XVIII International Conference on Water Resources
CMWR 2010
J. Carrera (Ed)
CIMNE, Barcelona 2010

INVERSE GROUNDWATER MODELING BASED ON B-SPLINE PARAMETERIZATION AND MODEL OPTIMIZATION

George Kourakos* **and Aristotelis Mantoglou†**

* National Technical University of Athens
e-mail: giorgk@gmail.com

† National Technical University of Athens
e-mail: mantog@central.ntua.gr, web page: http://www.ntua.gr

Key words: Inverse groundwater modeling, optimization, model calibration

Summary. A new method is developed for inverse groundwater modeling based on B-Spline parameterization of transmissivity. The objective is not only to obtain the values of model parameters but to determine the optimal number of parameters as well (optimal model complexity). The methodology is based on a B-Spline parameterization of transmissivity and a LLS transformation between models of different model complexities. This allows combination of models of different complexities using conventional GA operators.

1 INTRODUCTION

Inverse modeling is the practice of assigning appropriate values to model parameters so that the model can make good predictions of the model outputs. In groundwater modeling, the model output is often the hydraulic head and the most important calibration parameter is transmissivity. Parameterization methods are often employed, where transmissivity is specified as a two-dimensional function, controlled by only a few parameters. The interested reader may find a comprehensive review on inverse modeling techniques in Carrera et al. (2005)[1]

An important consideration in inverse groundwater modeling is the number of model parameters. This corresponds to the underlying model complexity. A model with too many parameters (complex model) may follow the errors in the data rather than the important dynamics of the physical system. A model with too few parameters on the other hand may not have sufficient flexibility to follow the system dynamics[2].

In this study we use B-Spline Surface (BSS) parameterization for transmissivity. BSS are controlled by a number of vertices known as Control Points (CP). One advantage of BSS parameterization is that each CP affects only a limited part (patch) of the surface; hence a change of the value of a CP does not affect the entire transmissivity field. The number of control points of BSS representation, corresponds to model parameters of the inverse model, (i.e. it is related to model complexity). In the proposed formulation, the number of control points is also a decision variable to be determined by optimization. Genetic Algorithms (GA) are employed to perform the search in the parameter space. Notice that the population of GA is composed by individuals of different number of parameters. A transformation procedure is applied, based on Linear Least Squares (LLS), which allows combination of models of

different parameters using conventional GA operators.

The proposed method is applied to a hypothetical confined aquifer. The results are good when the genetic algorithm stops early. However, if the algorithm is allowed to run for too long, it converges to models of highest allowable complexity.

2 FORMULATION OF INVERSE MODELING

The applications in Section 5 are based on a two-dimensional confined aquifer; however the method can be easily generalized to other inverse groundwater problems. The governing equation in 2D confined aquifer is written as:

$$S\frac{\partial \phi}{\partial t}=\frac{\partial}{\partial x}\left(T\,\frac{\partial \phi}{\partial x}\right)+\frac{\partial}{\partial y}\left(T\,\frac{\partial \phi}{\partial y}\right)+Q+N \tag{1}$$

where $\phi(x,y)$ is the piezometric head, Q are sources or sinks, N is groundwater recharge, S is storativity, $T=K\,b$ is transmissivity, and t represents time. Transmissivity is a two-dimensional function of spatial coordinates $T=f(x,y)$. In order to solve (1) we need to define boundary and initial conditions, and assign appropriate values for the unknown parameters. In practical applications the boundary - initial conditions are obtained by the aquifer geology, while storativity is not considered as unknown during calibration. Hence, the objective of inverse modeling is simplified to finding an appropriate two-dimensional distribution of transmissivity so that the piezometric heads predicted by the numerical model are close to the actual piezometric heads, measured by field work. This can be written as an optimization problem, where the objective is minimization of errors between predicted and desired heads

$$\min f_{err}=\frac{1}{N_t}\sum_{t=1}^{N_t}\left(\frac{1}{N_{obs}}\left(\hat{\boldsymbol{\varphi}}_t-\boldsymbol{\varphi}_t\right)^T\mathbf{C}^{-1}\left(\hat{\boldsymbol{\varphi}}_t-\boldsymbol{\varphi}_t\right)\right)^{1/2} \tag{2}$$

where $\hat{\boldsymbol{\varphi}}_t$ is the vector of the predicted piezometric heads at time t, $\boldsymbol{\varphi}_t$ the vector of the measured heads, $\mathbf{C}$ is a weight matrix, N_{obs} the number of observation points and N_t is the number of stress periods used in calibration. In the case of steady state: $N_t=1$. Objective (2) is a function of transmissivity $f_{err}=f(T)$ through (1). Using parameterization techniques, such as zonation, radial basis functions, pilot points, etc., the parameters that describe the distribution of transmissivity over the $x-y$ plane are the decision variables of the optimization problem.

