27	147.9	
Chattahoochee river, GA	30.433	8.570	0.50	1.23	744.4	
Chattahoochee river, GA	35.946	3.668	0.17	1.23	451.9	
Salt creek, NE	92.059	4.804	0.15	1.17	1694.9	
Difficult Run, VA*	48.600	3.672	0.30	1.38	115.2	
Bear creek, CO*	16.141	11.386	4.36	2.04	30.2	
Little Pincy Creek, MD	72.045	7.464	0.37	1.25	328.0	
Bayou Anacoco, LA*	38.889	4.786	1.05	1.76	185.4	
Comite river, LA*	41.467	6.249	0.26	1.35	67.5	
Tickfau river, LA	31.735	0.916	0.30	1.2	40.6	
Tangipahoe river, LA	33.479	3.862	0.49	1.29	248.7	
Tangipahoe river, LA	54.055	6.553	0.70	1.29	159.5	
Red river*	64.094	6.282	0.40	1.59	218.3	
Red river*	70.000	5.266	0.34	1.59	344.0	
Red river	41.639	7.018	0.30	1.59	187.7	
Red river	114.514	8.982	0.35	1.59	392.9	
Sabin river, LA	54.307	14.824	0.36	1.61	539.1	
Sabin river, LA*	70.088	17.840	0.61	1.61	520.0	
Sabin river, TX	17.671	2.351	0.40	1.52	87.8	
Sabin river, TX	19.840	1.068	0.28	1.52	53.7	
Sabin river, TX*	28.317	3.438	0.50	1.52	83.1	
Mississippi river, LA	40.624	12.641	0.37	1.73	210.2	
Mississippi river, MO	119.455	13.322	0.52	1.38	406.8	
Mississippi river, MO*	69.327	14.526	0.91	1.38	234.7	
Wind/Bighorn river, WY	50.538	6.143	0.70	1.15	165.0	
Wind/Bighorn river, WY*	30.354	9.150	0.58	1.15	167.1	
Colorado river, AZ	17.393	9.458	0.18	1.76	270.3	[26]
Colorado river, AZ	8.732	3.950	0.28	1.76	93.9	
Botna river*	20.500	4.800	0.33	1.12	69.8	[27]
Kogilnik river	5.750	7.145	0.05	1.31	23.6	
Byk river	18.214	6.824	0.40	1.23	106.3	

Note:*verification dataset

The new expression, i.e., (6), appears to have successfully been derived, with coefficient of correlation between measured and predicted values being equal to 0.92. It successfully predicted the highest three (K_f/HU_*) values of 1874.1, 1694.9 and 703.4 as 1878.7, 1100.3 and 792.5 respectively, and lowest three values of 23.6, 53.7 and 54.3 as 49.9, 79.52 and 83.0 respectively. Moreover, most of the predicted values by the new expression are distributed evenly about the ideal line, showing no bias for over or under prediction. Other expressions, on the other hand, either significantly overpredict or underpredict the measured values (Fig. 3).

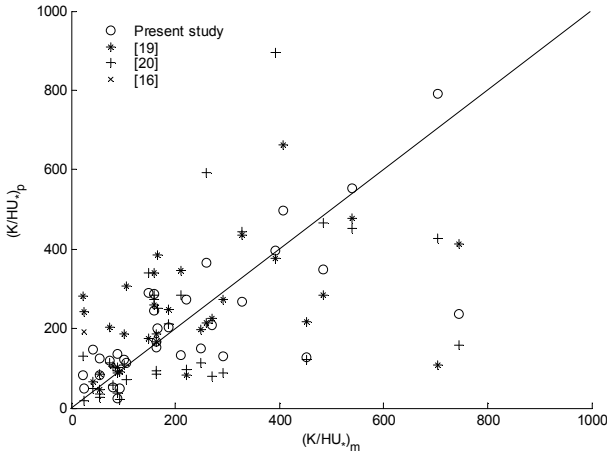

Fig. 3 Measured vs. predicted $K_f/(HU_*)$ (derivation dataset)

