Evaluation of simple statistical criteria to qualify a simulation

Fabrice Mesplé [a], Marc Troussellier [b,*], Claude Casellas [a], Pierre Legendre [c]

[a] *Département Sciences de l'Environnement et Santé Publique, URA-CNRS 1355, Faculté de Pharmacie, Avenue C. Flahault, F-34060 Montpellier, France*
[b] *Laboratoire d'Hydrobiologie Marine et Continentale, URA-CNRS 1355, Université de Montpellier II, F-34095 Montpellier cedex 05, France*
[c] *Département de Sciences Biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Que., Canada H3C 3J7*

Abstract

Statistical and deterministic simulation modelling rely on a complex process made of trials, errors, and gradual improvement of the simulations. The major problem is to be able to quantify the quality of the simulations in order to know if a modification of the concepts, the laws simulating the processes, or the parameters improve it. To try to quantify the quality of simulations using a mathematical criterion we focus on simple linear regression parameters: the values of the slope (a) and the y-intercept (b). The estimated values of these parameters differ depending on which kind of regression model (model I or II) is used. An artificial dataset illustrates that ordinary least-squares regression (OLS; model I regression) leads to results that are not those expected; but using major axis regression (MA; model II regression) instead of OLS leads to the correct answer. The value of a, when it significantly differs from 1, indicates a difference between observed and simulated values proportional to the values of the variable. The value of b, when it significantly differs from 0, indicates a systematic and constant difference between observations and simulations. Taking into account the values of a and b, we define four possible outcomes which allow, at first, to define the quality of a simulation without considering the coefficient of determination, r^2: (i) a n.s.d. (not significantly different from) 1 and b n.s.d. 0 (perfect agreement between observations and simulations), (ii) a n.s.d. 1 and b s.d. 0 (significant constant difference between observations and simulations) or a s.d. 1 and a s.d. 0 and b n.s.d. 0 (differences proportional to the values of the variable), (iii) a s.d. 1 and a s.d. 0 and b s.d. 0 (superimposition of a constant difference and a proportional difference), and (iv) a n.s.d. 0 (no relation between simulations and observations). The value of r^2 is used to rank two simulations pertaining to the same group. That classification of the quality of the simulations is applied to a real-data example: a simulation of the temporal change in chlorophyll a in a high-rate algal pond.

Keywords: Algae; Chlorophyll; Model comparison; Regression models

---

* Corresponding author. Fax: (+33) 67-14.37.19, E-mail: Troussel@Hydrobio.Univ-Montp2.fr

1. Introduction

Statistical and deterministic simulation modelling rely on a complex process made of trials, errors, and gradual improvement of the simulations. As "predictive models (calculation tools) are judged for their usefulness on how accurately they can predict aspects of the real world..." (Loehle, 1983), the major problem is to be able to quantify the quality of the simulations in order to know if a modification of the concepts, the laws simulating the processes, or the parameters would improve it.

Comparing modelling techniques is different from comparing simulations. Comparing several techniques requires to take into account the conceptual foundations of the models and to compare the models on the whole, i.e., considering all the simulated variables and other components such as the number of forcing variables, for instance. The criterion may be a scalar (i.e., a single figure) or better a vector since, according to Halfon and Reggiani (1978), several pieces of information are generally required to compare different models. Estimating a simulation quality, on the other hand, simply consists of comparing the observed and simulated patterns of a variable without questioning the model. It offers the advantage of allowing to compare simulations produced under different models, relying on different mathematical techniques (statistical and deterministic models, for instance: Keller, 1989), or having different numbers of parameters (Robinson, 1986).

This estimation of simulation quality may be carried out qualitatively, i.e., visually (Andersen et al., 1987; Smith and Putz, 1993), but that remains very subjective. "There are many models in the literature which are depicted together with such comparisons but on which it is hard to agree with the authors' optimism regarding similarity with reality". (Straškraba and Gnauck, 1985). It is more satisfying to try to quantify the quality of simulations with the help of a mathematical criterion even if its choice remains somewhat subjective.

Some tests and/or criteria allowing to validate a simulation are described in the literature. Tra-

ditional methods include ordinary least-squares (OLS) regression (Reckhow and Chapra, 1983; Costanza and Sklar, 1985; Straškraba and Gnauck, 1985; Robinson, 1986; Keller, 1989; Mayer and Butler, 1993; Smith and Rose, 1995), chi-square test (Straškraba and Gnauck, 1985; Costanza, 1989; Valentin and Coutinho, 1990) or t–test (Rose et al., 1988) and its nonparametric alternative the Mann–Whitney–Wilcoxon test (Reckhow and Chapra, 1983). More recent techniques may be found in the literature, such as the bootstrap (Halfon, 1989), the multiple resolution procedure proposed by Costanza (1989), the mean squared error of predictions (Wallach and Goffinet, 1989), or the modelling efficiency, regarded by Mayer and Butler (1993) as "...the best overall measure of agreement between observed and simulated values".

