# Using General linear model, Bayesian networks and Naive Bayes classifier for prediction of *Karenia selliformis* occurrences and blooms

**Wafa Feki-Sahnoun[1]*, Hasna Njah[2], Asma Hamza[1], Nouha Barraj[3], Mabrouka Mahfoudi[1], Ahmed Rebai[4], Malika Bel Hassen[3]**

[1] Institut National des Sciences et Technologies de la Mer, Centre de Sfax, Rue Madagascar, BP 1035, Sfax, CP 3018, Tunisie.

[2] Faculté des Sciences Économiques et de Gestion de Sfax, Route de l'Aéroport Km 4 Sfax 3018 Tunisie. Laboratoire de Multimedia, InfoRmation Systems and Advanced Computing Laboratory, Pôle technologique de Sfax, Route de Tunis Km 10 BP 242 CP 3021, Sfax, Tunisie

[3] Institut National des Sciences et Technologies de la Mer (INSTM), 28 rue 2 mars 1934, Salammbô 2025, Tunisie

[4] Centre de Biotechnologie de Sfax, Route Sidi Mansour Km 6, BP 1177, 3018 Sfax, Tunisie

∗Corresponding Author: **Wafa Feki-Sahnoun**

wafafeki@yahoo.fr

Tel: +216 20247428; fax: +216 74497989.

njah.hasna@gmail.com (H. Njah)

asma.hamza@instm.rnrt.tn (A. Hamza)

barraj.nouha@instm.rnrt.tn (N. Barraj)

mabroukamahfoudi@yahoo.fr (M. Mahfoudi)

ahmed.rebai@cbs.rnrt.tn (A. Rebai)

belhassen.malika@instm.rnrt.tn (M. Bel Hassen)

## Abbreviations

Generalized Linear Mixed-effect Model (GLMM)

*Karenia selliformis* (*K. selliformis*)

Analysis of Variance (ANOVA)

General Linear Model (GLM)

Bayesian Network (BN)

Naive Bayes classifiers (NB)

Tunisian National Meteorological Institute (INM)

Akaike Information Criterion (AIC)

Bayesian Information Criteria (BIC)

SamIam (Sensitivity Analysis, Modeling, Inference And More software)

Directed Acyclic Graph (DAG)

Conditional Probability Tables (CPTs).

Maximum likelihood (ML)

Maximum A Posteriori (MAP)

Evaporation (Evap)

Insolation (Insol)

Salinity (Sal)

Humidity (Humid)

Water Temperature (WatT)

## Research highlights

- Bayesian networks (BN) performed better than Naive Bayes and General linear model.
- All models incriminate high salinity in the blooms and site for the occurrences.
- In BN, all relationships are more explanatory due to their inter-dependencies.
- A non-linear behavior between salinity and species concentrations was observed.


## Abstract

The prediction of the dinoflagellate red tide forming *Karenia selliformis* is a relevant task to aid optimized management decisions in marine coastal water. The objective of the present study is to compare different modeling approaches for prediction of *Karenia selliformis* occurrences and blooms. A set of physical parameters (salinity, temperature and tide amplitude), meteorological constraints (evaporation, air temperature, insolation, rainfall, atmospheric pressure and humidity), sampling months and sampling sites are used. The model prediction included General Linear Model (GLM), Bayesian Network (BN) and the simplest BN type which is, Naive Bayes classifier (NB). The results showed that three models incriminated high salinity in *Karenia selliformis* blooms and the sampling sites, mainly Boughrara lagoon, in the occurrences. The BN performed better than linear models (NB and GLM) for both *Karenia selliformis* occurrences and blooms prediction. This later is related to the facts that BN considered the inter-independency between predictive variables and that the relationships between the variables and the outcome are often non-linear such us; the transition to bloom situations appeared to be triggered by a salinity threshold. This study is useful in the

management of this ecosystem so as to use the best disposal options in the early prediction of the toxic blooms.

**Keywords:** *Karenia selliformis*, Naive Bayes classifier, General linear model, Bayesian Network, hydro-meteorological parameters, the Gulf of Gabès.

## 1. Introduction

The harmful *Karenia* species are found throughout the world. They have been described as a result of investigations into extensive animal mortalities or human health problems. They have become the most studied species of harmful algae with extensive investigations on the physiology and bloom formation. *Karenia selliformis* (Hansen et al. 2004) has been the most abundant species causing severe harmful blooms in the Gulf of Gabès (Hamza and El Abed, 1994). Since the year 1990, *K. selliformis* blooms have occurred annually along the coast, and represented over 64% of the reported blooms in this area (Feki et al. 2008, 2013). It has been argued that their proliferation has been usually related to shellfish toxicity (Ben Naila et al. 2012; Marrouchi et al. 2009; Medhioub et al. 2010). Little is known about the effect of environmental factors on *K. selliformis* occurrences and blooms. Information on optimal growth of this species under controlled temperature and salinity was documented in culture experiments (Medhioub et al. 2009). Tentatively models have been developed to apprehend the effects of physical and meteorological variables on Ks occurrences and blooms in the Gulf of Gabès. A generalized linear mixed-effect model (GLMM) incriminated mainly water temperature and some nutrients in the spatiotemporal occurrences of *Karenia selliformis* (Feki et al. 2013). Bayesian network approach showed that the bloom can be predicted based on salinity threshold

95 (Feki-Sahnoun et al. 2017). However, the best performance model having the highest goodness
96 of fit on the prediction of *Karenia* blooms and occurrences still needs to be determined.
97 To date, a large variety of statistical models are available to analyze a relationship
98 between environmental variables and species distributions (i.e. cross validation criteria,
99 ANOVA) and one of the most popular is the General linear model (GLM) (e.g. McCulloch et
100 al. 2008). Typically, the biological and physical processes, generating this data, are highly
101 complex, resulting in multiple correlations/dependencies between covariates and also between
102 outcome variables. Standard statistical approaches have a limited ability to describe such
103 interdependent multi-factorial relationships. In the last decades, Bayesian Network's modelling
104 (BN) has been widely used in solving environmental problems (Borsuk et al. 2004, 2006;
105 Bromley et al. 2005; Feki-Sahnoun et al. 2017; Pollino et al. 2007; Smith et al. 2007; Zaffalon,
106 2005) to analyze multi-dimensional data. Among the BN models, the naïve Bayes (NB) model
107 appears to be the most popular (Aguilera et al. 2013; Fytilis and Rizzo, 2013; Markus et al.
108 2010; Ropero et al. 2014, 2015). Despite its linearity and simplicity, the NB has been found to
109 perform surprisingly well (Friedman et al. 1997) particularly in many complex situations
110 (Domingos and Pazzani, 1997; Zhang, 2004; Boets et al. 2015).
111 The present study compares results from Bayesian networks [BN, Bishop, 2006; Pearl,
112 1985)], naive Bayes classifier [NB, Friedman et al. 1997] and General linear model [GLM,
113 McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972] in the modeling of Ks
114 occurrences and blooms; which is based on a set of physical and meteorological variables. In
115 contrast to Feki-Sahnoun et al. (2017) who used the same dataset and focused only on the
116 Bayesian network, this research elaborated upon the comparison between the three models
117 based on a new threshold concentration for bloom and non-bloom conditions established using
118 the Feki-Sahnoun et al. (2017) output results. The choice of these models is due to the fact that
119 GLM analysis is a well-known method that allows predicting a target variable (in this case,

120 *Karenia selliformis*) conditionally on a set of variables by fitting a regression model using least
121 squares (Hastie et al. 2009). Many different models could be fitted by including different
122 subsets of the available observed variables and interaction terms between them. The challenge
123 in implementing this technique is the selection of the best subset of variables, as the inclusion
124 of correlated predictors may result in increased standard errors of the regression coefficients
125 which cause prediction's sensitivity to model changes (Burnham and Anderson, 2002).
126 Whereas GLMs are typically data-driven models techniques (Hastie and Tibshirani, 1990;
127 Madsen and Thyregod, 2011; McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972),
128 BNs can be based on expert knowledge and/or stakeholders (Aguilera et al. 2011; Cain et al.
129 2003; Chan et al. 2012; Haapasaari and Karjalainen, 2010; Haines-Young, 2011; McVittie et
130 al. 2015; Wang et al. 2009) in order to analyse how a system works, to seek information on the
131 appropriate probabilities, or to explore the usefulness of the decision process (Gonzalez-Redin
132 et al. 2016). BN is an alternative modeling approach applying different criteria for model
133 selection. This method consists of a graphical modeling tool that can be used to construct a
134 predictive model by factorizing the posterior density distribution function assuming a set of
135 stable conditional dependencies (Jensen, 1996; Lauritzen and Spiegelhater, 1988; Pearl, 1988).
136 Among BN models, naive Bayes classifier is the simplest; because of the fact that its variables
137 are conditionally independent given the class variable. It has the strength to not only provide a
138 prediction, but also the estimated probability associated with each possible outcome (Fernandez
139 et al. 2010).

140 The objective of the present study is to compare the efficiency of the linear (GLM, NB)
141 and non-linear (BN) models in order to predict *K. selliformis* occurrences and blooms in the
142 Gulf of Gabès using a set of meteorological (rainfall, atmospheric pressure, insolation, ...) and
143 physical (temperature, salinity, tide) variables. By comparing BN with GLM and NB using

identical data, the likely impact of such an analytical difference on the inferences was highlighted.

## 2. Material and methods

### 2.1. Study area

The present study was carried out in the Gulf of Gabès (Tunisia, Eastern Mediterranean Sea), in a wide continental shelf area extending from 9.5 to 12°E in longitude and from 33 to 35.5°N in latitude and sheltering various islands (Kerkennah and Djerba) and lagoons (Boughrara and El Bibane). The Boughrara lagoon is surrounded by solar saltern areas where net evaporation is very high and soil washing is negligible due to limited freshwater supplies coupled with scanty rainfall.

The bathymetry is characterized by a low slope; the 50-m depth is reached at a distance of 100 km (Fig. 1). The tide is semidiurnal, being among the highest in the Mediterranean and having a maximum range of about 2 m generated by resonance phenomena (Sammari et al. 2006). The climate is dry and sunny with strong easterly winds.

Despite being oligotrophic, this Gulf has the peculiarity of being highly productive and accounts for 65% of Tunisian fish production (DGPA, 2005–2009) and is a well-known habitat for marine turtles (Jribi et al. 2008; Lotze and Worm, 2009; Maffucci et al. 2006). Although it has huge importance for the local economy and wildlife conservation, the Gulf of Gabès has been subject to the strong pressures of urbanization, industry, fisheries, tourism and anthropogenic releases (Béjaoui et al. 2004; Ben Brahim et al. 2010; Rekik et al. 2012). It has been the subject of increased anthropogenic interference threatening the entire ecosystem.

### 2.2. Analysis process

166 The flowchart in Figure 1was followed in order to conduct a consistent analysis. The
167 analysis (Fig. 1) includes three major steps: Data preparation, Discretization and Model
168 Prediction.

169 **2.2.1. Data preparation**

170 Data were collected in the framework of the National Phytoplankton Monitoring Program
171 in the shellfish harvest areas. This program has been operating since 1995 and covers fifteen
172 sites with weekly measures (Fig. 2). The monitoring was performed on a regular schedule all
173 year round. Each site can include more than one sampling station totalizing thirty-two sampling
174 stations (Fig. 2). The temporal windows of the dataset used for the analysis; that covered the
175 period ranging from 1997 to 2007; are considered as the most consistent in terms of data
176 shortage and because the identification of the species was carried out by the same team of
177 scientists.

178 Water for phytoplankton identification (1l) was sampled with a sampling water-bottle.
179 Temperature and salinity were measured for each water sample using a Handheld
180 Multiparameter Instrument: WTW Multi 340i/SET. Samples were fixed with lugol (4%)
181 solution and phytoplankton was enumerated by an inverted microscope using the Utermöhl's
182 method (Utermöhl, 1958).

183 Physical and meteorological parameters were simultaneously collected. Meteorological
184 data were provided by the Tunisian National Meteorological Institute (INM) and consisted of
185 air temperature, rainfall, evaporation, insolation, humidity and atmospheric pressure collected
186 from 15 meteorological stations. The tide amplitude, difference between high and ebb tides
187 considered as the variable indicating the tide effect, was obtained from the tide gauge stations
188 located in the main Tunisian ports and operated by the Tunisian Hydrographic and
189 Oceanographic Services.

