

# THINKING OUTSIDE THE BOX - PREDICTING BIOTIC INTERACTIONS IN DATA-POOR ENVIRONMENTS

*DAVID BEAUCHESNE<sup>1\*</sup>, PHILIPPE DESJARDINS-PROULX<sup>2</sup>,  
PHILIPPE ARCHAMBAULT<sup>3</sup>, and DOMINIQUE GRAVEL<sup>2</sup>*

*\* email: [david.beauchesne@uqar.ca](mailto:david.beauchesne@uqar.ca)*

*<sup>1</sup> Université du Québec à Rimouski*

*<sup>2</sup> Université de Sherbrooke*

*<sup>3</sup> Université Laval*

September 13, 2016

# 1 Abstract

Large networks of ecological interactions, such as food webs, are complex to characterize, be it empirically or theoretically. The former requires exhaustive observations, while the latter generally requires ample data to be validated. We therefore wondered whether readily available data, namely empirically described interactions in a variety of ecosystems and taxonomy, could be combined to predict species interactions in data deficient ecosystems. To test this, we built a biotic interaction catalogue from a collection of 94 empirical food webs, detailed predator-prey interaction databases and interactions from the Global Biotic Interaction (GloBI) database. We use an unsupervised machine learning method to predict interactions between any given set of taxa, given pairwise taxonomic proximity and known consumer and resource sets found in the interaction catalogue. Initial results suggest that pairwise interactions can be predicted with high accuracy. Although results are seemingly dependent on the comprehensiveness of the catalogue rather than the taxonomic similarity between taxa, taxonomy was found to complement well the catalogue and to allow the algorithm to perform well when the amount of information available through the catalogue decreased. Given its high accuracy, this methodology could democratize the use of food webs and network level descriptors in remote location where empirical data is hard to gather. Network characteristics could then be efficiently evaluated and correlated to levels of environmental stressors in order to improve vulnerability assessments of ecosystems to global changes, opening promising avenues for further research and for management initiatives.

# 2 Introduction

Large networks of ecological interactions, such as food webs, are complex to characterize, be it empirically or theoretically. The former requires exhaustive observations, while the latter generally requires ample data to be validated. For this reason, studies focusing on communities of interacting species remain understudied even though we acknowledge the importance of considering the reticulated nature of complex networks (ref). When time is of the essence, the long term studies required quickly become impractical and the use of network level approaches relegated to the sideline (ref).

Alternatively, a currently evolving approach is to predict interactions using proxies such as functional traits, phylogenies and spatial distributions to infer biotic interactions (e.g. Gravel et al., 2013; Morales-Castilla et al., 2015). For example, multiple traits can play a significant role in community dynamics and influence the presence and intensity of biotic interactions, like the influence of body size on predator-prey interactions, a literal take on *big fish eats small fish* (Cohen et al., 2003; Brose et al., 2006; Gravel et al. 2013). However, the time required to gather the necessary data to apply those methods may still be restrictive, or the data be unavailable altogether, so much so that other methods have been developed to fill the gaps in knowledge (e.g. Schrodte et al. 2016).

We therefore wondered whether more readily available data could be used to infer interactions in data deficient ecosystems. There is an increasing amount of data available describing worldwide species interactions, some freely available through the Global Biotic Interactions (GloBI) database (Poelen et al. 2014). Another readily available piece of information on species is their taxonomy, through initiatives like the World Register of Marine Species (WoRMS; Bailly et al. 2016). More than simple nomenclature, evolutionary processes are thought to influence consumer-resource relationships so that taxonomically related species would be more likely to share similar types of both consumers and resources (Eklof et al. 2012; Morales-Castilla et al. 2015). Based on that assumption, taxonomy might be useful in predicting interactions for species lacking detailed information, but which have a taxonomically related species for which such information is available. The objective of this work is thus to combine empirical biotic interactions originating from a variety of ecosystems with taxonomic relatedness to predict interactions in data deficient ecosystems.

### 3 Methods

The objective of the methodology presented is to predict the interactions between all taxa within an arbitrary set  $S1$  using a set of taxa  $S0$  with empirically described interactions from which we can extract sets of consumers and resources and their taxonomy. We couple the use of empirical data with an unsupervised machine learning method to achieve this.