The objective of this work is not only to obtain the values of model parameters but to determine the optimal number of parameters as well (optimal model complexity). The model complexity depends on the number of model parameters, which depends on discretization in the parameter space. A parameterization based on B-Splines (BSS) is selected and, depending on discretization on the parameter space, models of various complexities are formed. These models are controlled by a grid of control points specifying the number of model parameters. A major advantage of BSS is that they are very flexible and can be used effectively in isotropic and anisotropic aquifers. In addition, based on a transformation discussed in section

4, models of different complexities can be evaluated simultaneously by the genetic algorithm. Next, some fundamental elements of BSS are presented.

3 B-SPLINE AQUIFER PARAMETERIZATION

A B-Spline surface (BSS) is usually defined by a rectangular grid of control points, although there are other varieties of BSS such as triangular[3] or multisided[4]. In this work we utilized uniform, quadratic B-Spline surfaces which are a particular class of BSS, however the analysis can easily be extended to cubic or higher degree BSS.

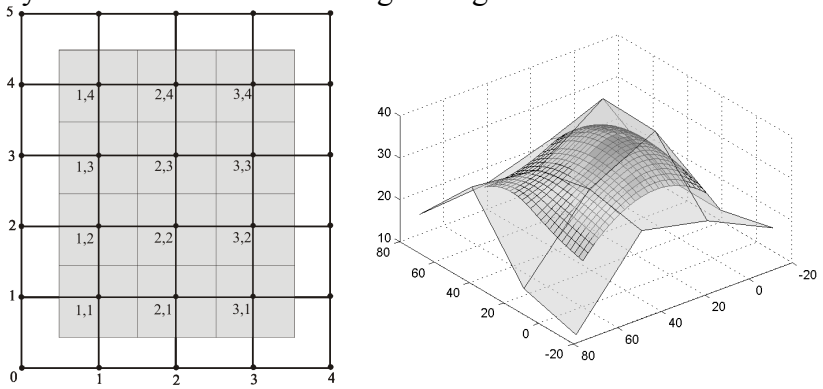

Figure 1. Left: A BSS controlled by 5×6 control points. Right: A BSS controlled by its characteristic polyhedron.

Figure 1 (left) illustrates a BSS (grey rectangles) controlled by a number of control points (CP). The CPs are organized in an array of dimensions $m\times n$ where m,n define the number of CPs in x,y direction respectively. The product $m\times n$ defines the number of parameters of BSS, i.e. the complexity of the model. As more CPs are used, more surface details can be captured by the BSS approximation. BSSs are constructed by a sequence of $s\times t$ patches, where $s\in[1,\, m-2]$ and $t\in[1,\, n-2]$. In the example of Figure 1, the control points are $m=5$ and $n=6$ (fig.1 follows the conventional BSS notation for CPs where the numbering starts from 0), while the number of patches are $m-2=3$ and $n-2=4$. Thus, the patch indexes are $s\in[1,\, 3]$ and $t\in[1,\, 4]$ respectively. For example the left lower patch (grey rectangle), is designated as $(1,1)$ while the top right patch is designated as $(3,4)$. For each patch (s,t), surface z is constructed by evaluating the following matrix equation:

$$z(u,w)=\tfrac{1}{2}\mathbf{U}\mathbf{M}_s\mathbf{P}_{s,t}\mathbf{M}_s^T\tfrac{1}{2}\mathbf{W}^T \tag{3}$$