Therefore, while several tests or criteria have been proposed in the literature, most are actually difficult to apply, either because they are not sufficiently detailed and documented, or the procedures to implement them are too heavy. This concurs with a statement by Wallach and Goffinet (1989): "The evaluation of model quality is obviously an essential aspect of the modelling activity. However, if one considers the modelling literature, one finds few indications as to how to evaluate a model".

In the present paper, we focus on simple linear regression because it is simple to use and it certainly represents the most widely used modelling technique in ecology. Simple linear regression between observed and computed data, and its associated parameters, are sufficient to check simulation quality but modellers have to consider carefully which simple linear regression model they have to use. First, we develop the concepts underlying the use of simple linear regression to quantify the fit of data computed by several deterministic or empirical (statistical) models. Secondly, we consider OLS, the widely-used model I method of simple linear regression, which, according to many authors (listed above), seems adequate to check simulation quality; we will show that it is not. It has been used for this purpose by many authors (e.g., Keller, 1989; Summers et al., 1991), including ourselves (Mesplé et

al., 1995b,1996). Thirdly, we suggest to use major axis regression (MA), one of the model II methods of simple linear regression, instead of OLS; and finally, we apply this method to examples bearing on real data.

2. The concepts underlying goodness of fit

As noted above, simple linear regression may be used to compare observed and simulated data and to quantify the quality of simulations, as well as its improvement or deterioration. The method is described in all textbooks of biological statistics, including Sokal and Rohlf (1995). The parameters we are interested in are: the slope (a) and the y-intercept (b) of the regression line, the 95% confidence intervals associated with these parameters (if one adopts the widely-used $\alpha = 0.05$ confidence level), and the coefficient of determination (r^2). A simulation perfectly reflecting natural variability would lead to the following linear regression comparing simulated to observed values: $X_{observed} = X_{simulated} + \varepsilon$ that is to say, the parameters of the regression equation $X_{observed} = a \cdot X_{simulated} + b + \varepsilon$ would be: $a = 1$ and $b = 0$, and the angle θ between the regression line and the x-axis would be 45° (Fig. 1a).

So, when comparing a 'good' simulation to observed data, the following should be found:
1. $a \neq 0$, i.e. the slope a is significantly different from 0;
2. $a = 1$, i.e. the slope a is not significantly different from 1;
3. $b = 0$, i.e. the intercept b is not significantly different from 0.

If a is not significantly different from 0 — in other words, if (1) is not true — then the observed data do not provide enough information to help evaluate the simulation, even though (2) may be true. If $a \neq 0$, it tells us something about the discrepancy between observations and simulations, showing on average whether the simulation overestimates ($a > 1$) or underestimates ($a < 1$) the observations proportionally to their value. In order to know if there is a relation between the observations and the simulations, we will test first whether a is significantly different from 0 (H_0: $a = 0$; H_1: $a > 0$) and, if so, whether it significantly differs from 1 (H_0: $a = 1$).

If a is found not to be significantly different from 0, there is no relation between simulations and observations and it is useless to consider b. This situation raises the problem of power, however. When a statistical null hypothesis is rejected, one knows what the probability is of tak-
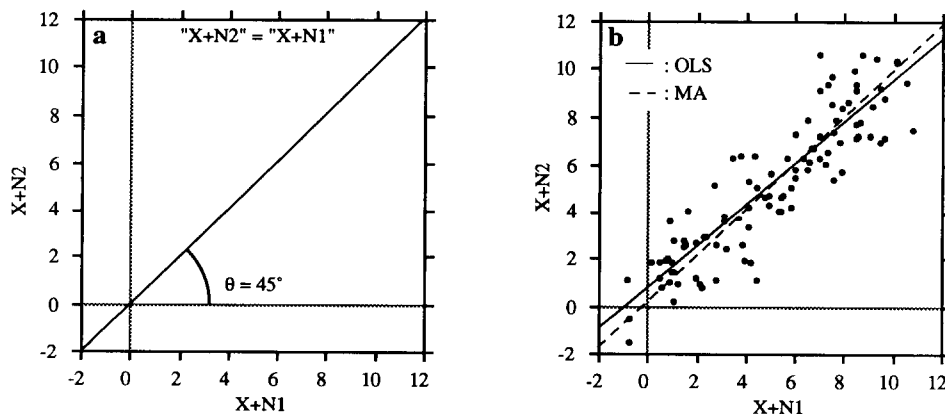