The toxic dinoflagellate *Karenia selliformis* abundance was considered as the investigated biological variable. The blooms generally define phytoplankton concentrations having more than $10^6$cells $L^{-1}$ (Lassus, 1988). In this study, a concentration of $10^5$ cells $L^{-1}$ was considered as indicative of *K. selliformis* blooms since this concentration limit was responsible for the red water discoloring in this area (Feki et al. 2008).

The dataset was filtered by removing the missing data i.e., the cases where some information about the studied factors was unavailable. The final dataset contains over 1900 observations.

### 2.2.2. Data discretization

The models were applied to discrete variables. The choice of the class number and class limits was already defined in a previous analysis (Feki-Sahnounet al. 2017).With regards to *Karenia selliformis*, two discretization schemes were chosen. The first one used presence/absence of the species to assess factors influencing the species occurrences whereas the second one used bloom/non-bloom forming species to assess the factor affecting the species abundance (Table 1).

### 2.2.3. Model Prediction

A set of linear and nonlinear models were used as prediction tools. For modeling the effects of multiple factors (physical, meteorological and spatiotemporal parameters) on *K. selliformis* occurrences and blooms, General linear model (GLM), Bayesian network (BN) and naive Bayes classifier (NB) were developed.

Adjusted coefficient of determination ($R^2$), Akaike (AIC) (Akaike, 1974) and Bayesian Information Criteria (BIC) (Schwartz, 1978) were used as model quality indicators to best predict the phenomena. $R^2$ is based on the ratio between the variance explained by the model to the total variance, while AIC and BIC are based on the likelihood of the data given the model.

GLM analysis was performed using the R package *MASS* (Venables and Ripley, 2002). BN and NB models used algorithms implemented in the bnlearn package for R (R Development Core Team, 2017; Scutari, 2010) and SamIam (Sensitivity Analysis, Modeling, Inference And More software) (Chan and Darwiche, 2004) used to visualize the results.

#### 2.2.3.1. General linear model

The univariate general linear models (GLM) allows categorical predictors to be included. It was performed with stepwise procedures. The best-approximating GLM model with the highest $R^2$ in combination with the lowest AIC obtained for each combination of studied factors were selected as the final model (Burnham and Anderson, 2002). A final model included marginally significant descriptors ($0.05 < P < 0.1$), if they explained deviance and/or AIC notably improved with the descriptors in the model. GLM models were validated using graphical tests implemented in R (plot function of a GLM object) to test residuals normality and independence .

#### 2.2.3.2. Bayesian Networks

Bayesian Networks (BNs) have two components: a qualitative one represented by a Directed Acyclic Graph (DAG) depicting dependence relationships between variables and a quantitative one represented by Conditional Probability Tables (CPTs).

Figure 3A shows a simple BN of five variables {U, V, W, X, Y} and four edges U->W, V->W, W->Y and X->Y. In a DAG, each node represents a variable of interest (i.e., U, V, W, X and Y, Fig. 3A). Each variable is expressed according to its possible values continuous or discrete (i.e., the values of X are x1 and x2). Nodes are connected to each other by edges (directed arcs) to represent dependency relations between them (e.g., in Fig. 3A, X and W are the parent of Y).

The CPT contain, for each possible value of the variable associated to a node, all the conditional probabilities with respect to all the values' combinations of the variables associated to the parent nodes. Generally, leaf nodes in a BN (nodes which have parents and that do not have children) are nodes that are used to make a decision.

Concerning structure learning, the used score is the Bayesian Information Criterion (BIC) (Schwartz, 1978) which is a penalized version of the log-likelihood. The search algorithm; that is opted for; is the Tabu search (Glover, 1989), which was considered as a fast and determinist algorithm.

For parameters' learning, a Maximum likelihood (ML) inference estimating the probability of an event based on its occurrence frequency in the dataset (Neapolitan, 2003; Pearl, 1988) was adopted.

#### 2.2.3.3. Naive Bayes classifier

Naive Bayes classifier (NB) networks are the simplest model among BN. The term 'Naïve' refers to the strong independence assumption between variables (Friedman et al. 1997). NB is graphically represented by a hierarchical structure where the class node is the parent of all attribute nodes (Duda et al. 2001; Friedman et al. 1997). Therefore, the NB is probabilistically defined by the conditional probabilities of each attribute given the class node (Cooper and Herskovits, 1992). Each node; that is representing a given attribute; is associated to Conditional Probability Table (CPT) containing, for each possible value of the corresponding attribute, all the conditional probabilities with respect to the class nodes values.

Figure 3B shows an example of a class which is noted as "Y" and a set of four variables {U, V, W, X}. The corresponding NB structure is composed of five nodes {U, V, W, X, Y} and four edges {Y->U, Y->V, Y->W and Y->X}. Each variable is expressed according to its possible values continuous or discrete (*i.e.*, the values of X are x1 and x2). The nodes are

connected to class "Y" by edges in order to represent the dependency relations to the class. Such a structure is very useful for determining the impact of the attributes' values on the class' value.

The Maximum A Posteriori (MAP) estimation (Kramer and Sorenson, 1988) for estimating the CPT of each node in the NB models was adopted.

The goal in this section is to determine the posterior probabilistic information. Thus, the model can be used to predict the impact (in terms of probability) of introducing evidence for certain variables (model a posteriori). For example, if the evidence (the observation that the species is present), P (species=presence) = 1, is set in the class variable, the density functions of the remaining environmental variables will be modified. In this way, an approximation to the most probable configuration for the species presence can be obtained.

## 3. Results

### 3.1. *Karenia selliformis* occurrences

#### 3.1.1. General linear model

The results of GLM analysis in Table 2 and from Equation (1) show that *Karenia selliformis* occurrence depends mainly on Site, Month, Rain, Evaporation, Insolation and Salinity.

High significant relationships were observed for sites belonging to the central and southern part of the Gulf of Gabès namely G2, G3, M1, M2, M3, M6, S5 and S6 ($P<0.1$).Late summer, autumn and winter were generally the periods of high frequency in this area ($P<0.1$). High salinity and medium evaporation levels were responsible for Ks occurrences ($P<0.1$) (Table 2, Eq. 3).

Sampling sites (M6 and S5) and evaporation (medium level) show a negative relationship with Ks occurrences. Whereas the relationships with G2, G3, M1, M2, M3, S6, from August to December and salinity (high level) are positive (Eq. 2).

284 The stepwise GLM model selection allowed decreasing the AIC from 2158.81 to 2144.24.
285 The retained GLM model has the following structure:
286 Ks= Intercept + Site + Month + Rain + Evap + Insol + Sal **(1)**
287 =-1.84 + 1.16(G2) + 1.22(G3) + 0.76(M1) + 1.44(M2) + 0.87(M3) + (-1.22) (M6) + (-
288 0.56) (S5) + 0.5(S6) + 0.54(Aug) + 0.52(Dec) + 0.94(Nov) + 1.31(Oct) + 0.54(Sep) + (-0.36)
289 (Evap b) + 0.74(Sal c) **(2)**
290 A numerical example of *Karenia selliformis* occurrences estimation in the M2 site and
291 during September is shown below:
292 Ks=-1.84 +1.16 (0) + 1.22 (0) + 0.76 (0) + 1.44 (1) + 0.87 (0) -1.22 (0) -0.56 (0) + 0.56 (0) +
293 0.54 (1) + 0.52 (0) + 0.94 (0) + 1.31 (0) +0.54 (0) + -0.36 (Evap b) + 0.74 (Sal c)
294 =-1.84 +1.44 + 0.54 - 0.36 (Evap b) + 0.74 (Sal c) **(3)**
295 From Eq. (3) and as shown in Table 2, the estimated values *Karenia selliformis*
296 occurrence is dependent upon the levels of evaporation and salinity.

297 **3.1.2. Naive Bayes classifier**

298 The resulting NB model is shown in Fig. 4. The introduction of the evidence "presence
299 of Ks" changes the probability distribution of the features because they are directly connected
300 with the Ks variable.
301 The marginal probability distributions of the NB model shows that it is likely to find the
302 Ks during October and July in the southern part of the Gabès' Gulf both in M2 and M3 sites
303 (Fig. 4A and B).
304 The model also suggests that it is likely to find the species where there is a low rainfall
305 and slight high air temperature levels. The remaining variables (humidity, evaporation,
306 insolation, tide amplitude, salinity, water temperature and atmospheric pressure) vary in the
307 medium levels (b Class) (Fig. 4B).

### 3.1.3. Bayesian networks

The BN linking the physical and meteorological variables and the presence/absence of the species shows (Fig. 5) that only the factor site influences the *Karenia selliformis* occurrence. CPTs show that the posterior probability of the species presence was high in sites M2, M3, G3, G2, G1, S6 and M1 (Figs. 5A and B), all of these sites belong to the central and southern part of the Gulf of Gabès.

### 3.1.4. Models comparison

The main difference of BN with respect to the NB is the relationship between the variables. In fact; this distinction increases the number of arcs in the structure and its complexity, but also improves the accuracy and expressivity of the model.

BN had not detected the associations founded on GLM and NB. Hence, while GLM and NB established a direct association between three and six variables, BN identified only site-dependent factor linked to the outcome.

M2 and M3 show posterior probability similar to the NB model (27.29 % and 10.51 % respectively) (Fig. 5B and 4B).In general, both NB and BN models show similar probabilities trends (Fig. 5 and 4). However, BN varies in the definition of relationships between the variables in the model, so that each variable is influenced, not only by the class variable Ks, but also by the variables directly connected with it in the network.

The prediction performance of the GLM model is moderate having adjusted $R^2$ value of 0.17 (Table 4). The NB and BN models appeared to have the best fit and performed almost equally well having an adjusted $R^2$ of 0.69 and 0.70 respectively, while BN was found to perform slightly better (Table 4).

## 3.2. *Karenia selliformis* blooms

### 3.2.1. General linear model

GLM shows that *Karenia selliformis* blooms mainly depends on Site, Month, Evaporation, Humidity, Water Temperature and Salinity (Table 3; Eq. (4)).

Humidity (Class b and c) shows a negative relationship with *Karenia selliformis* blooms, whereas evaporation (Class c) and water temperature (Class c) show positive correlations (Table 3, Eq. 5).

The GLM model selection, allows decreasing the AIC from 588.69 to 565.76.

The selected GLM model has the following structure:

Ks ~ Intercept + Site + Month + Evap + Humid + WatT + Sal **(4)**

Ks=22.54 + 1.59 (Evap c) -2.12 (Humid b) - 1.83 (Humid c) + 1.57 (WatT c) **(5)**

A numerical example of *Karenia selliformis* blooms estimation in the M2 site is shown below:

Ks=22.54 + 1.59 (Evap c) -2.12 (Humid b) - 1.83 (Humid c) + 1.57 (WatT c) **(6)**

From Eq. (6) and as shown in Table 3, the estimated value *Karenia selliformis* blooms are dependent upon the levels of Evaporation, Water temperature and Humidity.

### 3.2.2. Naive Bayes classifier

The model suggests that it is likely to find blooms not only when there is a high salinity and air temperature, but also when there are remaining variables that are situated in medium level (b class) (Fig. 6B). It can also be seen regarding bloom a slightly high conditional probability (38%) for low tide amplitude level (a Class). As Ks occurrences, NB model also shows that it is likely to find the bloom in M2 (85.29%) between the two-temporal scenario of summer and autumn (Fig. 6B).

### 3.2.3. Bayesian networks

354 The BN schema illustrating the physical, meteorological variables and the bloom/non-
355 bloom of the *K. selliformis* (Fig. 7) shows that salinity directly affects Ks blooms and represents
356 its single parent (Fig. 7). Salinity has site as parent, and site is dependent on tide, atmospheric
357 pressure, air temperature and month. CPTs show, that the probabilities of low salinity (a Class)
358 are very low and are not exceeding 6% (Fig. 7 A and B). It is worth noting that the absence of
359 blooms is high (70%) for medium salinity levels (b Class) particularly in M2 site (25%). (Fig.
360 7A). During Ks blooms, the probability to have high salinity (Class c) significantly increased
361 (80%). The highest probabilities are recorded in M2 (56%) (Fig. 7B). These results show that
362 the blooms were largely determined by salinity and were expected to occur in salinities higher
363 than 42.5, which are detected in site M2 (Fig. 7B).

364 **3.2.4. Models comparison**

365 BN do not detect the associations found on GLM and NB; except for salinity (Fig. 7).
366 Based on the prediction performance indicator, all models perform moderately with the better
367 performance for BN model ($R^2$=0.58) followed by the NB ($R^2$=0.54) and GLM models
368 ($R^2$=0.48) (Table 4).

