Determining threat status for data-limited fisheries based on catch-only stock assessment models

by

Lauren Weir

B.Sc. (Natural Resources Conservation), University of British Columbia, 2012

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Resource Management

Report No.: 660

in the
School of Resource and Environmental Management
Faculty of Environment

SIMON FRASER UNIVERSITY
Spring 2017

Approval

Name: **Lauren Weir**

Degree: **Master of Resource Management**

Report No.: **660**

Title: **Determining threat status for data-limited fisheries based on catch-only stock assessment models**

Examining Committee: **Chair:** **Gabrielle Pang**
Master of Resource Management Candidate

Andrew Cooper
Senior Supervisor
Adjunct Professor
Data Scientist
Seattle Children's Hospital

Brendan Connors
Supervisor
Adjunct Professor
Senior Systems Ecologist
ESSA Technologies

Date Defended/Approved: February 09, 2017

Abstract

Catch-only stock assessment methods have been developed to manage data-limited fisheries where only catch data is available. This research evaluated the ability of four catch-only stock assessment methods to correctly classify a stock of concern based on population trends. To accomplish this, true trends from simulated stocks and the trends produced by the models were used to classify stocks into threat categories based on percent change. ROC curves and PR curves were then used to test the effectiveness of the four models as classifiers. ROC curves indicated that the models performed well under most scenarios. However, the confusion matrices and PR curves revealed low precision values for all models. The high number of stocks falsely classified as threatened were masked in the ROC analysis by the imbalance of few threatened stocks compared to numerous non-threatened stocks. This is an important caveat, as it could lead to inappropriate threshold selection.

Keywords: ROC Curves; Precision Recall Curves; Class imbalance; Catch-only stock assessment; Data-limited fisheries; Population trends;

Table of Contents

List of Tables

List of Figures

List of Acronyms

AUC	Area Under the Curve
FAO	Fisheries and Aquaculture Organisation
IUCN	International Union for the Conservation of Nature
MAPE	Mean Absolute Potential Error
MPE	Mean Potential Error
ROC Curves	Receiver Operating Characteristic Curves
PR Curves	Precision Recall Curves

Chapter 1.

Introduction

Marine fish stocks are an important global resource. Global marine fisheries harvest has risen to an average of 80 million tonnes a year (FAO Fisheries and Aquaculture Organization 2016), and are relied on as a safety net to provide income and nutrition to many developing countries (Teh & Sumaila 2013). As of 2013, 144 nations have marine fisheries and it is estimated that approximately 260 million people are involved in global marine fisheries in both full and part time jobs, not including aquaculture or recreation (Teh & Sumaila 2013). Almost three billion people receive nearly 20% of their protein from marine fish (FAO Fisheries and Aquaculture Organisation 2012). These peoples reliance on marine fishes can be threatened by fishery collapses which can have immediate economic impacts and large ecological consequences and so it is important to constantly work towards sustainable fisheries management (Frank et al. 2005; Schrank 2005; Myers et al. 2007).

Effective fisheries management benefits significantly from formal stock assessment so as to be able to manage for a targeted biomass, yet many harvested fish stocks are not formally assessed (Kleisner et al. 2013). Fisheries stock assessments are important for evaluating different management actions designed to sustainably harvest or rebuild a stock (Punt & Hilborn 1997; Butterworth et al. 2010). Stock assessments require models that use ecological knowledge and mathematical equations to make predictions about the response of a fish population to different management actions (Punt & Hilborn 1997; Cooper 2006). The results from these assessments are then used to set harvest limits and inform policy action (Mace 2001; Worm et al. 2009). However, in 2012, over 80% of the global catch came from fish populations that are not assessed (Costello et al. 2012). In addition, of the 16% of harvested fisheries that had formal stock assessments, 58% were considered to be below the level of biomass capable of producing maximum sustainable yield (Ricard et al. 2012). More recently, Costello *et al* (2016) found that of the 4,713 stocks representing 78% of global catch, 91.6 % were unassessed(Costello et al. 2016). This is a concern, because these few assessed stocks may not suitably represent the fisheries that are not assessed. Most assessed fisheries represent fish species that are robust to harvesting over a long period of time, and may thus represent a biased subset of fished stocks (Froese et al. 2012).

Stock assessments require detailed information about fish populations, which is not always available and can be difficult to collect. Formal data-intensive stock assessment models include information such as catch data, mortality, age structure, stock recruitment relationships, catch per unit effort and other life history characteristics (Cooper 2006). It is the high operational costs associated with the collection of fisheries data that has been attributed to the low percentage of formal stock assessments (Costello et al. 2012; Ricard et al. 2012). Not only the costs, but also the length of time required to collect data and the quality of the data, as stock assessments depend on the quality of the data (Agnew et al. 2013). Unfortunately, it is not feasible to collect all the various data required for formal stocks assessments for all fisheries. Data-limited assessment methods were developed to deal with this issue, to manage fisheries sustainably and to better inform harvest policy.

For many exploited species, catch data is the only information available; as a result, assessment models have been developed that require varying degrees of information (Branch et al. 2011; Froese et al. 2012). However, even data poor models such as Depletion-Based Stock Reduction Analysis (DB-SRA) and Depletion-Corrected Average Catch (DCAC) still require information about natural mortality, fishing mortality at maximum sustainable yield (MSY), an estimate of depletion, and a long series of catch data (MacCall 2009; Dick & MacCall 2011). For many harvested fish stocks the only information available is catch data, which is information about the number or weight of fish caught annually(Branch et al. 2011; Rosenberg et al. 2014), hence several catch-only models have been developed that estimate biomass based only on a time series of catch data. The push for more data-limited models was driven in part by the United States of America Congress, which required all managed American fisheries to have accountability measures and requirements for setting annual catch limits by 2011, including the data-limited fisheries (Berkson et al. 2011). Interest in data-limited assessment models has continued since with publications of new models. Martell and Froese (2013) developed a catch only stock assessment method that uses only fish removal estimates and estimates for two parameters based on prior knowledge of fish stocks. Thorsen *et al.* (2013) created a stock assessment model that estimates fish biomass from catch data based on effort dynamics. As catch only models do not require the collection of additional data for fish stocks, they can provide a more cost-effective means to assess the health of fish populations. If effective, the catch-only models have the potential to be a powerful tool for fisheries management, particularly in developing countries where resources are more limited (Eggert & Greaker 2009; Le Manach et al. 2012).

Despite the widespread potential for using catch-based assessment methods, these data-limited methods have sparked significant debate in the scientific community over the accuracy of their assessments (Branch 2008; Branch et al. 2011; Daan et al. 2011; Froese et al. 2012; Agnew et al. 2013). For example, some researchers have found that catch trends considerably overestimated the number of stocks classified as overexploited and collapsed compared to trends in biomass from more data-intensive assessments, and thus failed to represent the true conditions of global fisheries (Branch et al. 2011). While using catch data to classify stocks into different statuses has been argued to be scientifically and theoretically unsound (Daan et al. 2011), others argue strongly for using the FAO catch data to classify stocks into various categories of exploitation (Froese et al. 2012). To quell the controversy, there is a clear need to fully evaluate the ability of catch-only models to quantify population abundance.

Many catch-only models have been evaluated based on their ability to estimate stock status and harvest rates, but not on their ability to estimate trends over time. Stock status is often described in terms of the ratio of the current population abundance to that which would produce maximum-sustainable-yield (B/Bmsy) (Worm et al. 2009; Branch et al. 2011; Rosenberg et al. 2014). This ratio is then used to determine whether a population is over, under, or sustainably exploited and used to set harvest rates (Mace 2001; Worm et al. 2009). The ability of data-limited stock assessment models to accurately estimate parameters such as harvest levels, stock abundance and fishing mortality on a year to year basis has been the dominant form of evaluation(Chen et al. 2005; Dick & MacCall 2011; Thorson et al. 2013; Rosenberg et al. 2014). While the ability of a model to accurately estimate stock status is an important feature, it only provides a snapshot in time of the status of the population. Population trends are informative because they can provide a long-term view of population size and can determine whether a population is declining or increasing; allowing managers to make inferences about the health of a population (Haro et al. 2000; Chamberlain et al. 2013). If a model cannot accurately estimate population abundance in a given year, but the trend of the model's estimates closely mirrors the true trend of the population, then that model may still provide useful information for fisheries management about the general health of a population (Figure 1). The population trends estimated from catch-only models have the potential to be used as indicators of population status and to classify populations into IUCN-like threat categories.

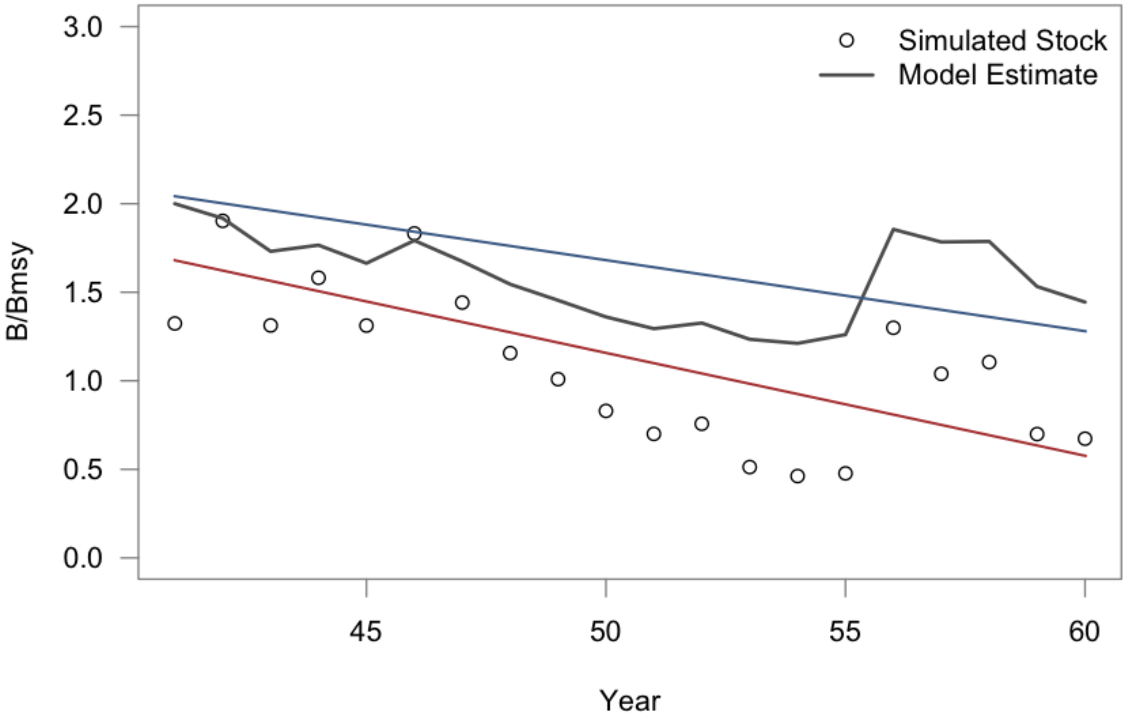

Figure 1 - An example of a comparison between the trend calculated from a model estimate (blue) versus the true underlying trend in B/Bmsy for the stock (red).

Receiver operating characteristics (ROC) curves can be used to visualize and evaluate the performance of classifiers (Fawcett 2006). ROC curves have been used across conservation, ecology, and physical sciences (Pearce et al. 2000; Baxter & Possingham 2011; Porszt et al. 2012). These curves have been proposed as an effective method to examine the tradeoffs between false positive rates and true positive rates in classification(Fawcett 2006). ROC curves typically examine a binary classification between a positive and a negative class. For the purposes of this study a positive is considered a threatened stock and a negative is a non-threatened stock. If a model correctly classifies a population as not threatened it would be a true negative, and it would be a true positive if a model correctly classifies a population as threatened. A false positive occurs when a model incorrectly classifies a stock as threatened and a false negative occurs when a model incorrectly classifies a stock as not threatened (Table 1). ROC curves plot the true positive rate against the false positive rate for a wide range of possible classification threshold values (Figure 2). This allows you to select a threshold to decide if a population is threatened or not based on the model's estimate of the population trend, by examining the rates of correct and misclassifications.

Table 1 - A generic confusion matrix representing the different possible classification outcomes. The predicated state represents the model estimates, and the true state is determined from the simulated stocks.

Confusion Matrix		True State	
		Threatened	**Not Threatened**
Predicted State	**Threatened**	True Positive	False Positive
	Not Threatened	False Negative	True Negative

A limitation of ROC analysis is that the ROC curve is insensitive to class imbalances (Fawcett 2006). A class imbalance occurs when one of the two classes, either positive or negative, is more prevalent. This can be an important consideration when dealing with rare event data, such as identifying endangered or critically endangered stocks. The cost of a misclassification varies depending on how many stocks are truly threatened versus not threatened. A small misclassification rate of many stocks will still be a large number of misclassifications compared to the same misclassification rate with a small number of stocks. So, with a class imbalance of many non-threatened stocks and few threatened stocks, there would be very different costs associated with a false positive versus a false negative. Hence, if class imbalances are not considered, there is a risk of coming to misleading and erroneously conclusions about classifier performance.

Precision recall (PR) curves have been proposed as an effective tool to evaluate classifier performance when there is class imbalance, particularly when the ratio of positive to negative classes is low (Davis & Goadrich 2006). These curves are commonly used in information retrieval and data mining, where there are few positive classes compared to negative classes (Dumais 1991; Järvelin & Kekäläinen 2000; Venna et al. 2010). PR curves plot precision (positive predictive value) against the true positive rate (recall) (Figure 2), visualizing the trade-off between decreasing precision as the true positive rate increases (Davis & Goadrich 2006). As precision examines both the number of true and false positives, it will reflect the ratio of threatened to non-threatened stocks, ensuring that any potential for over estimation of model performance due to class imbalances will be caught.