where z corresponds to transmissivity approximation. For quadratic B-Splines, vectors $\mathbf{U},\mathbf{W}$ contain parametric variables $\mathbf{U}=\left[u^2,u,1\right]$ and $\mathbf{W}=\left[w^2,w,1\right]$, where $u,w\in[0,1]$. Matrix $\mathbf{M}_s=[1,-2,1;\ -2,2,0;\ 1,1,0]$ is a basis transformation matrix and $\mathbf{P}_{s,t}$ is a 3x3 matrix that contains the coordinates of CPs with values depended on the particular patch s,t. Matrix $\mathbf{P}_{s,t}$ is a 3x3 matrix of the following format

$$\mathbf{P}_{s,t} = \begin{bmatrix} p_{s-1,t+1} & p_{s,t+1} & p_{s+1,t+1} \\ p_{s-1,t} & p_{s,t} & p_{s+1,t} \\ p_{s-1,t-1} & p_{s,t-1} & p_{s+1,t-1} \end{bmatrix} \tag{4}$$

where $p_{i,j}$ are the heights at the CPs. For $u = w = 0$, equation (3) returns the z coordinate of the left-bottom point of the patch being evaluated, while for $u = w = 1$ the value at top-right corner of the patch is returned. In order to generate the entire patch, we discretize parametric variables u, w, and (3) is evaluated for each pair of $u, w \in [0,1]$. More details of the methodology can be found in Mortenson (1997)[5]

The unknown parameters correspond to the heights at the CPs, i.e. correspond to $m \times n$ values $p_{i,j}$, $i \in [0, m-1]$ and $j \in [0, n-1]$. The locations of CPs over the $x-y$ plane are specified in an automated fashion depending on the number of control points m, n. Thus, the decision variables during calibration are limited to heights at the CPs as well as the number of parameters m, n.

4 GROUNDWATER CALIBRATION USING B-SPLINE SURFACES AND GENETIC ALGORITHMS

In typical groundwater calibration problems, the number of parameters (i.e. the complexity of the model) is usually specified by the modeler before optimization. Then, by minimizing the error, it is possible to find the optimum values of the decision variables for a fixed model complexity. Here, we develop a new approach, where the number of decision variables m, n are unknown and can change during optimization. The method is integrated within a GA optimization framework. Since the GA operators are designed to combine individuals with the same number of parameters, we apply a transformation among models of different decision variables without changing the basic shape of the surface described by the decision variables. Then, we can use conventional GAs to perform the search.

Next, the GA optimization method is discussed with reference to this transformation procedure. An initial population is first generated that spans the whole search domain. The population is produced by varying the number of parameters m, n as well as the values $p_{i,j}$ in (4). The population is evaluated and ranked based on the fitness function (2) and a selection operator is used to select the fittest individuals as parents. The parents are combined using a crossover operator and offspring are generated, which form the new population. The population is evaluated and the process is repeated until stopping criteria are met. After several generations, the algorithm is expected to converge to a near global optimum solution.

In this method, each individual represents parameters $p_{i,j}$, controlling the surface, which corresponds to the aquifer transmissivity. Since the model complexity represented by the number of parameters m, n is unknown to be determined by optimization, we apply a transformation which allows combination of individuals of different complexities. Further, the GA operators were modified in order to account for the particularities of BSS. In the modified

GA, the initial population is generated in two steps. First, the structural part (m, n) of each individual is generated. The m and n parameters are chosen randomly within range $m \in [3,\ldots,m^{\max}]$ and $n \in [3,\ldots,n^{\max}]$, where $m^{\max}$ and $n^{\max}$ is the maximum allowable parameter discretization along each directions x, y respectively. The minimum values of $m=3$, $n=3$ correspond to the simplest model with only one patch. Then the adjusting part is formed by generating $m \times n$ random numbers within range $\left[p^{\min}, p^{\max}\right]$, where $p^{\min}$ and $p^{\max}$ are the minimum and maximum values of parameters at the CPs. While $m^{\max}$ and $n^{\max}$ are used throughout the optimization procedure, the limits $p^{\min}$ and $p^{\max}$ are applied only during generation of the initial population.

The values m and n, define the location of CP over the x, y plane as follows. Let's suppose that the extend of the aquifer in the x direction is l. Then an interval distance dl is defined as: $dl = 2l/\left[(m-3)\cdot 2+2\right]$, where m is the number of control points in the x direction. Then, the coordinates of the control points span from $-dl/2$ to $l+dl/2$ with discretization distance of dl. A similar formula is used in the y direction. This formula has the effect of placing the surface at the center of the characteristic polyhedron.