369 **4. Discussion**

370 **4.1. Models comparison**

371 In computing the tree models, several strengths and constrains relative to one or other
372 models were raised. They are summarized in Table 5. One of these constraints was the exclusive
373 use of discrete variables which was considered as an important weakness of BNs (Landuyt et
374 al. 2013). Most environmental variables are continuous whilst GLM can use continues or
375 discrete data, Bayesian networks usually build the model over discrete domains, so that
376 continuous variables need to be first discretized (Uusitalo, 2007). Discretization implies
377 capturing only rough characteristics of the original distribution (Friedman and

Goldszmidt,1996) and loss of statistical information (Aguilera et al. 2011, 2010; Alameddine et al. 2011; Chen and Pollino, 2012; Hamilton et al. 2015; Jensen, 2001; Lucena-Moya et al. 2015; Meineri et al. 2015; Meyer et al. 2014; Nielsen and Jensen, 2007; Uusitalo, 2007). However, the discretization is beneficial in this study and the BN has succeeded in providing reliable results and generalized BNs which do not provide biased results. Thus, the discretization can mitigates the impact of the noisy data (incorrectly saved values / typos / wrong values).

The models also confirmed the role of salinity as a key factor for Ks bloom (Feki et al. 2008; Feki-Sahnoun et al.2017). The results of GLM analysis in Table 2 and from Equation (1) refutes the exclusion of salinity and the retention of water temperature from the generalized linear mixed-effect model (GLMM) established for the *K. selliformis* occurrences in the Gulf of Gabès (Feki et al. 2013). Therefore, BN, NB and GLM are likely to identify the same factors when associations are strong and highly significant. The advantage of BN, comparing to the two others models, is manifested in its ability to easy establish a direct association between bloom and salinity and between occurrence and sampling site. Indeed, BN identifies a network of inter-dependent factors linked to the outcome. It is more informative about potential causal pathways in the Ks occurrences and blooms than GLM and NB (Table 5). This potential advantage of the BN method would specifically be useful for observational studies with large number of variables, where causal and time relationships are often unknown (Pittavino et al. 2017). Moreover, in the retained GLM model (Table 2) some variables are interdependent such us insolation and evaporation both were related to salinity. Hence, the interdependencies between variables were revealed in BN that might not be discovered in GLM and NB, as the latter impose a linear relationship between covariates and the outcome. Thus, some hidden dynamic was still not detected while using GLM and NB when interaction terms were considered (Table 5). For instance; the relationships between tide, site and salinity (Fig. 7) are

403 explained by the fact that the species achieved its peak density in a high salinity and in a semi-
404 enclosed lagoon (M2), compared to the open coastline, incriminating the low water dilution
405 rate. In this case low tide amplitude, due to the weak water advection out of the lagoon, allows
406 to sustain the growth and the bloom's maintenance of the species (Feki-Sahnoun et al. 2017).
407 Hence, BN had an advantage over GLM and NB because it can disentangle the complex nature
408 of the data, stratifying further the presented internal mechanisms between the variables
409 (Pittavino et al. 2017).

410 The BN and NB performed almost similarly for both occurrences and blooms (Table 4).
411 The two models also yielded very close range of probabilities regarding all influencing
412 variables in the BN of Ks occurrences such as sampling site identified as the only parent (e.g.
413 given the evidence of presence: M2=27%) (Figs.4B and 5B) and salinity identified as the only
414 parent in the BN regarding Ks bloom (given the evidence of presence: c=79%) (Figs.6B and
415 7B). This could be in part explained by the reduced complexity of the studied dataset as shown
416 by Aguilera et al. (2010). Moreover, even performance indicators were rather close between the
417 two models, it was slightly higher for BN (Table 4). This could be explained by the
418 interdependency between variables considered in NB which might put more confidence on the
419 BN predicted probabilities (Boets et al. 2015). Indeed, BN models may better reflect the joint
420 probability distribution over the system's variables compared to a Naive Bayes classifier, as the
421 assumption of conditional independence among predictor variables, may not hold in reality
422 (Boets et al. 2015). In others words, while NB may predict occurrences or blooms better (Boets
423 et al. 2015), BN models may predict occurrence or blooms probabilities more accurately. This
424 may also explain the slightly weaker performance of the NB model compared to the BN model
425 to predict both presence/absence and bloom/non-bloom (Table 4). Although the number of arcs
426 in the structure and its complexity increases, no significant effect on model performance was
427 found and the BN improves usually the accuracy, flexibility and expressivity (Pearl, 1988).

Nevertheless, it is expected that; by increasing the number of states beyond the relevant range (more than three) can result in significant performance loss. Increasing the number of states would result in large CPTs wherein the probabilities that need to be estimated are conditional on very specific combinations of environmental conditions. This increases the chance of having no or only a limited number of data records available to learn these conditional probabilities. This would result in a lower predictive performance on an independent test dataset (Boets et al. 2015).

More importantly; the NB model also shows that the introduction of the evidence "presence of Ks" changes the probability distribution of several features mainly salinity, site, month, air temperature, tide amplitude compared to the evidence of the Ks being absent (Fig. 4). The same figure was observed on the introduction of evidence "bloom of Ks" compared to the evidence of the Ks being non -blooming (Fig. 6). However, the BN schemes demonstrated that introduction of the evidence "presence of Ks" or "bloom of Ks" do not change the probabilities except that for the parent of Ks (sampling site and salinity respectively) (Figs. 5 and 7). These outcomes arise from the facts that the NB may overamplify the weight of the evidence of each attribute on the class (Friedman et al. 1997).

The probability of salinity exceeding 42.5 (c class) was very high (above 70%) in *K. selliformis* blooms (Figs. 6 and 7). This result was explained by a non-linear behavior between salinity and species abundances, and it suggests that the transition to high biomasses appeared to be triggered by a salinity threshold (Feki-Sahnounet al. 2017). This circumstance, might advantage the use of BN as a powerful computational technique for modeling complex relationships in situations; where the proper form of the relation between the variables is unknown or nonlinear (Chen and Billings, 1992; Felipe et al. 2014; Schmitt and Brugere, 2013). The conventional techniques such as GLM and NB, although they are widespread within ecology, but they present some short comings related to the facts that the relationships between

variables in environmental sciences are often non-linear and that data rarely have normal errors (Chen and Billings, 1992).

The overall conclusion is that BN are the recommended models since it has less constrains than the others (Table 5) and provide the highest statistical performance (Table 4).

### 4.2. Ecological implications and management recommendations

The three models incriminated the variable site in Ks occurrence confirming the north-south gradient of Ks occurrence already pointed out in the Gulf (Feki et al. 2013). This is due to the influence of the site M2 (the Boughrara lagoon) which is totalizing to itself more than 27% of the species occurrences (Figs. 4B and 5B) and more than 85% of the species blooming (Fig. 6B). One of the direct implication of this result is the consideration of the site M2 as a "hot spot " for the species proliferation which might imply paying more attention to the variability of the physical and meteorological variables identified in the bloom forming in this specific location.

As it happens salinity appears as the parameter that effect directly Ks blooms. Nevertheless, GLM retained water temperature as an explicative factor in Ks Blooms (Table 3, Eq (4)) and NB showed an increase in the associated probabilities of the blooms regarding the highest temperature level (Fig. 6).Whereas no link was pointed out by BN between Ks bloom and water temperature (Fig. 7). The correlations between high water temperature level and Ks blooms were often highlighted in several ecosystems (Carreto et al. 2001; Clement et al. 2001; Uribe and Ruiz, 2001) and in the Gulf of Gabès (Dammak-Zouari et al. 2009) stressing the role of this parameter in the bloom enhancement. In the BN's variables interconnections, water temperature itself is affected by air temperature the one of Ks's parents (Fig. 7).This suggests that temperature on itself might not be the factor affecting Ks but it might rather impact on salinity which in turn influenced Ks. The mechanism by which salinity could affect the bloom

enhancement is not yet well identified: is this a purely physiological effect or could involve also physical conditions is still to be investigated in order to statute whether the apparent direct effect between Ks blooms and salinity was rather an association than a causality.

The identification of a salinity threshold, revealed by BN, is a relevant information for the prediction of the blooms in the studied ecosystem, and it can be used to design and set up an early warning system for Ks blooms based on a real time observation of salinity. One of the direct application of such a system is the suspension of shellfish exploitation during the suspected period of high salinity; pending the species abundance determination and the associated toxicity tests performed.

## Conclusions

In terms of predictive ability, BN performed better than linear models (NB and GLM) regarding Ks occurrences and blooms prediction, probably due to the existence of nonlinear relationships with the salinity key variable.

BN and NB performed quasi equally in terms of performance indicator. BN has an advantage over NB because of its capacity to capture and illustrate graphically the data's natural complexity more effectively. In BN, all relationships between variables are modeled, which appears to be more explanatory in the view of the inter-dependencies between variables studies. BN can work together with NB for pre-selection of variables inputs.

The three investigate models converge on the identification of salinity as a key variable for the prediction of Ks occurrences and blooms. The salinity is in turn site dependent with more that 55% of the bloom concentrated in Boughrara lagoon which suggests using this parameter for the control of the Ks bloom in this specific location identified as a hot spot area.

## Acknowledgments

500 This work was supported by the PASRI (L'Agence Nationale de Promotion de la Recherche
501 scientifique)/MOBIDOC (Mobilisation de docteur pour la réalisation de Travaux de Recherche
502 dans l'Entreprise) funded Project (post-doctoral grant for the first author). The authors wish to
503 thank Mr. Hamdi DKHIL, English Teacher in Franklin Center Sfax (Tunisia) for having edited
504 this Paper.

505 **References**

506 Aguilera, P.A., Fernandez, A., Fernandez, R., Rumi, R., Salmeron, A., 2011. Bayesian networks
507 in environmental modelling. Environ. Model. Softw. 26, 1376–1388.
508 Aguilera, P.A., Fernandez, A., Reche, F., Rumi, R., 2010. Hybrid Bayesian network classifiers:
509 application to species distribution models. Environ. Model. Softw. 25, 1630–1639.
510 Aguilera, P.A., Fernandez, A., Ropero, R.F., Molina, L., 2013. Groundwater quality assessment
511 using data clustering based on hybrid Bayesian networks. Stoch. Environ. Res. Risk Assess. 27,
512 435–447.
513 Akaike, H., 1974. A new look at the statistical identification model. IEEE Trans. Auto. Control.
514 19, 716-723.
515 Alameddine, I., Cha, Y.K., Reckhow, K.H., 2011. An evaluation of automated structure
516 learning with Bayesian networks: An application to estuarine chlorophyll dynamics. Environ.
517 Model. Softw. 26, 163–172.
518 Béjaoui, B., Rais, S., Koutitonsky, V., 2004. Modélisation de la dispersion du phosphogypse
519 dans le Golfe de Gabès. Bull. Inst. Natl. Sci. Technol. Mer de Salammbô 31, 103–109.
520 Ben Brahim, M., Hamza, A., Hannachi, I., Rebai, A., Jarboui, O., Bouain, A., Aleya, L., 2010.
521 Variability in the structure of epiphytic assemblages of *Posidonia oceanica* in relation to human
522 interferences in the Gulf of Gabès, Tunisia. Mar. Environ. Res. 70, 411–421.

523 Ben Naila, I., Hamza, A., Gdoura, R., Diogene, J., Iglesia, P., 2012. Prevalence and persistence
524 of gymnodimines in clams from the Gulf of Gabès (Tunisia) studied by mouse bioassay and
525 LC-MS/MS. Harmful Algae. 18, 56–64.
526 Bishop, C.M., 2006. Pattern Recognition and Machine learning, Springer, New York.
527 Boets, P., Landuyt, D., Everaert, G., Broekx, S., Goethals, P.L.M., 2015. Evaluation and
528 comparison of data-driven and knowledge-supported Bayesian Belief Networks to assess the
529 habitat suitability for alien Macroinvertebrates. Environ. Model. Softw. 74, 92–103.
530 Borsuk, M., Reichert, P., Peter, A., Schager, E., Burkhardt-Holm, P., 2006. Assessing the
531 decline of brown trout (*Salmo trutta*) in swiss rivers using Bayesian probability network. Ecol.
532 Model. 192, 224–244.
533 Borsuk, M., Stow, C., Reckhow, K., 2004. A Bayesian network of eutrophication models for
534 synthesis, prediction, and uncertainty analysis. Ecol. Model. 173, 219–239.
535 Bromley, J., Jackson, N., Clymer, O., Giacomello, A., Jensen, F., 2005. The use of Hugin to
536 develop Bayesian networks as aid to integrated water resource planning. Environ. Model.
537 Softw. 20, 231–242.
538 Burnham, K.P., Anderson D.R., 2002. Model Selection and Multimodel Inference: a Practical
539 Information-Theoretic Approach, 2nd ed. Springer Verlag, New York.
540 Cain, J.D., Jinapala, K., Makin, I.W., Somaratna, P.G., Ariyaratna, B.R., Perera, L.R., 2003.
541 Participatory decision support for agricultural management. A case study from Sri Lanka.
542 Agric. Syst. 76, 457–482.
543 Carreto, J.L., Seguel, M., Montoya, N.G., Clément, A., Carignan, M.O., 2001. Pigment profile
544 of the ichthyotoxic dinoflagellate *Gymnodinium* sp. from a massive bloom in southern Chile. J.
545 Plankton Res. 23, 1171–1175.

