ICES J. mar. Sci., 51: 439–445. 1994

GMDH algorithm as a tool for bivalve growth analysis and prediction

Michel R. Claereboudt

Claereboudt, M. R. 1994. GMDH algorithm as a tool for bivalve growth analysis and prediction. – ICES J. mar. Sci., 51: 439–455.

The question of whether growth in bivalves is predictable in terms of environmental conditions is addressed directly by trying to infer juvenile scallop growth from environmental data within and between two locations in the Baie des Chaleurs, Québec. Using models based on either self-organizing models – the group method of data handling (GMDH) algorithm – or on multilinear regressions, scallop growth was found to be predictable. GMDH models lead consistently to better predictions than multilinear regressions and could thus be a useful alternative tool in managing scallop fisheries and aquaculture. Temperature and food availability were the most prominent variables included in the GMDH models, emphasizing their importance as physical determinants of scallop growth.

Key words: GMDH, multilinear regression, growth, environmental effects.

Received 5 May 1993; accepted 14 April 1994

M. R. Claereboudt: Département de Biologie and GIROQ (Groupe Interuniversitaire de Recherches Océanographiques du Québec). Laval University, Québec G1K 7P4, Canada.

Introduction

The increasing importance of bivalves as aquaculture species has led to several attempts to model their growth rates (Grizzle and Lutz, 1989; Ross and Nisbet, 1990). For some species, such as *Mytilus edulis*, a large body of information collected both in the field and in laboratory experiments is available on their physiological ecology. Unfortunately, many of these data are unsuitable for modelling functional dependence as the experiments were not performed with modelling in mind. For other species, such as the pectinid *Placopecten magellanicus*, less information is available and the assessment of growth in natural environments requires several seasons of expensive field experiments. Modelling methods that could forecast the mean growth rate of scallops in a given environment would facilitate greatly the choice of "ideal" growth habitats for a species in an aquaculture context.

The fundamental hypothesis underlying most physiological models is that a set of "state" variables linked to "environmental" conditions completely determine vital rates such as growth or respiration (Ross and Nisbet, 1990). Many "environmental" variables have been identified that influence the growth rate of bivalves, for example temperature, food availability, food quality and size spectrum, fouling of the nets, and stocking density (MacDonald and Thompson, 1985; Wallace and Reisnes, 1985; Grant and Cranford, 1991; Lesser *et al.*, 1991; Côte *et al.*, 1993; Claereboudt *et al.*, 1994). This modelling approach is appealing because it is based on actual biological functions such as clearance rates, ingestion rates, respiration, and energy storage. Unfortunately, the mathematical form of the relations between these biological functions and environmental variables is often incomplete or difficult to express.

In statistical modelling, linear multivariate modelling is often used in dependence analysis (multilinear regressions), although the assumption that the processes involved are linear is rarely met in biology. Non-linear models based on a careful observation and description of the processes have been developed (Grizzle and Lutz, 1989). Nevertheless, they rely on an *a priori* knowledge of the structure of the system, which is often incomplete.

Self-organizing methods of modelling such as GMDH (group method of data handling) (Ivakhnenko, 1968) do not make assumptions about either the mathematical form of the relations or even the conceptual structure of the system being modelled. GMDH models extract the patterns and mimic the data and nothing else. The most evident drawback of GMDH models is that they do not show direct causality links between dependent and independent variables. Such analyses remain essentially empirical tools and as such have been applied to a wide variety of pattern recognition problems, including river-flow prediction (Ikeda, 1984), fisheries (Brooks and Probert, 1984), coral abundance (Green *et al.*, 1987), and economics (Scott and Hutchinson, 1984). Although

1054–3139/94/040439+07 $08.00/0

the GMDH algorithm has been strongly criticized (Green *et al.*, 1988), it remains useful in applications in which the exact form of the model is not known or varies from case to case (D. G. Green, Australian National University, Canberra, Australia, pers. comm.).

The present study examines the use of self-organizing modelling as a tool in analysing bivalve growth in a variable environment. Short-term growth rates of juvenile *Placopecten magellanicus* were related to environmental parameters using both the GMDH algorithm and the more traditional multilinear regression.