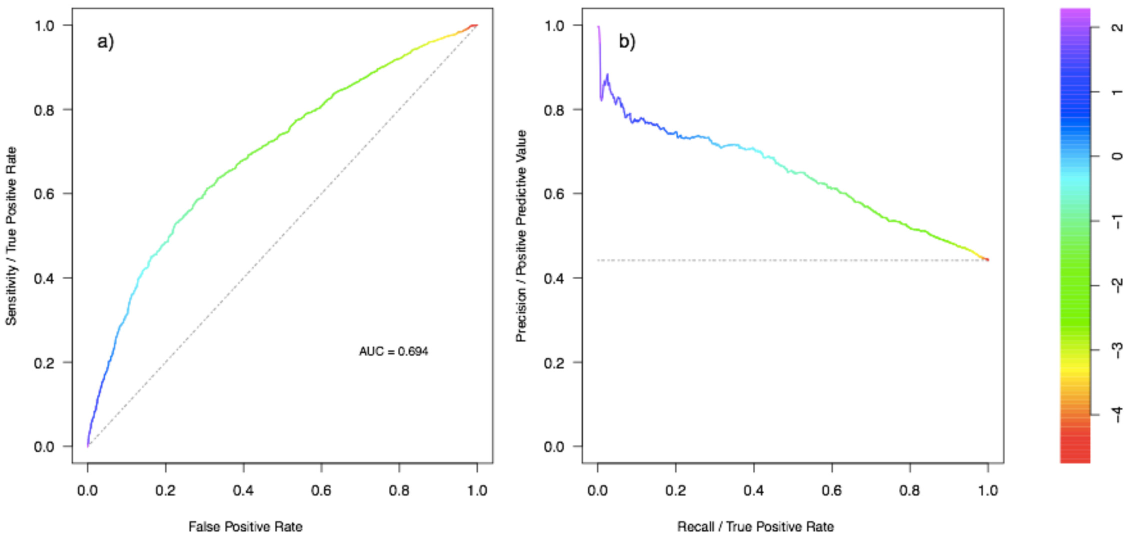

Figure 2 - An example of a ROC curve (a) and a corresponding PR curve (b). The ROC curve (a) plots the true positive rate against the false positive rate for a wide range of possible classification threshold values, as can be seen from the changing colour in the in the graph corresponding to the legend. Similarly, the PR Curve (b) plots the positive predictive value (precision) against the true positive rate (recall) for the same range of values. Looking at both figures you can then compare the precision of a specific threshold value compared to the true and false positive values on the ROC curve.

Building upon a United Nations Food and Agriculture Organization (FAO) report from 2014 (Rosenberg et al. 2014), four catch-only models were evaluated to assess how accurately these models estimated population trends over time. Rosenberg et al. (2014) evaluated the ability of four data-limited stock assessment models to determine stock status. The four models were chosen and adjusted to apply broadly to global fisheries and were tested on simulated stocks. The simulated stocks were designed to cover a broad range of life histories, allowing for an overall evaluation of these models that can be extrapolated to most fish populations (Rosenberg et al. 2014). Using a simulation framework for testing these models allows for reliable comparison of stock status between models, and allows for a robust evaluation of where models perform well or break down (Rosenberg et al. 2014). While the report assessed stock status estimates of the models, and found them to be robust (Rosenberg et al. 2014), analyses since have indicated that the models do not perform as well as initially thought (Andrew Cooper, personal communication).

The data from Rosenberg *et al*. (2014) was used to evaluate the ability of four catch-only models to estimate and classify population trends. The models' classification ability was then evaluated using both ROC and PR curves to ensure rigorous testing, as endangered and critically endangered stocks are rare events. This analysis provides an example of how using PR curves to account for class imbalance can correct for inaccurate conclusions derived from ROC curves. Evaluating the ability of the catch-only models to estimate population trends begins to address the knowledge gap that exists currently in the evaluation of the performance of catch only models.

Chapter 2. Methods

5760 simulated stocks with known relative abundance (current biomass versus biomass at maximum sustainable yield (B/B_{msy})) were used in this study along with the estimated B/B_{msy} produced by each of four catch only models created by Rosenberg *et al*. 2014. The simulated stocks were created to reflect global marine fisheries and had differing initial depletion, harvest rates, time series lengths and life history characteristics (Rosenberg et al. 2014). In addition, errors for catch reliability and recruitment variability were included(Rosenberg et al. 2014). Although it limits the inferences that can be drawn from the results to some extent, using simulated stocks instead of true fish stocks was necessary for a robust evaluation of the models' performance.

The models evaluated include one empirical model and three mechanistic models (Table 2). The empirical model is the modified log-linear panel regression model (mPRM) (Costello et al. 2012). The three mechanistic models were catch-MSY (CMSY), catch only model – sampling importance resampling (COM-SIR) and state-space catch only model (SSCOM), all of which are based on the Schaefer production model, with the latter two including harvest dynamics (Vasconcellos & Cochrane 2005; Martell & Froese 2013; Thorson et al. 2013). With the exception of the mPRM model, which requires basic life history classification of the fish species, as either demersal, large pelagic or small pelagic, all the models can be run solely with a series of catch data (Rosenberg et al. 2014). All four models are described in detail in Rosenberg *et al* 2014.

Table 2 - A comparison of the four different catch-only models used in this study.

	Empirical	**Mechanistic**		
Model	Modified panel regression (mPRM)	Modified catch-MSY (CMSY)	COM-SIR	SSCOM
Method	Log-linear regression model	Schafer model	Schafer model with harvest dynamics	Schafer model with harvest dynamics
Input	Catch, basic life history	Catch	Catch	Catch

Three different types of linear models were used to estimate both the true underlying population trend of the simulated stocks and the estimated trends produced by the data-limited models, henceforth referred to as the simulated trends and estimated trends, respectively. Both linear and robust regressions were used to estimate trends in relative abundance, as well as the non-parametric Sen's slope estimator. The three different types of linear models were used to examine if outliers or assumptions of normality affect the trend estimation and classification ability of the models. Robust regression was used as it is less sensitive to outliers than the linear regression, while the Sen's slope estimator was used as it is both insensitive to outliers and robust to small sample sizes(Wilcox 1998). Regressions were fit to both the true patterns of B/B_{msy} over time from the simulated stocks and the catch-only model estimates of B/B_{msy} for all stocks using the statistical program R (R Core Team 2014). There was no cross comparison between regression types at any stage of this analysis; the simulated and estimated trends were only compared within each regression type. To compare the effect of a long versus a short catch series on classifier performance, trends were calculated over both the whole time series and over the last ten-years for each stock.

Mean proportional error (MPE) and mean absolute proportional error (MAPE) were calculated to quantify bias of the model estimates. Proportional error is the estimated B/B_{msy} – the simulated B/B_{msy} over the simulated $B/B_{msy,}$ and when averaged across a set of values, is the average bias of the estimates. Additionally, absolute potential error was calculated to ensure that large errors evenly distributed between positive and negative biases were not missed. Regression trees were then used to examine any potential patterns or key factors resulting in larger or more biased errors. The factors considered were the different life histories, harvest rates, time series lengths, depletion rates, recruitment variability, error on catch, autocorrelation on recruitment residuals and the models themselves.

True and estimated trends were classified into three different conservation categories using thresholds based on the percent decline over time as per Criterion A of the International Union for the Conservation of Nature (IUCN), a widely recognized classification system (IUCN 2012). Three threat categories were selected as thresholds for this study: vulnerable (> 50% decline), endangered (> 70% decline) and critically endangered (> 90% decline). There were three separate binary classifications for each simulated stock based on the three different threat levels (e.g., vulnerable or worse vs.

not vulnerable, endangered or worse vs not endangered, critically endangered vs not critically endangered). Each stock was classified under each threshold based on the true simulated trend in B/B_{msy} and the estimated trend in B/B_{msy} from each of the four models. The simulated populations were classified based on the percent decline using the slope from the linear estimators over both the whole time series and the last ten years, representing a long and short time series respectively. For each simulated stock, there are 90 classification outcomes, 18 from each model and the "true" simulated outcomes, from the combination of linear model, time series length and threat threshold (Figure 3).

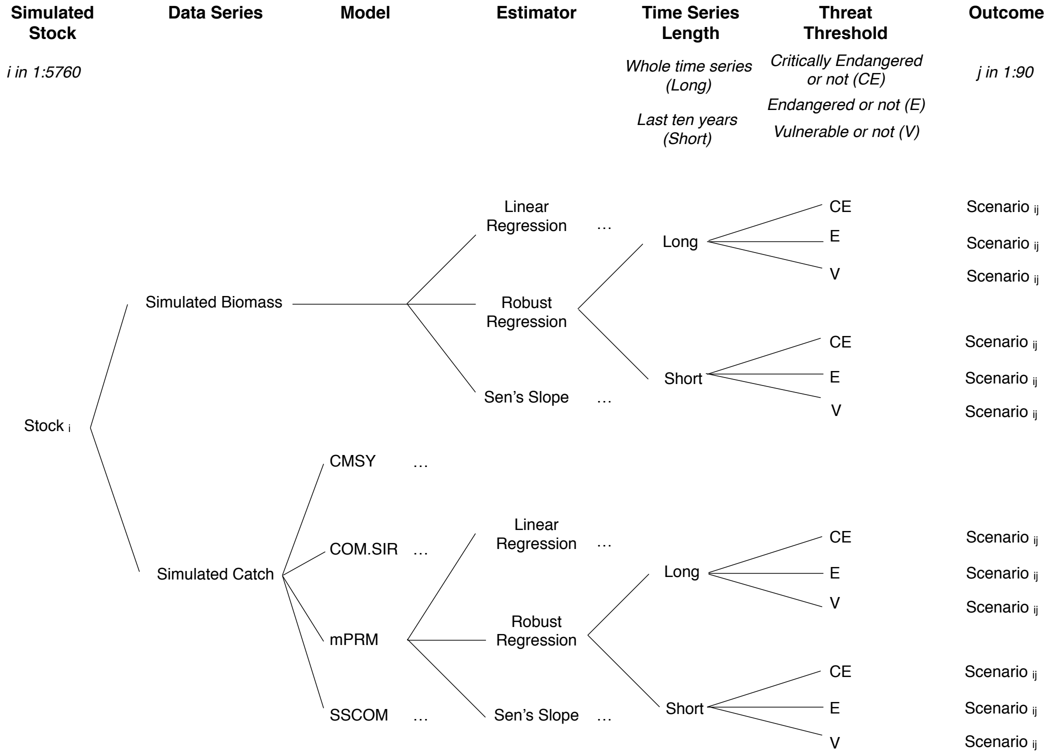

Figure 3 - A flow chart depicting the different scenarios that were examined for each of the simulated stocks.

The reliability of the model classifications based on the estimated trends was then quantified with ROC curves using the pROC package in R (Robin et al. 2011, R Core Team 2014). ROC curves examine the rates of true and false positives by varying the threshold used to classify the estimated trends as threatened or not, evaluating the ability of each model to correctly classify the stocks and identify the threshold minimizing the false positive rate and maximizing the true positive rate (Akobeng 2007; Porszt et al. 2012). The threshold identified that maximizes the area under the curve (AUC) represents the cut-off that would be used to identify a population as threatened or not for each threat category. In addition to computing the ROC curve and the AUC, the confusion matrix and common performance metrics were calculated for each scenario to thoroughly examine the classification performance of each model. The performance metrics included the false positive and negative rates, the true positive and negative rates, and precision. For the purposes of this study the true positive rate is the proportion of threatened stocks that are correctly classified as threatened by the model, and the false positive rate is the proportion of non-threatened stocks incorrectly classified as threatened. Whereas the true negative rate is the proportion of non-threatened stocks correctly classified as non-threatened, and the false negative rate is the proportion of threatened stocks that are incorrectly classified as not threatened by the model.

Precision Recall curves were computed to examine the change in precision and recall for different threshold values, using the PRROC package in R (Keilwagen et al 2014, R Core Team 2014). Recall is synonymous with the true positive rate and precision being the percent of true positives out of the total number of estimated positives. Precision Recall curves were used due to the sensitivity of the ROC curve to class skew and to ensure that the effects of class imbalance were not being missed in the ROC analysis (Dumais 1991; Järvelin & Kekäläinen 2000; Davis & Goadrich 2006; Fawcett 2006; Venna et al. 2010). This was important to include in our analysis, as endangered and particularly critically endangered stocks are arguably rare events.

Chapter 3. Results

The specific results from the Sen's slope estimator, linear and robust regressions differed, but regardless of the method, the general conclusions remain the same. For brevity, only the results of the robust regression will be presented here; the results from the linear regression and Sen's slope estimator are presented in Appendix 1.

The trends estimated by all the models underestimated the true trends in the data (see MPE and MAPE in Table 3). The regression trees detect any clear patterns in the biases in trends between the different life histories, harvest rates, time series lengths, depletion rates, recruitment variability, error on catch, autocorrelation on recruitment residuals or the models themselves.

Table 3 - The mean proportional error (MPE) and mean absolute proportional error (MAPE) for each of the data limited models. The error is the difference between the models' estimate of the simulated stock trend, and that of the true simulated stock. A negative MPE value indicates that the model estimates are underestimating the true simulated trend.

Model	Long Time Series		Short Time Series	
	MPE	MAPE	MPE	MAPE
CMSY	-0.512	4.493	-0.248	3.043
COM.SIR	-0.825	1.899	-1.071	1.723
mPRM	0.682	5.114	-1.501	3.837
SSCOM	-0.887	2.254	-0.778	2.476

The AUC values varied by model, threat threshold and time series length, but were high in most scenarios, suggesting reasonable to high classifier performance. Accordingly, the AUC values corresponded with mostly low false positive and negative rates and high true positive and negative rates. The AUC values declined as the threat level decreased from critically endangered to vulnerable. Here we present the ROC curves and confusion matrices at the optimal thresholds for the CMSY model (Figure 4, Table 4). CMSY was neither the best or worst performer, and is a representative example of the trends observed in all model performances. The remaining ROC and PR curves are available in Appendix 1.

