Available online at www.sciencedirect.com
ScienceDirect
Harmful Algae 6 (2007) 361–371
HARMFUL ALGAE
www.elsevier.com/locate/hal

Artificial neural network approaches to one-step weekly prediction of *Dinophysis acuminata* blooms in Huelva (Western Andalucía, Spain)

L. Velo-Suárez [a,*], J.C. Gutiérrez-Estrada [b,1]

[a] *Instituto Español de Oceanografía, Centro Oceanográfico de Vigo, Subida a Radiofaro 50-52, Cabo Estay, Canido, 36390 Vigo, Spain*
[b] *Dpto de Ciencias Agroforestales, Escuela Politécnica Superior, Campus Universitario de La Rábida, Universidad de Huelva, 21819 Palos de la Frontera, Huelva, Spain*

Received 12 September 2006; received in revised form 12 October 2006; accepted 7 November 2006

Abstract

On the Atlantic coasts of Andalucía, chronic spring–summer (March–June) diarrhetic shellfish poisoning (DSP) outbreaks are associated with blooms of *Dinophysis acuminata*, Claparède and Lachmann. Artificial neural networks (ANNs) have been successfully used to model primary production and have recently been tested for the prediction of harmful algae blooms. In this study, we evaluated the performance of feed forward ANN models trained to predict *D. acuminata* blooms. ANN models were trained and tested using weekly data (5 previous weeks) of *D. acuminata* cell counts from eight stations of the Andalucía HAB monitoring programme in the coasts of Huelva between 1998 and 2004. Principal component analysis (PCA) were previously carried out to find out possible similarities within time series from each zone with the aim of reducing the number of areas to model. Our results show that ANN models with a low number of input variables are able to reproduce trends in *D. acuminata* population dynamics.

Keywords: Algal blooms; *Dinophysis acuminata*; Forecast; Models; Artificial neural networks; Phytoplankton

1. Introduction

Diarrhetic shellfish poisoning (DSP) is a gastro-intestinal disease resulting from ingestion of shellfish contaminated with dinoflagellates toxins. This disease threatens public health and shellfish industries, due to its worldwide occurrence, high morbidity rate and the long duration of shellfish toxicity (Lee et al., 1989).

To comply with the European policies, The Andalucía Government (South Spain) implemented a monitoring programme for the control of toxic phytoplankton and marine biotoxins in shellfish in Andalucía bivalve harvesting areas. Since 1994, when this programme was established, recurrent blooms of *Dinophysis acuminata* have been associated with seasonal DSP outbreaks along the Atlantic coast of Andalucía (Fernández et al., 2003).

Although the relationship between phytoplankton and environmental variables has been extensively studied, the causality and dynamics of algal blooms are very complex and not entirely understood (Lee et al., 2003). Nowadays, the prediction of these events remains a very difficult task whereas modelling techniques have

* Corresponding author. Tel.: +34 986 492111.
E-mail addresses: lourdes.velo@vi.ieo.com, lourdes.velo@vi.ieo.es (L. Velo-Suárez), juanc@uhu.es (J.C. Gutiérrez-Estrada).
[1] Tel.: +34 959 217528; fax: +34 959 217528.

1568-9883/$ – see front matter
doi:10.1016/j.hal.2006.11.002

greatly advanced in recent years. The uncertainty in the kinetic rates and the complexity of deterministic models has slowed down their effective use as forecasting tools. Even so, Sarkar and Chattopadhayay (2003) modeled phytoplankton dynamics using mathematical coupled physical–biological models, but the results obtained could not be applied to specific algae populations. Although deterministic models contribute to a better understanding of planktonic food webs, their predictive potential still remains relatively low.

An alternative modelling approach, based on heuristic models is currently in progress as a valuable predictive tool in ecological and environmental sciences (Scardi, 1996). The computational or artificial neural networks (CNNs or ANNs) can be classified as 'black box' type models. An ANN is a non-linear mathematical structure capable of representing the complex non-linear processes that relate the inputs to the outputs of a system. The most important advantage of an ANN approach over classical ones is its capability to deal with uncertain information and incomplete or inconsistent data. Therefore, ANNs are promising for the modeling of harmful algae blooms dynamics. In this way, some studies on ANN modeling of algal blooms have been carried out. They have been tested in limnological systems (Recknagel et al., 1997; Yabukana et al., 1997; Karul et al., 2000), rivers (Whitehead et al., 1997; Maier et al., 1998) and coastal systems (Barciela et al., 1999; Belgrano et al., 2001; Lee et al., 2003).

In this paper, a study of ANN modelling to predict *D. acuminata* blooms in the Huelva coast (South Spain) is presented. Thus, ANNs are introduced and applied as a new model type in this specie. This model could be used as a new forecasting tool which complements current phytoplankton analysis inside the HABs monitoring program carried out in the study area.

2. Material and methods

2.1. Study area

The study was carried out on Huelva coast, southwest Spain. This area is delimited by the Guadiana River in the west and the Guadalquivir River in the east (Fig. 1).

Huelva coast is formed by large sandy beaches (145 km long) only interrupted by the presence of estuarine mouths. The regional oceanography of the littoral zone of Huelva is influenced by the North Atlantic geostrophic current, which flows to the east within this zone. Tidal regime, wave action and fluvial discharge control the hydrodynamic processes. The tidal regime is mesotidal (Borrego et al., 1993). Dominant waves associated with the Atlantic circulation come from the southwest. Fluvial sediment and run-off from the nearby land are important during December and January when rainfall is highest.