#### 3.1 Biotic interaction catalogue

We built a biotic interaction catalogue to serve as a set of taxa  $S0$  with empirically described interactions. The empirical data used to construct the interaction catalogue was gathered in two successive steps. The first consisted of gathering data from a collection of 94 empirical food webs in marine and coastal ecosystems from which we extracted pairwise taxa interactions (Brose et al. 2005; Kortsch et al. 2015; GlobalWeb database). We also used a detailed predator-prey interaction database describing trophic relationships between  $XX$  predators and their prey (Barnes et al. 2008). From these datasets, only interactions between taxa at the taxonomic scale of the family or higher were selected for inclusion in the catalogue.

As empirical food webs are vastly dominated by non-interactions, these datasets yielded a highly skewed distribution of interactions vs non-interactions. To counterbalance this, the second step of data compilation consisted of extracting observed interactions from the Global Biotic Interaction (GloBI) database (ref), which describes binary interactions for a wide range of taxa worldwide. We extracted all interactions available on GloBI for species belonging to the families of taxa identified through step 1. Interactions were extracted using the rGloBI package in R (ref). As for step 1, only interactions between taxa at the taxonomic scale of the family or higher were retained

The nomenclature used between datasets and food webs varied substantially. Taxa names thus had to be verified, modified according to the scientific nomenclature and validated. This process was performed using the Taxize package in R (ref) and manually verified for errors. The same package was used to extract the taxonomy of all taxa for which interactions were obtained in previous steps. The complete R code and data used for the catalogue is available at [https://github.com/david-beauchesne/Interaction\\_catalog](https://github.com/david-beauchesne/Interaction_catalog).

### 3.2 Unsupervised machine learning

We use the  $K$ -nearest neighbor (KNN) algorithm (ref) to predict pairwise interactions for a set of taxa  $S$ . The KNN algorithm predicts missing entries or proposes additional entries by a majority vote based on the  $K$  nearest (*i.e.* most similar) entries. In this case, taxa are described by a set of resources when considered as a consumer, a set of consumers when considered as a resource and their taxonomy (*i.e.* kingdom, phylum, class, order, family, genus, species). Similarity between taxa was evaluated using the Tanimoto similarity measure (ref), which compares two vectors with  $i$  elements based on the number of elements they share and contain:

$$\text{tanimoto}(\mathbf{x}, \mathbf{y}) = \frac{\sum_i x_i \wedge y_i}{\sum_i x_i \vee y_i}, \quad (1)$$

where  $\wedge$  is bitwise *and*, while  $\vee$  is the bitwise *or* operators. Adding a weighing scheme, we can measure the similarity using two different sets of vectors with  $i$  and  $j$  elements, respectively.

$$\text{tanimoto}_t(\mathbf{x}, \mathbf{y}, w_t) = w_t \text{tanimoto}(\mathbf{x}_i, \mathbf{y}_i) + (1 - w_t) \text{tanimoto}(\mathbf{x}_j, \mathbf{y}_j), \quad (2)$$

where  $w_t$  is the weight given to vector  $i$ ,  $\mathbf{x}_i$ ,  $\mathbf{y}_i$  are the resource or consumer sets of the two taxa and  $\mathbf{x}_j$  and  $\mathbf{y}_j$  are the vectors for the taxonomy of two taxa. When  $w_t = 0$  only resource or consumer sets are used to compute similarity, while  $w_t = 1$  solely uses taxonomy.

### 3.3 Predicting interactions, Biotic predictor algorithm, Two-way Tanimoto algorithm, Feng shui name algorithm, Find a name for the algorithm

The XXX algorithm is built on a series of logical steps that ultimately predicts a candidate resources list  $C_R$  for each taxon in  $S1$  (Figure 2). For all consumer taxa  $T_C$  in  $S1$ , the algorithm first verify whether it has empirical resources  $T_R$  listed in the catalogue. When it does, if  $T_R$  are also in  $S1$ , they are added as predicted resources for  $T_C$ . This corresponds to what we refer to as the catalogue contribution to resource predictions. Two taxa in  $S1$  that are known to interact through the catalogue are automatically assumed to interact in  $S1$ .