Notwithstanding what seemed to be high classifier performance, two factors indicated that the ROC analysis was not fully capturing the models' performance: the optimal thresholds and the number of true positives relative to false positives. The optimal thresholds for maximizing classifier performance were almost always lower than the threat threshold (e.g. for a short time series with SSCOM, a stock should be considered critically endangered [i.e., has declined by more than 90%] if there is an estimated decline of 37%). However, not all the threat thresholds were logically meaningful. For CMSY the optimal threat threshold for classifying a population as vulnerable over a long time series was a 13% increase (Appendix 1). In the confusion matrices, there was only one scenario under which there were more true positives than false positives, despite low false positive rates. These two factors suggested that the ROC analysis was not robustly evaluating classifier performance.

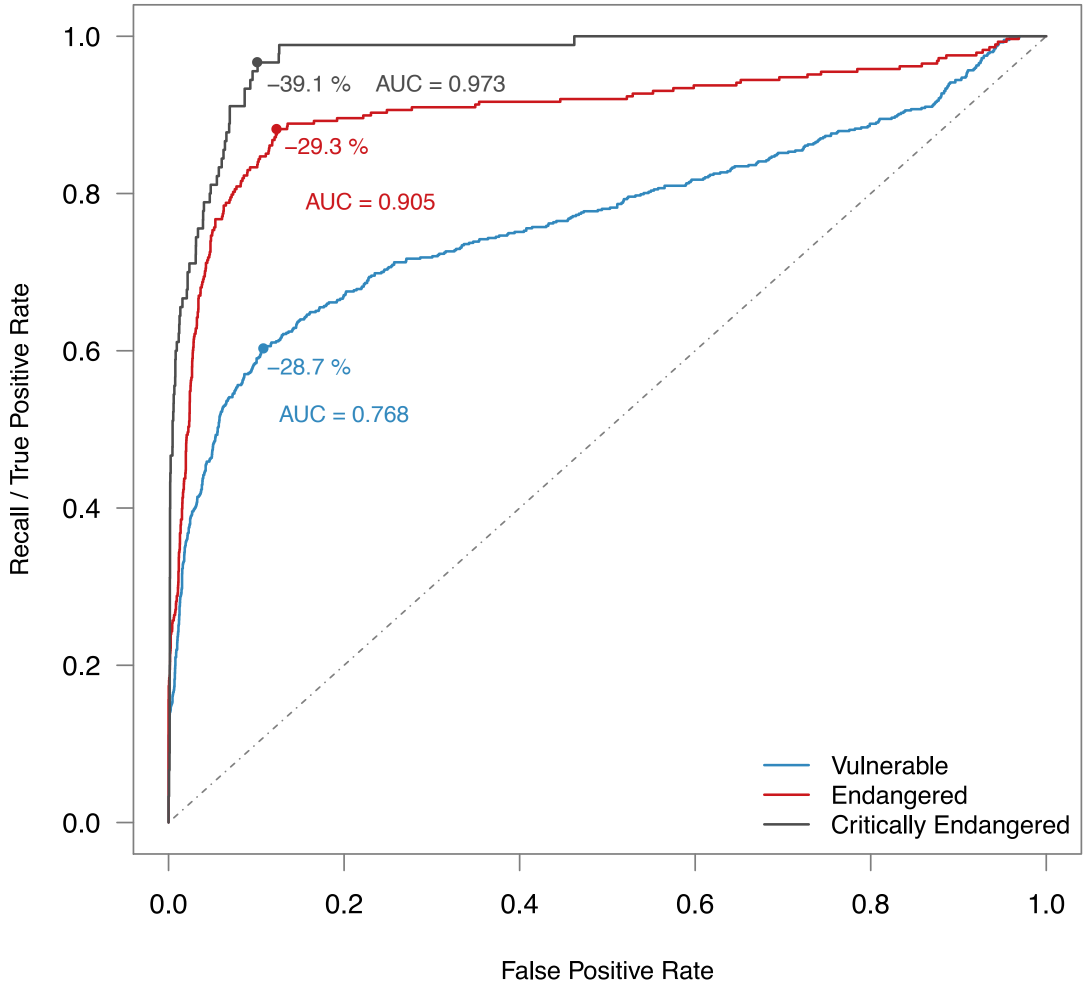

Figure 4 - **The ROC curve for CMSY at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds, representing percent change over ten years, and the AUC values are presented on the figure, with the performance of a random classifier represented by the dashed grey line.**

Table 4 - The confusion matrices for CMSY at the three different optimal thresholds for a) Vulnerable, b) Endangered and c) Critically Endangered. The class imbalance indicated in the tables is the ratio of threatened to non-threatened stocks.

a)

At the optimal threshold -28.7 %		True Status	
		Vulnerable	Not Vulnerable
Predicted Status	Vulnerable	389	527
	Not Vulnerable	258	4348
Class Imbalance = 1:8		647	4875

b)

At the optimal threshold -29.3 %		True Status	
		Endangered	Not Endangered
Predicted Status	Endangered	254	646
	Not Endangered	34	4588
Class Imbalance = 1:18		288	5234

c)

At the optimal threshold -39.1 %		True Status	
		Critically Endangered	Not Critically Endangered
Predicted Status	Critically Endangered	87	552
	Not Critically Endangered	3	4880
Class Imbalance = 1:60		90	5432

The PR Curves varied by model, threat threshold, and time series length. However, they all indicated that while the models perform better than random classifiers, there is still considerable room for improvement, with precision, also known as positive predictive value, dropping rapidly as the true positive rate increases (Figure 5). Precision at the optimal threshold (the threshold that minimizes false positives and maximizes true positives) declines as the threat level increases from vulnerable to critically endangered, with a maximum value of 0.57 at the optimal threshold identified by the ROC analysis. The highest precision for all models occurs when the threat threshold is vulnerable over the whole time series for all models. The lowest precision is associated with the highest true positive rates. This means that while the models are identifying a high proportion of the threatened stocks at the optimal thresholds, they are also incorrectly identifying non-threatened stocks as threatened, despite the low false positive rates. These changes observed in both curves corresponded to changes in the ratio of threatened to non-threatened stocks.

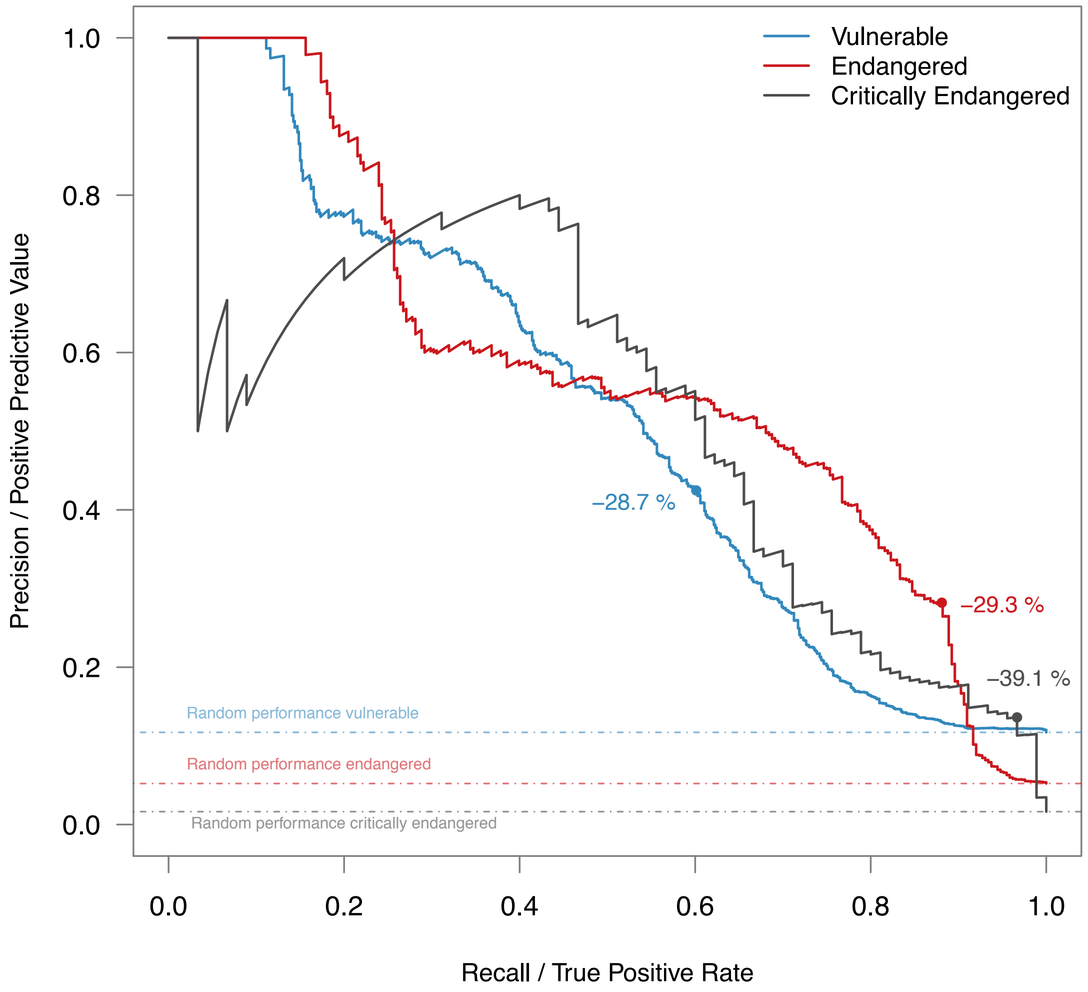

Figure 5 - **The PR curve for CMSY at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds identified in the ROC analysis, representing percent change over the whole time series, are presented on the figure. As random performance in PR curves varies with the class skew, the three corresponding random classifier performances are depicted by the dashed lines.**

In all scenarios, the number of threatened stocks (positive classes) is considerably lower than the number of non-threatened stocks (negative classes), creating a class imbalance. At the lowest class imbalance ratio, and for the vulnerable threat threshold, there is still approximately double the number of non-threatened stocks compared to threatened stocks. The largest class imbalances reached as high as 60 (or greater) non-threatened stocks for each threatened stock. All the largest class imbalances occurred when the threat level was critically endangered.

The class imbalance between threatened and non-threatened stocks masked the number of false positive classifications by all models. This pattern was consistent for all models. Interestingly, under specific circumstances, SSCOM could be potentially useful to identify stocks that are not critically endangered, despite the class imbalance, as it had a false negative rate of 0 (Table 5).

Table 5 - The confusion matrix for SSCOM at the optimal threshold for the critically endangered threat level. The class imbalance indicated in the tables is the ratio of threatened to non-threatened stocks.

At the optimal threshold (-0.368)		True Status	
		Critically Endangered	Not Critically Endangered
Predicted Status	Critically Endangered	92	435
	Not Critically Endangered	0	5233
Class Imbalance = 1:62		92	5668

Chapter 4. Discussion

The results from the ROC curve and the AUC values suggested that the four catch-only stock assessment models could be used to classify populations into threat categories in certain scenarios. However, upon careful inspection, the cases of high AUC values and low false positive rates were misleading regarding the model's classification ability. With only a few exceptions there were more false positives than true positives in all confusion matrices. In several scenarios, there were more than four times the number of false positives. However, the false positive rate remained low because ROC curves are insensitive to class imbalance and consequently did not reflect the high number of false positives due to the low ratio of positive to negative classes (Fawcett 2006).

Due to the class imbalance, the trade-offs between decreasing the false positive rate and increasing the true positive rate are unequal. For example, with a false positive rate of 0.1, 10% of 5000 non-threatened stocks are misclassified (500 false positives) and with a true positive rate of 0.9, 90% of 250 threatened stocks are classified correctly (225 true positives). While these rates would indicate high classifier performance, there are still more than double the number of false positives compared to true positives (500 false positives to 225 true positives). When ROC curves examine the classification rates, having a low false positive rate and a high true positive rate are weighted evenly, when in reality the costs associated with those rates are different when there is class skew. With the current skew, one false negative affects the true positive rate more than one false positive effects the false positive rate, hence minimizing the number of false negatives is where classifier performance is maximized. The class imbalance increased as the threat level increased from vulnerable to critically endangered. This occurs as there are fewer vulnerable stocks than non-vulnerable stocks and even fewer endangered and critically endangered stocks. When the ratio between the positive and negative classes approached 1:1, the false positive rate increased and the overall performance from the ROC curves indicated that the models are not particularly accurate classifiers.

Precision, the probability of a stock being of concern given the model classified the stock as of concern, was a better indicator of model performance. PR curves showed that the thresholds maximizing AUC from the ROC curve did not result in high precision. Precision decreased as the level of the threat threshold increased, due to the increasing class imbalance, the opposite of the ROC analysis, which increased with the increasing threat threshold. The highest precision values occurred when the trend was calculated over a long time series, versus a short time series, as it resulted in a lower class imbalance. It has been previously established in other fields that PR curves are more appropriate to use with imbalanced data sets (Davis & Goadrich 2006; He & Garcia 2009). Accordingly, Precision recall analysis is a well-established method in information retrieval, an area of research that often deals with imbalanced data sets (Lesk 1969; Dumais 1991; Hull 1993; Gao et al. 2001; Lafferty & Zhai 2001; Zhai & Lafferty 2001; Shah et al. 2002; Alzahrani et al. 2012)

When using ROC curves, it is important to not only look at the four main performance metrics and AUC values, but to also examine the confusion matrix and less commonly used performance metrics. The true and false positive and negative rates can be misleading and it is imperative to also carefully examine confusion matrices for class imbalance. Precision and the negative predictive value are two performance metrics that can be used to evaluate classifiers that will be affected by class imbalance and hence are more representative of classifier performance under conditions of class skew. Using PR curves in conjunction with ROC curves is a straightforward way to visualize important trade-offs when evaluating classifier performance.

When using ROC curves in ecology, where you do not always expect even class ratios, it is important to understand this limitation of ROC curves as it could lead to false confidence in a model's classification ability. Imbalances in the positive and negative classes overinflate the AUC values and are not reflected accurately by the true positive and the false positive rates. The high number of negative classes masked the number of false positives relative to the number of true positives. Relying solely on the rates from the ROC analysis would have led to the conclusion that the models were very accurate classifiers in some cases. Examining the confusion matrices and PR curves revealed the shortfalls of the models.