546 Chan, H., Darwiche, A., 2004. Sensitivity analysis in Bayesian networks: From single to
547 multiple parameters, in: Chickering, M., Halpern, J. (Eds.), Proceedings of the 20th conference
548 on Uncertainty in Artificial Intelligence. AUAI Press, Arlington, VA, pp. 67–75.
549 Chan, T.U., Hart, B.T., Kennard, M.J., Pusey, B.J., Shenton, W., Douglas, M.M., Valentine, E.,
550 Patel, S., 2012. Bayesian network models for environmental flow decision making in the Daly
551 river, northern Territory, Australia. River Res. Appl. 28, 283–301.
552 Chen, S., Billings, S.A., 1992. Neural networks for nonlinear dynamic system modelling and
553 identification. Internat. J. Control. 56, 319–346.
554 Chen, S.H., Pollino, C.A., 2012. Good practice in Bayesian network modelling. Environ.
555 Model. Softw. 37, 134–145.
556 Clement, A., Miriam, S., Arzul, G., Guzman, L., Alarcon, C., 2001. Widespread outbreak of a
557 haemolytic, ichthyotoxic *Gymnodinium* sp. in southern Chile, in: Hallegraeff, G.M., Blackburn,
558 S.I., Bolch, C.J., Lewis, R.J. (Eds.), Harmful Algal Blooms 2000. IOC of UNESCO, Paris, pp.
559 66–69.
560 Cooper, G.F., Herskovits, E., 1992. A Bayesian method for the induction of probabilistic
561 networks from data. Mach. Learn. 9, 309–347.
562 Dammak-Zouari, H., Hamza, A., Bouain, A., 2009. Gymnodiniales in the Gulf of Gabès
563 (Tunisia). Cah. Biol. Mar. 50, 153–170.
564 DGPA., 2005–2009. Direction Générale de la pêche et de l'aquaculture. Ministère de
565 l'agriculture, Tunisie, annuaire statistique.
566 Domingos, P., Pazzani, M., 1997. On the optimality of the simple bayesian classifier under
567 zero-one loss. Mach. Learn. 29, 103–130.
568 Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, second ed. Wiley Interscience,
569 New York.

570 Feki-Sahnoun, W., Hamza, A., Njah, H., Barraj, N., Mahfoudi, M., Rebai, A., Bel Hassen, M.,
571 A., 2017. Bayesian network approach to determine environmental factors controlling *Karenia*
572 *selliformis* occurrences and blooms in the Gulf of Gabès, Tunisia. Harmful Algae. 63, 119–132.
573 Feki, W., Hamza, A., Bel Hassen, M., Rebai, A., 2008. Les efflorescences phytoplanctoniques
574 dans le golfe de Gabès (Tunisie) au cours de dix ans de surveillance (1995–2005). Bull. Inst.
575 Natl. Sci. Tech. Oceanogr. Peche Salammbo. 35, 105–116.
576 Feki, W., Hamza, A., Frossard, V., Abdennadher, M., Hannachi, I., Jacquot, M., Bel Hassen,
577 M., Aleya, L., 2013. What are the potential drivers of blooms of the toxic dinoflagellate *Karenia*
578 *selliformis*? A 10-year study in the Gulf of Gabès, Tunisia, southwestern Mediterranean Sea.
579 Harmful Algae. 23, 8–18.
580 Felipe, V.P., Okut, H., Gianola, D., Silva, M.A., Rosa, G.J., 2014. Effect of genotype
581 imputation on genome-enabled prediction of complex traits: an empirical study with mice data.
582 BMC Genet. 15, 149.
583 Fernandes, J.A., Irigoien, X., Goikoetxea, N., Lozano, J.A., Inza, I., Pérez, A., Bode, A., 2010.
584 Fish recruitment prediction, using robust supervised classification methods. Ecol. Model. 221,
585 338–352.

586

587 Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. Mach. Learn.
588 29, 131–163.
589 Friedman, N., Goldszmidt, M., 1996. Discretization of continuous attributes while learning
590 Bayesian networks, in: Saitta, L. (Ed.), Proceedings of the Thirteenth International Conference
591 on Machine Learning. CA: Morgan Kaufmann, San Francisco, pp. 157–165.
592 Fytilis, N., Rizzo, D.M., 2013. Coupling self-organizing maps with a Naïve Bayesian classifier:
593 stream classification studies using multiple assessment data. Water Ressour. Res. 49, 7747–
594 7762.

595 Glover, F., 1989. Tabu Search—Part I. ORSA J. Comput. 1, 190–206.
596 Gonzalez-Redin, J., Luque, S., Poggio, L., Smith, R., Gimona, A., 2016. Spatial Bayesian belief
597 networks as a planning decision tool for mapping ecosystem services trade-offs on forested
598 landscapes. Environ. Res. 144,15–26.
599 Haapasaari, P., Karjalainen, T.P., 2010. Formalizing expert knowledge to compare alternative
600 management plans: sociological perspective to the future management of Baltic salmon stocks.
601 Mar. Policy 34, 477–486.
602 Haines-Young, R., 2011. Exploring ecosystem service issues across diverse knowledge
603 domains using Bayesian Belief Networks. Prog. Phys. Geogr. 35, 681–699.
604 Hamilton, S.H., Pollino, C.A., Jakeman, A.J., 2015. Habitat suitability modelling of rare species
605 using Bayesian networks: model evaluation under limited data. Ecol. Model. 299, 64–78.
606 Hamza, A., El Abed, A., 1994. Les eaux colorées dans le golfe de Gabès: Bilan de six ans de
607 surveillance (1989–1994). Bull. Inst. Natl. Sci. Tech. Oceanogr. Peche Salammbo. 21, 66–72.
608 Hansen, G., Erard-Le Denn, E., Daugbjerg, N., Rodriguez, F., 2004. *Karenia selliformis*
609 responsible for the fish-kills in the gulf of Gabès, Tunisia 1994. Harmful algal Blooms. Program
610 and Abstracts of the 11th International Conference: Cape Town. Communication Ifremer 2004.
611 Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning, Springer,
612 New York.
613 Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall/CRC.
614 Jensen, F.V., 1996. An introduction to Bayesian Networks, Springer Verlag, New York.
615 Jensen, F.V., 2001. Bayesian Networks and Decision Graphs, Springer-Verlag, New York.
616 Jribi, I., Echwikhi, K., Bradai, M.N., Bouain, A., 2008. Incidental capture of sea turtles by
617 longlines in the Gulf of Gabès (South Tunisia): a comparative study between bottom and surface
618 longlines. Sci. Mar. 72, 337–342.

Kramer, S., Sorenson, H.W., 1988. Recursive Bayesian estimation using piece-wise constant approximations. Automatica. 24, 789–801.

Landuyt, D., Broekx, S., D'Hondt, R., Engelen, G., Aertsens, J., Goethals, P.L.M., 2013. A review of Bayesian belief networks in ecosystem service modelling. Environ. Model. Softw. 46, 1–11.

Lassus P., 1988. Plancton toxique et plancton d'eaux rouges sur les côtes européennes, Ed. IFREMER, Paris.

Lauritzen, S.L., Spiegelhalter, D.J., 1988. Local computations with probabilities on graphical structures and their applications to expert systems. J. R. Stat. Soc. 50, 157–224.

Lotze, H.K., Worm, B., 2009. Historical baselines for large marine animals. Trends Ecol. Evol. 24, 254–262.

Lucena-Moya, P., Brawata, R., Kath, J., Harrison, E., ElSawah, S., Dyer, F., 2015. Discretization of continuous predictor variables in Bayesian networks: an networks is NP-hard. Artif. Intell. 60, 141–153.

Madsen, H., Thyregod, P., 2011. Introduction to General and Generalized Linear Models. Chapman & Hall/CRC.

Maffucci, F., Kooistra, W.H.C.F., Bentivegna, F., 2006. Natal origin of loggerhead turtles, *Caretta caretta*, in the neritic habitat off the Italian coasts, Central Mediterranean. Biol. Conserv. 127, 183–189.

Markus, M., Hejazi, M.I., Bajcsy, P., Giustolisi, O., Savic, D.A., 2010. Prediction of weekly nitrate-N fluctuations in a small agricultural watershed in Illinois. J. Hydroinformatics. 12, 251–261.

Marrouchi, R., Benoit, E., Kharrat, R., Molgó, J., 2009. Gymnodimines: a family of phycotoxins contaminating shellfish, in: Benoit, E., Goudey-Perrière, F., Marchot, P., Servent,

643 D. (Eds.), Toxins and Signalling. Châtenay-Malabry, SFET Editions, Collection Rencontres en
644 Toxinologie, France, pp. 79–83.
645 McCulloch, C.E., Searle, S.R., Neuhaus, J.M., 2008. Generalized, Linear, and Mixed Models,
646 2nd edition, Hoboken, New Jersey: John Wiley & Sons.
647 McCullagh, P., Nelder, J., 1989. Generalized Linear Models, Second Edition, Boca Raton:
648 Chapman and Hall/CRC.
649 McVittie, A., Norton, L., Martin-Ortega, J., Siameti, I., Glenk, K., Aalders, I., 2015.
650 Operationalizing an ecosystem services-based approach using Bayesian Belief Networks: an
651 application to riparian buffer strips. Ecol. Econ. 110, 15–27.
652 Medhioub, A., Medhioub, W., Amzil, Z., Sibat, M., Bardouil, M., Ben Neila, I., Mezghani, S.,
653 Hamza, A., Lassus, P., 2009. Influence on environmental parameters on *Karenia selliformis*
654 toxin content in culture. Cah. Biol. Mar. 50, 333–342.
655 Medhioub, W., Guéguen, M., Lassus, P., Bardouil, M., Truquet, P., Sibat, M., Medhioub, N.,
656 Soudant, P., Kraiem, M., Amzil, Z., 2010. Detoxification enhancement in the gymnodimine
657 contaminated grooved carpet shell, *Ruditapes decussatus* (Linné). Harmful Algae. 9, 200–207.
658 Meineri, E., Dahlberg, C.J., Hylander, K., 2015. Using Gaussian Bayesian Networks to
659 disentangle direct and indirect associations between landscape physiography, environmental
660 variables and species distribution. Ecol. Model. 313, 127–136.
661 Meyer, S.R., Johnson, M.L., Lilieholm, R.J., Cronan, C.S., 2014. Development of a
662 stakeholder-driven spatial modeling framework for strategic landscape planning using Bayesian
663 networks across two urban–rural gradients in Maine, USA. Ecol. Model. 291, 42–57.
664 Neapolitan, R., 2003. Learning Bayesian networks, Prentice Hall, New York.
665 Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized Linear Models. J. R. Stat. Soc. A, 135,
666 370–384.