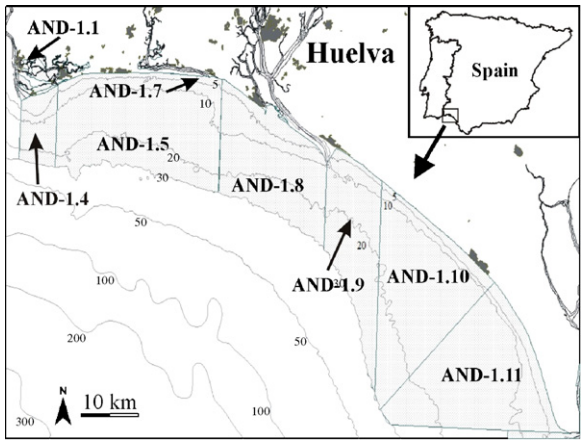

Fig. 1. Study area showing the location of Huelva bivalve mollusc harvesting zones.

2.2. Data source and data pre-treatment

The present study was performed with *D. acuminata* concentrations data (cells/L) in Huelva from the Andalucía HAB monitoring program. Time series from 1998 to 2004 were used for forecasting purposes.

Water samples were collected weekly at AND-1.1, AND-1.4, AND-1.5, AND-1.7, AND-1.8, AND-1.9, AND-1.10 and AND-1.11 bivalve mollusc harvesting areas (Fig. 1). At each site, water samples for qualitative and quantitative analysis were taken using a 20 μm mesh-net and a weighted plastic hose 10 m long (Lindahl, 1986), respectively. Qualitative samples were preserved with 4% formaldehyde and quantitative ones with acid lugol's solution. Qualitative analysis was performed in order to identify harmful species while *D. acuminata* cells were counted using 25 mL from quantitative samples in sedimentation chambers (Utermöhl, 1958).

After phytoplankton analysis, all *D. acuminata* records were gathered in a database. Statistical analyses were carried out to find out possible similarities among time series from each zone. The aim of the application of these statistical methods was to weight up the possibility of merge closely related data sets and reduce areas to model. The relationship among zones was calculated with correlation indexes for the whole period. Besides that, a principal components analysis

(PCA) was performed on the whole data set (1998–2003) and each annual time series individually.

2.3. Artificial neural network

ANNs are mathematical models inspired by the neural architecture of the human brain. A neurone or node is a simple non-linear unit. Neurones collect inputs from single or multiple sources and produce an output. Interconnecting many of these single nodes in a known layer configuration creates a artificial neural model.

Each node j receives incoming signals from every node i in the previous layer. Associated with each incoming signal (x_i) there is a weight (W_{ji}). The effective incoming signal (I_j) to node j is the weighted sum of all the incoming signals:

$$I_j = \sum_{i=1}^{q} x_i W_{ji} \tag{1}$$

The effective incoming signal, I_j, is passed through an activation function (sometimes called a transfer function) to produce the outgoing signal (y_j) of the node j. In this study, the linear function ($y_j = I_j$) will be used in the output layer and the sigmoid non-linear function will be used in the hidden layers:

$$y_j = f(I_j) = \frac{1}{1 + \exp(-I_j)} \tag{2}$$

To determine the set of weights a corrective-repetitive process called 'learning' or 'training' is performed. This training forms the interconnections between neurons, and it is accomplished using known inputs and outputs (training sets or patterns). The strength of these interconnections is adjusted using an error convergence technique so a desired output will be produced for a given input. The training method used is a standard back-propagation proposed by Rumelhart et al. (1986). Data of time series from 1998 to 2003 were used for model calibration while data from 2004 were use for model validation.

In the study presented in this paper, the learning process was controlled by the method of internal validation (about 25% of calibration data was held back and used to test the error at the end of each epoch) (Tsoukalas and Uhrig, 1997; Gutiérrez-Estrada et al., 2004). The weights are updated at the end of each epoch. The number of epochs with the smallest error of the internal validation indicates the weights to select.

Only weekly *D. acuminata* concentrations (cells/L) were considered as input variables. The independent variable in each week t (output variable of the ANN) was the concentration in that week. Input variables of the ANN were *D. acuminata* concentrations in 5 previous weeks ($t-1$, $t-2$, $t-3$, $t-4$ and $t-5$). They were selected using the results provided by a previous autocorrelation and partial autocorrelation analysis. The numbers of hidden layers and nodes in the hidden layers were determined by trial and error. ANNs with two hidden layers and 2–10 hidden nodes were successively trained based on the calibration data set. In this study, the following network parameters were used: learning rate = 0.01 and momentum term = 0.3. The ANN having the best performance when applied to the validation set, within a pool of six repetitions, was selected (Anctil and Rat, 2005). ANN models were implemented using STATISTICA 6.0 (Statsoft Inc., 1984–2002).

2.4. Error terms: model selections

To assess the performance of the ANN during the validation phase and therefore to identify the best model, several measures of accuracy were applied, as there is not a unique and more suitable performance evaluation test (Legates and McCabe, 1999; Abrahart and See, 2000). The correlation between observed and predicted *Dinophysis* concentration was expressed by means of the correlation coefficient R. The coefficient of determination (R^2) describes the proportion of the total variance in the observed data that can be explained by the model. Other measures of variances applied were the percent standard error of prediction (%SEP) (Ventura et al., 1995), the coefficient of efficiency (E_j) (Nash and Sutcliffe, 1970; Kitanidis and Bras, 1980) and the average relative variance (ARV) (Griñó, 1992). These four estimators are unbiased ones employed to see how far the model is able to explain the total variance of the data.

In addition, it is advisable to quantify the error in the same units than the variables. These error terms are representative of the size of a "typical" error and easier to understand. These measures, or absolute error measures, included the square root of the mean square error (RMSE) and the mean absolute error (MAE), given by:

$$\mathrm{RMSE} = \left(\frac{\sum_{i=1}^{N} (Q_t - \hat{Q}_t)^2}{N} \right)^{1/2} = \mathrm{MSE}^{1/2}, \qquad \mathrm{MAE} = \frac{\sum_{i=1}^{N} |Q_t - \hat{Q}_t|}{N} \tag{3}$$

where Q_t is the observed *Dinophysis* concentration at the time step t, $\hat{Q}_t$ the estimated *Dinophysis* concentra-

tion at the same time step t and N is the total number of observations of the validation set.

