Fitting a Deterministic PZN Model to Remote-sensed Chlorophyll Data using a Genetic Algorithm and Self-organizing Map

R.A. Cropp, A.J. Gabric and R.D. Braddock.

Faculty of Environmental Science, Griffith University, Nathan, Queensland, Australia, 4111.
(R.Cropp@mailbox.gu.edu.au)

Abstract: A genetic algorithm (GA) was used to fit a deterministic phytoplankton-zooplankton-nutrient (PZN) model to a time series of remote-sensed chlorophyll. A self-organizing map (SOM) was used to visualise the fitness landscape of the parameter space searched by the GA. These techniques complement each other, as the GA operates by breeding more offspring in regions of high fitness, and the SOM recognises these clusters in high-dimensional space and maps them to low-dimension space for easy visualisation. The fitness landscape of this optimisation was found to be extremely rugged and precipitous, with several areas of high fitness that contained multiple optima. A three-stage optimisation process located an optimum, but doubt remains that it is a global optimum.

Keywords: Genetic algorithm; Self-organizing map; Data fitting; *PZN* model.

1. INTRODUCTION

Satellite ocean colour imagery is currently the only source of long-term time series of chlorophyll in many areas of the world. Fitting deterministic population models to this data has several advantages over empirical models that do not consider the underlying ecological processes. Deterministic models can estimate ecosystem parameters, and may provide parsimonious descriptions of data [Solow, 1995].

Genetic algorithms (GAs) are nonlinear optimisation techniques based on the principles of Darwinian natural selection [Holland, 1975; Mitchell, 1997], and have been used successfully to solve difficult optimisation problems. GAs evolve a population of binary 'chromosomes', comprised of the concatenated parameter values. GAs evaluate the 'fitness' of each point according to a defined fitness function, which then reproduce in proportion to their fitness, with mutation and exchange of genetic material producing a fitter population. GAs do not require any derivative information of the problem being optimised, but converge to an optimum by sampling more intensely near regions of high fitness. Clusters of samplings in the parameter space should therefore indicate regions of high fitness.

A self-organizing map (SOM) is an unsupervised neural net that positions prototype vectors on a regular low-dimensional grid in an ordered fashion, making it a powerful visualization tool [Kohonen, 1997]. SOMs look for similarities between vectors, and can be efficient methods for identifying clusters of similar data. They are therefore a useful adjunct to GAs that generate clusters of data in regions of high fitness, and sparse data in regions of low fitness.

We use a GA to fit a deterministic PZN (P=phytoplankton, Z=zooplankton, N=nutrient) model to a time series of remote-sensed chlorophyll values in the Southern Ocean. A SOM is used to assess the 'topography' of the 'fitness landscape' to give some indication of the likelihood that we have found a global optimum.

2. METHODS

The PZN model includes five biological rate parameters, an available nutrient concentration, and two parameters describing a physical forcing on the phytoplankton growth rate.

$$\frac{dP}{dt} = fk_4\left(\frac{N}{N+k_5}\right)P - k_1PZ \qquad (1)$$

$$\frac{dZ}{dt} = k_1(1-k_3)PZ - k_2Z \quad (2)$$

$$\frac{dN}{dt} = k_1k_3PZ + k_2Z - fk_4\left(\frac{N}{N+k_5}\right)P \quad (3)$$

$$f = \frac{1}{2}\left([1+k_7]+[1-k_7]\cos\left(\frac{2\pi(t-k_8)}{365}\right)\right) \quad (4)$$

It is clear from equations (1-3) that the model is closed with respect to mass (*No = P+Z+N*). The physical forcing (*f*) represents a gross approximation of the effect of seasonal changes in sea surface temperature, irradiance and mixed layer depth on phytoplankton growth rates. Mixed layer depth is primarily a function of temperature profiles in the upper ocean and wind stress on the ocean surface.

The model's parameter space is centred on parameter values reported for marine plankton systems, with ranges defined as ± 50% of the reported values (Table 1). The parameter values have been converted to a fixed 20 m deep mixed layer, approximating depths viewed by remote optical sensors [Robinson, 1995]. The model behaviour within this parameter space has been investigated by Cropp and Gabric [2001].