Evaluating whether model performance is poor, acceptable or good will depend on the scenarios where the model will be used and the costs of any errors. While these models have a low precision, there is still potential for using them. The models have low false negative rates, meaning they could still be used in instances where it is extremely important to identify all populations that are potentially declining past a certain threshold. False negatives are arguably worse than false positives in fisheries, making it more important in certain scenarios to find all stocks of concern than avoid false positives (Peterman 1990). However, which error is considered more concerning would depend on the population in question, the opinions of managers, and economic considerations. The performance metrics indicate that the models are still more accurate than a 50/50 guess. If there are no other options, besides a manager's best guess, these models could still be valuable. We do not know what a manager's AUC value would be; their opinions could be biased towards keeping fisheries open or by shifting baselines (Pauly 1995). Implementation errors by managers can lead to overexploitation (Patrick et al. 2013). For short time series, SSCOM is the most promising, as it has the lowest false negative rates and more realistic thresholds indicated from the ROC analysis for the three threat categories. CMSY could be used for longer time series, although the thresholds are less logical and the false negative rate is higher. Both models perform better over a short time series as oscillations in biomass are more likely to occur in a long time series and may not be accurately reflected when fitting the data to a line. These models would still require additional testing on data rich stocks to confirm that these patterns persist when using non-simulated stocks.

While evaluating data limited assessment methods' performance by the ability to classify stocks into categories is not novel, stocks have generally been classified into statuses across a range from developing to recovering based on catch or relative biomass estimates (Froese & Kesner-Reyes 2002; Kleisner and Pauly 2011, Carruthers *et al.* 2012; Kleisner *et al.* 2013). Although not applied to data-limited stocks previously, using the IUCN criterion for setting threat thresholds to classify extinction risk has been widely used in other studies (Cheung et al. 2005; Dulvy et al. 2005, 2006; Porszt et al. 2012; Hornborg et al. 2013; d'Eon-Eggertson et al. 2014; Visconti et al. 2016). Data limited models have been assessed for the ability to accurately estimate B/B_{msy}, F_{msy}, overfishing limits (OFL), harvest levels and adhere to management rules, either compared to full stock assessment methods or a true simulated condition (Dick & MacCall 2011; Wetzel & Punt 2011; Carruthers et al. 2012, 2014; Cope 2013; Martell & Froese 2013; Thorson et al. 2013). To our knowledge this is the first time that data limited models have been evaluated based on the ability to estimate population trends.

Accurate methods of fisheries assessment are required to allow managers to increase the sustainability of fisheries globally, however, current methods are not always an option or reliable. This raises the question, what else can be done with data-limited stocks? What methods provide a better than 50/50 chance of correctly classifying a stock as threatened or not? SSCOM could be an additional fisheries management tool, as additional support or opposition for action, if the model is used with the knowledge of its caveats. Further research is being conducted to determine what additional information will have the largest effect on model performance.

References

Agnew DJ, Gutiérrez NL, Butterworth DS. 2013. Fish catch data: Less than what meets the eye. Marine Policy **42**:268–269.

Akobeng AK. 2007. Understanding diagnostic tests 3: Receiver operating characteristic curves. Acta paediatrica **96**:644–647.

Alzahrani S, Vasile P, Salim N, Abraham A. 2012. Using Structural Information and Citation Evidence to Detect Significant Plagiarism Cases in Scientific Publications. Journal of the American Society for Information Science and Technology **63**:286–312.

Baxter PWJ, Possingham HP. 2011. Optimizing search strategies for invasive pests: Learn before you leap. Journal of Applied Ecology **48**:86–95.

Berkson J et al. 2011. Calculating Acceptable Biological Catch for Stocks That Have Reliable Catch Data Only (Only Reliable Catch Stocks – ORCS). NOAA Technical Memorandum NMFS-SEFSC-616:56.

Branch T A. 2008. Not all fisheries will be collapsed in 2048. Marine Policy **32**:38–39.

Branch T A., Jensen OP, Ricard D, Ye Y, Hilborn R. 2011. Contrasting Global Trends in Marine Fishery Status Obtained from Catches and from Stock Assessments. Conservation Biology **25**:777–786.

Butterworth DS, Johnston SJ, Brandao A. 2010. Pretesting the Likely Efficacy of Suggested Management Approaches to Data-Poor Fisheries. Marine and Coastal Fisheries **2**:131–145. Available from http://dx.doi.org/10.1577/C08-038.1.

Carruthers TR, Punt AE, Walters CJ, MacCall A, McAllister MK, Dick EJ, Cope J. 2014. Evaluating methods for setting catch limits in data-limited fisheries. Fisheries Research **153**:48–68. Available from http://dx.doi.org/10.1016/j.fishres.2013.12.014.

Carruthers TR, Walters CJ, McAllister MK. 2012. Evaluating methods that classify fisheries stock status using only fisheries catch data. Fisheries Research **119–120**:66–79. Available from http://dx.doi.org/10.1016/j.fishres.2011.12.011.

Chamberlain DE, Austin GE, Green RE, Hulme MF, Burton NHK. 2013. Improved estimates of population trends of Great Cormorants Phalacrocorax carbo in England and Wales for effective management of a protected species at the centre of a human–wildlife conflict. Bird Study **60**:335–344. Available from http://www.tandfonline.com/doi/abs/10.1080/00063657.2013.798258.

Chen Y, Kanaiwa M, Wilson C. 2005. Developing and evaluating a size-structured stock assessment model for the American lobster, Homarus americanus, fishery. New Zealand Journal of Marine and Freshwater Research **39**:645–660. Available from http://repositorio-aberto.up.pt/handle/10216/68026.

Cheung WWL, Pitcher TJ, Pauly D. 2005. A fuzzy logic expert system to estimate intrinsic extinction vulnerabilities of marine fishes to fishing. Biological Conservation **124**:97–111.

Cooper A. 2006. A guide to fisheries stock assessment: from data to recommendations:44. Available from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+Guide+to+Fisheries+Stock+Assessment+From+Data+to+Recommendations#0%5Cnhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+guide+to+fisheries+stock+assessment:+from+data+to+recommend.

Cope JM. 2013. Implementing a statistical catch-at-age model (Stock Synthesis) as a tool for deriving overfishing limits in data-limited situations. Fisheries Research **142**:3–14. Available from http://dx.doi.org/10.1016/j.fishres.2012.03.006.

Costello C et al. 2016. Global fishery futures under contrasting management regimes. Proceedings of the National Academy of Sciences of the United States of America:in review.

Costello C, Ovando D, Hilborn R, Gaines SD, Deschenes O, Lester SE. 2012. Status and Solutions for the World's Unassessed Fisheries. Science **338**:517–520.

d'Eon-Eggertson F, Dulvy NK, Peterman RM. 2014. Reliable identification of declining populations in an uncertain world. Conservation Letters **0**:1-11. Available from http://dx.doi.org/10.1111/conl.12123.

Daan N, Gislason H, Pope JG, Rice JC. 2011. Apocalypse in world fisheries? the reports of their death are greatly exaggerated. ICES Journal of Marine Science **68**:1375–1378.

Davis J, Goadrich M. 2006. The Relationship Between Precision-Recall and ROC Curves. Page 223-240 International Conference on Machine Learning (ICML). ACM Press, New York, NY. Available from http://portal.acm.org/citation.cfm?doid=1143844.1143874.

Dick EJ, MacCall AD. 2011. Depletion-Based Stock Reduction Analysis: A catch-based method for determining sustainable yields for data-poor fish stocks. Fisheries Research **110**:331–341. Available from http://dx.doi.org/10.1016/j.fishres.2011.05.007.

Dulvy NK, Jennings S, Goodwin NB, Grant A, Reynolds JD. 2005. Comparison of threat and exploitation status in North-East Atlantic marine populations. Journal of Applied Ecology **42**:883–891.

Dulvy NK, Jennings S, Rogers SI, Maxwell DL. 2006. Threat and decline in fishes: an indicator of marine biodiversity. Canadian Journal of Fisheries and Aquatic Sciences **63**:1267–1275.

Dumais S. 1991. Improving the retrieval of information from external sources. Behavior Research Methods, Instruments, & Computers **23**:229–236.

Eggert H, Greaker M. 2009. Environment for Development Effects of Global Fisheries on Developing Countries. Working Papers in Economics 393, University of Gothenburg, Department of Economics.

FAO Fisheries and Aquaculture Organisation. 2012. The State of World Fisheries and Aquaculture 2012. Food and Agriculture Organisation of the United Nations. Rome.

FAO Fisheries and Aquaculture Organization. 2016. The State of World Fisheries and Aquaculture. Fisheries and Aquaculture Organisation of the United Nations. Rome.

Fawcett T. 2006. An introduction to ROC analysis. Pattern Recognition Letters **27**:861–874.

Frank KT, Petrie B, Choi JS, Leggett WC. 2005. Trophic cascades in a formerly cod-dominated ecosystem. Science **308**:1621–1623.

Froese R, Kesner-Reyes K. 2002. Impact of fishing on the abundance of marine species. Journal of Marine Science **12**:1–12.

Froese R, Zeller D, Kleisner K, Pauly D. 2012. What catch data can tell us about the status of global fisheries. Marine Biology **159**:1283–1292.

Gao J, Nie J-Y, Xun E, Zhang J, Zhou M, Huang C. 2001. Improving query translation for cross-language information retrieval using statistical models. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01). ACM, New York, NY, USA, 96-104. Available from https://doi.org/10.1145/383952.383966

Haro A, Richkus W, Whalen K, Hoar A, Busch W-D, Lary S, Brush T, Dixon D. 2000. Population Decline of the American Eel: Implications for Research and Management. Fisheries **25**:7–16.

He H, Garcia EA. 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**:1263–1284.

Hornborg S, Svensson M, Nilsson P, Ziegler F. 2013. By-catch impacts in fisheries: Utilizing the IUCN red list categories for enhanced product level assessment in seafood LCAs. Environmental Management **52**:1239–1248.

Hull D. 1993. Using Statistical Testing in the Evaluation of Retrieval Experiments. In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '93), Robert Korfhage, Edie Rasmussen, and Peter Willett (Eds.). ACM, New York, NY, USA, 329-338. Available from http://dx.doi.org/10.1145/160688.160758

IUCN. 2012. IUCN Red List Categories and Criteria. Version 3.1. 2nd edition. Gland, Switzerland and Cambridge, UK.

Järvelin K, Kekäläinen J. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. n Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '00). ACM, New York, NY, USA, 41-48. Available from http://dx.doi.org/10.1145/345508.345545

Keilwagen J, Grosse I, Grau J. 2014. Area under Precision-Recall Curves for Weighted and Unweighted Data. PLoS ONE 9(3): e92209. Available from https://doi.org/10.1371/journal.pone.0092209

Kleisner K, Zeller D, Froese R, Pauly D. 2013. Using global catch data for inferences on the world's marine fisheries. Fish and Fisheries **14**:293–311.

Lafferty J, Zhai C. 2001. Document language models, query models, and risk minimization for information retrieval. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01). ACM, New York, NY, USA, 111-119. Available from https://doi.org/10.1145/383952.383970

Le Manach F, Gough C, Harris A, Humber F, Harper S, Zeller D. 2012. Unreported fishing, hungry people and political turmoil: The recipe for a food security crisis in Madagascar? Marine Policy **36**:218–225.

Lesk ME. 1969. Word-word associations in document retrieval systems. American Documentation **20**:27–38. Available from http://doi.wiley.com/10.1002/asi.4630200106 (accessed January 10, 2017).

MacCall AD. 2009. Depletion-corrected average catch: A simple formula for estimating sustainable yields in data-poor situations. ICES Journal of Marine Science **66**:2267–2271.

Mace PM. 2001. A new role for MSY in single-species and ecosystem approaches to fisheries stock assessment and management. Fish and Fisheries **2**:2–32.

Martell S, Froese R. 2013. A simple method for estimating MSY from catch and resilience. Fish and Fisheries **14**:504–514.

Myers RA, Baum JK, Shepherd TD, Powers SP, Peterson CH. 2007. Cascading Effects of the Loss of Apex Predatory Sharks from a Coastal Ocean. Science:1846–1850.

Patrick WS, Morrison W, Nelson M, González Marrero RL. 2013. Factors affecting management uncertainty in U.S. fisheries and methodological solutions. Ocean and Coastal Management **71**:64–72. Available from http://dx.doi.org/10.1016/j.ocecoaman.2012.11.002.

Pauly D. 1995. Anecdotes and the shifting baseline syndrome of fisheries. Trends in Ecology & Evolution **10**:430. Available from http://dx.doi.org/10.1016/S0169-5347(00)89171-5.

Pearce J, Ferrier S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling **133**:225–245.

Peterman RM. 1990. Statistical Power Analysis can Improve Fisheries Research and Management. Canadian Journal of Fisheries and Aquatic Sciences **47**:2–15.

Porszt EJ, Peterman RM, Dulvy NK, Cooper AB, Irvine JR. 2012. Reliability of Indicators of Decline in Abundance. Conservation Biology **26**:894–904.

Punt AE, Hilborn R. 1997. Fisheries stock assessment and decision analysis : the Bayesian approach **63**:35–63.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Avilable from http://www.R-project.org/.

Ricard D, Minto C, Jensen OP, Baum JK. 2012. Examining the knowledge base and status of commercially exploited marine species with the RAM Legacy Stock Assessment Database. Fish and Fisheries **13**:380–398.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics **12**:1471–2105.

Rosenberg AA et al. 2014. Developing New Approaches To Global Stock Status.

Schrank WE. 2005. The Newfoundland fishery: Ten years after the moratorium. Marine Policy **29**:407–420.

Shah U, Finin T, Joshi A, Cost RS, Matfield J. 2002. Information retrieval on the semantic web. Proceedings of the eleventh international conference on Information and knowledge management:461–468.