Otherwise, the algorithm passes to what we refer to as the predictive contribution to resource predictions, with candidate resources for  $T_C$  identified with the KNN algorithm. If  $T_R$  are also in  $S1$ , K most similar resource  $T'_R$  are identified in  $S1$  to add to  $C_R$ . Then for all  $T_C$  in  $S1$ , the algorithm identifies K most similar consumer  $T'_C$  in  $S0$  and extracts their resource sets. As before, if those resources are found in  $S1$  they are added to  $C_R$ , otherwise K most similar resources  $T'_R$  are identified in  $S1$  to add to  $C_R$ . A simple working example is presented at Figure 1. Other parameters are used in the algorithm, but not presented here for the sake of message clarity. A more comprehensive mathematical description of the algorithm and the parameters used is however available through Figure 2 and the complete R code and data used for the algorithm is available at [https://github.com/david-beauchesne/Predict\\_interactions](https://github.com/david-beauchesne/Predict_interactions).

<i>S0</i> taxa ID	taxonomy	resource	consumer	<i>S1</i> taxa ID
$T_1$	$\{a, b, c\}$	$\{T_2, T_3, T_{12}\}$	$\{T_4\}$	$T_1$
$T_2$	$\{e, f, g\}$		$\{T_1, T_5\}$	$T_9$
$T_3$	$\{i, j, k\}$		$\{T_5\}$	$T_{10}$
$T_4$	$\{m, n, o\}$	$\{T_1, T_5\}$		$T_{11}$
$T_5$	$\{a, b, d\}$	$\{T_8, T_9\}$	$\{T_4\}$	$T_{12}$
$T_6$	$\{i, q, r\}$	$\{T_2, T_8\}$	$\{T_4\}$	
$T_7$	$\{e, f, h\}$		$\{T_1, T_6\}$	
$T_8$	$\{s, t, u\}$		$\{T_5, T_6\}$	
$T_9$	$\{s, t, v\}$		$\{T_5\}$	
$T_{10}$	$\{i, j, l\}$			
$T_{11}$	$\{m, n, p\}$			
$T_{12}$	$\{q, r, s\}$		$\{T_1\}$	

Similarity measured with set of resources/consumers

	$T_1$	$T_2$	$T_3$	$T_4$	$T_5$	$T_6$	$T_7$	$T_8$	$T_9$	$T_{10}$	$T_{11}$	$T_{12}$
$T_1$	-	0	0	0	0, 1	0.3, 1	0	0	0	0	0	0
$T_2$	0	-	0, 0.5	0	0	0	0, 0.3	0, 0.3	0, 0.5	0	0	0, 0.5
$T_3$	0	0	-	0	0	0	0	0, 0.5	0, 1	0	0	0
$T_4$	0	0	0	-	0	0	0	0	0	0	0	0
$T_5$	0.5	0	0	0	-	0.3, 1	0	0	0	0	0	0
$T_6$	0	0	0.2	0	0	-	0	0	0	0	0	0
$T_7$	0	0.5	0	0	0	0	-	0, 0.3	0	0	0	0, 0.5
$T_8$	0	0	0	0	0	0	0	-	0	0	0	0
$T_9$	0	0	0	0	0	0	0	0.5	-	0	0	0
$T_{10}$	0	0	0.5	0	0	0.2	0	0	0	-	0	0
$T_{11}$	0	0	0	0.5	0	0	0	0	0	0	-	0
$T_{12}$	0	0	0	0	0	0.5	0	0.2	0.2	0	0	-

Similarity measured with taxonomy

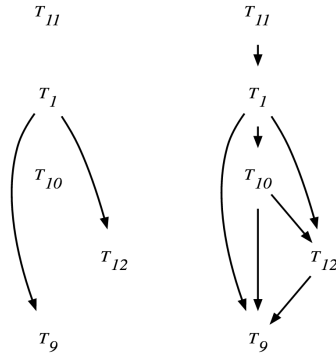


Figure 1: Simple example of logical steps through the algorithm

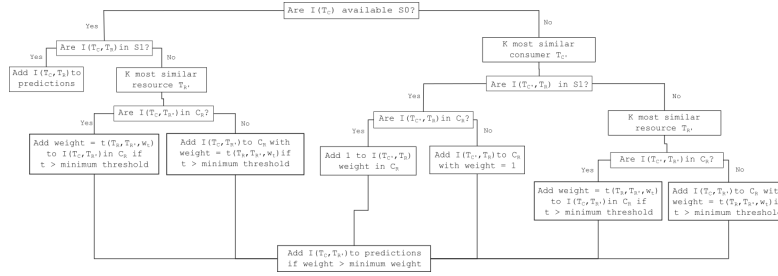


Figure 2: Description of the logical steps used by the algorithm to suggest a list of candidate resources ( $C_R$ ) for each taxa ( $T_C$ ) in  $S1$