Once N_{pop} individuals are generated, they are evaluated based on the fitness criterion (2). After evaluation, N_{pop} pairs of parents are selected for mating to produce next generation. Since the model structure controlled by m, n is also a decision variable, models of different complexities are evaluated concurrently and two parents may have different model structures. A method to combine them is developed below.

Let's suppose that two parents a and b, with m^a, n^a and m^b, n^b CPs respectively, are chosen to produce an offspring with m^o, n^o CPs. Note that each individual may be composed by different number of CPs. Also the offspring may have different CP than its parents (i.e. $m^a \neq m^b \neq m^o, n^a \neq n^b \neq n^o$). In order to combine the two parents, we apply a transformation where both parents are converted to individuals with the same number of control points as the offspring, (i.e. m^o, n^o) without altering essentially the shape of their initial surface.

Let's suppose that surface S_a defined by the values of matrix $\mathbf{P}_{s,t}$ corresponding to the elements m^a, n^a of parent a. In order to rebuild surface S_a controlled by a different number of CPs (m^o, n^o), one needs to calculate new values of matrix $\mathbf{P}^{a,o}_{s,t}$ based on m^o, n^o CPs. The notation $\mathbf{P}^{a,o}_{s,t}$ represents the BSS matrix, which describes the surface S_a of parent a using the discretization of offspring o. In this transformation, the $p^{a,o}_{i,j}$ values corresponding to the offspring control points, must be evaluated.

By extending (3) we obtain for surface S_a based on the new offspring parameterization:

$$\begin{aligned} z^{a,o}(u,w) = {} & U_1 W_1 \, p^{a,o}_{s+1,t-1} + U_1 W_2 \, p^{a,o}_{s+1,t} + U_1 W_3 \, p^{a,o}_{s+1,t+1} + U_2 W_1 \, p^{a,o}_{s,t-1} + U_2 W_2 \, p^{a,o}_{s,t} \\ & + U_2 W_3 \, p^{a,o}_{s,t+1} + U_3 W_1 \, p^{a,o}_{s-1,t-1} + U_3 W_2 \, p^{a,o}_{s-1,t} + U_3 W_3 \, p^{a,o}_{s-1,t+1} \end{aligned} \tag{5}$$

where $U_1 = \frac{1}{2}(u^2 - 2u + 1)$, $U_2 = \frac{1}{2}(-2u^2 + 2u + 1)$, $U_3 = \frac{1}{2}u^2$, $W_1 = \frac{1}{2}(w^2 - 2w + 1)$, $W_2 = \frac{1}{2}(-2w^2 + 2w + 1)$, $W_3 = \frac{1}{2}w^2$. The $p_{i,j}^{a,o}$ values for the new CPs, are estimated so that for different pairs of u, w, the $z^{a,o}$ values obtained by (5) match the points z^a of the original surface S_a. This is possible by generating values for a set of pairs between (u, w) and z^a for a dense grid covering the entire surface, and matching these values with the values produced by equation (5) based on the offspring parameterization. Let $\mathbf{Z}^a$ be the values corresponding to the original (parent) surface and $\mathbf{Z}^{a,o}$ the values corresponding to the trasnformed surface at the same set of points. Equation (5) is linear and can be written in a vector –matrix form as: $\mathbf{Z}^{a,o} = \mathbf{A} \cdot \mathbf{P}^{a,o}$. Matrix $\mathbf{A}$ depends on terms U, W of new (offspring) parameterization and matrix $\mathbf{P}^{a,o}$ contains the $p_{i,j}^{a,o}$ values at the new CPs. Then the original (parent) values are expressed as $\mathbf{Z}^a = \mathbf{A} \cdot \mathbf{P}^{a,o} + \mathbf{e}$, where $\mathbf{e}$ is an error vector. The $p_{i,j}^{a,o}$ values of the new parameterization can now be calculated by a linear least squares fit that minimizes the mean squared error of $\mathbf{e}$. According to the LLS method, the values $\mathbf{P}^{a,o}$ of the new CPs are calculated by the following well known expression: $\mathbf{P}^{a,o} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Z}^a$. The number of points selected for fitting, depends on the number of CPs of the parent. When the parent is more complex, more points are selected for matching in the LLS transformation.