Fig. 1. (a) Regression line reflecting a perfect simulation ($a = 1$, $b = 0$); (b) ordinary least-squares regression line (OLS) ($a = 0.86$, $b = 0.77$) and major axis (MA) regression line ($a = 0.96$, $b = 0.29$) calculated with the artificial dataset.

ing the wrong decision; this is given either by the α significance level of the test, or by the computed probability associated with the test statistic. When the null hypothesis cannot be rejected, however, one does not know the probability of making a mistake if one concludes that, therefore, the null hypothesis must be true. The probability of this kind of error (type II error) is usually referred to as β. Its one-complement, $1-\beta$, is called the power of a test. The power of a test is the probability of successfully rejecting the null hypothesis when indeed it should be rejected. The importance of power analysis in ecological studies has been stressed by Peterman (1990). For the most usual tests of significance, power can be determined by looking up published tables, as in Kraemer and Thiemann (1987) or Cohen (1988). In the present case, it is easy to determine the power of the one-tailed test of the correlation coefficient (r) since published tables are available. A test of $r = 0$ is exactly the same as a test of $a = 0$. A statement about power, or about the probability of a type II error, should be made whenever one cannot reject the null hypothesis and wants to conclude that this hypothesis should therefore be accepted, in order to draw conclusions about the quality of the simulations. Power of this test can be increased in a variety of ways: by increasing α (the probability of a type I error), by increasing the sample size n, by using a one-tailed instead of a two-tailed test (this is always the case with the present problem, however), and by increasing the smallest value of correlation one wants to be able to detect (ex. $r = 0.20$).

The intercept parameter b can be tested (H_0: $b = 0$) by calculating its 95% confidence interval (or confidence interval, C.I., at some other preselected significance level) and verifying that it includes the value 0. If n is high enough to allow rejecting the null hypothesis $a = 0$, then it is likely to provide enough power as well to the test of significance of the intercept b.

Five outcomes are possible all together:

1. $(a \approx 1, b \approx 0)$. The model may be regarded as perfectly simulating, on average, the observations;
2. $(a \approx 1, b \neq 0)$. The simulation overestimates or underestimates, on average, the observations;

3. $(a \neq 1, b \approx 0)$. The gap between the simulated and observed curves is proportional to the values of the respective points;
4. $(a \neq 1, b \neq 0)$. Superimposition of outcomes (2) and (3);
5. $a \approx 0$. No relation between simulations and observations.

Coefficients of determination (r^2) are useful to discriminate among simulations that belong to the same of the first four groups, and indicate which one is to be preferred.

If the results of a simulation classify it in the first group $(a \approx 1, b \approx 0)$, then that simulation is better than another one that would be in the other four groups. If the results put a simulation in group 4 $(a \neq 1, b \neq 0)$, then that simulation is worse than another one that would be in groups 1 to 3. The worst of all is group 5, where the simulation results are linearly unrelated to the observations. There is, however, some ambiguity concerning the two intermediate groups (2 and 3). They cannot be arranged along a linear scale since a simulation pertaining to the second group $(a \approx 1, b \neq 0)$ has a systematic and constant difference with the observations, while a simulation of the third group $(a \neq 1, b \approx 0)$ shows a difference proportional to the variable's value. Each of these differences brings an information about the weaknesses of the model and we cannot state that a simulation pertaining to group 2 is better or worse than a simulation pertaining to group 3.

3. Ordinary least-squares regression (OLS)

This method is the most popular of the simple linear regression methods, because it is easy to compute, taught in elementary courses of statistics, and implemented in all statistical software packages. But besides the assumption that the variables under study are linearly related, model I regression also makes the assumption that the independent variable is controlled (i.e. the values along the abscissa are known a priori by the observer or experimenter), or measured without error. It may seem that any given simulation generates a unique series of values, without error; but since the forcing variables or functions, the

parameters, and sometimes even the functions describing the processes are estimated with error, one should realize that the simulated values obtained from a model possess an error. Modellers are certainly aware of this fact because one of the steps of the modelling procedure is to perform a sensitivity analysis in order to know how the model reacts to variations in the forcing variables, the parameters or the functions; its objectives are (i) to determine if the model could be used in other comparable conditions (other ecosystems, other seasons) described mathematically by other values of the forcing variables, functions and parameters, and (ii) to make sure that the model is not too sensitive to the uncertainty associated with those forcing variables, functions and parameters. According to Jørgensen (1994) "The relative change in the parameter value is chosen based on our knowledge of the certainty of the parameters". Reckhow and Chapra (1983) propose to evaluate a type of confidence interval of a simulation, computing it many times with parameters issued from their confidence intervals.