### 3.4 Algorithm prediction accuracy

We used the most extensive and taxonomically detailed datasets included in the catalogue (**ref**) to assess the prediction accuracy of the algorithm. Testing accuracy of a particular dataset was done by first removing from the catalogue all pairwise interacting originating from that dataset. Accuracy was evaluated using three different statistics:

1.  $Score_y$  is the fraction of interactions correctly predicted:

$$Score_y = \frac{a}{a + c} \quad (3)$$

2.  $Score_{-y}$  is the fraction of non-interactions correctly predicted:

$$Score_{-y} = \frac{d}{b + d} \quad (4)$$

3. TSS, The True Skilled Statistics (TSS) evaluated prediction success by considering both true and false predictions, returning a value ranging from 1 (prefect predictions) to -1 (inverted predictions; **ref**):

$$TSS = \frac{(ad - bc)}{(a + c)(b + d)} \quad (5)$$

where a is the number of links predicted (1) and observed (1), b is the number predicted (1) but not observed (0), c is the number predicted absent (0) but observed (1) and d is the number of predicted absent (0) and observed absent (0). These three statistics give a different perspective on prediction accuracy, focusing in turn on true interactions and non-interactions, and on both true and false predictions.

We evaluated the three statistics for the complete algorithm and for the catalogue and the predictions individually to evaluate their respective contribution to the algorithm predictive accuracy. Multiple  $w_t$  values were also tested to

evaluate whether taxa similarity measured as a function of resource/consumer sets or taxonomy contributed more significantly towards increased predictive accuracy. The same was done with multiple  $K$  values.

Finally, we evaluated the influence of the comprehensiveness of the catalogue on prediction accuracy. We selected the arctic food web from Kortsch et al. (2015). This food web was selected as it is highly detailed taxonomically and because empirical data remains available for most of its taxa after its exclusion from the catalogue. We iteratively and randomly ( $n = 50$ ) removed a percentage of empirical data describing the food web taxa from the catalogue before generating new predictions with the algorithm. We also tested  $w_t$  values of 0.5 and 1 to evaluate whether taxonomic similarity could support predictive accuracy in cases when empirical data for species in  $S1$  in the catalogue is unavailable.

## 4 Results

### 4.1 Biotic interaction catalogue

The data compilation process allowed us to build an interaction catalogue composed of 276708 pairwise interactions (interactions = 72110; non-interactions = 204598). A total of 9712 taxa (Superfamily = 15; Family = 591; Subfamily = 29; Tribe = 8; Genus = 1972; Species = 7097) are included in the catalogue, 4159 of which have data as consumers and 4375 as resources.

### 4.2 Algorithm predictive accuracy

The overall predictive accuracy of the algorithm ranges between 80% to almost 100% in certain cases (Figure 3), except for one food web (*i.e.* **ref**). Both interactions and non-interactions are well predicted by the algorithm. TSS scores are lower than  $Score_y$  and  $Score_{-y}$  due to misclassified interactions and non-interactions. This can also be observed through the effect of varying  $K$  values, which increases the number of potential candidate resources for each taxa in the predictive portion of the algorithm. Prediction accuracy increases for interactions, while it decreases for non-interactions, as  $K$  values increase. This means that non-interactions are misclassified as interactions in the process of increasing  $Score_y$ , consequently decreasing  $Score_{-y}$ .

Similarity being predominantly measured with resource/consumer sets ( $w_t$  closer to 0) yielded better predictions than when measured with taxonomy ( $w_t$  closer to 1; Figure 3). Resource/consumer sets therefore appears to serve as a better predictor of similarity between taxa for interactions predictions. It is nonetheless interesting to note that although the predictive contribution of the algorithm decreases as  $w_t$  increases, an increased mean and decreased variability values for the TSS and  $Score_y$  statistics is also observed (Figure 3)). This suggests that while using taxonomy for similarity measurements yields lower predictive accuracy, it may also complement the catalogue contribution by predicting interactions not captured through empirical data, effectively increasing