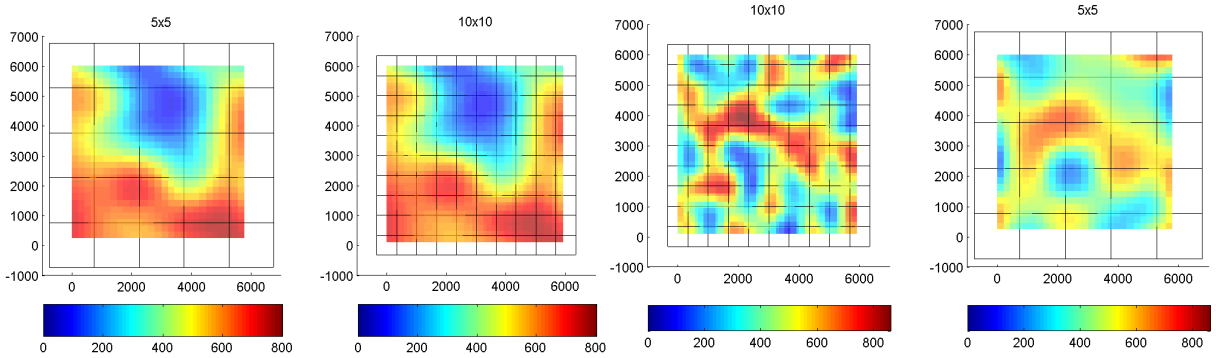

Figure 2. Left: Transformation of a surface from 5x5 to 10x10.
Right: Transformation of a surface form 10x10 to 5x5.

Figure 2 illustrates the transformation of a surface using the proposed method. Notice that when the number of CPs is increased (fig.2, left) the new surface is almost identical to the original. On the other hand, when the number of CPs is decreased, there is a noticeable loss of information, however the general characteristics of the initial surface are retained (fig.2, right).

In order to combine two parents a and b, with $\mathbf{P}^a$ and $\mathbf{P}^b$ CPs, we first convert them into vectors $\mathbf{P}^{a,o}$ and $\mathbf{P}^{b,o}$ of the offspring parameterization. While the new number of parameters of parents a and b are now different than their initial number, the resultant transmissivity maps are akin to the original ones S_a and S_b. The resultant complexity is now equal to $m^o \times n^o$ for both parents, therefore using conventional GA operators, the parents exchange

their transformed genes to produce offspring. After crossover, each individual undergoes mutation. In this step, random perturbations are introduced for each individual with a very small probability. The new population is evaluated and the optimization terminates when stopping criteria are met.

5 APPLICATIONS AND CONCLUSION

The proposed methodology is applied to a hypothetical aquifer of orthogonal shape, where it is assumed that there are $N_{obs} = 40$ observation points distributed in the aquifer. Detailed description of the hypothetical example can be found in Tamer Ayvaz et al. (2007)[6]. As discussed above, the model parameters (model complexity) varied among solutions during optimization, while the maximum allowable number of control points was set equal to $m^{\max} = n^{\max} = 11$. The goal is minimization of observation error (2). The model complexity is expressed here as the number of parameters of the model, however, in general other measures of model complexity could be selected.

In order to test the efficiency of the method, three different transmissivity maps are generated. The first test is based on an isotropic aquifer, the second is anisotropic and the last has a zonation partition. The data used in calibration are generated by the aquifer model and are assumed to be error free. The 40 observation points are randomly uniformly distributed over the aquifer.