IV. Verification of the New Expression

The comparison of the new and other reported expressions for dimensionless longitudinal dispersion coefficient in free-flowing zones of rivers, i.e. K_f/HU_*, is accomplished using 20 measured datasets of Table I. They were not used for deriving the expression. The comparison models considered here are [16], [19] and [20]. The performance indices used for comparison of models are coefficient of correlation (CC), root mean square error (RMSE), discrepancy ratio (DR) and accuracy which are defined as

$$CC = \frac{\left[\sum_{i=1}^{N}\left(\frac{K_f}{HU_*}\right)_p\left(\frac{K_f}{HU_*}\right)_m - \sum_{i=1}^{N}\left(\frac{K_f}{HU_*}\right)_p\sum_{i=1}^{N}\left(\frac{K_f}{HU_*}\right)_m\right]}{N S_p S_m} \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left[\left(\frac{K_f}{HU_*}\right)_p - \left(\frac{K_f}{HU_*}\right)_m\right]^2}{N}} \quad (8)$$

$$DR = \log\frac{\left(\frac{K_f}{HU_*}\right)_p}{\left(\frac{K_f}{HU_*}\right)_m} \quad (9)$$

$$\text{Accuracy} = 1 \text{ (if DR lies between -0.2 to 0.2), or, } 0 \quad (10)$$

where $(K_f/HU_*)_p$ and $(K_f/HU_*)_m$ are predicted and measured dimensionless longitudinal dispersion coefficient in free-flowing river zone, respectively, and S_p and S_m are standard deviations in the predicted and the measured values, respectively. From (9), DR=0 suggests exact matching between measured and predicted values, otherwise, there is either overprediction (DR>0) or underprediction (DR<0).

Fig. 4 shows predicted vs. measured values of (K_f/HU_*) by all considered models for the verification dataset. It shows the maximum number of predicted values by the new expression closer to the measured values. The largest three values of the parameter, i.e., 1130.0, 520.0 and 344.0 are successfully predicted by the new expression as 868.9, 512.6 and 278.3. However, deviations are shown larger in case of the other expressions. Table II summarizes performance indices of the considered models. It shows the predictive accuracy of the newly derived expression higher than other models, RMSE being the least and CC, the largest for all datasets, i.e., derivation, verification and whole datasets. When extreme values, i.e., (W/H) > 50, are neglected, the performance of all models improve with their RMSE reducing significantly. The greatest improvement is seen in [16] implying its inadequacy for large and shallow rivers.

DR range, another performance indicator, is found superior in case of the new expression, while other considered expressions have tendency towards overprediction.

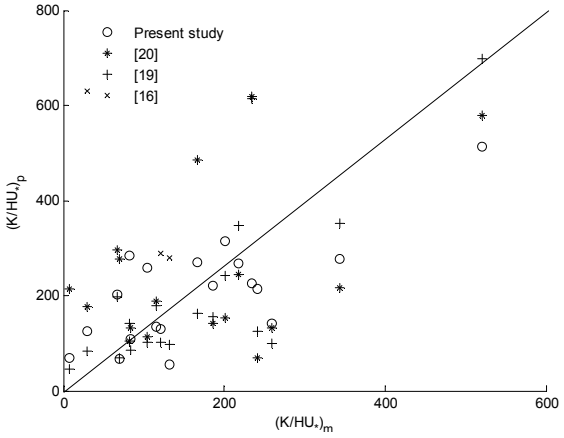

Fig. 4 Measured vs. predicted $K_f/(HU_*)$ (verification dataset)

If accuracy of a model can be defined as the percentage of the predicted values lying between ± 40% of the measured values, i.e., discrepancy ratio lying between -0.2 and 0.2, then it can be observed from Table II that the proposed expression has predicted the parameter accurately in 60 % of the cases, the highest among all the comparing models. Accuracies of [20], [19] and [16] are estimated as 35%, 35%, and 0 % respectively (Fig. 5).