A consequence of the errors, associated with the observations and the simulations as well, is that a model I regression is not appropriate for our purpose; this is illustrated using an artificial dataset. We generated a variable X containing 100 random values drawn from a uniform distribution $[0, 10]$ and two other variables, $N1$ and $N2$, made of random numbers drawn from a normal distribution $N(0, 1)$. We combined these variables to create two new variables $X + N1$ and $X + N2$. OLS regression produces the following results; the cloud of points and the regression line are shown in Fig. 1b:

$$(X + N2) = 0.86(X + N1) + 0.77 + \varepsilon.$$

The angle between the regression line and the x-axis is $40.9°$.

The 95% confidence interval of the slope a is $[0.78, 0.95]$; it does not include 0 nor 1. That the slope is significantly different from 0 is confirmed by the coefficient of determination: $r^2 = 0.814$, $P < 0.0001$.

The confidence interval of the y-intercept is $[0.27, 1.28]$; it does not include 0. While one would have expected a regression line with a slope equal
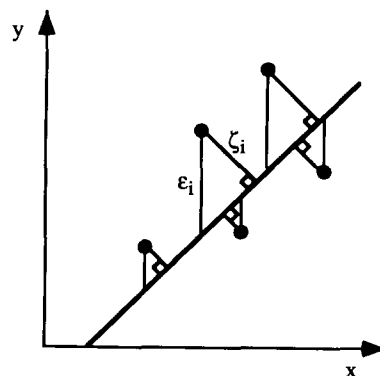

Fig. 2. Illustration of the differences between the ordinary least-squares regression (OLS) and the major axis (MA) regression. The OLS regression line minimizes the ε_i when the MA regression line minimizes the ζ_i.

to, or near 1, and a y-intercept equal to, or near 0, the results are different. What went wrong? Simply that we used the wrong model of linear regression. Instead of OLS, we should have used major axis regression (MA), one of the model II regression techniques.

4. Major axis regression (MA)

According to Sokal and Rohlf (1995), when both variables are subject to error and are in the same physical units, major axis regression is to be used. The major axis is actually the first principal component of the scatter of points in Fig. 1b. While the OLS regression line is that which minimizes the sum of squares of the vertical residuals (ε_i) between the observed values and the regression line, the MA regression line minimizes the sum, over all points, of the squared Euclidean distances (ζ_i) of the points to the regression line (Fig. 2). The parameters of the major axis are simple to compute; the method is described in Sokal and Rohlf (1995, Box 15.6) for instance.

Using the same data set as above, MA regression gives the following results (Fig. 1b):

$$(X + N2) = 0.96(X + N1) + 0.29 + \zeta$$

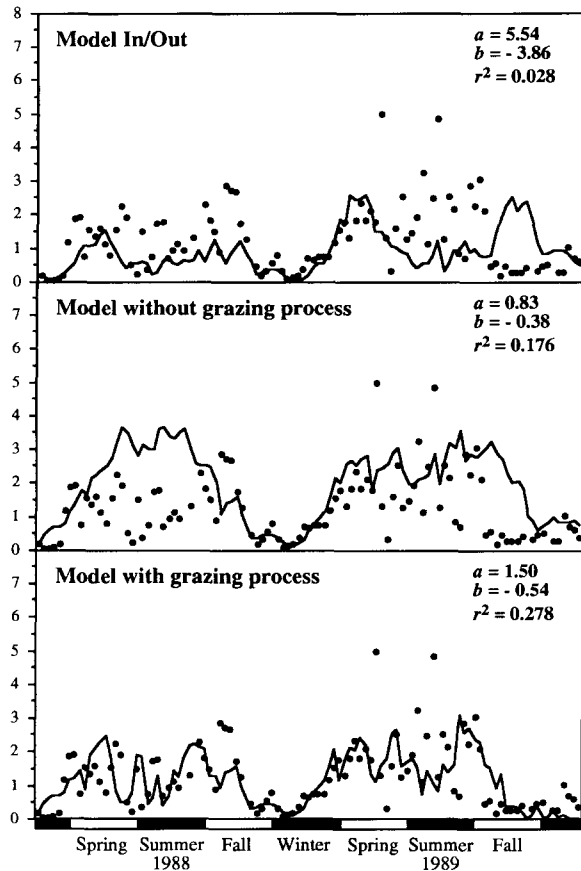The angle between the regression line and the x-axis is $43.8°$.