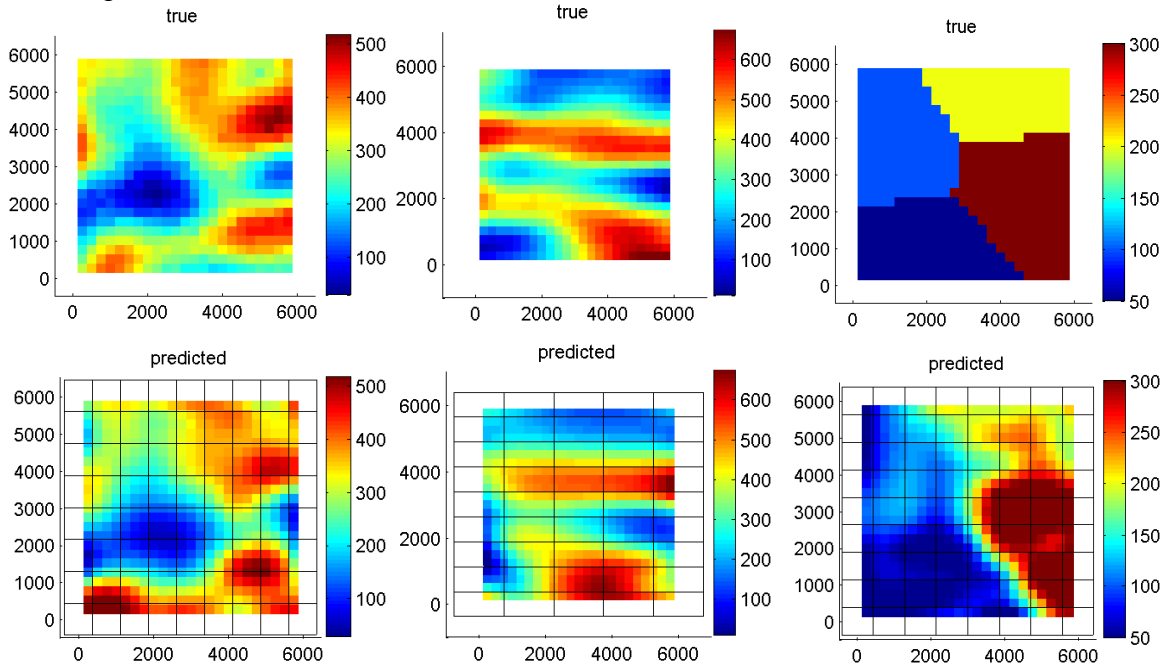

Figure 3 Comparison of true vs. predicted hydraulic field for isotropic, anisotropic and zone aquifer.

Figure 3 summarizes the results of the three tests. In all three cases the estimated transmissivity is similar to the true field. The complexity of the optimum solution in the

isotropic case was $m=10, n=9$. The positions of the CPs of the estimated field are also plotted (correspond to intersections of the grid lines). Notice that in the isotropic case the algorithm converged to a solution with $m \approx n$. In the anisotropic case, the algorithm converged to a solution with $m=6,\ n=10$. It is interesting that the algorithm converged automatically to a solution with $m < n$, which is in accordance with the stronger anisotropy along y than the on x axis. The model complexity in the last zonation application was found equal to $m=9,\ n=10$.

Notice that the algorithm did not converge to the most complex solution in all three cases. This is possibly because optimization terminated before full convergence of the algorithm (the population size was set equal to 100 and the optimization terminated after 400 generations, i.e. the maximum number of function evaluations, is equal to 40000). Notice that if the algorithm is allowed to run for too long, it is expected to converge to models of the highest allowable complexity. Such models would minimize observation error, but risk memorizing the data and will not have a good predictive power, particularly in cases of significant measurement error.

In order to avoid over fitting the data with too complex models, early stopping of the algorithm is advised. Alternatively one could use additional criteria penalizing overly complex models[7].

REFERENCES

[1] J. Carrera, A. Alcolea, A. Medina, J. Hidalgo and L.J. Slooten, "Inverse problem in hydrology", *J. Hydrogeol.,* 13, 206-222 (2005).

[2] A. Mantoglou, "On optimal model complexity in inverse modeling of heterogeneous aquifers", *Journal of Hydraulic Research*, 43(5) : 574-583, (2005).

[3] G. Greiner and H. Seidel, "Modeling with Triangular B-Splines." *IEEE Computer Graphics and Applications,* 14(2), 56-59 (1994).

[4] K. Kato, "Generation of n-Sided Surfaces Patches with Holes" *Computer Aided Design,* 23(10), 676-683 (1991).

[5] M. E. Mortenson, "Geometric Modeling", John Wiley & Sons, New York (1997)

[6] M. Tamer Ayvaz, H. Karahan and M. M. Aral, "Aquifer parameter and zone structure estimation using kernel-based fuzzy c-means clustering and genetic algorithm" *Journal of Hydrology*, 343, 240-253 (2007).

[7] M. Ye, P. D. Meyer, and S. P. Neuman, "On model selection criteria in multimodel analysis", *Water Resour. Res.*, 44, 3, doi:10.1029/2008WR006803 (2008).