Teh LCL, Sumaila UR. 2013. Contribution of marine fisheries to worldwide employment. Fish and Fisheries **14**:77–88.

Thorson JT, Minto C, Minte-Vera CV, Kleisner KM, Longo C, Jacobson L. 2013. A new role for effort dynamics in the theory of harvested populations and data-poor stock assessment. Canadian Journal of Fisheries and Aquatic Sciences **70**:1829–1844. Available from http://www.nrcresearchpress.com/doi/abs/10.1139/cjfas-2013-0280.

Vasconcellos M, Cochrane K. 2005. Overview of World Status of Data-Limited Fisheries: Inferences from Landings Statistics. Page 958in C. Gélinas, M. Fortier, C. Viens, L. Fillion, and K. Puntillo, editors.Fisheries Assessment and Management in Data-Limited Situations. Alaska Sea Grant College Program, University of Alaska Fairbanks.

Venna J, Peltonen J, Nybo K, Aidos H, Kaski S, Aidos H, Nybo K, Peltonen J. 2010. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. The Journal of Machine Learning Research **11**:451–490. Available from http://www.scopus.com/inward/record.url?eid=2-s2.0-77949507946&partnerID=tZOtx3y1.

Visconti P et al. 2016. Projecting Global Biodiversity Indicators under Future Development Scenarios. Conservation Letters **9**:5–13.

Wetzel CR, Punt AE. 2011. Model performance for the determination of appropriate harvest levels in the case of data-poor stocks. Fisheries Research **110**:342–355.

Wilcox RR. 1998. A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic. Biometrical Journal **3**:261–268.

Worm B et al. 2009. Rebuilding global fisheries. Science (New York, N.Y.) **325**:578–585.

Zhai C, Lafferty J. 2001. Model-based feedback in the language modeling approach to information retrieval. In Proceedings of the tenth international conference on Information and knowledge management (CIKM '01), Henrique Paques, Ling Liu, and David Grossman (Eds.). ACM, New York, NY, USA, 403-410. DOI: http://dx.doi.org/10.1145/502585.502654

Appendix

Supplemental Figures

CMSY

Short Time Series

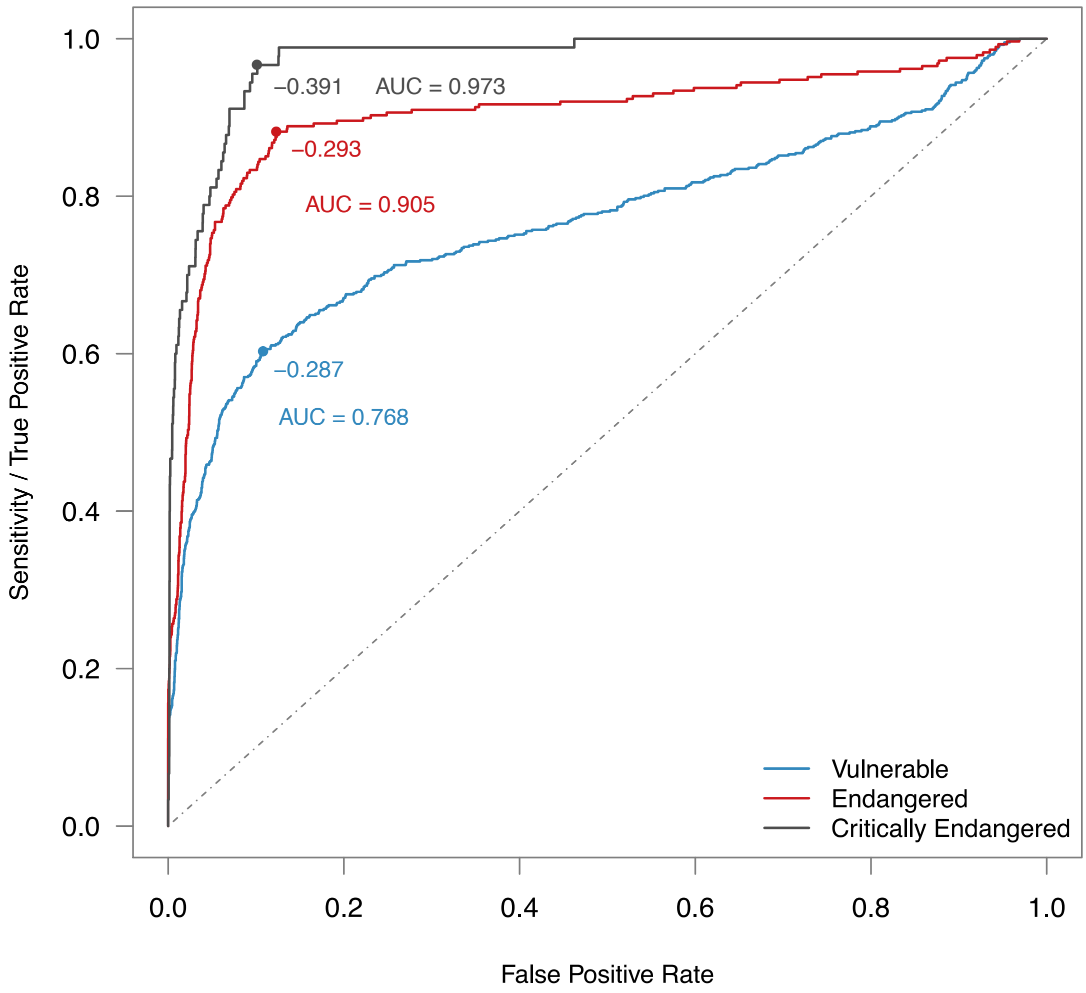

Figure A1 – The ROC curve for CMSY at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds, representing percent change over ten years, and the AUC values are presented on the figure, with the performance of a random classifier represented by the dashed grey line.

Table A1 - The confusion matrices for CMSY at the three different optimal thresholds for a) Vulnerable, b) Endangered and c) Critically Endangered over the last ten years. The class imbalance indicated in the tables is the ratio of threatned to non-threatened stocks.

a)

At the optimal threshold -28.7 %		True Status	
		Vulnerable	Not Vulnerable
Predicted Status	Vulnerable	389	527
	Not Vulnerable	258	4348
Class Imbalance = 1:8		647	4875

b)

At the optimal threshold -29.3 %		True Status	
		Endangered	Not Endangered
Predicted Status	Endangered	254	646
	Not Endangered	34	4588
Class Imbalance = 1:18		288	5234

c)

At the optimal threshold -39.1 %		True Status	
		Critically Endangered	Not Critically Endangered
Predicted Status	Critically Endangered	87	552
	Not Critically Endangered	3	4880
Class Imbalance = 1:60		90	5432

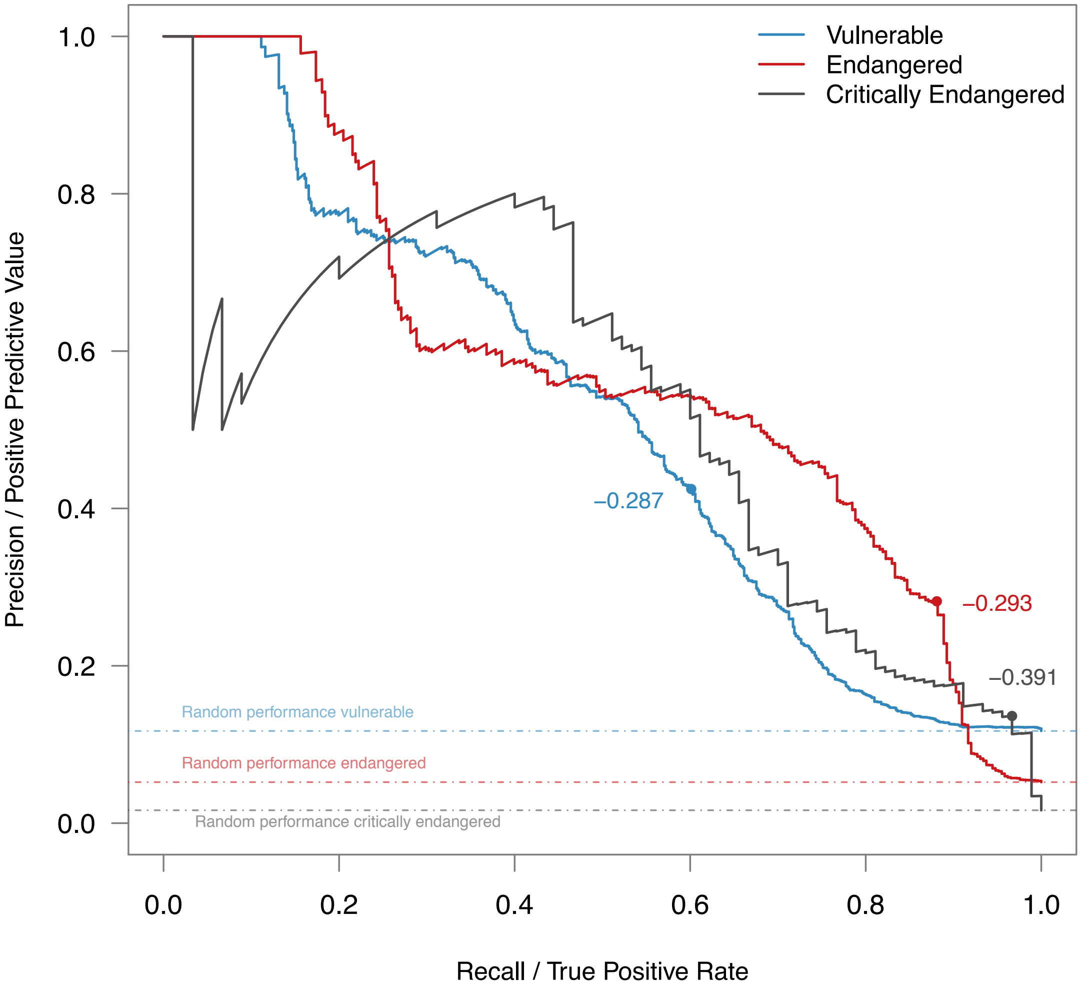

Figure A2 - **The PR curve for CMSY at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds identified in the ROC analysis, representing percent change over the whole time series, are presented on the figure. As a random performance in PR curves varies with the class skew, the three corresponding random classifier performances are depicted by the dashed lines.**

Long Time Series

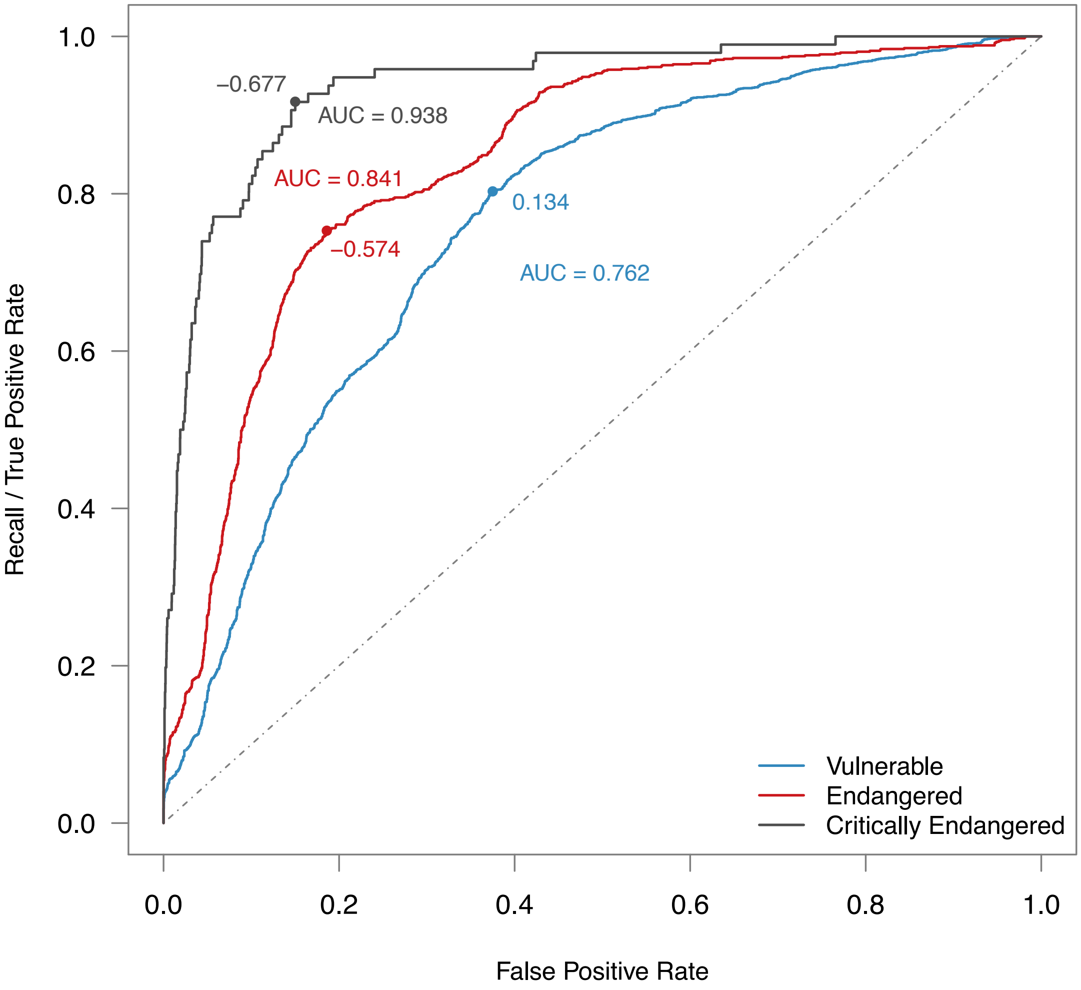

Figure A3 - The ROC curve for CMSY at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds, representing percent change over the whole time series, and the AUC values are presented on the figure, with the performance of a random classifier represented by the dashed grey line.