TABLE II
PERFORMANCE INDICES OF MODELS

	Model	Whole datasets		Derivation datasets		Verification datasets				
		CC	RMSE	CC	RMSE	CC	RMSE	RMSE (W/H>50 ignored)	DR Range	Accuracy (%)
K_f/HU_*	Present	0.92	142.0	0.92	233.8	0.92	104.1	80.1	-0.37 to 0.94	60
	C-S	0.43	333.6	0.41	384.0	0.47	218.9	53.2	-0.58 to 0.76	35
	C-Y-S	0.16	373.5	0.40	377.8	-0.11	365.9	257.9	-1.28 to 1.44	35
	Pedersen	0.10	274648	-0.02	113134.7	0.33	430160.7	57805.7	-0.33 to 3.87	0

Note: Pedersen = [16], C-S = [19] and C-Y-S= [20].

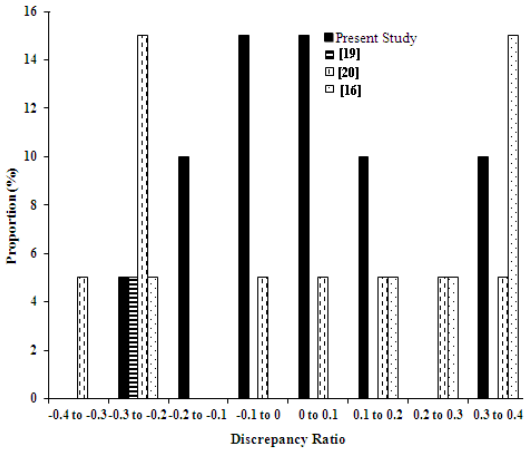

Fig. 5 DR values by models (verification dataset)

VI. CONCLUSION

The presence of transient storage zones in rivers significantly modifies dispersion of pollutants/solutes spilled into them. Such retention domains may keep these pollutants for considerable period of time before releasing them slowly into the main flowing zones of rivers. After vertical and cross-sectional mixing of the pollutant is complete, its concentration variation in the flowing direction is largely dependent on the longitudinal dispersion coefficient in the free-flowing zone of the river. Measured values of K_f for every reach of a river is not possible as it requires concentration sampling at various points on the river reach. For those reasons, many investigators have given empirical expressions for K_f. However, the evaluation carried out in this study finds them inadequate. In the present work, by implementing genetic programming on published dispersion data, a new expression for dimensionless longitudinal dispersion coefficient in free-flowing zone of rivers has been derived. The expression uses few hydraulic and geometric characteristics of a river, i.e., stream width, stream sinuosity, mean flow velocity, mean flow depth, shear velocity and Peclet number. These physical quantities can be reasonably estimated or directly measured. The performance of the proposed expression was compared against [16], [19], and [20]. Based on various performance indices, the new expression was found superior to all other considered models with CC between the measured and the predicted values being the highest and RMSE, the lowest. About sixty percent of prediction by the new expression was found to lie between ± 40% of the measured values, whereas, the predictive accuracy of other models were far less. When extreme values, i.e., (W/H) > 50, are neglected, prediction from all models improved but the most significant improvements were seen in [16] and [19] indicating their inadequacy for wide and shallow rivers. On the other hand, the new model did not show much improvement suggesting its suitability even for extreme rivers.