Fig. 3. Phytoplankton in pond A: comparison of observed (points) and simulated data (solid lines). a and b: respectively major axis slope and y-intercept of the MA regression of simulated on observed data (partly from Mesplé et al., 1995a).

5. Real-data example

Let us apply the proposed criteria, forming the couple (a, b) calculated from MA regression, to a concrete example: a deterministic model simulating the temporal change in chlorophyll a concentration in a high-rate algal pond (Mesplé et al., 1995a). Simulations are conducted under three models, respectively called 'in/out' (changes in chlorophyll a concentration depend only on inflow and outflow, which are related in turn to the mean water residence time in the pond), 'without grazing' (phytoplankton growth and death added, but not grazing), and 'with grazing' (grazing effect added). The models were calibrated with the data set from pond A, and then validated using
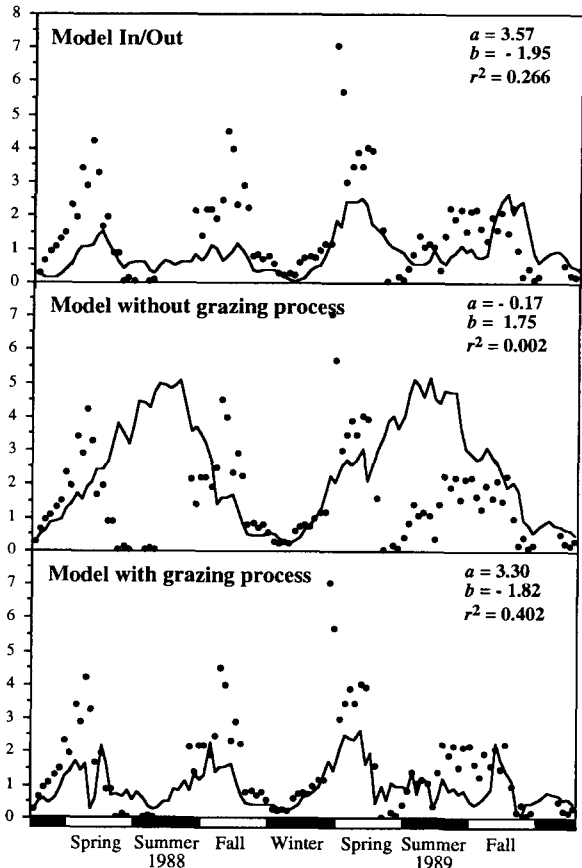

Fig. 4. Phytoplankton in pond B: comparison of observed (points) and simulated data (solid lines). a and b: respectively major axis slope and y-intercept of the MA regression of simulated on observed data (partly from Mesplé et al., 1995a).

The 95% confidence interval of the slope a is [0.87, 1.06]; it does not include the value 0 but it includes 1. The first part of this statement, that a is significantly different from 0, is confirmed by the correlation coefficient: $r = 0.897$, $P < 0.0001$.

The intercept value b follows from the slope and the coordinates of the centroid of the distribution. Therefore, the confidence interval of b can be computed as in OLS regression. For this test data set, the confidence interval of the y-intercept is estimated to be [−0.21, 0.80]; it does include the value 0.

the dataset from pond B (both sample sizes, n equals 105). Those ponds were studied in parallel: the only difference between them was that mean water residence time in pond A varied from 2 to 12 days (shortest residence time in summer in order to limit the development of zooplankton) when mean water residence time in pond B was fixed and equal to 8 days. The results of the simulations are shown in Fig. 3 for pond A and in Fig. 4 for pond B. The results of the MA regression of simulated on observed values are presented in Table 1. They lead to the following observations:

5.1. Pond A (calibration of the models)

Chlorophyll a concentrations of the 'in/out' model are not correlated to the observed concentrations ($a \approx 0$), but they are ($a \neq 0$) for the other two models (with and without grazing). The value of the slope (a) can, in the model without grazing, be considered equal to 1 but not in the model with grazing. The value of the y–intercept (b) increases dramatically from –3.86 ('in/out'

model) to –0.38 (model without grazing) and –0.54 (model with grazing); it can never be considered to be equal to 0.