Table A2 - The confusion matrices for CMSY at the three different optimal thresholds for a) Vulnerable, b) Endangered and c) Critically Endangered over the whole time series. The class imbalance indicated in the tables is the ratio of threatned to non-threatened stocks.

a)

At the optimal threshold 13.4 %		True Status	
		Vulnerable	Not Vulnerable
Predicted Status	Vulnerable	1512	1363
	Not Vulnerable	373	2274
Class Imbalance = 1:1.9		1885	3637

b)

At the optimal threshold -57.4 %		True Status	
		Endangered	Not Endangered
Predicted Status	Endangered	658	870
	Not Endangered	216	3778
Class Imbalance = 1:5.3		874	4648

c)

At the optimal threshold -67.7 %		True Status	
		Critically Endangered	Not Critically Endangered
Predicted Status	Critically Endangered	87	811
	Not Critically Endangered	9	4615
Class Imbalance = 1:56.5		96	5426

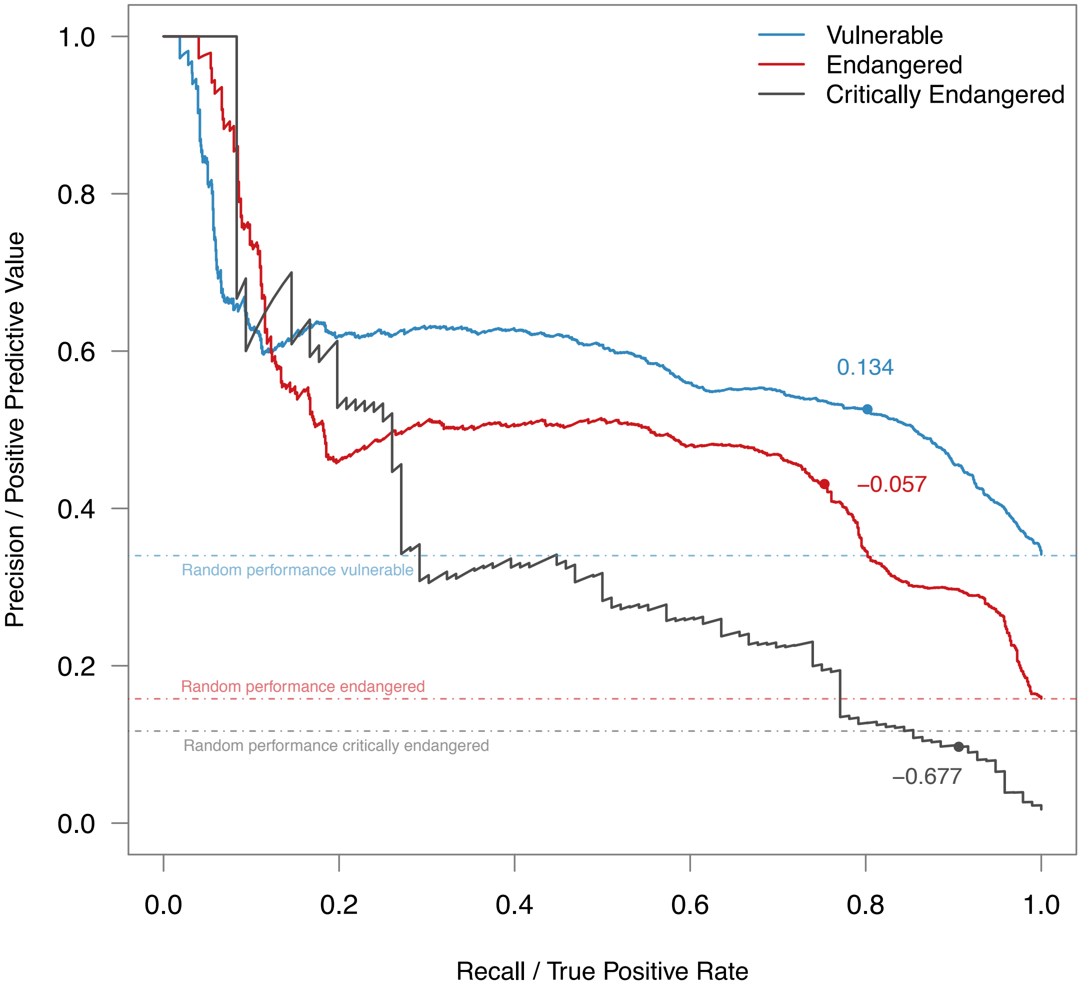

Figure A4 - **The PR curve for CMSY at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds identified in the ROC analysis, representing percent change over the whole time series, are presented on the figure. As a random performance in PR curves varies with the class skew, the three corresponding random classifier performances are depicted by the dashed lines.**

COM-SIR

Short Time Series

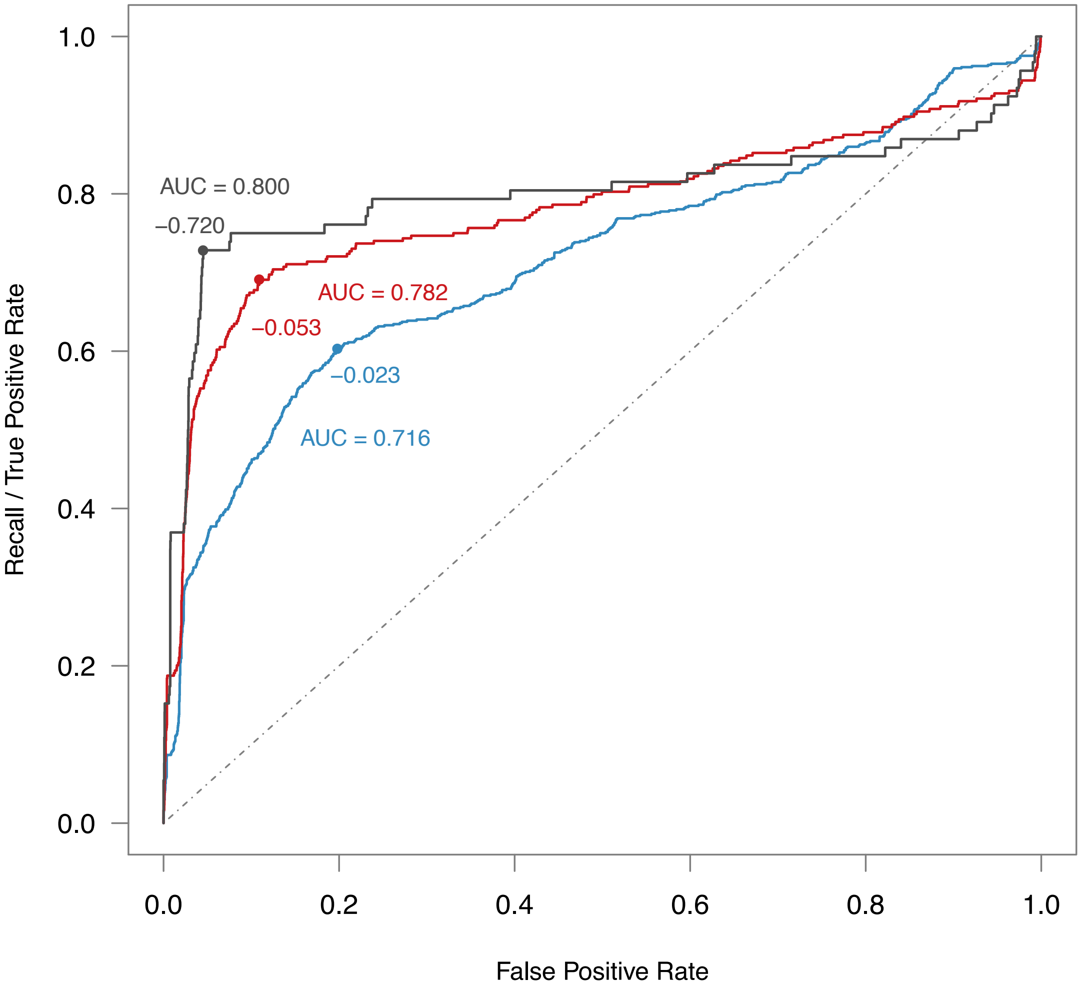

Figure A5 - The ROC curve for COM-SIR at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds, representing percent change over ten years, and the AUC values are presented on the figure, with the performance of a random classifier represented by the dashed grey line.

Table A3 - The confusion matrices for COM-SIR at the three different optimal thresholds for a) Vulnerable, b) Endangered and c) Critically Endangered over the last ten years. The class imbalance indicated in the tables is the ratio of threatned to non-threatened stocks.

a)

At the optimal threshold -2.3 %		True Status	
		Vulnerable	Not Vulnerable
Predicted Status	Vulnerable	413	991
	Not Vulnerable	279	4072
Class Imbalance = 1:7.3		692	5063

b)

At the optimal threshold -5.3 %		True Status	
		Endangered	Not Endangered
Predicted Status	Endangered	210	599
	Not Endangered	94	4852
Class Imbalance = 1:17.2		304	5234

c)

At the optimal threshold -72.0 %		True Status	
		Critically Endangered	Not Critically Endangered
Predicted Status	Critically Endangered	67	257
	Not Critically Endangered	25	5406
Class Imbalance = 1:61.5		92	5663

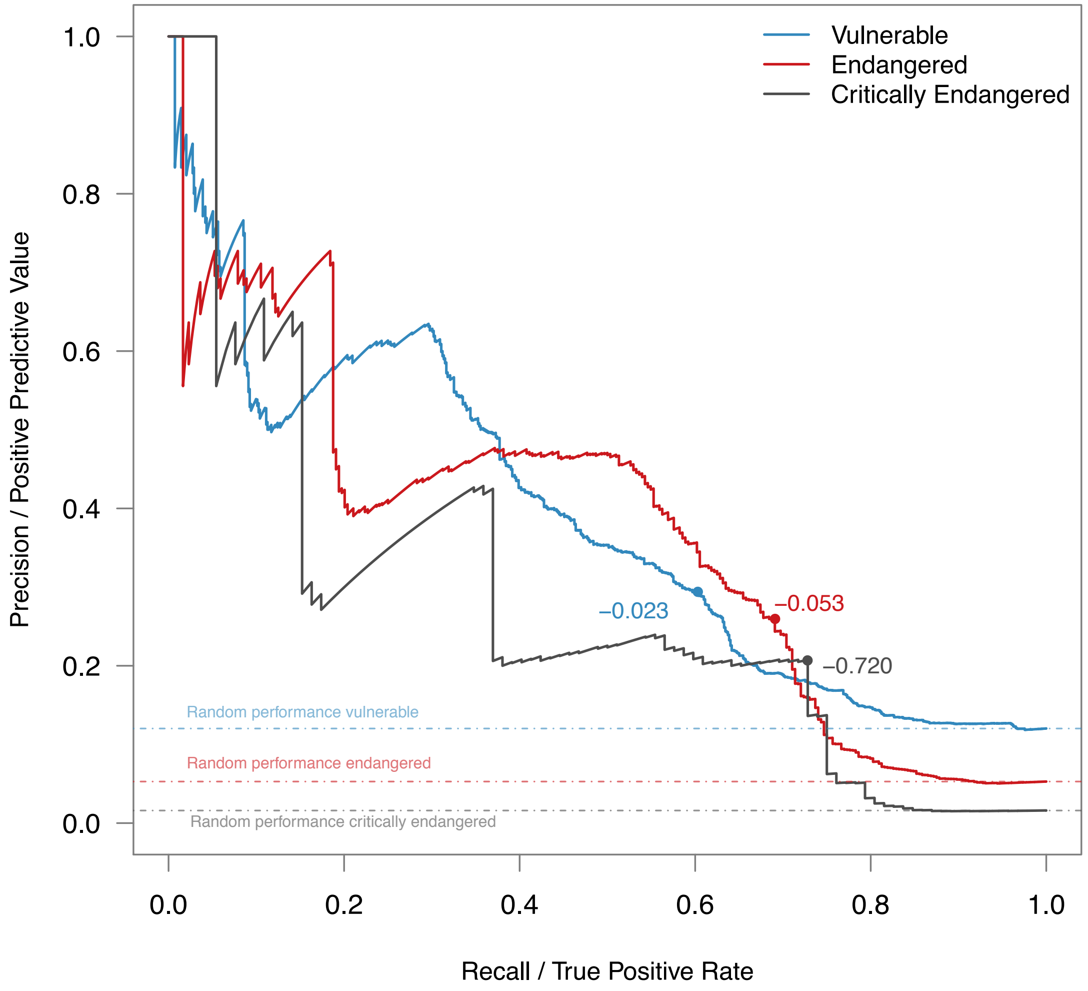

Figure A6 – The PR curve for COM-SIR at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds identified in the ROC analysis, representing percent change over the last ten years, are presented on the figure. As a random performance in PR curves varies with the class skew, the three corresponding random classifier performances are depicted by the dashed lines.