Table 1. Parameter values for the *PZN* model.

Par.	Description	Min.	Max.
k_1	*Z* grazing rate^1	0.003	0.009
k_2	*Z* higher predation^2	0.025	0.075
k_3	*Z* assimilation eff.^{3}	0.200	0.600
k_4	*P* growth rate^2	0.450	1.350
k_5	*P* half-sat. constant^4	137	414
k_6	*No* (total nutrient)^4	500	3,000
k_7	Minimum forcing^3	0.200	0.700
k_8	Day of forcing max	0	365

^{1} m^2 mg at N^{-1} day^{-1}; ^{2} day^{-1}; ^{3} dimensionless
^{4} mg at N m^{-2}.

The steady state phytoplankton population size is given by:

$$P_{eq} = \frac{k_2}{k_1(1-k_3)} \quad (5)$$

The resilience of the model steady state, defined as the negative of the real part of the dominant eigenvalue of the linearised system [DeAngelis, 1980] is given by:

$$Res = \frac{k_4}{2}\left(\frac{k_5}{(N_{eq}+k_5)^2}\right)P_{eq} \quad (6)$$

The study region was the area of the Southern Ocean at 45-50^{0}S, 123-145^{0}E. A large sampling area was selected to minimise the influence of water from different masses on the time series. Our study region usually contains water from the Subtropical Convergence Zone and the Antarctic Polar Front, which mix in the region [Griffiths et al., 1999]. The maximum sea surface temperature in the study region occurs around January 17 (simulation day 116). Monthly mean wind speeds vary between 11 ms^{-1} in January to 7 ms^{-1} in June. Mixed layer depths vary from approximately 100 m in summer to about 500 m in winter. The euphotic zone depth is fairly constant throughout the year at about 90 m, and dissolved nitrate have been measured in the range 30-160 mg at N m^{-3} [Gabric et al., 1995].

A time series of chlorophyll concentrations was derived from three years of SeaWiFS 'weekly' (8 day) Standard Mapped Images. These images each have 14,000 pixels in the study region. The value for each point in the time series is the average of the cloud-free pixels in the study region for that 'week'. Cloud is a major problem for satellite measurement of ocean colour in the Southern Ocean; the number of cloud-free pixels in the study region ranged from 12 % in winter to 69 % in summer. The time series was commenced at the austral vernal equinox (September 23), just prior to the phytoplankton spring bloom.

Least squares estimators (7) are commonly used as maximum likelihood estimators when fitting deterministic models to time series data [Fasham and Evans, 1995; Solow, 1995].

$$L.S.E. = \sum_{i=1}^{N}\left[y_i - y\left(t_i, \underset{\sim}{k}\right)\right]^2 \quad (7)$$

Press et al. [1997] suggest minimising a χ^2 statistic (8) if the standard deviations of the data points are not constant.

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{y_i - y\left(t_i, \underset{\sim}{k}\right)}{\sigma_i}\right)^2 \quad (8)$$

This statistic allows an estimate of the statistical significance of the fit. Press et al. [1997] note however that large standard deviations σ_i associated with the data points y_i can cause a poor fit to be statistically significant.

The satellite data set has large variance (coefficients of variation of 16-88%) due to the effects of cloud and the large area study region, suggesting that a χ^2 statistic alone may be misleading. Consequently, a least squares and a χ^2 estimator were implemented, and compared, as measures of fit

The initial optimisations using the GA searched the parameter space investigated by Cropp and Gabric [2001]. Each point in parameter space visited by the GA was used to integrate the PZN model for two years. The fitness was calculated from the goodness of fit of the second year of the integration to the data. This reduced the influence of transient dynamics on the fitting procedure. The unforced steady state values of P, Z and N for each parameter set were used as initial conditions for the model integrations.