The percent standard error of prediction, %SEP, is defined by:

$$\mathrm{SEP} = \frac{100}{\bar{Q} \times \mathrm{RMSE}} \tag{4}$$

where $\bar{Q}$ is the average of the observed *Dinophysis* concentration of the validation set. The principal advantage of %SEP is its non-dimensionality that allows forecasts given by different models to be compared on the same basis. The coefficient of efficiency E_j and the average relative variance ARV are used to see how the model explains the total variance of the data and represent the 'proportion' of the variation of the observed data considered by the model. E_j and ARV are given by:

$$E_j = 1.0 - \frac{\sum_{i=1}^{N} (Q_t - \hat{Q}_t)^j}{\sum_{i=1}^{N} (Q_t - \bar{Q})^j}, \qquad \mathrm{ARV} = \frac{\sum_{i=1}^{N} (Q_t - \hat{Q}_t)^2}{\sum_{i=1}^{N} (Q_t - \bar{Q})^2} = 1.0 - E_2 \tag{5}$$

For a perfect match, the values of R^2 and of E_j should be close to one and those of %SEP and ARV close to 0.

Also the persistence index, PI, was used for the model performance evaluation (Kitanidis and Bras, 1980):

$$\mathrm{PI} = 1 - \frac{\sum_{i=1}^{N} (Q_t - \hat{Q}_t)^2}{\sum_{i=1}^{N} (Q_t - Q_{t-L})^2} \tag{6}$$

where Q_{t-L} is the observed *Dinophysis* concentration at the time step t–L and L is the lead-time. In all the simulations, L was set equal to 1, since only 1-week ahead forecasts were performed. A PI value of one reflects a perfect adjustment between predicted and observed values.

Another index used to identify the best linear and non-linear models was the Akaike Information Criterion (AIC) (Qi and Zhang, 2001) given by:

$$\mathrm{AIC} = \log(\mathrm{MSE}) + \frac{2m}{N} \tag{7}$$

In the previous equation m is the number of parameters of the model. The first term of the second member measures the goodness-of-fit of the model to the data, while the second term sets a penalty for the model over-parameterization. The optimal model is selected when AIC is minimized.

3. Results

The results provided by the correlation analysis are shown in Table 1. Broadly, correlations are significant with $p < 0.01$, although generally not very high. Pearson indices estimated for the 1998–2003 period showed that close zones are higher correlated (AND-1.4, AND-1.5 = 0.70) than far ones (AND-1.10, AND-1.1 = 0.19), but these results did not indicate any positive and clear association among zones.

PCA results are shown in Table 2. Data of annual analysis indicates high variability in the number of components extracted and variance explained. In the 1998 data, the first component extracted accounts for 85% of the variance. In contrast, in the 1999 and 2002 data, three factors are needed to explain ca. 70% of the variance.

PCA analysis for the whole data set extracted two components which explained 68% of the total variation. Fig. 2 shows the biplot of the projection of the initial zones in the plane defined by these components. AND-1.4, AND-1.5, AND-1.7 and AND-1.9 were significantly

Table 1
Pearson indexes calculated for each sampling zone and the first and second factors (F) extracted by global PCA

	AND-1.1	AND-1.4	AND-1.5	AND-1.7	AND-1.8	AND-1.9	AND-1.10	AND-1.11	*F*1	*F*2
AND-1.1	1									
AND-1.4	0.75**	1								
AND-1.5	0.56**	0.70**	1							
AND-1.7	0.50**	0.67**	0.65**	1						
AND-1.8	0.24**	0.25**	0.29**	0.47**	1					
AND-1.9	0.35**	0.35**	0.44**	0.43**	0.73**	1				
AND-1.10	0.19**	0.18**	0.20**	0.19**	0.44**	0.65**	1			
AND-1.11	0.12*	0.26**	0.18**	0.25**	0.29**	0.33**	0.38**	1		
*F*1	0.70**	0.79**	0.76**	0.79**	0.66**	0.76**	0.54**	0.48**	1	
*F*2	0.44**	0.48**	0.38**	0.25**	0.46**	0.47**	0.62**	0.37**	0	1

* $p < 0.05$ (bilateral).
** $p < 0.01$ (bilateral).

Table 2
Eigenvalues and principal components extracted by annual and global PCA

Year	Eigenvalues			
	Component	Total	%Variance explained	%Accumulated
1998	1	6.812	85.147	85.147
	2	0.555	6.942	92.089
1999	1	3.274	40.921	40.921
	2	1.794	22.426	63.347
	3	1.064	13.304	76.650
2000	1	5.917	73.966	73.966
	2	0.938	11.726	85.692
2001	1	4.638	57.976	57.976
	2	1.453	18.164	76.140
2002	1	2.228	27.850	27.850
	2	2.076	25.944	53.795
	3	1.077	13.462	67.257
2003	1	4.686	58.579	58.579
	2	1.749	21.866	80.445
1998–2003	1	3.861	48.265	48.265
	2	1.601	20.008	68.273

($R > 0.7$) correlated with the first component while the second component separated Western from Eastern zones. Using the information provided by PCA analysis, data from AND-1.4, AND-1.5 and AND-1.7 were merged to form a single area.

Twelve networks were calibrated for each hidden layer neural architecture (5-2-2-1; 5-4-2-1; 5-4-4-1; 5-6-4-1; 5-6-6-1; 5-8-6-1; 5-8-8-1; 5-10-8-1; 5-10-10-1). Therefore, 648 models were calibrated and trained, 108 for each sampling area.