If a correlation were present between the observations and the simulations using the 'in/out' model, we would expect it to be at least of medium size, i.e. $r = 0.30$ ($r^2 = 0.09$) (Cohen, 1988). With $n = 105$, the one-tailed test of such a correlation, at significance level $\alpha = 0.05$, would give us power of about 93%, or a probability of a type II error of 7% (Cohen, 1988). The observed value of $r = 0.167$ ($r^2 = 0.028$) is much smaller than 0.30, which allows us to state that the data really support the null hypothesis ($r = 0$, $a = 0$), with a probability of type II error not exceeding 7%.

5.2. Pond B (validation of the models calibrated with pond A data)

In the results of the 'in/out' model, the slope of the regression line (a) significantly differs from 0 and 1 and the y-intercept differs from 0. Modelling the evolution of chlorophyll a concentra-

Table 1
Results of the major axis regression of simulated chlorophyll a on observed values for the two ponds; a = slope, b = y-intercept of the regression line ($X_{observed} = a \cdot X_{simulated} + b + \zeta$)

Chlorophyll a simulation	Parameters	Values	Lower limit (95%)	Upper limit (95%)	Conclusions
POND A					
in/out	a	5.54	−38.80	2.48	$a \approx 0$
model	b	−3.86	−4.19	−3.53	$b \neq 0$
	r^2	0.02756	–	–	–
without	a	0.83	0.52	1.26	$a \neq 0$; $a \approx 1$
grazing model	b	−0.38	−0.74	−0.02	$b \neq 0$
	r^2	0.17594	–	–	–
with grazing	a	1.50	1.11	2.11	$a \neq 0$; $a \neq 1$
model	b	−0.54	−0.83	−0.25	$b \neq 0$
	r^2	0.27736	–	–	–
POND B					
in/out	a	3.57	2.68	5.23	$a \neq 0$; $a \neq 1$
model	b	−1.95	−2.36	−1.56	$b \neq 0$
	r^2	0.26596	–	–	–
without	a	−0.17	−1.96	0.96	$a \approx 0$
grazing model	b	1.75	1.25	2.25	$b \neq 0$
	r^2	0.00154	–	–	–
with grazing	a	3.30	2.66	4.31	$a \neq 0$; $a \neq 1$
model	b	−1.82	−2.22	−1.43	$b \neq 0$
	r^2	0.40237	–	–	–

tions without considering the grazing process deteriorates simulation quality, as there is no longer a correlation between simulated and observed data. As discussed above, power on the test of significance of $r = 0.045$, or $r^2 = 0.002$, is very good, so that we can state that the data actually support the null hypotheses $r = 0$ and $a = 0$; since our alternative hypothesis was one-tailed (we expected to find a positive relation, if any), finding a negative slope gives us further confidence for accepting the null hypothesis. Adding the grazing process allows to find again a correlation (even though the slope of the regression line is different from 1) and a y-intercept (b) that are different from 0.

Considering pond A, the best simulation (model without grazing) is in the group ($a \approx 1$, $b \neq 0$) with $a = 0.83$ and $b = -0.38$; the 'in/out' model gives $a \approx 0$, and the model with grazing is in the group ($a \neq 1$, $b \neq 0$). Thus, the maximum theoretical simulation quality ($a \approx 1$, $b \approx 0$) is not reached: adding growth and death processes to