The GA was configured with a population of 30 individuals, each represented as one 80-bit chromosome composed of the eight parameters. The population was evolved for 50 generations, using initial and final mutation and crossover probabilities of 0.010-0.005 and 0.750-0.975 respectively. The probabilities were varied during the simulation according to a power law. The reproductive success of individuals was implemented using sigma-scaled Monte Carlo selection to prevent premature convergence of the GA [Mitchell, 1997]. Sigma scaling controls the relative reproductive success of individuals within the population. It maintained the probability of reproductive success of an individual with a fitness two standard deviations below the mean at 3 times that of an individual two standard deviations above the mean.

The optimisation process was implemented in three stages. Initially the GA was configured to undertake an extensive search of the parameter space, so that an appreciation of the sensitivity of the estimators to the various parameters could be obtained. Each estimator was implemented 10 times by the GA. The coefficients of variation of the mean parameter values of the optimum parameter set found in each implementation were used as indicators of the sensitivity of the estimators to each of the parameters.

The second stage involved holding the sensitive parameters constant at the values obtained in the first stage, and searching for optima controlled by the other parameters. These optima were resolved in the third stage, in which further-restricted, adaptive parameter spaces were defined and evolved by the GA. This approach allowed us explore the parameter space and infer the topography of the fitness landscape while fitting the *PZN* model to the data.

3. RESULTS

The initial simulations indicated that both estimators were similarly sensitive to the parameters, and estimated similar values for the most influential parameters (k_1, k_2, k_3, k_7 and k_8 - Table 2). χ^2 values less than 16 are significant at p = 0.001.

Table 2. Average values and coefficients of variation (%) of average optimum fit of ten optimisations.

Par.	L.S.E.		χ^2	
	value	c.v.	value	c.v.
k_1	0.0035	13.1	0.0036	13.1
k_2	0.0656	12.0	0.0641	10.9
k_3	0.5263	9.8	0.5309	10.5
k_4	0.8876	33.2	0.9095	23.2
k_5	285	28.3	301	19.3
k_6	886	31.8	1575	49.1
k_7	0.3187	14.7	0.4089	25.3
k_8	229	3.7	216	9.1
P_{eq}	39.48	0.96	37.74	1.2
Fitness	1180	14.1	3.78	21.7

It is obvious from Table 2 that the optimisation procedure is really fitting the model equilibrium phytoplankton value to the average data phytoplankton value (39.37). The variation in k_1, k_2, and k_3 is a result of there being many combinations of these parameters in the vicinity of the optimum values that result in $P_{eq} \approx 39.37$.

The resilience of the model steady state is a property that might also be expected to influence the fitting process. Models with low resilience parameter sets will exhibit larger 'predator-prey' cycles, for longer duration, in the model integrations. These may influence the fitness estimation, as only one year of integration is allocated for these oscillations to decay. Analysis of the GA output revealed a negative correlation ($r = -0.489$, $p < 0.001$) between the L.S.E. estimator and the maximum resilience (i.e. regions of good fit (high fitness) were usually associated with high resiliences). This correlation was not however evident within the regions of high fitness.

The second stage of the optimisation process was implemented on a reduced parameter space of only five dimensions, with k_1, and k_3 at their minimum and maximum values respectively, and *k2*, selected so that $P_{eq} = 39.37$ (Table 3).

These data indicate that k_8 is highly influential once k_1, k_2, and k_3 are 'optimised'. GA runs to find the optimum parameter sets were then implemented on a four dimensional parameter space, after fixing k_8 =225 and holding k_1, k_2, and k_3 as before. Figure 1 shows that over the majority of the parameter space the L.S.E. and χ^2 estimators are highly correlated.

Table 3. Average values and coefficients of variation (%) of average optimum fit of ten optimisations with k_1, k_2, and k_3 fixed.