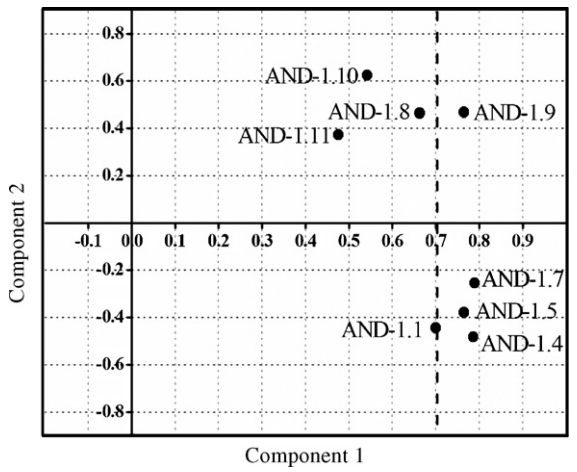

Fig. 2. Biplot of the projection of the initial zones in the plan defined by principal components.

All cases were tested in order to identify the ANN model with the best performance. The best networks for each zone were selected based on accuracy measures presented in Section 2.4. For AND-1.11 zone, the best results obtained were for the model with four nodes per hidden layer (case 3) [ANN(5-4-4-1)]. For AND-1.10 area, the best network had two hidden layers with eight and six nodes, respectively (case 2) [ANN(5-8-6-1)]. For AND-1.9, the best performance was achieved with the neural architecture of four nodes in the first and second layers (case 5) [ANN(5-4-4-1)]. For AND-1.8 zone, the network selected had 10 neurons in the two hidden layers (case 5) [ANN(5-10-10-1)]. For the group composed of AND-1.7, AND-1.5 and AND-1.4 areas, the best validation results were obtained for the configuration 5-10-10-1 (case 5). Finally, in AND-1.1 zone, the best model had two hidden layers with 10 and 8 nodes, respectively (case 4) [ANN(5-10-8-1)] (Figs. 3 and 4).

The best model calibrated and validated for AND-1.11 time series stood out for its correlation values between observed and estimated data ($N = 46$ and $R^2 = 0.96$) and close adjustment from line 1:1. In this case, %SEP, E and ARV are the most outstanding error terms from the rest of the validated networks (%SEP = 20.69; $E = 0.96$; ARV = 0.04). The model selected for AND-1.9 reached the highest PI (PI = 0.84) and the lowest AIC (AIC = 10.47). All the estimated error terms are shown in Figs. 3 and 4.

The scatterplot between observed and forecast concentrations of the best models from AND-1.8, group (AND-1.7, AND-1.5 and AND-1.4) and AND-1.1 showed the highest deviations from line 1:1, while AND-1.11 and AND-1.9 denoted a very close adjustment. Calibrated models with time series from AND-1.8, group and AND-1.1 show a general tendency to underestimate high concentrations while the model selected for AND-1.10 zone overestimate those concentrations.

An range error analysis was carried out (Table 3). Ranges were established according to monitoring thresholds. This analysis was not applied to AND-1.1 data because they did not reach 500 cells/L in the validation period. Almost all best results of the accuracy measures occurred in the range from 500 to 2500 cells/L (with the exception of models calibrated with the AND-1.8 and group data). In general, the error terms became worse as the concentration was lower. Networks of AND-1.11, AND-1.10 and AND-1.9 showed high values of PI in up ranges while models from AND-1.8 and group had negative values of this error term. Although R^2 and %SEP values pointed towards the good