Methods

Data set

At two locations along the north shore of Baie des Chaleurs (Québec, Canada), at Grande-Rivière, situated at the entrance to the bay, and at Gascons, ≃40 km inside the bay, we monitored the growth of juvenile scallops in pearl nets from mid-June to late October 1991, and concurrently measured the development of fouling on the nets (Claereboudt *et al.*, 1994). The 1.5-year-old scallops were individually marked with small (2 × 3 mm) tags glued to the inferior valve with 5-min epoxy and placed in pearl nets suspended from a long line at 9, 15, and 21 m below the surface. The scallops initially measured 20 to 38 mm in shell height and were stocked at densities of 28 individuals per net. At approximately monthly intervals, the scallops were collected by Scuba and transferred to seawater aquaria. Their shell height was measured to the nearest 0.1 mm using electronic calipers and they were then returned to the grow-out site in their original pearl net. For each individual, the daily growth rate between consecutive samplings was calculated. To manage the unavoidable individual variability in growth and size of the scallops and to reduce the cost of handling numerous pearl nets, scallops were grown in four pearl nets per experimental condition: two with high fouling development and two with low fouling development (changed at each sampling). This procedure resulted, unfortunately, in a strong risk of pseudoreplication between individual scallops; therefore, all growth increments from a single pearl net were averaged (Hurlbert, 1984).

At each site, water temperature was continuously recorded from June to November at the three experimental depths using Ryan thermographs, and weekly samples of the water column at the experimental depths were analysed for seston concentration and chlorophyll *a* content in two size classes (cells<5 µm and cells >5 µm). Samples of water were filtered on GF/F fibreglass filters and chlorophyll *a* concentration was determined by fluorometry after 24-h extraction in 90% acetone. Total seston was estimated by the difference in mass of dried fibreglass GF/F filters before and after filtering 250-ml water samples. NaCl was removed by flushing the filters with isotonic ammonium formiate. The total fouling of the pearl nets was estimated by subtracting the mass of a new clean net from the mass of each of four fouled nets immersed at the experimental depths for 1, 2, 3, and 4 months, respectively.

Data preprocessing

Correlation between all pairs of variables was first tested to avoid possible combinations of variables that would cause singular matrices to arise during the GMDH procedure. Further, the dataset was examined for possible serious violations of the assumptions of normality and homoscedasticity of the data. A non-normal distribution of the residuals from the multilinear model indicated the need for a transformation. The square-root transformation was applied to the growth data and improved the distribution of the residuals. The environmental data were averaged over the corresponding period of scallop growth.

Model

GMDH can be best described as a non-parametric learning algorithm (Green *et al.*, 1988). The GMDH procedures generate polynomials of extremely high degree of the original independent variables to mimic the variations of the dependent variable. It is generally not possible to expand the model in full polynomial form since this could require literally thousands of terms. However, the model is well represented as a network of simple submodels structured in the form of a pyramid. (Fig. 1) and leading to successively better estimated values of the dependent variable, y. The reference function $R(x_1, x_2)$ is a simple function that provides the basic "building block" of the GMDH algorithm. Each submodel uses the same reference function with a different set of values for the parameters. In our model, the reference function was a polynomial of the form (Farlow, 1984b; Green *et al.*, 1988):

$$y = ax_1^2 + bx_2^2 + cx_1x_2 + dx_1ex_2 + f.$$

The pyramid of reference functions is created by iteration (one for each level of the pyramid) as follows. The data points are first distributed randomly into either a learning set or a checking set. In the first iteration, each pair of independent variables is tested sequentially. For each pair x_i, x_j, the reference function is fitted to the learning set by a multiple regression which sets the six parameters (a, b, c, d, e, and f). The expected values of y are then calculated for the checking set and the coefficient of determination, D, is computed as follows:

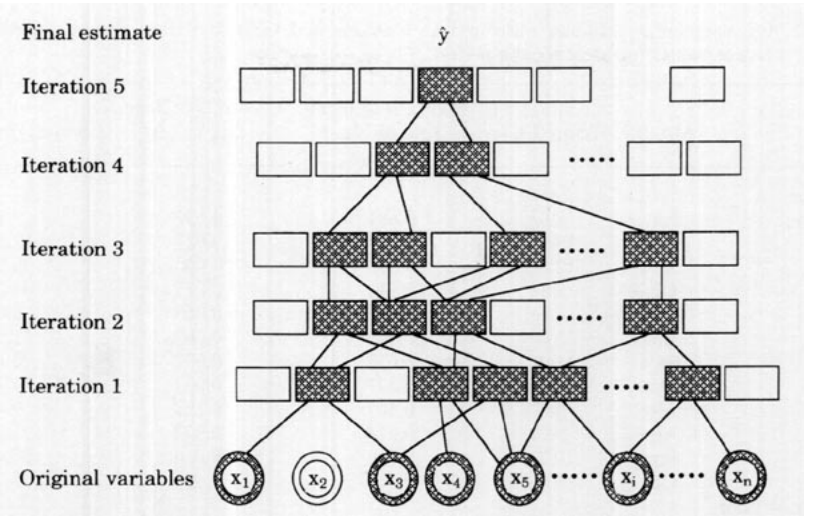

Figure 1. Schematic diagram a GMDH model. Each box represents a submodel (polynomial of degree 2). The shaded boxes represent the polynomials contributing to the final model.

$$D^2 = \frac{\sum_{i=nt+1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=nt+1}^{n} y_i^2},$$

where $\hat{y}_i$ is the predicted value associated with the actual value y_i of the dependent variable; n is the total number of observations, and nt the number of observations in the learning set. If D is less than some predetermined value (the regularity criterion), the fitted reference function for the pair i, j is stored, otherwise it is discarded. In our model, the regularity criterion was adapted at each iteration to keep the 20 polynomials that had the lowest D^2. This procedure allowed a shorter processing time without decreasing the GMDH performance (Green *et al.*, 1988). When all pairs of variables have been processed, the first iteration is complete and a maximum of $(k/2) = k(k-1)/2$ regression polynomials are stored, where k = the number of variables. In the next iteration, the original training data is replaced by the predicted values given by each one of the stored reference functions. The iterative process continues until no further improvement in the predicted values can be achieved. The reference function of the last iteration that gives the best fit (the minimum D^2) is chosen as the final submodel at the top of the pyramid. As many of the polynomial blocks saved during the iterative process do not contribute to the final model (the white blocks in Fig. 1), they can now be discarded and the remaining polynomials form the pyramidal structure of the GMDH model (shaded in Fig. 1). Even though the procedure is simple, the computing time can become exceedingly long, especially in small computers. The nine iterations of our models, including five variables and 20 polynomials at each iteration, took approximately 35 min on a Macintosh SE/30. A complete description of the basic model and some improvements can be found in Farlow (1984b) and Green *et al.* (1988), a FORTRAN listing of the algorithm in Farlow (1984a), and a critical review of the model's behaviour in Green *et al.* (1988).

A GMDH program was written in PROGRAPH for the Macintosh environment. This graphic programming environment was chosen for its object-oriented structure based on dataflow rather than variables, allowing an easy representation of the network structure of the GMDH algorithm as well as a variable-sized stack implementation. All procedures were computed in double precision floating-point arithmetic.

Modelling experiments

Three modelling experiments were conducted. The first compared the modelling abilities of GMDH with that of a standard multilinear regression in terms of distribution of errors and goodness of fit. In the second, we tested the role of each variable in the model resulting from the first trial. Although the variables that affect the prediction of the growth rate are described by the model itself, the next question to be addressed is the relative importance of the different variables. For example, does a typical fluctuation in temperature have more or less effect than a comparable fluctuation in fouling abundance or in suspended particulate matter. Even though the coefficients of the polynomials are entirely determined, the complexity of the resulting model forbids the direct examination of these coefficients. The importance of a given variable in the model was thus assessed using the standard fluctuation method (Brooks and Probert, 1984). Standard fluctuations of each of the variables

Table 1. Environmental variables measured at Gascons and Grande Rivière at the experimental depths. Average values for each growth period.