the 'in/out' model have improved the simulation quality, but adding grazing process has reduced it. In pond B, increasing the number of processes from the simple 'in/out' model to the model without grazing decreases the simulation quality as we go from ($a \neq 1$, $b \neq 0$) to ($a \approx 0$); Jørgensen (1992) underlines this point: "It has often been argued that a more complex model should be able to account more accurately for the reactions of a real system, but this is not necessarily true. Additional factors are involved. A more complex model contains more parameters and increases the level of uncertainty, because parameters have to be estimated either by more observations in the field, by laboratory experiments, or by calibrations, which again are based on field measurements. Parameter estimations are never completely without errors, and the errors are carried through into the model and will thereby contribute to its uncertainty". Going up one more step in the complexity of the model (from the model without grazing to the model with grazing)
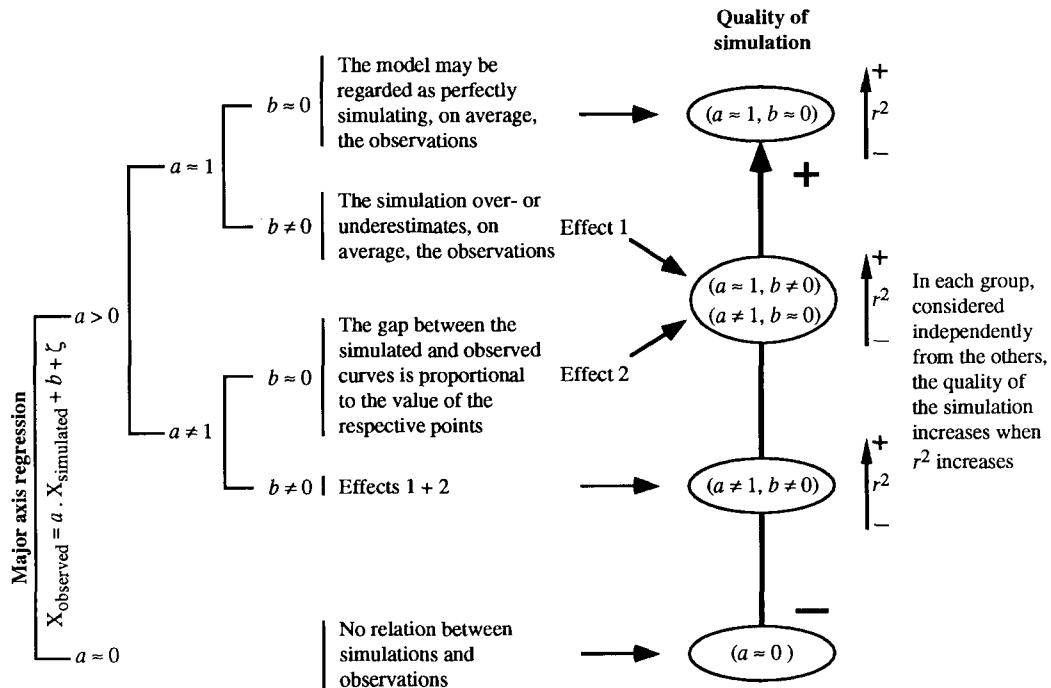

Fig. 5. The five possible outcomes of a major axis simple linear regression model: $X_{observed} = a \cdot X_{simulated} + b + \zeta$ and the four groups (in the ellipses) allowing to determine the quality of a simulation ($\approx$: not significantly different from).

increases the simulation quality, as it falls in the same group than the 'in/out' model ($a \neq 1$, $b \neq 0$) with a greater coefficient of determination ($r^2 =$ 0.402 vs $r^2 = 0.266$). The conclusion is that, as the development of zooplankton increases with increasing residence time, it becomes more and more important to take into account grazing process in the model.

In order to synthesize what is presented above and to provide guidelines, one can refer to Fig. 5. Determining the quality of a simulation using major axis linear regression requires two steps:

- First, consider a. There is a hierarchy of possibilities: $a = 0$ or $a \neq 0$; if $a \neq 0$, then $a \approx 1$ or $a \neq 1$.
- Secondly, consider b when $a \neq 0$. There are two possibilities: $b \approx 0$ or $b \neq 0$.

Simulation results fall in one of the four groups defined by the various combinations of a and b.

Comparing the quality of two simulations requires a third step if they fall in the same group. In that situation, the value of the coefficient of determination allows to grade the simulations; the higher the value of r^2, the better the quality of the simulation.

6. Conclusion

We have attempted to measure simulation quality only, and not model quality; these are two different things. An advantage of considering simulations only is that it is possible to compare simulations coming from different kinds of models (deterministic and statistical, for example); a drawback is that our measures of simulation quality are not related to model complexity. In any case, future ecological models calibrated or validated against field data using linear regression should (i) use major axis regression instead of ordinary least-squares regression, and (ii) be presented with some objective criterion of simulation quality, like the couple (a, b) proposed in this paper. But, even in a perfect simulation, Bard (1974) (referred to in Halfon, 1985) insists on the fact that " ...whereas a lack of fit constitutes a strong grounds for rejecting or at least amending the model, a good fit does not prove that the

model is correct. A good fit merely establishes the fact that there is no reason to reject the model on the basis of the data on hand. In fact, no amount of data can ever prove a model; all that we can hope is that it does not disprove it". Meanwhile, Thom (1979) holds a somewhat opposite point of view, stating that if a model empirically simulates observations with good accuracy, even without our understanding why, there is no reasons to forego using that unexplained agreement in pragmatic ways.

Acknowledgements

References

Andersen, V., Nival, P. and Harris, R.P., 1987. Modelling of a planktonic ecosystem in an enclosed water column. J. Mar. Biol. Assoc. UK, 67: 407–430.

Bard, Y., 1974. Non Linear Parameter Estimation. Academic Press, New York, NY, 341 pp.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd edition. Lawrence Erlbaum Assoc., Hillsdale, NJ, 567 pp.