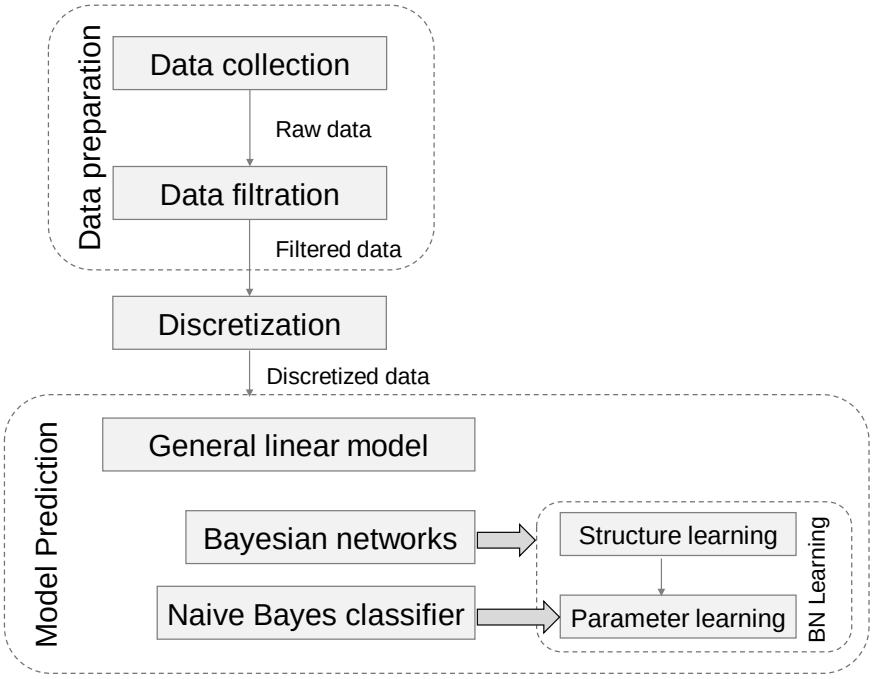
667 Nielsen, T.D., Jensen, F.V., 2007. Bayesian Networks and Decision Graphs, second ed.
668 Springer, New York.
669 Pearl, J., 1985. A model of self-activated memory for evidential reasoning In Proceedings of
670 the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA:
671 University of California, pp. 329–334.
672 Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference,
673 first ed. Morgan Kaufmann, San Mateo, CA.
674 Pittavino, M., Dreyfus, A., Heuer, C., Benschop, J., Wilson, P., Collins-Emerson, J., Torgerson,
675 P.R., Furrer, R., 2017. Comparison between Generalized Linear Modelling and Additive
676 Bayesian Network; Identification of Factors associated with the Incidence of Antibodies against
677 *Leptospira interrogans* sv *Pomona* in Meat Workers in New Zealand. Acta Trop. 173, 191–
678 199.
679 Pollino, C., White, A., Hart, B., 2007. Examination of conflicts and improved strategies for the
680 management of an endangered Eucalypt species using Bayesian networks. Ecol. Model. 201,
681 37–59.
682 R Development Core Team, 2017. R: A Language and Environment for Statistical Computing.
683 Vienna, Austria: the R Foundation for Statistical Computing. ISBN 3-900051-07-0.
684 http://www.R-project.org/(accessed 25.07.17).
685 Rekik, A., Drira, Z., Guermazi, W., Elloumi, J., Maalej, S., Aleya, L., Ayadi, H., 2012. Impacts
686 of an uncontrolled phosphogypsum dumpsite on summer distribution of phytoplankton,
687 copepods and ciliates in relation to abiotic variables along the near-shore of the southwestern
688 Mediterranean coast. Mar. Pollut. Bull. 64, 336–346.
689 Ropero, R.F., Aguilera, P.A., Fernandez, A., Rumí, R., 2014. Regression using hybrid Bayesian
690 networks: modelling landscape-socioeconomy relationships. Environ. Model. Softw. 54, 127–
691 137.

692 Ropero, R.F., Aguilera, P.A., Rumí, R., 2015. Analysis of the socioecological structure and
693 dynamics of the territory using a hybrid Bayesian network classifier. Ecol. Model. 311, 73–87.
694 Sammari, C., Koutitonsky, V.G., Moussa, M., 2006. Sea level variability and tidal resonance in
695 the Gulf of Gabès, Tunisia. Cont. Shelf. Res. 26, 338–350.
696 Schmitt, LHM., Brugere, C., 2013. Capturing Ecosystem Services, Stakeholders' Preferences
697 and Trade-Offs in Coastal Aquaculture Decisions: A Bayesian Belief Network Application.
698 PLoS ONE. 8, e75956.
699 Schwarz, G., 1978. Estimating the dimension of a model, Ann. Stat. 6, 461–464.
700 Scutari, M., 2010. Learning Bayesian Networks with the bnlearn R Package. J. Stat. Softw.
701 35,1–22.
702 Smith, C., Howes, A., Price, B., McAlpine, C., 2007. Using Bayesian belief network to predict
703 suitable habitat of an endangered mammal-the Julia Creek dunnart (*Sminthopsis douglasi*).
704 Biol. Conserv. 139, 333–347.
705 Uribe, J.C., Ruiz, M., 2001. *Gymnodinium* brown tide in the Magellanic fjords, southern Chile.
706 Rev. Biol. Mar. Oceanogr. 36, 155–164.
707 Utermöhl, H., 1958. Zur vervollkommnung der quantitativen phytoplankton-methodik. Mitt.
708 Int. Ver. Theor. Angew. Limnol. 9, 1–38.
709 Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental
710 modelling. Ecol. Model. 203, 312–318.
711 Venables, W.N., Ripley, B.D., 2002. Random and Mixed Effects, in: Modern Applied Statistics
712 with S. Statistics and Computing. Springer, New York, NY.
713 Wang, Q.J., Robertson, D.E., Haines, C.L., 2009. A Bayesian network approach to knowledge
714 integration and representation of farm irrigation: 1. Model development: knowledge integration
715 of farm irrigation, 1. Water Resour. Res. 45, W02409.

716 Zaffalon, M., 2005. Credible classification for environmental problems. Environ. Model.
717 Softw. 20, 1003–1012.
718 Zhang, H., 2004. The optimality of naive Bayes, in: Proceedings of the 17th International Florida
719 Artificial Intelligence Research Society Conference. AAAI Press, Florida, USA.

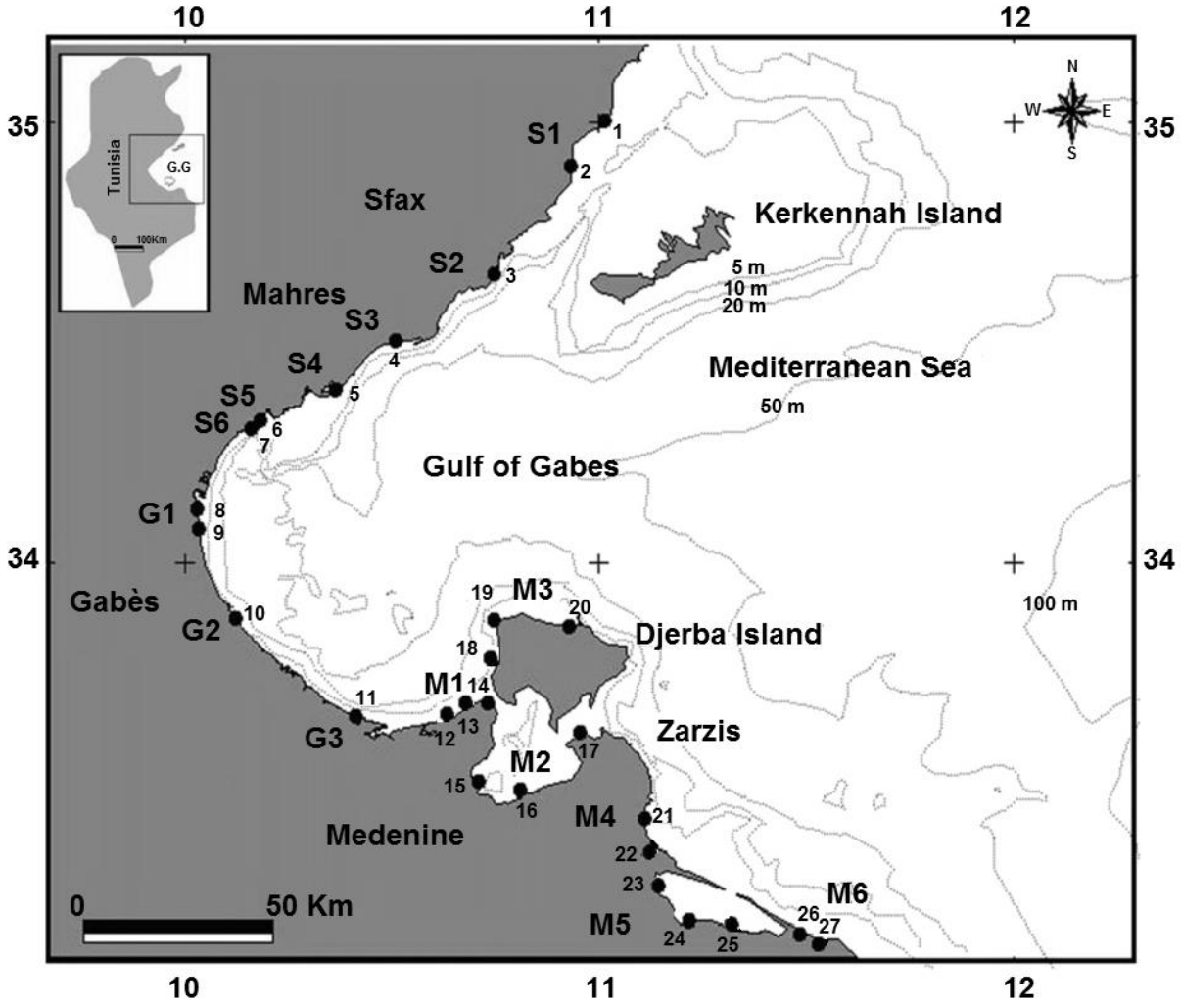
720 **Figures**

721 **Fig.1.** Framework of the proposed approach describing the three major steps: Data preparation,
722 Discretization and Model Prediction.




723

724 **Fig. 2.** Geographical map focusing on the monitoring network of phytoplanktonic sampling
725 stations in the Gulf of Gabès, Tunisia: 15 sampling sites and 28 sampling locations from 1997
726 to 2007. S1 (1-2), S2 (3), S3 (4), S4 (5), S5 (6), S6 (7), G1 (8-9), G2 (10), G3 (11), M1 (12-
727 13), M2 (14-15-16-17), M3 (18-19-20), M4 (21-22), M5 (23-24-25), M6 (26-27).




728

729

730

731

732

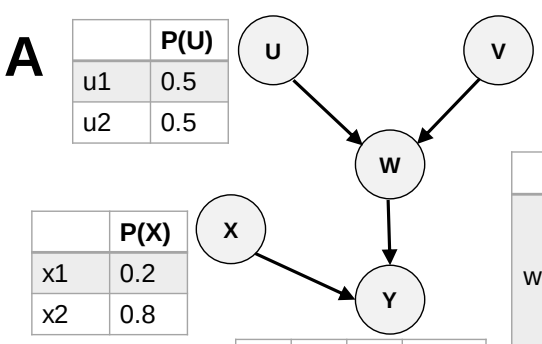
733

734

735

736

**Fig.3.** An example of A) a simple Bayesian Network and B) a Naïve Bayesian network




| | P(U) |
|---|---|
| u1 | 0.5 |
| u2 | 0.5 |

| | P(V) |
|---|---|
| v1 | 0.6 |
| v2 | 0.4 |

| | P(X) |
|---|---|
| x1 | 0.2 |
| x2 | 0.8 |

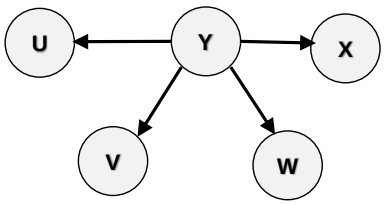
| | U | V | P(W) |
|---|---|---|---|
| w1 | u1 | v1 | 0.2 |
| | u1 | v2 | 0.1 |
| | u2 | v1 | 0.05 |
| | u2 | v2 | 0.3 |
| w2 | u1 | v1 | 0.06 |
| | u1 | v2 | 0. 2 |
| | u2 | v1 | 0.04 |
| | u2 | v2 | 0.05 |

| | X | W | P(Y) |
|---|---|---|---|
| y1 | x1 | w1 | 0.1 |
| | x1 | w2 | 0.2 |
| | x2 | w1 | 0.04 |
| | x2 | w2 | 0.2 |
| y2 | x1 | w1 | 0.06 |
| | x1 | w2 | 0.3 |
| | x2 | w1 | 0.05 |
| | x2 | w2 | 0.05 |