Long Time Series

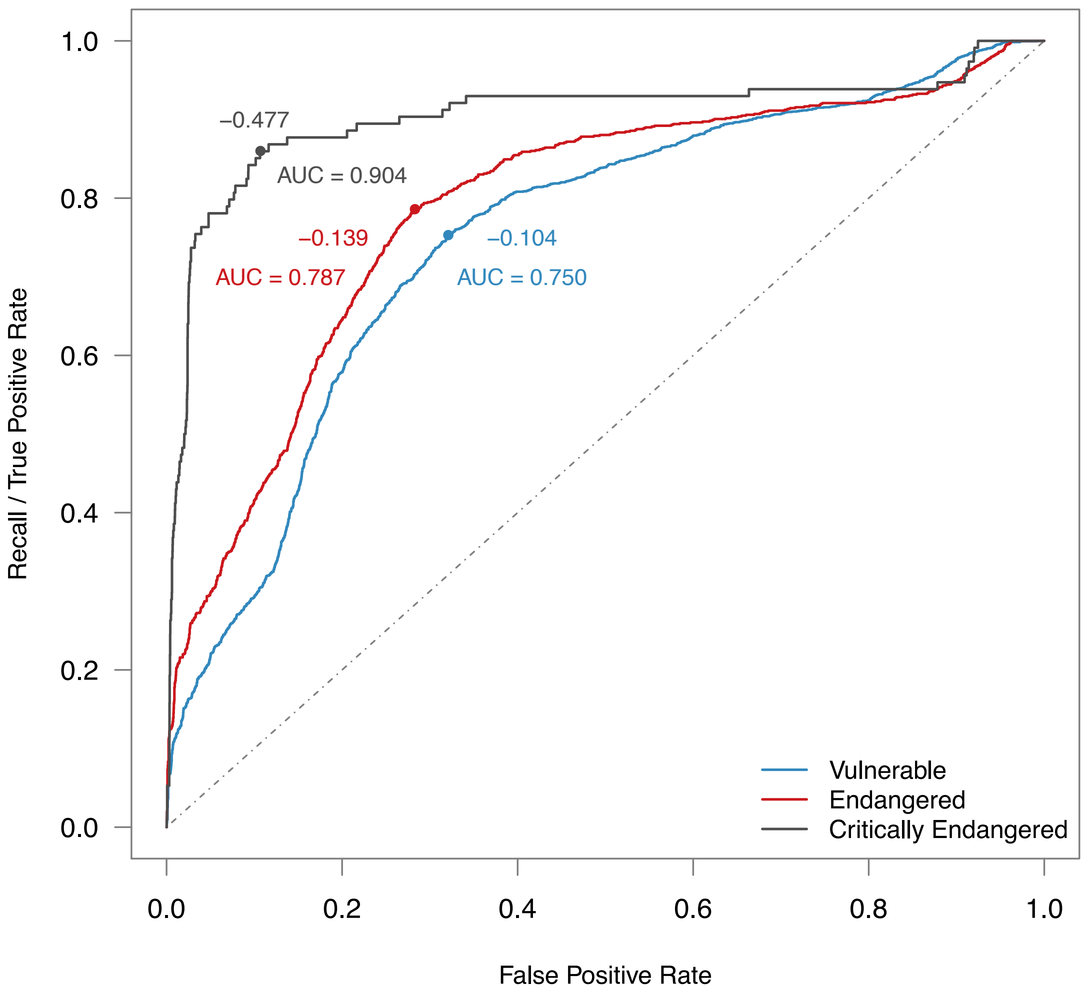

Figure A7 - **The ROC curve for COM-SIR at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds, representing percent change over the whole time series, and the AUC values are presented on the figure, with the performance of a random classifier represented by the dashed grey line.**

Table A4 - The confusion matrices for COM-SIR at the three different optimal thresholds for a) Vulnerable, b) Endangered and c) Critically Endangered over the whole time series. The class imbalance indicated in the tables is the ratio of threatned to non-threatened stocks.

a)

At the optimal threshold -10.4 %		True Status	
		Vulnerable	Not Vulnerable
Predicted Status	Vulnerable	1509	1209
	Not Vulnerable	493	2544
Class Imbalance = 1:1.9		2002	3753

b)

At the optimal threshold -13.9 %		True Status	
		Endangered	Not Endangered
Predicted Status	Endangered	733	1363
	Not Endangered	203	3456
Class Imbalance = 1:5.1		936	4819

c)

At the optimal threshold -47.7%		True Status	
		Critically Endangered	Not Critically Endangered
Predicted Status	Critically Endangered	98	601
	Not Critically Endangered	16	5040
Class Imbalance = 1:60		114	5641

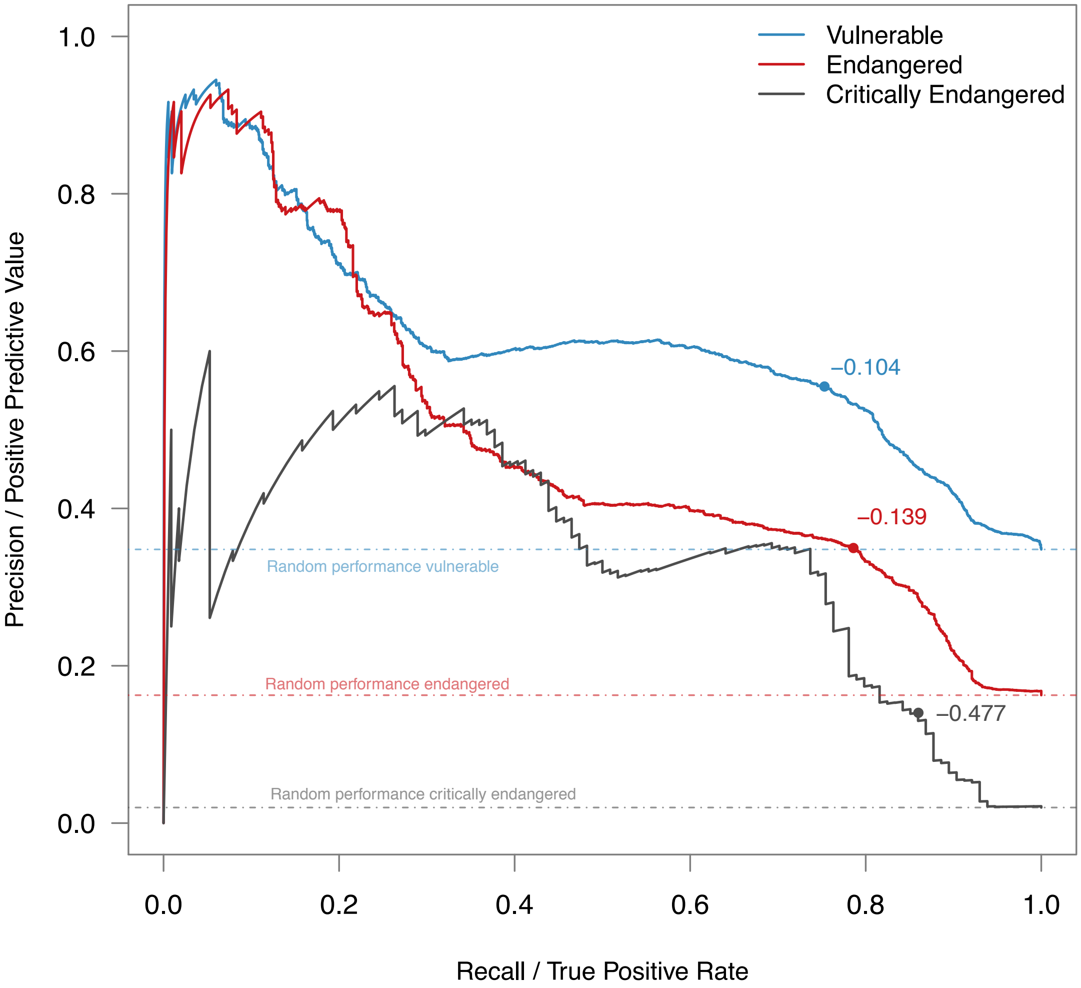

Figure A8 - **The PR curve for COM-SIR at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds identified in the ROC analysis, representing percent change over the whole time series, are presented on the figure. As a random performance in PR curves varies with the class skew, the three corresponding random classifier performances are depicted by the dashed lines.**

mPRM

Short Time Series

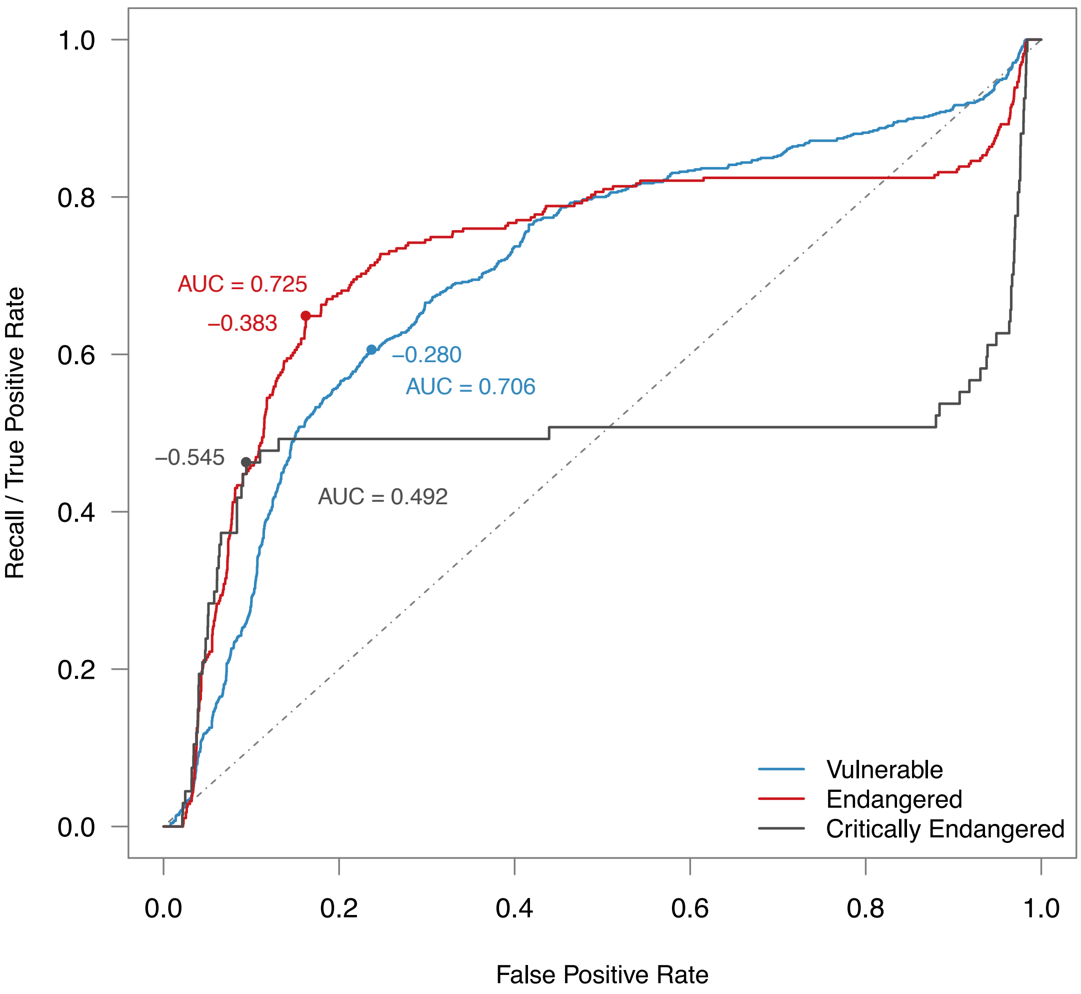

Figure A9 - The ROC curve for mPRM at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds, representing percent change over ten years, and the AUC values are presented on the figure, with the performance of a random classifier represented by the dashed grey line.

Table A5 - The confusion matrices for mPRM at the three different optimal thresholds for a) Vulnerable, b) Endangered and c) Critically Endangered over the last ten years. The class imbalance indicated in the tables is the ratio of threatned to non-threatened stocks.

a)

At the optimal threshold -28.0 %		True Status	
		Vulnerable	Not Vulnerable
Predicted Status	Vulnerable	415	1197
	Not Vulnerable	270	3853
Class Imbalance = 1:7.4		685	5050

b)

At the optimal threshold -38.3 %		True Status	
		Endangered	Not Endangered
Predicted Status	Endangered	181	886
	Not Endangered	98	4570
Class Imbalance = 1:19.6		279	5456

c)

At the optimal threshold -54.5 %		True Status	
		Critically Endangered	Not Critically Endangered
Predicted Status	Critically Endangered	31	536
	Not Critically Endangered	36	5132
Class Imbalance = 1:84.6		67	5668

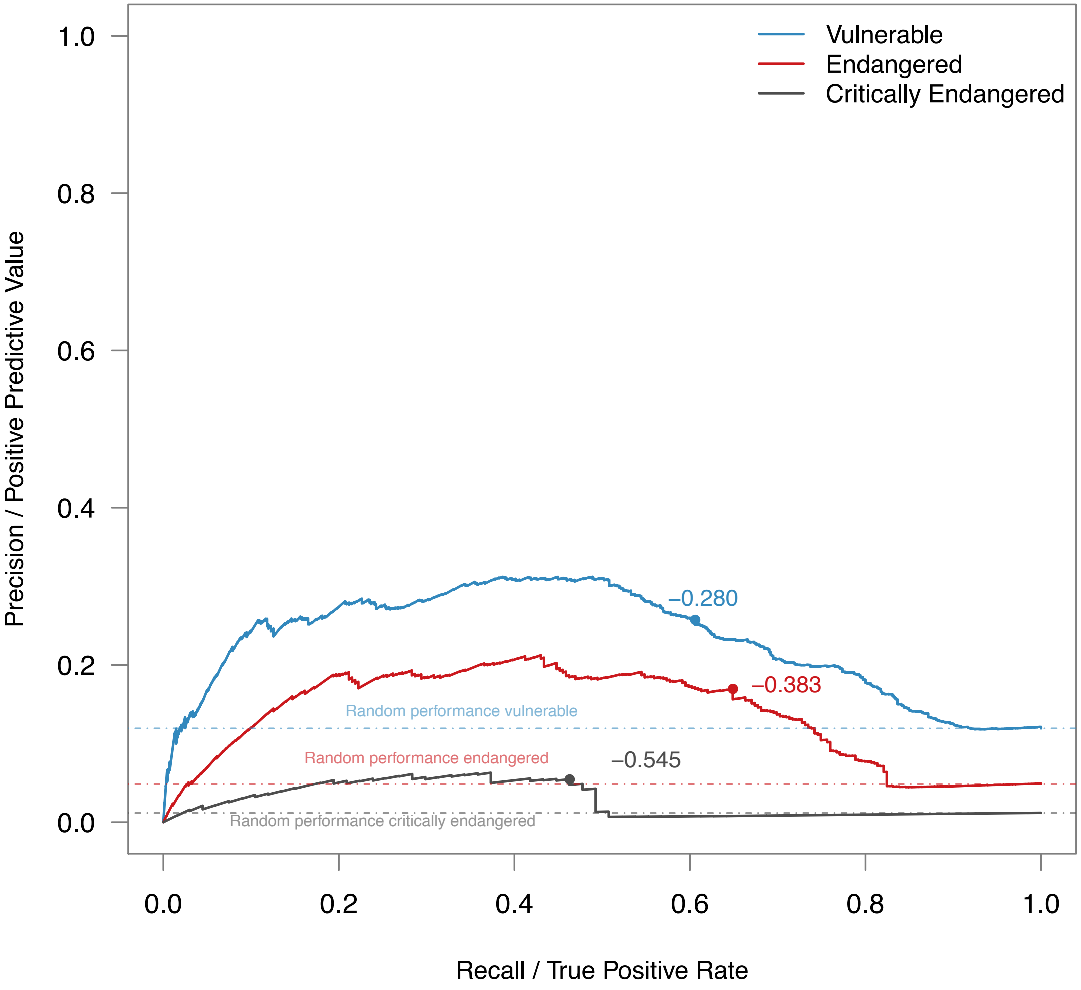

Figure A10 - The PR curve for mPRM at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds identified in the ROC analysis, representing percent change over the last ten years, are presented on the figure. As a random performance in PR curves varies with the class skew, the three corresponding random classifier performances are depicted by the dashed lines.