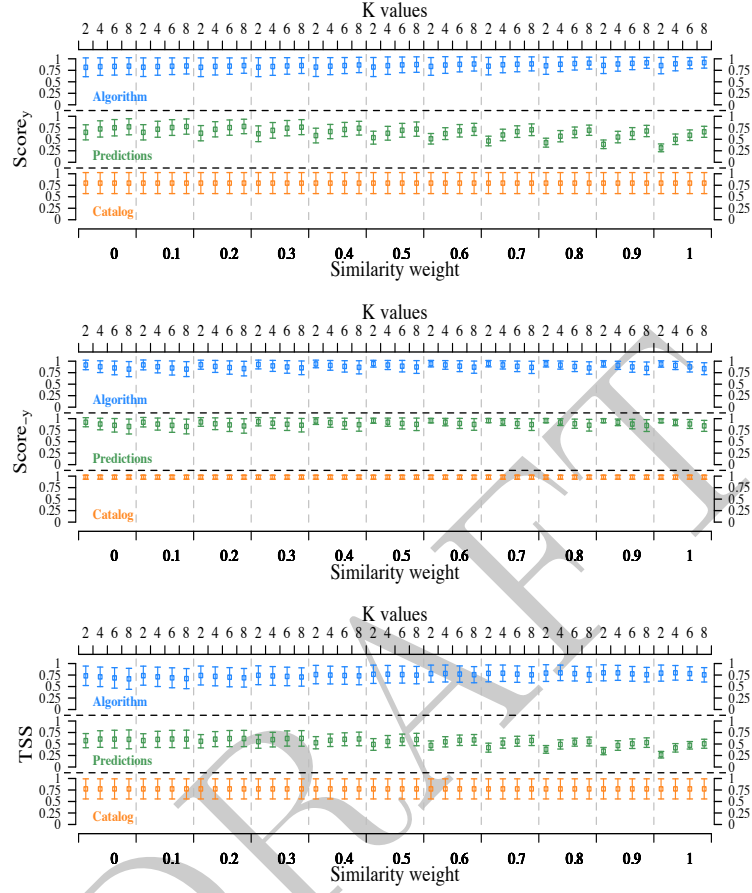


Figure 3: The graph presents the three statistics as a function of trait weight, which varies between 0 and 1. A weight of 0 means that similarity is measured only using set of resources for each taxa, while a weight equal to 1 means that similarity is based solely on taxonomy. We present 6 food webs with over 50 taxa each and the Barnes et al. (2008) dataset.

the predictive accuracy of the complete algorithm.

The partitioning of the catalogue and predictive portions of the algorithm shows that it is dependent on the comprehensiveness of the catalogue for high prediction accuracy (Figures 3, 4). As the amount of empirical data available in the catalogue decreases so does the overall accuracy of the algorithm (Figures 4). The predictive contribution of the algorithm however slows down the decrease in the prediction efficiency of the algorithm. Prediction accuracy still remains around 75% with only 40% of  $S1$  taxa found in the catalogue (Figures 4). Furthermore, the use of taxonomy for similarity measurements is more efficient