Depth	Month	Temp. (°C)	Chloro. a <20 µm (mg m^{-3})	Chloro. a >20 µm (mg m^{-3})	Seston (mg l^{-1})
Gascons					
9 m	July	8.0	0.684	0.032	9.1
	Aug	13.5	0.867	0.027	12.3
	Sept	12.0	0.650	0.050	30.3
	Oct	8.1	0.715	0.077	31.2
15 m	July	7.5	0.627	0.068	8.8
	Aug	12.0	0.449	0.062	11.7
	Sept	11.4	0.617	0.027	28.4
	Oct	8.2	0.477	0.037	31.1
21	July	7.1	0.567	0.127	9.4
	Aug	9.3	0.214	0.092	11.2
	Sept	10.2	0.366	0.014	27.4
	Oct	8.1	0.304	0.024	3.1
Grande Rivière					
9 m	July	7.16	0.773	0.027	9.4
	Aug	11.34	0.748	0.074	11.2
	Sept	9.31	0.639	0.307	27.4
	Oct	8.85	0.542	0.058	3.1
15 m	July	7.01	0.507	0.034	9.4
	Aug	11.18	0.690	0.079	11.2
	Sept	8.43	0.496	0.206	27.4
	Oct	8.30	0.428	0.034	3.1
21 m	July	6.89	0.241	0.041	9.4
	Aug	11.06	0.633	0.084	11.2
	Sept	8.02	0.353	0.105	27.4
	Oct	7.09	0.325	0.037	3.1

Table 2. Pearson's correlation coefficients between the variables used in the modelling experiments.

Variable	Fouling	Temp.	Chl. a <20 µm	Chl. a >20 µm	Seston	Growth
Fouling	—	0.22	0.32	0.01	0.38	−0.059
Temperature	—	—	0.57	−0.47	0.18	0.53
Chlorophyll a <20 µm	—	—	—	−0.67	−0.09	0.19
Chlorophyll a >20 µm	—	—	—	—	−0.06	−0.66
Seston	—	—	—	—	—	−0.51
Growth	—	—	—	—	—	—

were introduced in the original data set and the results compared with that of the unmodified data. A standard fluctuation of a variable was defined as ± 2.5% of the total variance recorded for that variable in the data set. The importance of each variable was assessed by the mean-squared errors between the predicted values computed from the modified data set and the predicted values computed from the unmodified data set. The larger the mean-squared error, the more "impact" that variable had on the predicted growth values.

In the last experiment, the predictive abilities of GMDH were compared to that of a standard multilinear regression (Zar, 1984). Both models were trained on one set of data (the growth and environmental data collected at Gascons) and then tested for goodness of fit with predictions of growth based on environmental data collected at Grand Rivière.

Results

General results

The range of environmental conditions is shown in Table 1 for the four periods during which scallop growth was measured. Since all correlations between variables were low, values ranging from 0.01 to −0.67 (Table 2), all variables were included in the data sets used in the model.

The predictions of scallop growth given by a GMDH model are presented along with those given by

Table 3. Analysis of variance (ANOVA) of the multilinear regression of the shell height increments of juvenile giant scallops (*Placopecten magellanicus*) to various environmental variables from June to October 1991. The non-significant factors (F<4) and interactions (P>0.2) were successively removed from the model.

Source of variation	d.f.	Mean squares	F	P
Temperature	1	0.035	24.3	0.0001
Chlorophyll *a* <20 µm	1	0.022	15.5	0.0002
Chlorophyll *a* >20 µm	1	0.065	44.9	0.0001
Total seston	1	0.050	34.6	0.0001
Temp × seston	1	0.094	65.4	0.0001
Chlorophyll *a* <20 µm × seston	1	0.021	14.5	0.0003
Residual	71	0.001		

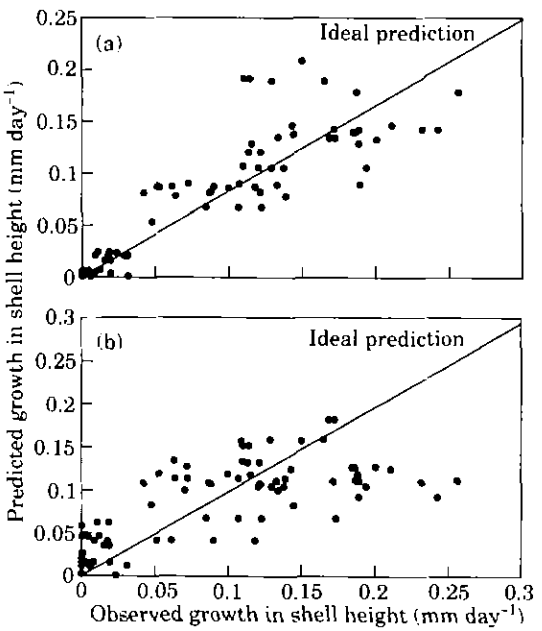

Figure 2. Actual and ideal (lines) predictions of scallop growth of: (a) a GMDH model and (b) of a multilinear regression computed from the same dataset.