B

| | Y | P(U) |
|---|---|---|
| u1 | y1 | 0.25 |
| | y2 | 0.75 |
| u2 | y1 | 0.45 |
| | y2 | 0.55 |
| u3 | y1 | 0.3 |
| | y2 | 0.7 |

| | P(Y) |
|---|---|
| y1 | 0.5 |
| y2 | 0.5 |

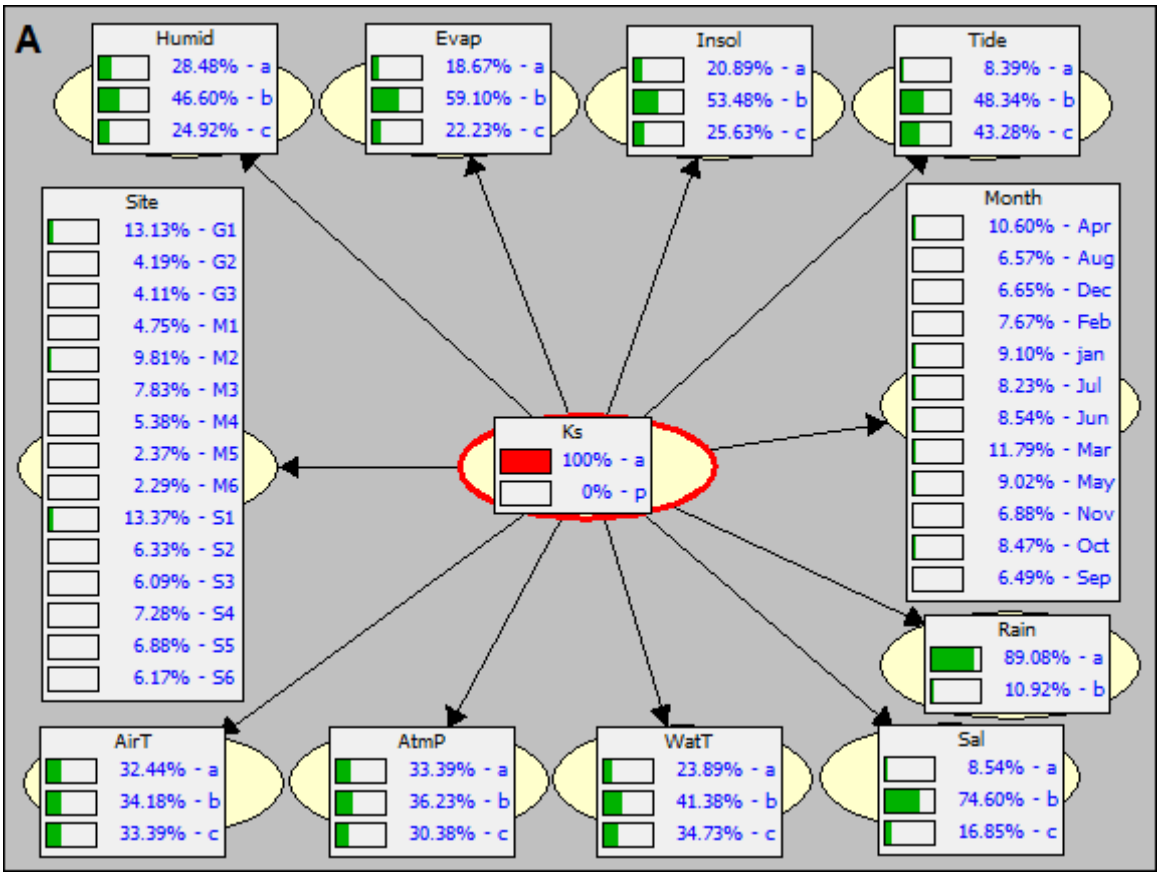



| | Y | P(x) |
|---|---|---|
| x1 | y1 | 0.25 |
| | y2 | 0.75 |
| x2 | y1 | 0.45 |
| | y2 | 0.55 |

| | Y | P(v) |
|---|---|---|
| v1 | y1 | 0.25 |
| | y2 | 0.75 |
| v2 | y1 | 0.45 |
| | y2 | 0.55 |

| | Y | P(w) |
|---|---|---|
| w1 | y1 | 0.25 |
| | y2 | 0.75 |
| w2 | y1 | 0.45 |
| | y2 | 0.55 |

741 **Fig.4.** Naïve Bayesian network plot depicting the relationship between the hydro-
742 meteorological parameters, spatio-temporal factors and the presence (A) and absence (B) of
743 *Karenia selliformis* in the Gulf of Gabès.



744

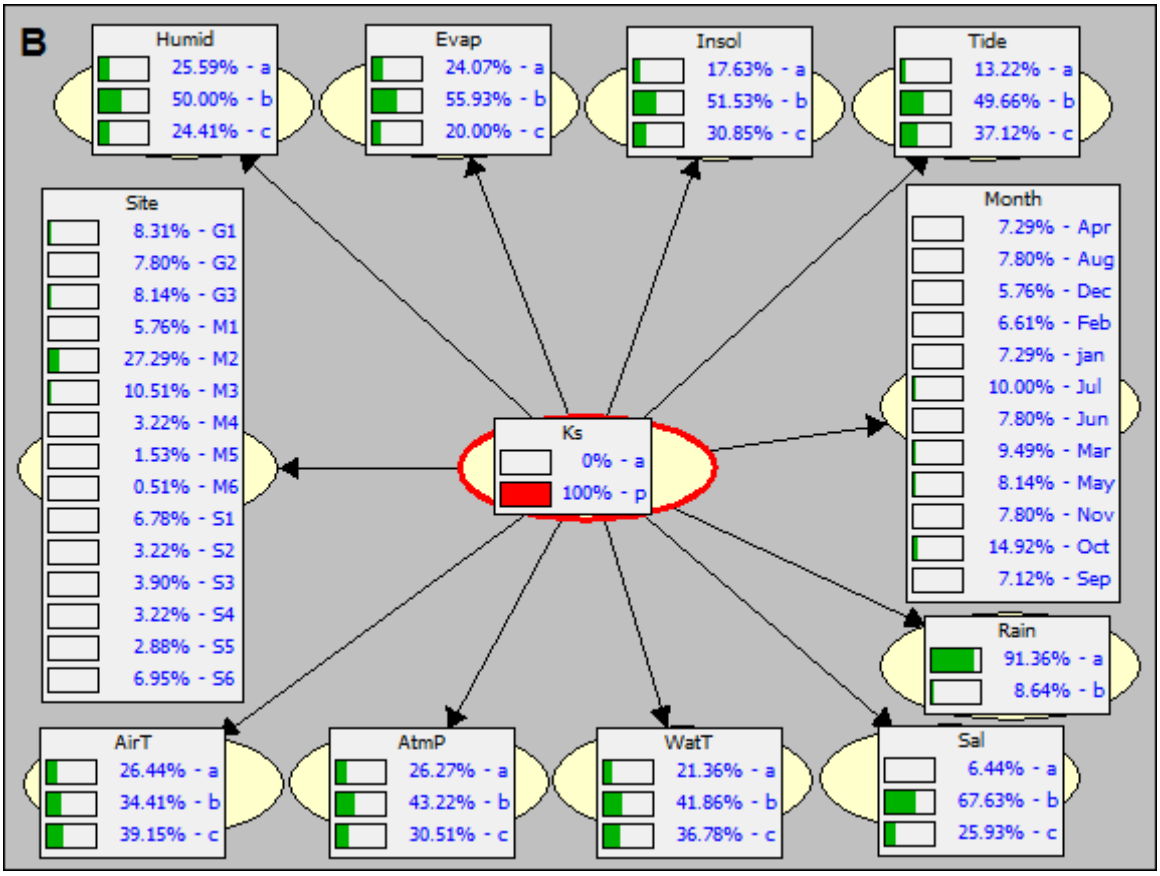
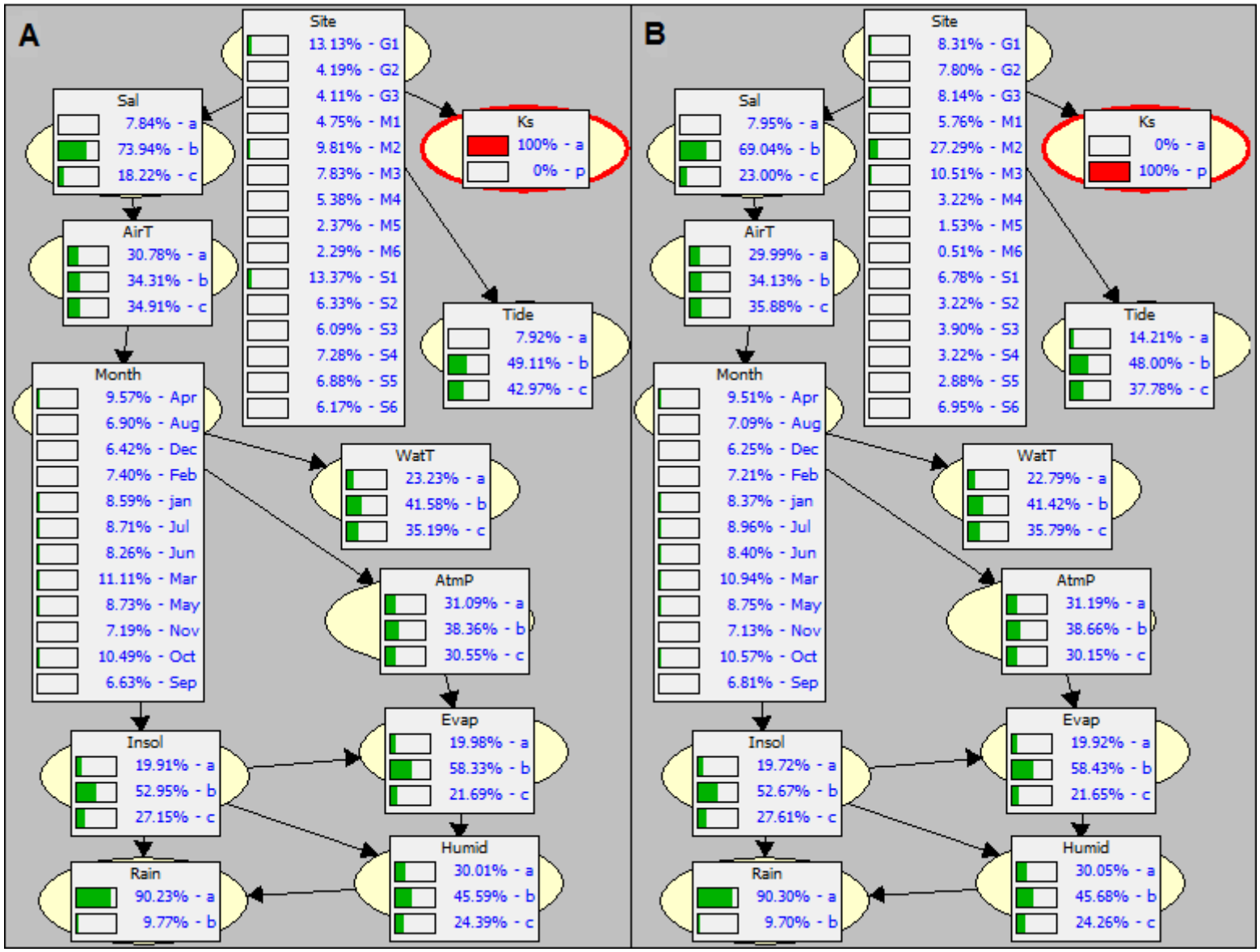
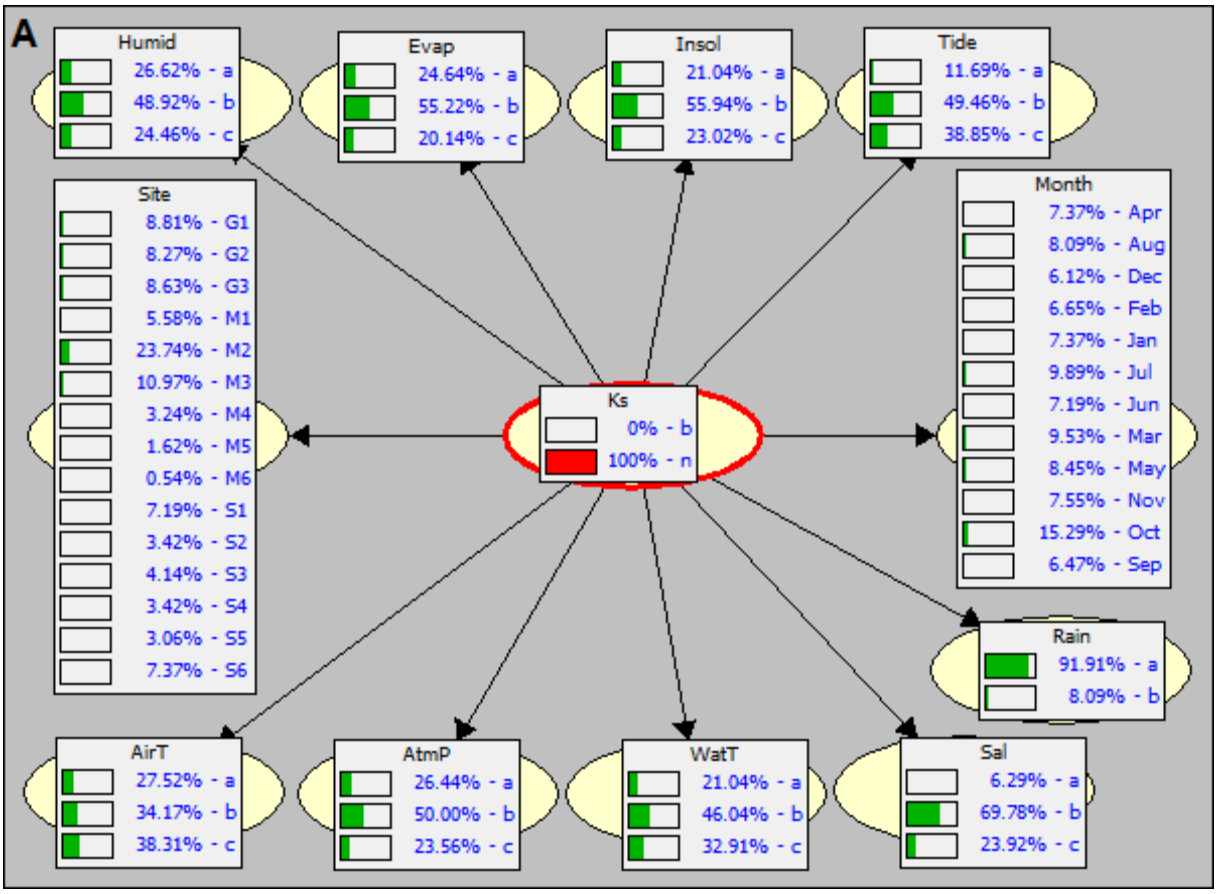