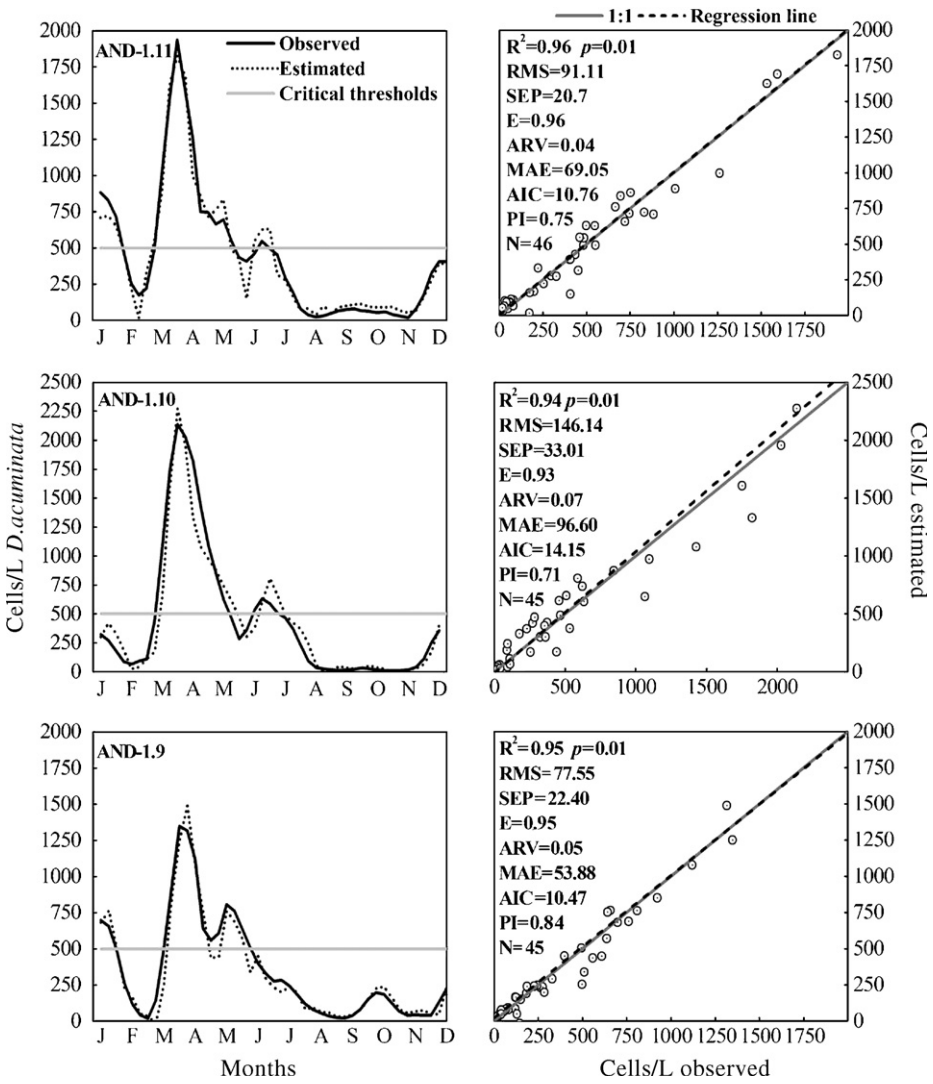

Fig. 3. One-step-ahead prediction of weekly concentrations for the validation period and scatterplot between observed and forecast concentrations for the models with the best performance of AND-1.11, AND-1.10 and AND-1.9 sampling zones. The critical threshold is 500 cells/L and was established by the Andalucía HAB monitoring programme.

performance of the model, other measures (PI) suggested the opposite (poor performance of the same model) in those areas.

The model with the best performance calibrated for the zone AND-1.11 was used to test all the validation series of the rest of the zones. Fig. 5 shows validation results from zones AND-1.10 and the group formed by AND-1.4, AND-1.5 and AND-1.7. For AND-1.10, the best performance was achieved with the model trained with its own calibration data instead of the results from the model calibrated with AND.-1.11. Otherwise, in the group zone (AND-1.4, AND-1.5 and AND-1.7), the best results obtained were for the model trained with AND-1.11 calibration data set.

4. Discussion

In recent years, a wide range of modelling techniques has been used to characterize and forecast algal blooms and primary production. Recently, the HABES (Harmful Algae Bloom Expert System) project has developed an expert system which comprised models based on fuzzy modelling techniques for several selected harmful algae species. A part of this expert system was created to forecast DSP events caused by *D. acuminata* and *Dinophysis acuta* blooms (Blauw et al., 2006; Johansen et al., 2004). The fuzzy logic scheme for this study was mainly based on the hypothesis that there was an onset of offshore populations of *Dinophysis* that are advected

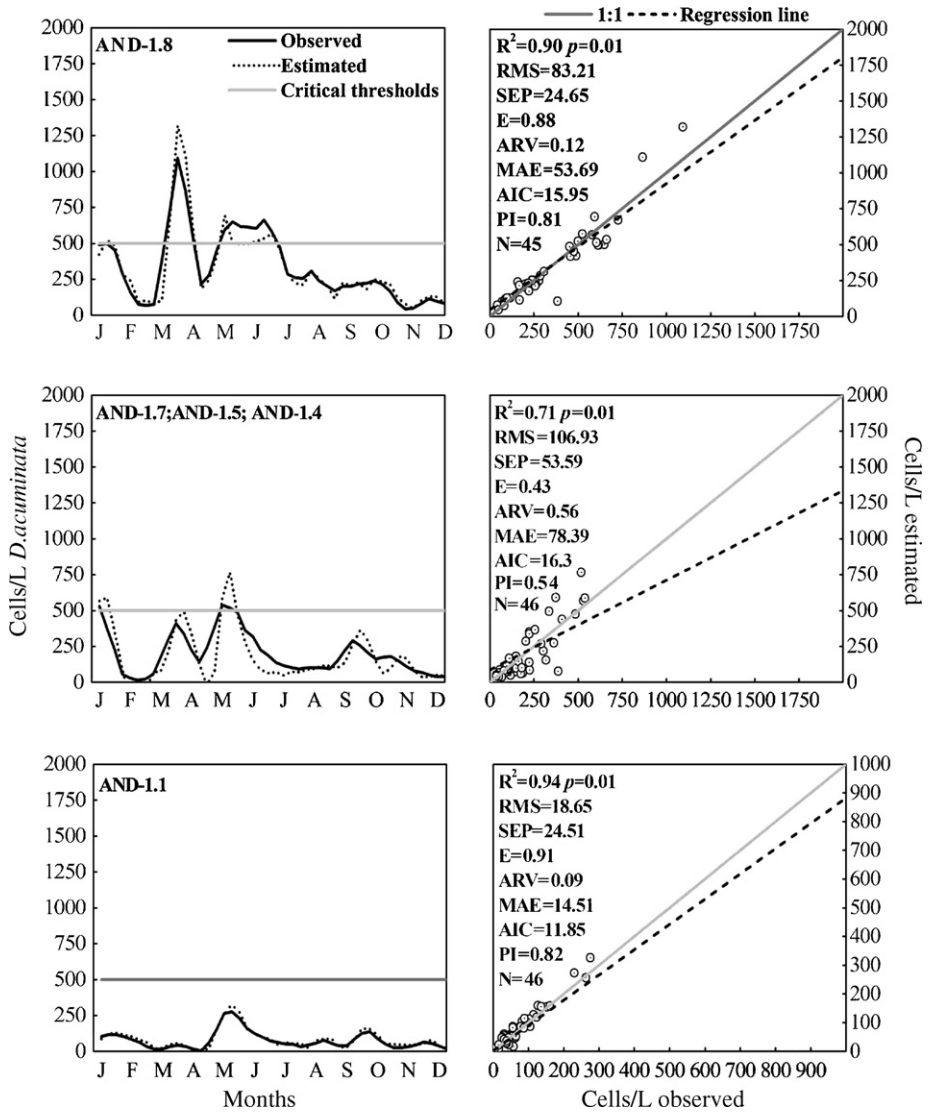

Fig. 4. One-step-ahead prediction of weekly concentrations for the validation period and scatterplot between observed and forecast concentrations for the models with the best performance of AND-1.8, the group AND-1.7, AND-1.5 and AND-1.4, and AND-1.1. The critical threshold is 500 cells/L and was established by the Andalucía HAB monitoring programme.