Par.	L.S.E.		χ^2	
	value	c.v.	value	c.v.
k_1	0.0030	-	0.0030	-
k_2	0.0472	-	0.0472	-
k_3	0.6000	-	0.6000	-
k_4	1.0097	29.6	1.0073	28.6
k_5	252	34.5	286	28.4
k_6	896	18.5	1261	48.2
k_7	0.3680	20.5	0.4296	18.6
k_8	229	1.7	224	2.3
Fitness	973	15.0	3.03	16.7

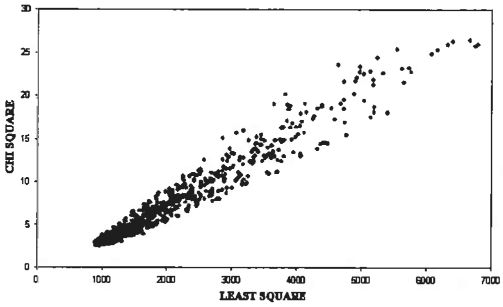

Figure 1. Correlation between sum squares and chi squared of all points visited in the reduced parameter space (r = 0.985, p << 0.001, n = 6000).

The region of high fitness (low L.S.E. and low χ^2) in Figure 1 is expanded in Figure 2. This strongly suggests that there are at least two near-equal L.S.E. fitness peaks in the reduced parameter space that can be distinguished by their χ^2 estimators. Figure 2 also implies that in this region of high fitness, the estimators are not necessarily positively correlated.

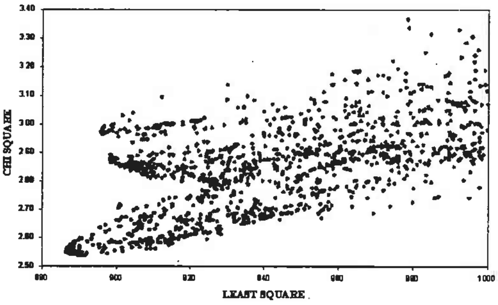

Figure 2. Correlation between sum squares and chi squared of points with L.S.E. < 1,000 in the reduced parameter space (r = 0.553, p << 0.001, n = 4438).

The parameter sets of the peaks, [A] evidenced by low L.S.E. and low χ^2, and [B] with low L.S.E. and relatively high χ^2, were obtained from inspection of the GA data (Table 4).

Table 4. Parameter values of peaks [A] and [B] (k_1, k_2, k_3 and k_8 were fixed).

Par.	A	B
k_1	0.0030	0.0030
k_2	0.0472	0.0472
k_3	0.6000	0.6000
k_4	0.5520	0.703
k_5	156	177
k_6	595	693
k_7	0.342	0.342
k_8	225	225
L.S.E.	890	899
χ^2	2.54	2.86

A SOM of this data (Figure 3) clearly reveals clustering of sampling points. The clustering of vectors in the distance matrix (top left figure, dark regions indicate a high density of points), correspond with the Fitness mapping (bottom right, dark indicates highest fitness). All maps have the same axes, so points in the same place in each of the maps represent the same data vectors. The other four maps are of the parameters k_4, k_5, k_6 and k_7 (left to right, top to bottom).

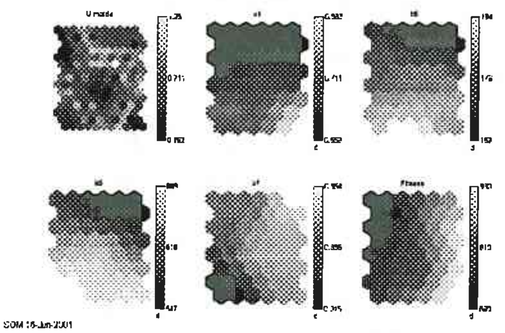

Figure 3. SOM of best 3,000 data from Fig 2.

The SOM clearly shows three fitness peaks (clusters of dark pixels in the top left, centre and bottom right of the upper left map), for which fitness values of 890, 900 and 930 respectively can be inferred from the fitness (bottom right) map. This SOM is of the highest fitness region of parameter space only. The SOM indicates that there may be several small distinct regions of high fitness in the reduced parameter space, as indicated by the individual dark pixels in the distance matrix plot. The 'fitness landscape' of this optimisation problem appears to be highly undulating, suggesting it will be difficult to find a global optimum.

The final stage of the optimisation procedure was implemented by using the GA to search reduced, adaptive parameter spaces, of the original eight dimensions, around each of the peaks A and B. The new parameter spaces were defined as ± 5% of the parameter value of each peak. The parameter space was updated after each iteration to ± 5% of the new parameter values for each new optimum found by the GA, until the GA had converged. This approach allowed the GA to traverse a 'ridge' in the area of high fitness around each peak until it arrived at the optimum for that region, even if it lay outside the initial parameter space.