REFERENCES

[1] Dunk, M. J., McMath, S. M. and J. Arikans, "A new management approach for the remediation of polluted surface water outfalls to improve river water quality", *Water and Environment Journal*, 22, 2008, pp. 32-41.
[2] Hays, J. R., Krenkel, P. A., and K. B. J. Schnelle, *Mass transport mechanisms in open-channel flow*, Sanitary and water resources engineering department of civil engineering technical report 8, 1967, Van-derbilt University, Nashville, Tennessee.
[3] Thackston, E. L. and K. B. Schnelle, "Predicting effects of dead zones on stream mixing", *J. of the Sanitary Engineering Division*, 96, 1970, SA2, pp. 319-331.
[4] Bencala, K. E. and R. A. Walters, "Simulation of Solute Transport in a Mountain Pool-and-Riffle Stream", *Water Resources Research*, 19, 1983, pp. 718-724.
[5] Runkel, R. L., Bencala, K. E., Broshears, R. E. and S. E. Chopra, "Reactive solute transport in streams, 1, development of an equilibrium-based model" *Water Resources Research*, 32, 1996, pp. 409-418.
[6] A. Worman, "Comparison of models for transient storage of solutes in small streams." *Water Resources Research*, 36, 2000, 455-468.
[7] Seo, I. W. and D. Yu, "Modeling solute transport in pool-and-riffle streams." *Water Engineering Research*, 1, 2000, 171-185.
[8] Boxall, J. B, Guymer, I. and A. Marion, "Locating Outfalls on Meandering Channels to Optimise Transverse Mixing." *Water and Environment Journal*, DOI: 10.1111/j.1747-6593.2002.tb00394.x.
[9] Rowiński, P. M., and A. Piotrowski, "Estimation of parameters of transient storage model by means of multi-layer perceptron neural networks." *Hydrological Sciences Journal*, 53, 2008, 165-178.
[10] Piotrowski, A. P., Rowinski, P. M. and J. J. Napiorkowski, "Estimation of parameters of models of pollutant transport in rivers depending on data availability", *Proc. of 33rd IAHR Congress: Water Engineering for a Sustainable Environment*, 2009, 1179-1186.
[11] H. M. Azamathulla, "Genetic programming for predicting longitudinal dispersion coefficients in streams." *Water Resour. Manage.*, 25, 2011, 1537-1544.
[12] R. R. Sahay, "Predicting Transient Storage Model Parameters of Rivers by Genetic Algorithm". *Water Resour. Manage.*, 26, 2012, pp. 3667-3685.
[13] J. W. Elder," The dispersion of a marked fluid in turbulent shear flow", *J. of Fluid Mechanics*, 5, 1959, pp. 544-560.
[14] B. H. Fischer, "The mechanics of dispersion in natural streams", *J. of Hydraulic Engineering*, 93, 1967, 187-216.
[15] M. K. Bansal, "Dispersion in natural streams" *J. of Hydraulic Division*, 97, 1971, pp. 1867-1886.
[16] F. B. Pedersen, Prediction of Longitudinal Dispersion in Natural Streams, Series Paper 14, 1977, Technical University of Denmark, Lyngby.
[17] A. Okubo, "Effect of shoreline irregularities on stream wise dispersion in estuaries and other embayments." *Netherlands J. of Sea Res.*. 6, 1973, 213–224.

[18] Nordin, C. F. and G. V. Sabol, Empirical data on longitudinal dispersion, *US Geological Survey Water Resources*, Investigation Report, 1974, 20–74.
[19] Cheong, T. S. and I. W. Seo, "Parameter estimation of the transient storage model by a routing method for river mixing processes." *Water Resources Research*, 39, 2003, pp. 1074–1084.
[20] Cheong, T. S., Younis, B. A. and I. W. Seo," Estimation of key parameters in model for solute transport in rivers and streams", *Water Resources Management*, 21, 2007, pp. 1165–1186.
[21] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MIT Press, 1992.
[22] Langdon, W. B. and R. Poli, *Foundations of Genetic Programming*. Springer-Verlag, 2001.
[23] Searson, D. P., Leahy, E. D. E. and, M. J. Willis, GPTIPS: An Open Source Genetic Programming Toolbox For Multigene Symbolic Regression. *Proc. of International Multi-conference of Engineers and Computer Scientists*, Hongkong, 1, March 17-19, 2010.
[24] H. B. Fischer, Method for predicting dispersion coefficients in natural streams, with applications to lower reaches of the Green and Duwamish Rivers, *Washington. Professional Paper 582-A, U.S. Geological Survey*, Washington, D.C, 1968b.
[25] Godfrey, R. G, and B. J. Frederick, "Stream dispersion at selected sites." Professional Paper 433-K, U.S. Geological Survey, Washington, D.C, 2007.
[26] B. Graf, ''Observed and predicted velocity and longitudinal dispersion at steady and unsteady flow, Colorado River, Glen Canyon Dam to Lake Mead." Water Resources Bulletin31, 1995, pp. 265–281.
[27] Czernuszenko, W., Rowinski, P. M. and A. Sukhodolov, "Experimental and numerical validation of the dead-zone model for longitudinal dispersion in rivers", *J. of Hydraulic Research*, 36, 1998, pp. 69–280.

Rajeev Ranjan Sahay is Professor in Civil Engineering Department of Birla Institute of Technology, Mesra (India). He has been working in the field of water resources for the last twenty five years.