into inshore areas resulting in DSP in mussels. The factors included were wind speed and wind direction, both for creation of stratification and thereby offshore populations, but also inflow probability to inshore areas. This study concluded that offshore stratification in combination with advective transport to the coastal zone was too complex to model with fuzzy logic. The resulting knowledge rules for the model were chosen based on bibliography and previous surveys. Also, Barciela et al. (1999) developed an analytical model to predict primary production in an embayment affected by upwelling. These authors reported that these types of models are difficult to calibrate and implement. The results that emerged from that study indicated that analytical models contribute to a better understanding of planktonic food webs but, although reproducing the main patterns of large-scale variability, its short-medium time predictive potential is low due to the large uncertainty associated with parameter estimation. In contrast, the relationships among trigger bloom factors and biological parameters are previously unknown in ANN studies. In the case of variables related with high non-linearity, ANN models seem to have better performance than analytical models. Over the last decade, some works on ANN modelling techniques (Belgrano et al., 2001; Lee et al., 2003)

Table 3
Range values

Sampling zones	Range (cells/L)	R^2	RMS	SEP	E	ARV	MAE	PI	AIC
AND-1.11	Mean	0.94	110.16	25.03	0.94	0.06	73.18	0.79	11.1
	S.D.	0.02	23.41	5.32	0.03	0.03	11.23	0.04	0.39
	(0, 250)	0.44	53.76	66.44	0.17	0.83	40.32	−0.04	11.97
	(250, 500)	0.56	99.25	24.27	−0.76	1.76	68.36	0.13	15.86
	>500	0.92	123.67	12.62	0.75	0.25	110.66	0.68	15.35
AND-1.10	Mean	0.95	131.27	29.65	0.94	0.06	87.89	0.79	13.88
	S.D.	0.03	33.90	7.66	0.03	0.03	23.70	0.06	0.52
	(0, 250)	0.79	62.31	107.48	−0.18	1.18	39.66	−0.42	16.81
	(250, 500)	0.11	130.46	36.35	−2.20	3.20	104.73	−0.04	28.54
	>500	0.86	232.95	20.11	0.84	0.16	186.72	0.56	25.36
AND-1.9	Mean	0.92	109.69	32.50	0.79	0.21	70.97	0.75	16.48
	S.D.	0.03	19.83	5.88	0.07	0.07	12.06	0.05	0.38
	(0, 250)	0.69	42.22	43.03	0.61	0.39	26.55	0.08	10.69
	(250, 500)	0.42	102.71	28.30	−0.20	1.20	71.70	−1.11	20.69
	>500	0.91	108.21	13.29	0.84	0.16	96.86	0.53	15.52
AND-1.8	Mean	0.86	109.69	32.50	0.79	0.21	70.97	0.75	16.48
	S.D.	0.05	19.83	5.88	0.07	0.07	12.06	0.05	0.38
	(0, 250)	0.80	32.19	22.61	0.76	0.24	25.18	0.16	22.18
	(250, 500)	0.52	86.36	24.85	0.12	0.88	47.87	0.35	35.58
	>500	0.84	129.21	19.31	0.96	0.04	109.43	−0.27	36.39
AND-1.7, AND-1.5, AND-1.4	Mean	0.82	74.73	37.45	0.71	0.29	55.38	0.66	15.51
	S.D.	0.07	21.95	11.00	0.17	0.17	14.03	0.09	0.60
	(0, 250)	0.68	42.05	34.09	0.60	0.40	34.28	−0.14	17.17
	(250, 500)	0.47	57.43	16.31	0.16	0.84	46.54	−0.59	17.80
	>500	0.55	49.97	9.45	−31.94	32.94	42.15	−10.64	17.52
AND-1.1	Mean	0.84	26.89	35.35	0.79	0.21	20.20	0.74	12.50
	S.D.	0.88	9.06	11.91	0.15	0.15	6.82	0.10	0.62

S.D.: standard deviation. Range study was carried out with estimated data set of the model with best performance. Means and S.D. were calculated among all repetitions of models calibrated and validated with best performance architecture.

have demonstrated that these kinds of models are well suited for algal blooms prediction.

Lee et al. (2003) performed an optimisation of the inputs variables of the ANN models. Their final results suggested that the simpler the network, the better result was obtained. The network that consisted of time-lagged Chl-a only as input was considered the optimal network. The conclusions of this study agree with the results obtained for *D. acuminata* concentration forecasting in Huelva area and suggest that algal biomass at time t is determined from what happened during previous weeks. In this way all ecological information appeared to be embodied in the algal concentration itself. The addition of numerous inputs, such as water temperature or solar irradiation, may present the neural network noise rather than useful information.

Ranges analysis shows that models calibrated with AND-1.11, AND-1.10 and AND-1.9 time series achieved good accuracy measures in the highest range. This range fits with critical thresholds established by the HAB monitoring programme in the study area. In AND-1.8 and the group formed by zones AND-1.7, AND-1.5 and AND-1.4, the error terms R^2, %SEP, RMSE and AIC calculated on range >500 showed a good performance, with values of %SEP lower than 10 and values of R^2 higher than 0.8. However, such measures do not give a real indication of the neural network performance in that range. The poor results obtained of the error terms PI (both models), ARV and E (group case) suggested that it is necessary to carry out a multicriteria assessment with different evaluation procedures and at different intervals of concentrations. Similar conclusions were obtained by Legates and McCabe (1999).