The convergence characteristics of the GA in this final optimisation (Figure 4) suggest that there are in fact two distinct optima in this small region of parameter space. In both cases, the GA converges linearly toward the optimum, always maintaining a distinction between the two peaks. The adaptive parameter spaces used would have allowed the two peaks to converge to the same optimum.

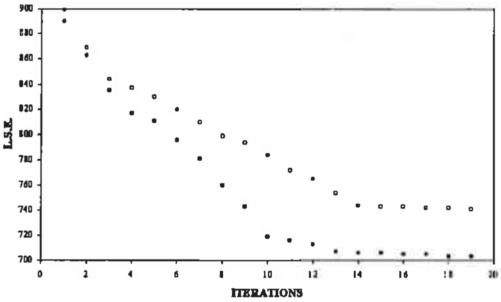

Figure 4. Convergence of GA in final optimisation parameter spaces (Peak A is represented by solid circles; Peak B by open circles).

The parameter values of the optima in Figure 5 are listed in Table 5:

Table 5. Parameter values of optima converged to from peaks [A] and [B].

PAR	A	B
k_1	0.00303	0.00279
k_2	0.03285	0.03426
k_3	0.72456	0.68926
k_4	0.71363	0.87319
k_5	153	157
k_6	660	769
k_7	0.41730	0.40427
k_8	228	230
L.S.E.	703	741
χ^2	2.28	2.74

The parameter sets of the two optima differ primarily in the values of k_4 (maximum phytoplankton growth rate) and k_6 (total nutrient). Although they have different parameter sets, and different fitness, there is little difference in fit between the actual curves (Figure 5). The different parameter values that generate the curves do however have ecological significance, indicating that similar chlorophyll signals may be generated by different species of phytoplankton.

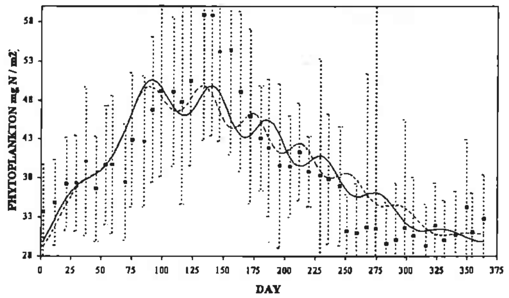

Figure 5. Best least squares fits (Peak A is solid line; Peak B is dotted line). Error bars are ± ½ standard deviation.

A final parameter space was defined by the values of the optima in A and B (i.e. the area of parameter space between A and B was searched). Searches for minimum L.S.E. and χ^2 estimators revealed multiple optima between A and B, and a SOM (Figure 6) confirmed that the fitness landscape is highly undulating. The highest fitness regions (dark areas of the rightmost two maps, bottom row) clearly contain multiple clusters of points (top left map).

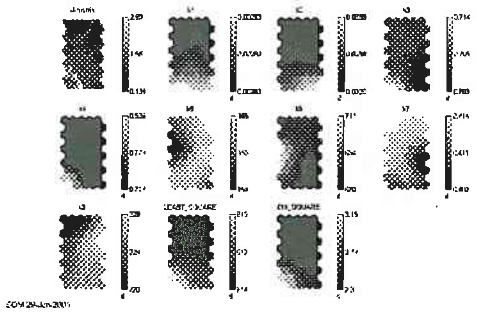

Figure 6. SOM of final parameter space (as for Figure 3, except that all parameters are shown, and both L.S.E. and χ^2 estimators are included as fitness measures).

A search for poor fits in this space revealed L.S.E. values up to 2,300, revealing that there are deep, precipitous 'valleys' of low fitness between the peaks.

4. CONCLUSIONS

This optimisation process has included extensive simulation to determine the topography of the fitness landscape associated with the fitting of the *PZN* model to remote sensed satellite data. The SOM has demonstrated its useful for visualising the topography of the fitness landscape. This problem has been demonstrated to have an extremely rugged fitness landscape, including some precipitous ravines, as nearby peaks of fitness can have deep valleys between them.