Costanza, R., 1989. Model goodness of fit: a multiple resolution procedure. Ecol. Model., 47: 199–215.

Costanza, R. and Sklar, F.H., 1985. Articulation, accuracy and effectiveness of mathematical models: a review of freshwater wetland applications. Ecol. Model., 27: 45–68.

Halfon, E., 1985. Is there a best model structure ? III. Testing the goodness of fit. Ecol. Model., 27: 15–23.

Halfon, E., 1989. Probabilistic validation of computer simulations using the bootstrap. Ecol. Model., 46: 213–219.

Halfon, E. and Reggiani, M.G., 1978. Adequacy of ecosystem models. Ecol. Model., 4: 41–50.

Jørgensen, S.E., 1992. Integration of Ecosystem Theories: A Pattern. Kluwer Academic Publishers, Dordrecht, 383 pp.

Jørgensen, S.E., 1994. Fundamentals of Ecological Modelling, 2nd edition. Elsevier, Amsterdam, 628 pp.

Keller, A.A., 1989. Modeling the effects of temperature, light, and nutrients on primary productivity: an empirical and a mechanistic approach compared. Limnol. Oceanogr., 34: 82–95.

Kraemer, H.C. and Thiemann, S., 1987. How Many Subjects? Statistical Power Analysis in Research. Sage Publications, Newbury Park, CA, 120 pp.

Loehle, C., 1983. Evaluation of theories and calculation tools in ecology. Ecol. Model., 19: 239–247.

Mayer, D.G. and Butler, D.G., 1993. Statistical validation. Ecol. Model., 68: 21–32.

Mesplé, F., Casellas, C., Troussellier, M. and Bontoux, J., 1995a. Some difficulties in modelling chlorophyll *a* evolution in a high rate algal pond (HRAP) ecosystem. Ecol. Model., 78: 25–36.

Mesplé, F., Troussellier, M., Casellas, C. and Bontoux, J., 1995b. Difficulties in modelling phosphate removal in a high rate algal pond. Water Sci. Technol., 31: in press.

Mesplé F., Casellas, C., Troussellier, M. and Bontoux, J., 1996. Modelling orthophosphates evolution in a high rate algal pond. Ecol. Model. (in press).

Peterman, R.M., 1990. Statistical power analysis can improve fisheries research and management. Can. J. Fish. Aquat. Sci., 47: 2–15.

Reckhow, K.H. and Chapra, S.C., 1983. Confirmation of water quality models. Ecol. Model., 20: 113–133.

Robinson, J.A., 1986. Approaches and limits to modeling microbiological processes. In: F. Megušar and M. Gantar (Editors), Perspectives in Microbial Ecology, Proceedings of the Fourth International Symposium on Microbial Ecology. Slovene Society for Microbiology, Ljubljana, pp. 20–29.

Rose, K.A., Swartzman, G.L., Kindig, A.C. and Taub, F.B., 1988. Stepwise iterative calibration of a multi-species phytoplankton–zooplankton simulation model using laboratory data. Ecol. Model., 42: 1–32.

Smith, D.W. and Putz, G., 1993. The simulation of river concentrations of coliform bacteria using a transverse mixing model. Environ. Technol., 14: 1117–1130.

Smith, E.P. and Rose, K.A., 1995. Model goodness-of-fit analysis using regression and related techniques. Ecol. Model., 77: 49–64.

Sokal, R.R. and Rohlf, F.J., 1995. Biometry, 3rd edition. W.H. Freeman and Company, New York, NY, 887 pp.

Straškraba, M. and Gnauck, A.H., 1985. Freshwater Ecosystems: Modelling and Simulation. Elsevier, Amsterdam, 313 pp.

Summers, J.K., Kayzak, P.F. and Weisberg, S.B., 1991. A water quality model for a river receiving paper mill effluents and conventional sewage. Ecol. Model., 58: 25–54.

Thom, R., 1979. Modélisation et scientificité. In: P. Delattre et M. Thellier (Editors), Actes du Colloque "Élaboration et Justification des Modèles: Application en Biologie", Tome I. Maloine, Paris, pp. 21–29.

Valentin, J.L. and Coutinho, R., 1990. Modelling maximum cholorophyll in the Cabo Frio (Brazil) upwelling: a preliminary approach. Ecol. Model., 52: 103–113.

Wallach, D. and Goffinet, B., 1989. Mean squared error of prediction as a criterion for evaluating and comparing system models. Ecol. Model., 44: 299–306.