multilinear regression computed with the same set of environmental data set Figure 2. The ANOVA table of the multilinear regression (Table 3) indicates that even though most interaction terms were highly non-significant, there were interaction effects between seston and both temperature and large particles. The goodness of fit (multiple correlation coefficient r^2) was greater for the GMDH model (r^2=0.796) than for the multilinear regression (r^2=0.654). The GMDH method did considerably better at modelling low growth rates. In addition, the frequency distribution of the signed errors (predicted value–observed value) followed a normal distribution in the GMDH model (Kolmogorov–Smirnov test p=0.24), but could not be considered as normal (Kolmogorov–Smirnov test p=0.0085) in the multilinear regression (Fig. 3).

Importance of variables

The GMDH algorithm included all five variables (i.e. chlorophyll *a* in large cells and in small cells, total seston, temperature, and fouling) in the model and converged after nine iterations. Figure 4 illustrates the residual mean-squared error generated by standard fluctuations of each variable independently. This technique indicates that temperature and total seston rank first and second in importance in the model.

Predictive abilities

By comparing predictions of growth from an environmental data set for which the model had not been trained beforehand, it is possible to assess its general predictive abilities. Figure 5 compares the predictive abilities of a GMDH model with that of a multilinear regression built on the same reduced data set. GMDH had better predictive power than the multilinear regression. This ability is reflected by the better multiple correlation coefficent (r^2) calculated between observed and predicted values in the GMDH model (r^2=0.77) compared to the multilinear regression (r^2=0.54). In particular, the multilinear regression was unable to reproduce the large range of growth recorded in the field data.

Discussion

In general terms, scallop growth is predictable in terms of environmental conditions and the GMDH models result in coefficients of determination >0.7 between predicted and actual growth values for juvenile scallops. This method of forecasting the growth of a bivalve given a set of environmental parameters appears to be a valuable tool in aquaculture feasibility studies. GMDH does not attempt to identify the causes of the variation observed in the growth rate of scallops in the field but simply to simulate and correlate this variation with environmental variables. As Green *et al.* (1988) pointed out, GMDH models have a tendency to become highly

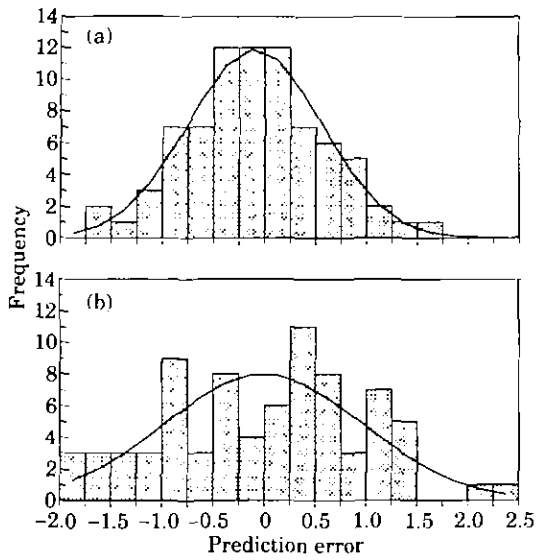

Figure 3. Observed distribution and fitted normal distribution of the signed errors from predictions of: (a) a GMDH and (b) a multilinear regression model computed on the same dataset.

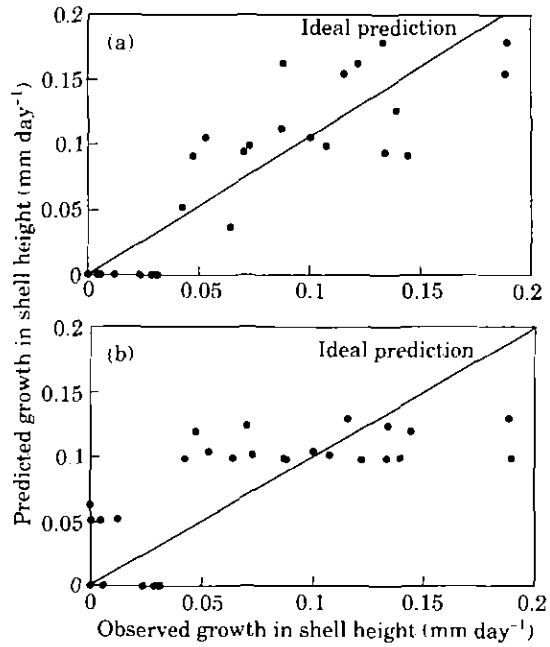

Figure 5. Actual and ideal (lines) predictions of scallop growth of: (a) a GMDH model and (b) a multilinear regression computed from the same dataset. The model was trained with data at Gascons and used to predict growth rates at Grande Rivière.