Quality data are necessary to achieve accurate predictions. High frequency errors occur mostly due to sampling, phytoplankton counting methods, standardization process carried out and the particular

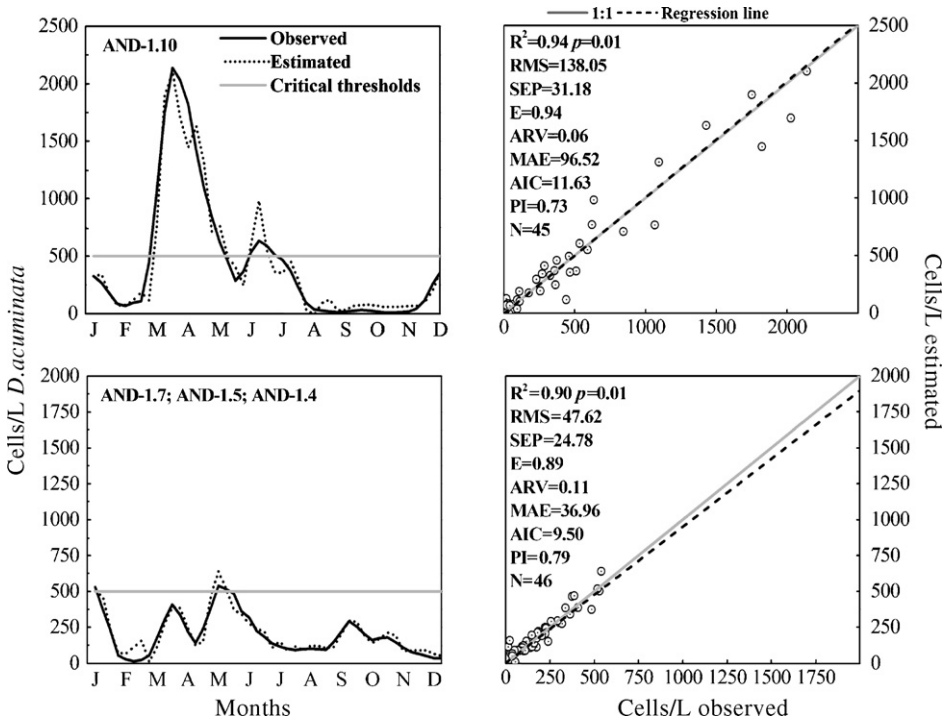

Fig. 5. One-step-ahead prediction of weekly concentrations for the validation period (zones AND-1.10 and the group AND-1.7, AND-1.5 and AND-1.1) and scatterplot between observed and forecast concentration for the model with the best performance of AND-1.11. The critical threshold is 500 cells/L and was established by the Andalucía HAB monitoring programme.

characteristics of each zone. In general, models calibrated with AND-1.11, AND-1.10 and AND-1.9 time series have better performance than networks calibrated with AND-1.8, group and AND-1.1 data sets. Good results from validation of group (AND-1.4, AND-1.5 and AND-1.7) data with AND-1.11 best model showed that time series from those zones (AND-1.4, AND-1.5 and AND-1.7) could have lost information when they were merged.

The hydrological characteristics of each zone have not been taken into account in this study. The Huelva coast is the northern boundary of the Gulf of Cádiz. The Gulf of Cádiz has been the focus of fieldwork and research but most of the literature deals with only the deep water mass circulation. There is evidence for the existence of a stable anticyclonic circulation of warm water in the central and southern part of the Gulf. Folkard et al. (1997) found that this temporal evolution of surface circulation patterns in the Gulf of Cádiz was significantly correlated with the zonal wind in the region. When prevailing winds in the Gulf are from the west, superficial and coastal waters move into the centre of the Gulf. Strong westerlies force cold surface waters around Cape S. Vicente to travel east along the coast. In contrast, easterlies move surface waters from the centre of the Gulf inshore. Surface currents and therefore plankton distribution, are wind induced in the study area. Therefore, the addition of wind data as a new key variable to the net could be clearly beneficial and improve ANN results. It is necessary to carry out a thorough analysis to determine which these key variables are. High variability among PCA yearly results could be due to different meteorological conditions among different years. Runoff and fluvial discharge may be taken into account too. Seasonal variability of salinity and nutrient concentration due to river discharge and rainfall could explain differences among plankton populations of harvesting zones. Zones AND-1.1, AND-1.7 are located in estuarine mouths and AND-1.11 is closed to the mouth of the Guadalquivir River.

PCA results for the whole period separated two regions within the study area delimited by Piedras River mouth. AND-1.11, AND-1.10 and AND-1.9 zones are partially overlapped to ‘‘coastal strip’’ area defined by Vargas et al. (2002). This area shows some oddities related to primary production and presents the highest temperatures and nutrient concentrations in summer than the rest of the basin.

5. Conclusions

Our results show that ANN models have a great generalisation power. Models predictions significantly fit trends in *D. acuminata* observed data in validation periods. ANN are, in this case, an effective harmful algae bloom forecasting tool and could be integrated in HABs monitoring programmes carried out in the study area.

Further studies are necessary to apply multivariate models and incorporate new key input variables. Better results may be expected after pruning non-useful connections and using new kind of algorithms more appropriated for time series forecasting.

In spite of the results obtained, it is necessary to frame the utility of the model (its forecast capacity) in the context of a short-medium time period. In this framework, the model can be used to estimate the *D. acuminata* concentration values when the real value cannot be obtained, to estimate interpolated data between two consecutives samples and to simulate different scenarios. Also, the model may be used to detect important changes in the ecosystem organisation because a lack of fit between observed and estimated data will indicate that new pattern must be incorporated to the model and therefore the model should be recalibrated and revalidated.

Acknowledgements

The authors are grateful to the Agriculture and Fisheries Department of Andalucía Government and to the Public Company for Agrarian and Fishing Development, for the access to their phytoplankton database. All colleagues participating in the monitoring program are acknowledge for fruitful co-operation. In special to Luz Mamán Menéndez as head of the Laboratorio de Control de Calidad de los Recursos Pesqueros (L.C.C.RR.PP.), Idelfonso Márquez Pascual and Manuel Aguilar Perea. Thanks, also, to P. Culverhouse and P. Gentien for their valuables comments.