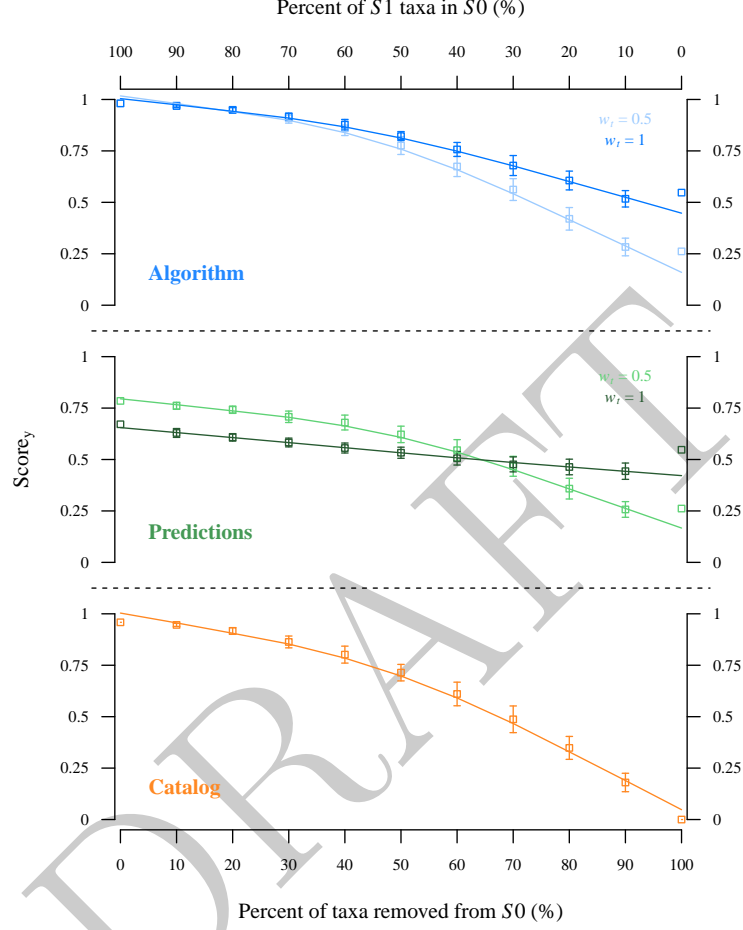


Figure 4: Graph presenting predictive accuracy as a function of the amount of information available in the catalogue. The arctic food web from Kortsch et al. (2015) was used for this, as it is highly detailed and because almost all taxa found in it had information in the catalogue even when not included in the catalogue. A random percentage of taxa in the web was iteratively removed from the catalogue ( $n = 50$ ) before predicting interactions with the XXX algorithm.

as empirical data becomes scarcer and no different than resource/consumer sets for the complete algorithm when ample data is available (Figures 4).

## 5 Discussion

### 5.1 Algorithm accuracy

#### 5.1.1 Misclassified non-interactions

Interactions vs non-interactions, misclassified non-interactions: real? Could be that non-interactions are not actually observed. Can we really trust  $Score_{\neg y}$  then?

The algorithm could simply perform poorly in predicting non-interactions, classifying them instead as interactions. The original empirical food webs could also be the ones doing poorly at observing non-interactions. Indeed, we assumed that non-interactions in empirical food webs meant that there were no interactions between taxa. However, most of the empirical webs had a strong focus attributed to higher order consumer species and very little attention given to other taxa (*e.g.* benthic invertebrates). The catalogue of interactions, on the other hand, has a much broader focus than the original empirical webs. It therefore very well may be that wrongly classified non-interactions could indeed be real but unobserved interactions in natural systems.

Furthermore, the algorithm is currently coded to ignore non-interactions in the similarity measurements. Maybe it would make sense to do the same in the interpretation of the results itself. Or a third vector to the similarity measurements could be added that considers non-interactions as a set of resources not consumed by consumers (which we have in the data). While this could make sense, more thought needs to be given to non-interactions stemming from empirical food webs and their actual value as observed data. If we think them valuable, then extra logical steps could be added to the algorithm to decrease probabilities of taxa being proposed as candidate resources by the KNN algorithm when they were observed to not interact in food webs.

#### 5.1.2 Usefulness of taxonomy in topology prediction

Interestingly, as the proportion of *S1* taxa in the catalogue decreases, the similarity measured with taxonomy only performs better, with predictions performed with  $w_t = 1$  surpassing those made with  $w_t = 0.5$  as percent of taxa removed from catalogue increases. Taxonomy could thus be highly useful in data poor environments that also lack data in the catalogue.

While taxonomy plays a significant role in influencing food web structure (Eklof et al. 2015), it does not account for certain traits such as body size. Even though we advocate for as simple an approach as possible, we believe that in cases where the comprehensiveness of the catalogue does not encompass a significant portion of *S1*, traits could be added as a third set in the similarity measurement.

## **5.2 Perspectives**

### **5.2.1 Democratizing the use of network level predictors**

### **5.2.2 Management**

### **5.2.3 Applied research in theoretical ecology**

### **5.2.4 Adding interaction strength to predictions**

## **6 Acknowledgements**

We thank the Fond de Recherche Québécois Nature et Technologie (FRQNT) and the Natural Science and Engineering Council of Canada (CRSNG) for financial support. This project is also supported by Québec Océan, the Quebec Centre for Biodiversity Science (QCBS), and the Notre Golfe and CHONeII networks. We also wish to thank K. Cazelles for constructive comments and suggestions.