**Fig.5.** Bayesian network plot depicting the relationship between the hydro-meteorological parameters, spatio-temporal factors and the presence (A) and absence (B) of *Karenia selliformis* in the Gulf of Gabès.


A
Site
13.13% - G1
4.19% - G2
4.11% - G3
4.75% - M1
9.81% - M2
7.83% - M3
5.38% - M4
2.37% - M5
2.29% - M6
13.37% - S1
6.33% - S2
6.09% - S3
7.28% - S4
6.88% - S5
6.17% - S6
Sal
7.84% - a
73.94% - b
18.22% - c
Ks
100% - a
0% - p
AirT
30.78% - a
34.31% - b
34.91% - c
Tide
7.92% - a
49.11% - b
42.97% - c
Month
9.57% - Apr
6.90% - Aug
6.42% - Dec
7.40% - Feb
8.59% - jan
8.71% - Jul
8.26% - Jun
11.11% - Mar
8.73% - May
7.19% - Nov
10.49% - Oct
6.63% - Sep
WatT
23.23% - a
41.58% - b
35.19% - c
AtmP
31.09% - a
38.36% - b
30.55% - c
Evap
19.98% - a
58.33% - b
21.69% - c
Insol
19.91% - a
52.95% - b
27.15% - c
Humid
30.01% - a
45.59% - b
24.39% - c
Rain
90.23% - a
9.77% - b
B
Site
8.31% - G1
7.80% - G2
8.14% - G3
5.76% - M1
27.29% - M2
10.51% - M3
3.22% - M4
1.53% - M5
0.51% - M6
6.78% - S1
3.22% - S2
3.90% - S3
3.22% - S4
2.88% - S5
6.95% - S6
Sal
7.95% - a
69.04% - b
23.00% - c
Ks
0% - a
100% - p
AirT
29.99% - a
34.13% - b
35.88% - c
Tide
14.21% - a
48.00% - b
37.78% - c
Month
9.51% - Apr
7.09% - Aug
6.25% - Dec
7.21% - Feb
8.37% - jan
8.96% - Jul
8.40% - Jun
10.94% - Mar
8.75% - May
7.13% - Nov
10.57% - Oct
6.81% - Sep
WatT
22.79% - a
41.42% - b
35.79% - c
AtmP
31.19% - a
38.66% - b
30.15% - c
Evap
19.92% - a
58.43% - b
21.65% - c
Insol
19.72% - a
52.67% - b
27.61% - c
Humid
30.05% - a
45.68% - b
24.26% - c
Rain
90.30% - a
9.70% - b

757 **Fig.6.** Naïve Bayesian network plot depicting the relationship between the hydro-
758 meteorological parameters, spatio-temporal factors and the non-bloom (A) and bloom (B) of
759 *Karenia selliformis* in the Gulf of Gabès.



760

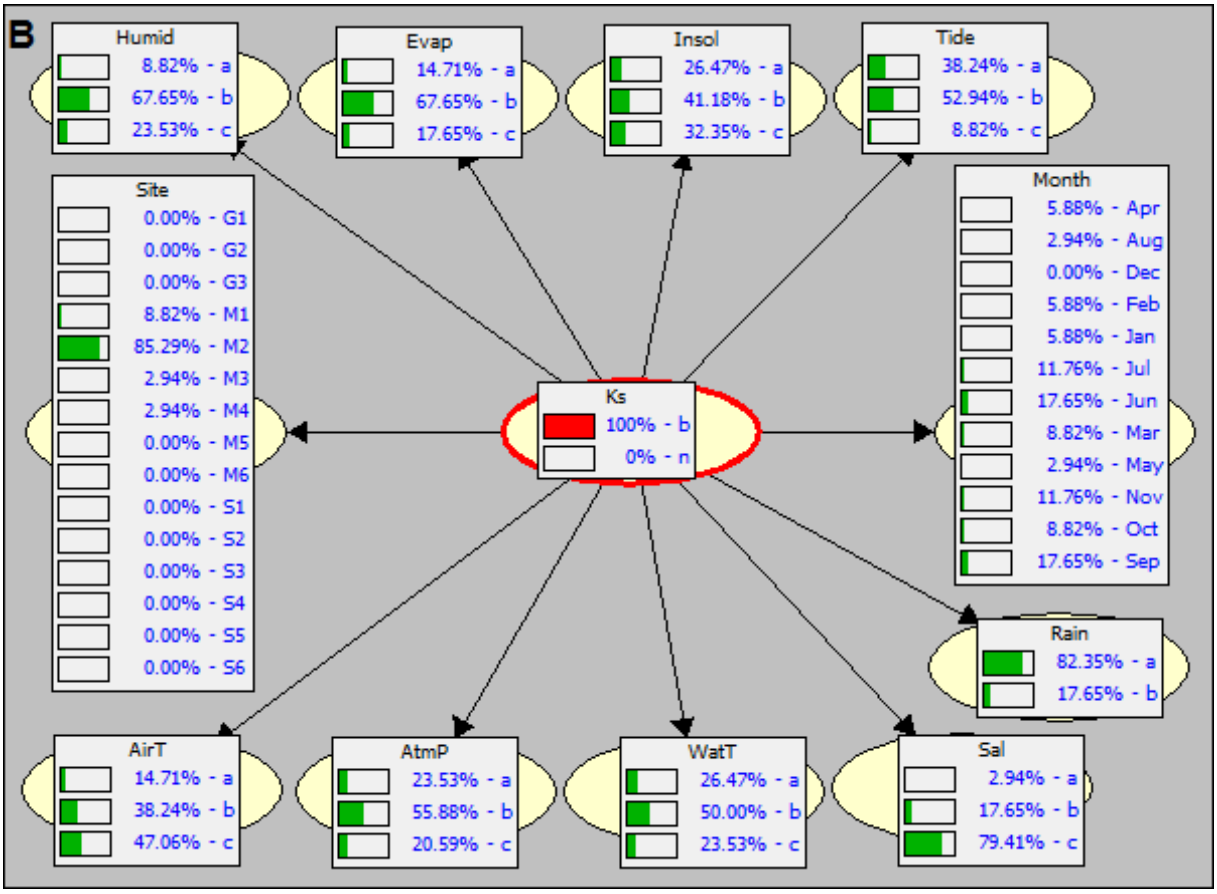
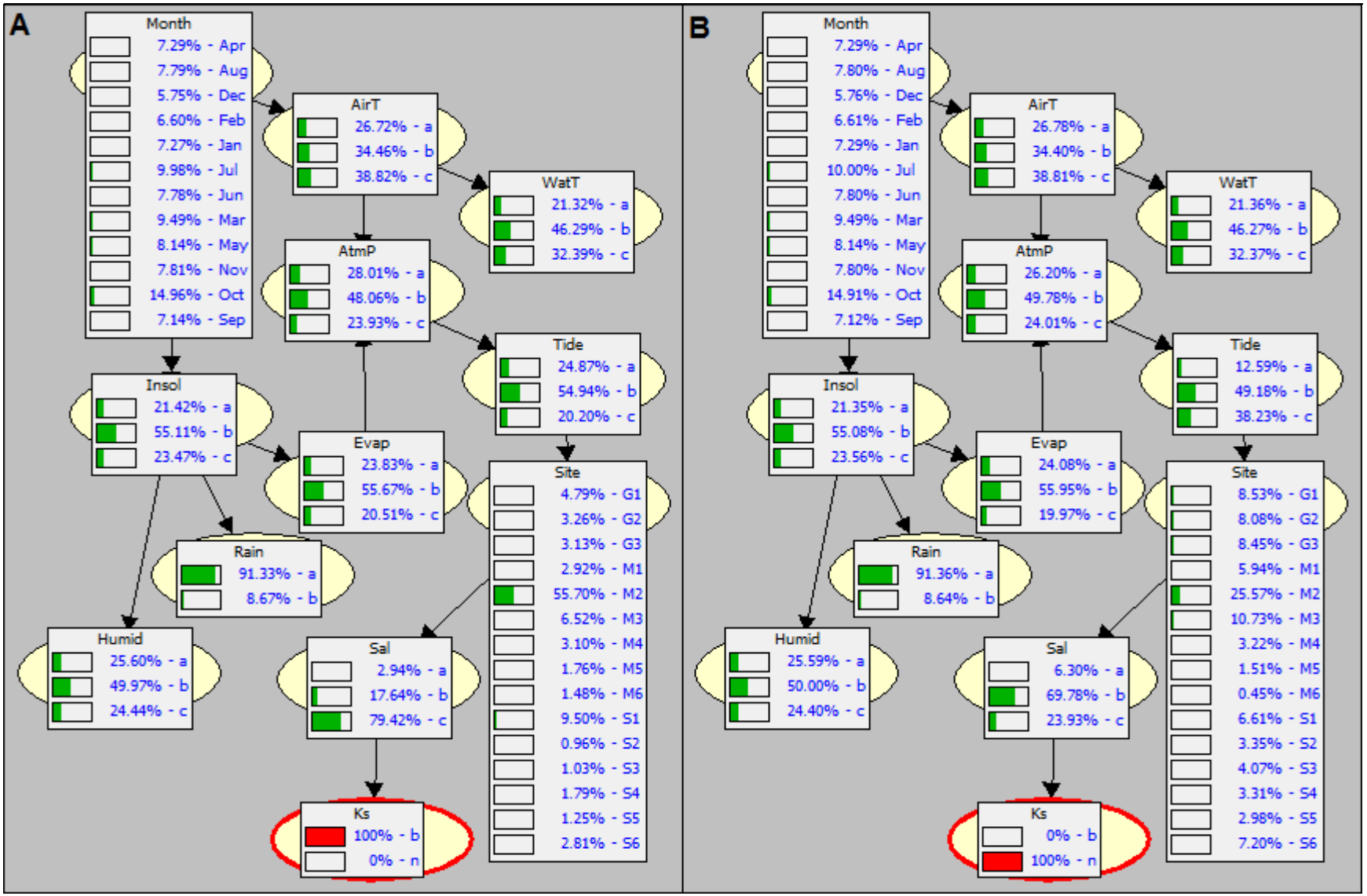



**Fig.7.** Bayesian network plot depicting the relationship between the hydro-meteorological parameters, spatio-temporal factors and the bloom (A) and non-bloom (B) of *Karenia selliformis* in the Gulf of Gabès.