References

Abrahart, R.J., See, L., 2000. Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments. Hydrol. Process. 14, 2157–2172.

Anctil, F., Rat, A., 2005. Evaluation of neural network streamflow forecasting on 47 watersheds. J. Hydrologic. Engrg. 10 (1), 85–88.

Barciela, R., García, E., Fernández, E., 1999. Modelling primary production in a coastal embayment affected by upwelling using dynamic ecosystem model and artificial neural networks. Ecol. Model. 120, 199–211.

Belgrano, A., Malmgren, B., Lindahl, O., 2001. Application of artificial neural networks (ANN) to primary production time-series data. J. Plankton Res. 23, 651–658.

Blauw, A.N., Anderson, P., Estrada, M., Johansen, M., Laanemets, J., Peperzak, L., Purdie, D., Raine, R., Batear, E., 2006. The use of fuzzy logic for data analysis and modelling of European harmful algal blooms: results of the HABES project. Afr. J. Mar. Sci. 28 (2), 365–369.

Borrego, J., Morales, J.A., Pendón, J.G., 1993. Holocene filling of an Estuarine Lagoon along the Mesotidal Coast of Huelva: the Piedras River mouth, southwestern Spain. J. Coastal Res. 9 (1), 242–254.

Fernández, R., Mamán, L., Jaén, D., Márquez, I., 2003. Control y seguimiento del fitoplancton tóxico en las costas andaluzas durante los años 2001 y 2002. VIII. Reunión Ibérica Sobre Fitoplancton Tóxico y Biotoxinas. Universidad de la Laguna, pp. 99–107.

Folkard, A.M., Davies, P.A., Fiúza, A.F.G., Ambar, I., 1997. Remotely sensed sea surface thermal patterns in the Gulf of Cádiz and Strait of Gibraltar: variability, correlations, and relationships with the surface wind field. J. Geophys. Res. 102, 5669–5683.

Griñó, R., 1992. Neural networks for univariate time series forecasting and their application to water demand prediction. Neural Netw. World 2 (5), 437–450.

Gutiérrez-Estrada, J.C., de Pedro-Sanz, E., López-Luque, R., Pulido-Calvo, I., 2004. Comparison between traditional methods and artificial neural networks for ammonia concentration forescasting in an eel (*Anguilla anguilla* L.) intensive rearing system. Aquacult. Eng. 31, 183–203.

Johansen, M., Hansen, G., Lindahl, O., Raine, R., Purdie, D., Blauw, A., 2004. Integrated results on *Dinophysis*, HABES. Final Report. WWW page, http://www.habes.net.

Karul, C., Soyupak, S., Cilesiz, A., Akbay, N., Germen, E., 2000. Case studies on the use of neural networks in eutrophication modelling. Ecol. Model. 134, 145–152.

Kitanidis, P.K., Bras, R.L., 1980. Real time forecasting with a conceptual hydrological model. 2. Applications and results. Water Resour. Res. 16 (6), 1034–1044.

Lee, J., Igarashi, T., Fraga, S., Dahl, E., Hovgaard, P., Yasumoto, T., 1989. Determination of diarrhetic shellfish toxins in various dinoflagellates species. J. Appl. Phycol. 1, 147–151.

Lee, J., Huang, Y., Dickman, M., Jayawardena, A., 2003. Neural network modelling of coastal algal blooms. Ecol. Model. 159, 179–201.

Legates, D.R., McCabe Jr., G.J., 1999. Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35 (1), 233–241.

Lindahl, O., 1986. A dividable hose for phytoplankton sampling. In Report of the working group on phytoplankton and management of their effects. International Council for the Exploration of the Sea. C.M.1986/L: 26, annex 3.

Maier, H., Dandy, G., Burch, M., 1998. Use of artificial neural networks for modelling cyanobacteria *Anabaena* spp. in the River Murray, South Australia. Ecol. Model. 105, 257–272.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models. I. A discussion of principles. J. Hydrol. 10, 282–290.

Qi, M., Zhang, G.P., 2001. An investigation of model selection criteria for neural network time series forecasting. Eur. J. Oper. Res. 132, 666–680.

Recknagel, F., French, M., Harkonen, P., Yabukana, K., 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecol. Model. 96, 11–28.

Rumelhart, D.E., Hinton, G.E., Willians, R.J., 1986. 'Learning' representations by backpropagation errors. Nature 323, 533–536.

Sarkar, R., Chattopadhayay, J., 2003. Occurrence of planktonic blooms under environmental fluctuations and its possible control mechanism-mathematical models and experimental observations. J. Theor. Biol. 224, 501–516.

Scardi, M., 1996. Artificial neural networks as empirical models of phytoplankton production. Mar. Ecol. Prog. Ser. 139, 289–299.

Tsoukalas, L.H., Uhrig, R.E., 1997. Fuzzy and Neural Approaches in Engineering. Wiley Interscience, New York, 587 pp.

Utermöhl, H., 1958. Zur vervollkommnung der quantitativen photoplankton-methodik. Mitt. Int. Ver. Limnol. 9, 1–38.

Vargas, J.M., García-Lafuente, J., Delgado, J., Criado, F., 2002. Seasonal and wind-induced variability of sea surface temperature patterns in the Gulf of Cádiz. J. Mar. Syst. 38 (3/4), 205–219.

Ventura, S., Silva, M., Pérez-Bendito, D., Hervás, C., 1995. Artificial neural networks for estimation of kinetic analytical parameters. Anal. Chem. 67 (9), 1521–1525.

Whitehead, P., Howard, A., Arulmani, C., 1997. Modelling algal growth and transport in rivers: a comparison of time series analysis, dynamic bass balance and neural network techniques. Hydrobiologia 349, 39–46.

Yabukana, K., Hosomi, M., Murakami, M., 1997. Novel application of backpropagation artificial neural network formulated to predict algal bloom. Water Sci. Technol. 36 (5), 89–99.