unstable outside the range of the source data. Therefore, if these models are used in forecasting applications they should be tested thoroughly to ensure adequate extrapolations outside that training range. In its present stage of development, our model was able to simulate successfully growth value from environmental data at Grande Rivière, for which it was not trained, and performed considerably better than a multilinear model in the same conditions, suggesting a better ability to extract relevant patterns in the structure of the data. However, most of the environmental parameters in Grande Rivière were within the range recorded at Gascons. In order to test whether our GMDH model would succeed in predicting growth in an environment outside the range recorded in our summer field studies, we ran the model with environmental values similar to those in the Baie des Chaleurs but with abnormally low temperature values (actual values -2°C). In many cases, GMDH lead to unacceptably high or low values of growth (mean error >150%). This confirms its inability to extrapolate outside the range of the original source data. The use of a larger database (including variables reflecting food quality and wider environmental variations, for instance), as well as the introduction of a better algorithm in the processing of the data (Ivakhnenko, 1984), should provide considerable improvements in the forecasting abilities of GMDH models in the case of scallop growth. It

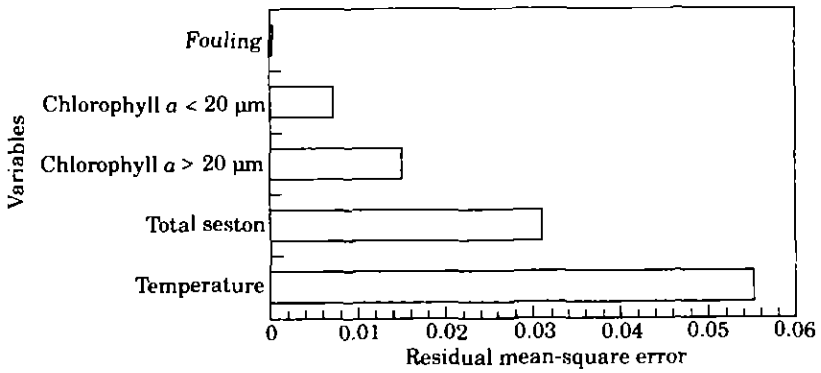

Figure 4. Relative importance of the five environmental variables on the prediction of growth rate expressed as residual mean-squared error generated by standard fluctuations of these variables.

may be likely that statistical models which include higher-order interactions, changes in variables, or specially designed non-linear terms could perform as well or even better than GMDH. However, the functional relationships that are required in the design of such models are not known and may vary from species to species.

GMDH models provide new methods for data analysis. Since the GMDH algorithm indirectly tests for all relationships, including non-linear ones, it may provide ecologist with a mean of objectively testing the importance or the non-importance of some variables in a complex system. Its prime limitation is its inability to represent the actual contribution of each independent variable to values of the dependent variable. However, sensitivity analysis can be performed (Brooks and Probert, 1984) and our GMDH models reflect the documented dependency of bivalve growth on both temperature and food availability (MacDonald and Thompson, 1985; Ciocco, 1991). Furthermore, in many systems, the complexity and the number of all possible interactions between known variables is so high that ecological theory does not provide a method to reduce those variables to a manageable level. Since GMDH tests also for non-linear relationships, it represents an alternative non-linear approach to factor analysis.

The GMDH algorithm is a valid tool to analyse the complex interactions between the environment and the growth of bivalves and seems to be superior to at least the simple multilinear regression. To be applicable on a large geographical scale, it must be based on a wide database representing the extreme environmental variations in which bivalves grow.

Acknowledgements

I thank J. H. Himmelman and D. Monti for their helpful comment in improving the manuscript. This project was supported by a Center of Excellence of Canada grant to the Ocean Production Enhancement Network.

References

Brooks, H. A., and Probert, T. H. 1984. Let's ask GMDH what effect the environment has on fisheries. *In* Self-organizing methods in modeling. GDMH type algorithms, pp. 169–178. Ed. by S. J. Farlow. Marcel Dekker, New York and Basel.