Although the optimum *PZN* model fits the data quite well (with a highly significant χ^2, but recall Press et al's [1997] warning) it fails to capture the amplitude of the summer phytoplankton bloom. The seasonal forcing derived for the model also lags the irradiance and temperature forcings by 138 and 112 days respectively. The model seasonal forcing includes these factors plus the effects of mixed layer depth, which correlates positively with temperature and negatively with wind stress. As the maximum wind stress in the study region occurs at about the same time as the temperature maximum, the cycle of mixed layer depths is not obvious, although it is generally shallow in summer and deep in winter in the sub-Antarctic ocean [Gabric et al, 1995].

The nature of the fitness landscape precludes us from any confidence that the optimum we have found is a global one. The rugged topography of the fitness landscape reduces the effectiveness of the GA, as the resolution of the parameters in the GA coding must be very precise, and the searching procedure exhaustive to detect the narrow peaks.

Our optimisations suggest that the most effective approach to fitting a deterministic *PZN* model to our data may lie in modifying the fitness landscape to reduce its ruggedness. The implementation of known variations in mixed layer depth would allow us to exclude physical forcings from the optimisation problem altogether. The question of whether a more complex biological model, with commensurately more parameters to be optimised, would provide a less rugged fitness landscape is moot.

We have attempted to fit the model to the data using dynamics of the model near its steady state. An alternative would be to attempt to fit a transient dynamic of the model to the data. This would require the model initial conditions to be included as parameters. The fitness landscapes of the steady state and transient models are likely to be substantially different, with there being no reason to suspect that the transient landscape would be less rugged.

5. ACKNOWLEDGEMENTS

The authors would like to thank the SeaWiFS Project (Code 970.2) and the Distributed Active Archive Centre (Code 902) at the Goddard Space Flight Centre, Greenbelt, MD 20771, for the production and distribution of the satellite data respectively. These activities are sponsored by NASA's Mission to Planet Earth Program.

The SOM was provided as a toolbox for Matlab 5 by the SOM Toolbox Team, Helsinki University of Technology, (http://www.cis.hut.fi/projects/somtoolbox/).

6. REFERENCES

Cropp, R.A. and A.J. Gabric, Ecosystem resilience: Do ecosystems maximise resilience?, *Ecology*, (in press), 2001.

DeAngelis, D.L., Energy flow, nutrient cycling and ecosystem resilience, *Ecology*, 61, 764-771, 1980.

Fasham, M.J.R. and G.T. Evans, The use of optimization techniques to model marine ecosystem dynamics at the JGOFS station at 47^0 N 20^0 W, *Philosophical Transactions of the Royal Society of London B*, 348, 203-209, 1995.

Gabric, A.J., G.P. Ayers and G.C. Sander, Independent marine and atmospheric model estimates of the sea-air flux of dimethylsulfide in the Southern Ocean, *Geophysical Research Letters*, 22(24), 3521-3524, 1995.

Griffiths, F.B., T.S. Bates, P.K. Quinn, L.A. Clementson and J.S. Parslow, Oceanographic context of the First Aerosol Characterisation Experiment (ACE 1): A physical, chemical and biological overview, *Journal of Geophysical Research*, 104 (D17), 21649-21671, 1999.

Holland, J.H., *Adaptation in Natural and Artificial Systems*, University of Michigan Press, 211 pp, Ann Arbor, 1975

Kohonen, T., *Self-Organizing Maps*, Springer, 426 pp., Berlin, 1997.

Mitchell, M., *An Introduction to Genetic Algorithms*, MIT Press, 208 pp., Cambridge, Massachusetts, 1997.

Press, W.H., S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipes in Fortran 77*, Cambridge University Press, 933 pp., Cambridge, 1997.

Robinson, I.S., *Satellite Oceanography*, John Wiley, 455 pp, Chichester, 1995.

Solow, A.R., Fitting population models to time series data. in Powell, T.M. and J.H. Steele (eds.) *Ecological Time Series*, Chapman and Hall, 20-27, New York, 1995.