A
Month
7.29% - Apr
7.79% - Aug
5.75% - Dec
6.60% - Feb
7.27% - Jan
9.98% - Jul
7.78% - Jun
9.49% - Mar
8.14% - May
7.81% - Nov
14.96% - Oct
7.14% - Sep
AirT
26.72% - a
34.46% - b
38.82% - c
WatT
21.32% - a
46.29% - b
32.39% - c
AtmP
28.01% - a
48.06% - b
23.93% - c
Tide
24.87% - a
54.94% - b
20.20% - c
Insol
21.42% - a
55.11% - b
23.47% - c
Evap
23.83% - a
55.67% - b
20.51% - c
Site
4.79% - G1
3.26% - G2
3.13% - G3
2.92% - M1
55.70% - M2
6.52% - M3
3.10% - M4
1.76% - M5
1.48% - M6
9.50% - S1
0.96% - S2
1.03% - S3
1.79% - S4
1.25% - S5
2.81% - S6
Rain
91.33% - a
8.67% - b
Humid
25.60% - a
49.97% - b
24.44% - c
Sal
2.94% - a
17.64% - b
79.42% - c
Ks
100% - b
0% - n
B
Month
7.29% - Apr
7.80% - Aug
5.76% - Dec
6.61% - Feb
7.29% - Jan
10.00% - Jul
7.80% - Jun
9.49% - Mar
8.14% - May
7.80% - Nov
14.91% - Oct
7.12% - Sep
AirT
26.78% - a
34.40% - b
38.81% - c
WatT
21.36% - a
46.27% - b
32.37% - c
AtmP
26.20% - a
49.78% - b
24.01% - c
Tide
12.59% - a
49.18% - b
38.23% - c
Insol
21.35% - a
55.08% - b
23.56% - c
Evap
24.08% - a
55.95% - b
19.97% - c
Site
8.53% - G1
8.08% - G2
8.45% - G3
5.94% - M1
25.57% - M2
10.73% - M3
3.22% - M4
1.51% - M5
0.45% - M6
6.61% - S1
3.35% - S2
4.07% - S3
3.31% - S4
2.98% - S5
7.20% - S6
Rain
91.36% - a
8.64% - b
Humid
25.59% - a
50.00% - b
24.40% - c
Sal
6.30% - a
69.78% - b
23.93% - c
Ks
0% - b
100% - n

771

772

773 **Table 1**

774 The biological and hydro-meteorological parameters discretized into intervals.

775 Rainfall (Rain), Evaporation (Evap), Air temperature (AirT), Insolation (Insol), Humidity (Humid), Atmospheric pressure (AtmP), Tide amplitude
776 (Tide), Water temperature (WatT), Salinity (Sal) and *Karenia selliformis* (Ks).

777

| **Rain (mm)** | **Evap (mm)** | **AirT (°C)** | **Insol (h)** | **Humid (%)** | **AtmP (Pa)** | **Tide (m)** | **WatT (°C)** | **Sal** | **Ks** | **Ks** |
|---|---|---|---|---|---|---|---|---|---|---|
| a=0 | a<20 | a<10 | a<8 | a<52 | a<1011 | a=0 | a<14 | a<37 | a=absent | b=bloom |
| b>0 | 20≤b≤50 | 10≤b≤30 | 8≤b≤10 | 52≤b≤80 | 1011≤b≤1019 | 0≤b≤1.5 | 14≤b<29 | 37≤b<42.5 | p=present | n=non-bloom |
| | c>50 | c>30 | c>10 | c>80 | c>1019 | c>1.5 | c≥29 | c≥42.5 | | |

778

**Table 2**

Model accounting for the observed variation of *Karenia selliformis* occurrences in the Gulf of Gabès according to the results of the general linear regression model analyses (GLM) with stepwise both selections of variables.

| | AIC |
|---|---|
| Ks ~ Site + Month + Rain + Evap + AirT + Insol + Humid + AtmP + Tide + WatT + Sal | 2158.81 |
| Ks ~ Site + Month + Rain + Evap + AirT + Insol + AtmP + Tide + WatT + Sal | 2154.98 |
| Ks ~ Site + Month + Rain + Evap + Insol + AtmP + Tide + WatT + Sal | 2151.16 |
| Ks ~ Site + Month + Rain + Evap + Insol + Tide + WatT + Sal | 2147.46 |
| Ks ~ Site + Month + Rain + Evap + Insol + WatT + Sal | 2144.66 |
| Ks ~ Site + Month + Rain + Evap + Insol + Sal | 2144.24 |

| | Variable | Estimate | Std. error | z value | P value |
|---|---|---|---|---|---|
| (Intercept) | -1.84 | 0.36 | -5.09 | 0 | *** |
| SiteG2 | 1.16 | 0.27 | 4.35 | 0 | *** |
| SiteG3 | 1.22 | 0.26 | 4.6 | 0 | *** |
| SiteM1 | 0.76 | 0.28 | 2.71 | 0.01 | ** |
| SiteM2 | 1.44 | 0.23 | 6.36 | 0 | *** |
| SiteM3 | 0.87 | 0.24 | 3.66 | 0 | *** |
| SiteM4 | -0.1 | 0.32 | -0.31 | 0.75 | |
| SiteM5 | -0.1 | 0.43 | -0.23 | 0.82 | |
| SiteM6 | -1.22 | 0.64 | -1.9 | 0.06 | . |
| SiteS1 | -0.32 | 0.25 | -1.26 | 0.21 | |
| SiteS2 | -0.11 | 0.31 | -0.36 | 0.72 | |
| SiteS3 | 0.16 | 0.3 | 0.53 | 0.6 | |

| | | | | | |
|---|---|---|---|---|---|
| SiteS4 | -0.29 | 0.31 | -0.95 | 0.34 | |
| SiteS5 | -0.56 | 0.32 | -1.76 | 0.08 | . |
| SiteS6 | 0.5 | 0.26 | 1.94 | 0.05 | . |
| MonthAug | 0.5 | 0.28 | 1.77 | 0.08 | . |
| MonthDec | 0.52 | 0.3 | 1.77 | 0.08 | . |
| MonthFeb | 0.45 | 0.28 | 1.61 | 0.11 | |
| Monthjan | 0.41 | 0.28 | 1.47 | 0.14 | |
| MonthJul | 0.23 | 0.28 | 0.84 | 0.4 | |
| MonthJun | 0 | 0.28 | 0.01 | 1 | |
| MonthMar | 0.3 | 0.26 | 1.18 | 0.24 | |
| MonthMay | 0.2 | 0.27 | 0.77 | 0.44 | |
| MonthNov | 0.94 | 0.28 | 3.37 | 0 | *** |
| MonthOct | 1.31 | 0.26 | 5.06 | 0 | *** |
| MonthSep | 0.54 | 0.28 | 1.92 | 0.05 | . |
| Rainb | -0.31 | 0.21 | -1.46 | 0.15 | |
| Evapb | -0.36 | 0.15 | -2.47 | 0.01 | * |
| Evapc | -0.28 | 0.19 | -1.48 | 0.14 | |
| Insolb | -0.08 | 0.16 | -0.5 | 0.62 | |
| Insolc | 0.37 | 0.23 | 1.63 | 0.1 | |
| Salb | 0.32 | 0.22 | 1.48 | 0.14 | |
| Salc | 0.74 | 0.26 | 2.85 | 0 | ** |

**Table 3**

Model accounting for the observed variation of *Karenia selliformis* blooms in the Gulf of Gabès according to the results of the general linear regression model analyses (GLM) with stepwise both selections of variables.

| | AIC |
|---|---|
| Ks ~ Site + Month + Rain + Evap + AirT + Insol + Humid + AtmP + Tide + WatT + Sal | 232.86 |
| Ks ~ Site + Month + Rain + Evap + AirT + Humid + AtmP + Tide + WatT + Sal | 229.04 |
| Ks ~ Site + Month + Rain + Evap + Humid + AtmP + Tide + WatT + Sal | 226.91 |
| Ks ~ Site + Month + Rain + Evap + Humid + Tide + WatT + Sal | 224.22 |
| Ks ~ Site + Month + Rain + Evap + Humid + WatT + Sal | 222.14 |
| Ks ~ Site + Month + Evap + Humid + WatT + Sal | 220.15 |

| | Variable | Estimate | Std. error | z value | P value |
|---|---|---|---|---|---|
| (Intercept) | 22.54 | 3495.00 | 0.01 | 0.99 | |
| SiteG2 | -0.66 | 5103.00 | 0.00 | 1.00 | |
| SiteG3 | -0.54 | 5085.00 | 0.00 | 1.00 | |
| SiteM1 | -19.80 | 3495.00 | -0.01 | 1.00 | |
| SiteM2 | -20.45 | 3495.00 | -0.01 | 1.00 | |
| SiteM3 | -18.18 | 3495.00 | -0.01 | 1.00 | |
| SiteM4 | -18.56 | 3495.00 | -0.01 | 1.00 | |
| SiteM5 | -0.14 | 10050.00 | 0.00 | 1.00 | |
| SiteM6 | -0.26 | 16060.00 | 0.00 | 1.00 | |
| SiteS1 | 0.12 | 5264.00 | 0.00 | 1.00 | |
| SiteS2 | -0.04 | 6711.00 | 0.00 | 1.00 | |
| SiteS3 | -1.74 | 6627.00 | 0.00 | 1.00 | |

| | | | | | |
|---|---|---|---|---|---|
| SiteS4 | -0.82 | 6673.00 | 0.00 | 1.00 | |
| SiteS5 | -0.28 | 7086.00 | 0.00 | 1.00 | |
| SiteS6 | -0.58 | 5319.00 | 0.00 | 1.00 | |
| MonthAug | 0.43 | 1.52 | 0.28 | 0.78 | |
| MonthDec | 19.68 | 3962.00 | 0.01 | 1.00 | |
| MonthFeb | 0.65 | 1.36 | 0.48 | 0.63 | |
| MonthJan | 0.98 | 1.26 | 0.78 | 0.44 | |
| MonthJul | 0.90 | 1.20 | 0.76 | 0.45 | |
| MonthJun | -0.37 | 1.08 | -0.34 | 0.73 | |
| MonthMar | 1.17 | 1.18 | 0.99 | 0.32 | |
| MonthMay | 1.88 | 1.40 | 1.34 | 0.18 | |
| MonthNov | -1.74 | 1.10 | -1.59 | 0.11 | |
| MonthOct | 1.03 | 1.23 | 0.84 | 0.40 | |
| MonthSep | -0.86 | 1.17 | -0.73 | 0.46 | |
| Evapb | 0.54 | 0.72 | 0.75 | 0.45 | |
| Evapc | 1.59 | 0.87 | 1.83 | 0.07 | . |
| Humidb | -2.12 | 0.89 | -2.40 | 0.02 | * |
| Humidc | -1.83 | 0.97 | -1.88 | 0.06 | . |
| WatTb | 0.40 | 0.74 | 0.54 | 0.59 | |
| WatTc | 1.57 | 0.90 | 1.75 | 0.08 | . |
| Salb | 0.41 | 1.38 | 0.30 | 0.77 | |
| Salc | -1.02 | 1.39 | -0.74 | 0.46 | |

**Table 4**

Adjusted $R^2$ values for the three models: General linear model (GLM). Bayesian networks (BN) and naive Bayes classifier (NB)

| | Adjusted $R^2$ | |
|---|---|---|
| | Occurrence | Bloom |
| BN | 0.70 | 0.58 |
| NB | 0.69 | 0.54 |
| GLM | 0.17 | 0.48 |

**Table 5**

Comparative analysis of the main fundamentals of the three predictive models: BN, NB and GLM. The lines separate models that do not have the same characteristics.

| GLM | NB | BN |
|---|---|---|
| Can handle continues or discrete data | Exclusive use of discrete variables which often results in a significant information loss | |
| Not informative about potential causal pathways between predictive variables | | More informative about potential causal pathways and able to easy establish direct associations between variables |
| The interdependencies between variables were not discovered | | The interdependencies between variables were revealed |
| The relationships between variables and outcome are linear | | Can handle nonlinear relationships between variables and outcome |
| | Less confidence on the predicted probabilities | More confidence on the predicted probabilities and may better reflect the joint probability distribution over the system's variables |
| | The number of arcs reduced the model performance | The number of arcs has no significant effect on model performance and the BN |

| | | |
|---|---|---|
| | | improves usually the accuracy, flexibility and expressivity |
| | The introduction of the evidence changes the probability distribution of several variables and amplify the weight of the evidence of each attribute on the class | The introduction of the evidence do not change the probabilities of variables except that for the parent of the class |