Ciocco, N. F. 1991. Differences in individual growth rate among scallop (*Chlamys tehuelcha* (d'Orb.)) populations from San José Gulf (Argentina). Fisheries Research, 12: 31–42.

Claereboudt, M. R., Bureau, D., Côté, J., and Himmelman, J. H. 1994. Fouling development and its effect on the growth of juvenile giant scallops (*Placopecten magellanicus*) in suspended culture. Aquaculture, 121: 327–342.

Côte, J., Himmelman, J. H., Claereboudt, M. R., and Bonardelli, J. 1993. Influence of density and depth on the growth of juvenile giant scallop (*Placopecten magellanicus*, Gmelin, 1791) in suspended culture in the Baie des Chaleurs. Canadian Journal of Fisheries and Aquatic Sciences, 50: 1857–1869.

Farlow, S. J. 1984a. A Fortran program for the GMDH algorithm. *In* Self-organization methods in modelling. GMDH type algorithms, pp. 277–289. Ed. by S. J. Farlow. Marcel Dekker, New York and Basel.

Farlow, S. J. 1984b. The GMDH algorithm. *In* Self-organization methods in modelling. GMDH type algorithms, pp. 1–24. Ed. by S. J. Farlow. Marcel Dekker, New York and Basel.

Grant, J., and Cranford, P. J. 1991. Carbon and nitrogen scope for growth as a function of diet in the sea scallop *Placopecten magellanicus*. Journal of the Marine Biology Association of the UK, 71: 437–450.

Green, D. G., Bradbury, R. H., and Reichelt, R. E. 1987. Patterns of predictability in coral reef community structure. Coral Reefs, 6: 27–34.

Green, D. G., Reichelt, R. E., and Bradbury, R. H. 1988. Statistical behaviour of the GDMH algorithm. Biometrics, 44: 49–69.

Grizzle, R. E., and Lutz, R. A. 1989. A statistical model relating horizontal seston fluxes and bottom sediment characteristics to growth of *Mercenaria mercenaria*. Marine Biology, 102: 95–105.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs, 54: 187–211.

Ikeda, S. 1984. Nonlinear prediction models for river flows and typhon precipitation by self-organizing methods. *In* Self-organizing methods in modelling. GMDH algorithms, pp. 149–167. Ed. by S. J. Farlow. Marcel Dekker, New York and Basel.

Ivakhnenko, A. G. 1968. Group method of data handling—a rival of the method of stochastic approximation. Sov. Autom. Control, 3: 43–71.

Ivakhnenko, A. G. 1984. Past, present and future of GMDH. *In* Self-organizing methods in modeling: GMDH type algorithms, pp. 105–117. Ed. by S. J. Farlow. Marcel Dekker, New York and Basel.

Lesser, M. P., Shumway, S. E., Cucci, T., Barter, J., and Edwards, J. 1991. Size specific selection of phytoplankton by juvenile filter-feeding bivalves: a comparison of the sea-scallop *Placopecten magellanicus* (Gmelin, 1791) with *Mya arenaria* Linnaeus, 1758 and *Mytilus edulis* Linnaeus, 1758. *In* Proceedings of the 7th International Pectinid Workshop, Vol. 1, pp. 341–346. Portland, Maine.

MacDonald, B. A., and Thompson, R. J. 1985. Influence of temperature and food availability on the ecological energetics of the giant scallop *Placopecten magellanicus*. I. Growth rates of shell and somatic tissue. Marine Ecology Progress Series, 25: 279–294.

Ross, A. H., and Nisbet, R. M. 1990. Dynamic models of growth and reproduction of the mussel *Mytilus edulis* L. Functional Ecology, 4: 777–787.

Scott, D. E., and Hutchinson, C. E. 1984. An application of the GMDH algorithm to economic modeling. *In* Self-organizing methods in modeling: GMDH type algorithms, pp. 243–255. Ed. by S. J. Farlow. Marcel Dekker, New York and Basel.

Wallace, J. C., and Reisnes, T. G. 1985. The significance of various environmental parameters for growth of the Iceland scallop, *Chlamys islandica* (Pectinidae) in hanging culture. Aquaculture, 44: 229–242.

Zar, J. H. 1984. Biostatistical analysis. Prentice-Hall, Englewood Cliffs, New Jersey, 718 pp.