Long Time Series

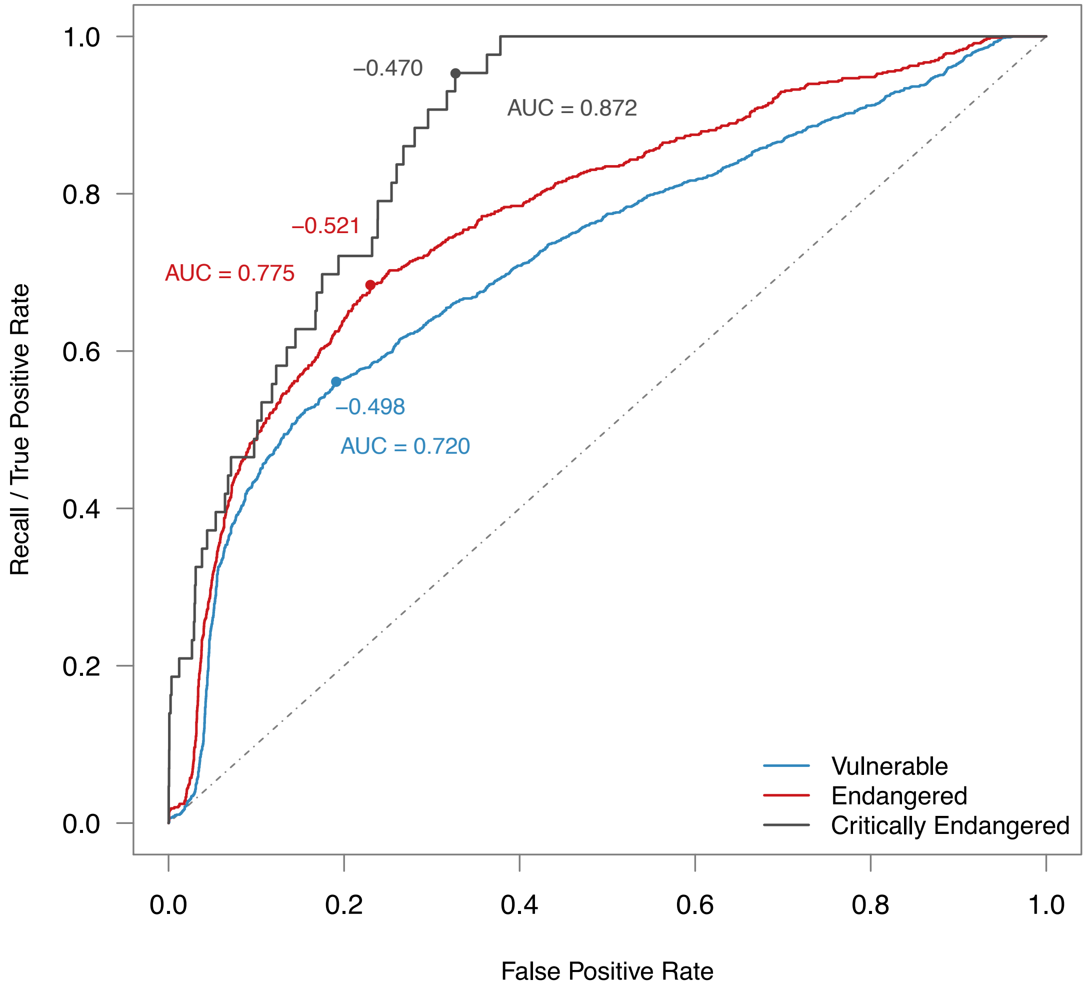

Figure A11 - The ROC curve for mPRM at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds, representing percent change over the whole series, and the AUC values are presented on the figure, with the performance of a random classifier represented by the dashed grey line.

Table A6 - The confusion matrices for mPRM at the three different optimal thresholds for a) Vulnerable, b) Endangered and c) Critically Endangered over the whole time series. The class imbalance indicated in the tables is the ratio of threatned to non-threatened stocks.

a)

At the optimal threshold -49.8 %		True Status	
		Vulnerable	Not Vulnerable
Predicted Status	Vulnerable	995	755
	Not Vulnerable	779	3206
Class Imbalance = 1:2.2		1774	3961

b)

At the optimal threshold -52.1 %		True Status	
		Endangered	Not Endangered
Predicted Status	Endangered	476	1162
	Not Endangered	220	3877
Class Imbalance = 1:7.2		696	5039

c)

At the optimal threshold -47.7 %		True Status	
		Critically Endangered	Not Critically Endangered
Predicted Status	Critically Endangered	41	1859
	Not Critically Endangered	2	3833
Class Imbalance = 1:132.4		43	5692

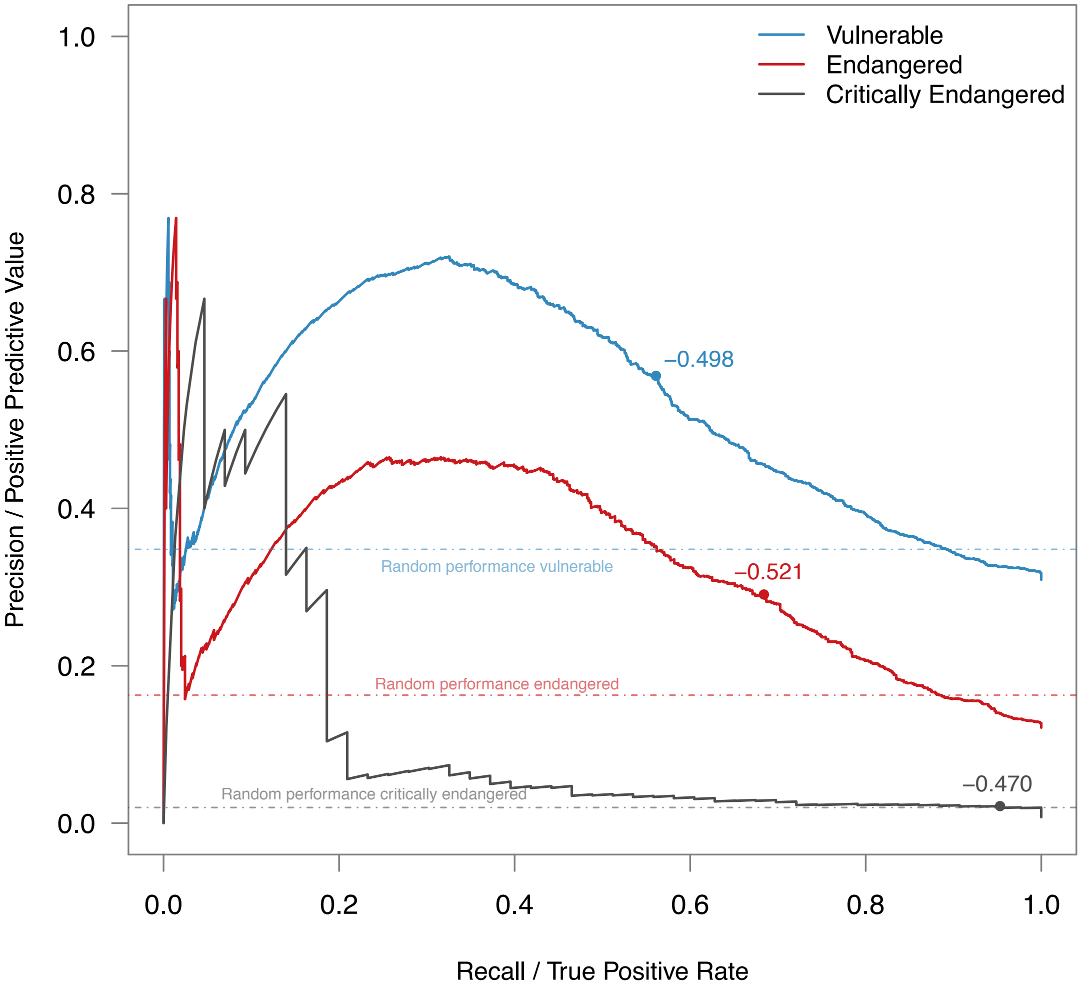

Figure A12 - The PR curve for mPRM at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds identified in the ROC analysis, representing percent change over the whole time series, are presented on the figure. As a random performance in PR curves varies with the class skew, the three corresponding random classifier performances are depicted by the dashed lines.

SSCOM

Short Time Series

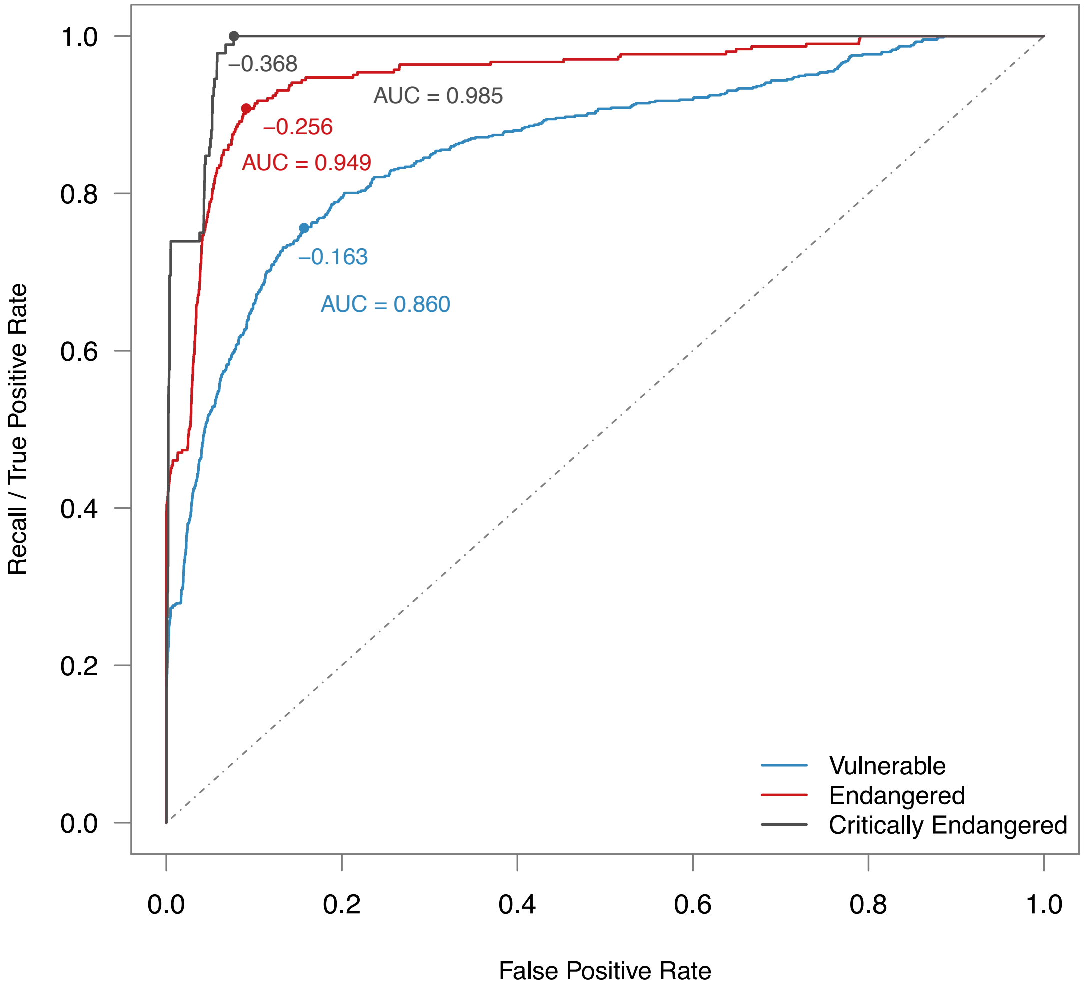

Figure A13 - The ROC curve for SSCOM at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds, representing percent change over ten years, and the AUC values are presented on the figure, with the performance of a random classifier represented by the dashed grey line.

Table A7 - The confusion matrices for SSCOM at the three different optimal thresholds for a) Vulnerable, b) Endangered and c) Critically Endangered over the last ten years. The class imbalance indicated in the tables is the ratio of threatned to non-threatened stocks.

a)

At the optimal threshold -16.3 %		True Status	
		Vulnerable	Not Vulnerable
Predicted Status	Vulnerable	523	796
	Not Vulnerable	169	4272
Class Imbalance = 1:8		692	5068

b)

At the optimal threshold -25.6 %		True Status	
		Endangered	Not Endangered
Predicted Status	Endangered	275	497
	Not Endangered	29	4959
Class Imbalance = 1:17.9		304	5456

c)

At the optimal threshold -36.8 %		True Status	
		Critically Endangered	Not Critically Endangered
Predicted Status	Critically Endangered	92	435
	Not Critically Endangered	0	5233
Class Imbalance = 1:60		92	5668

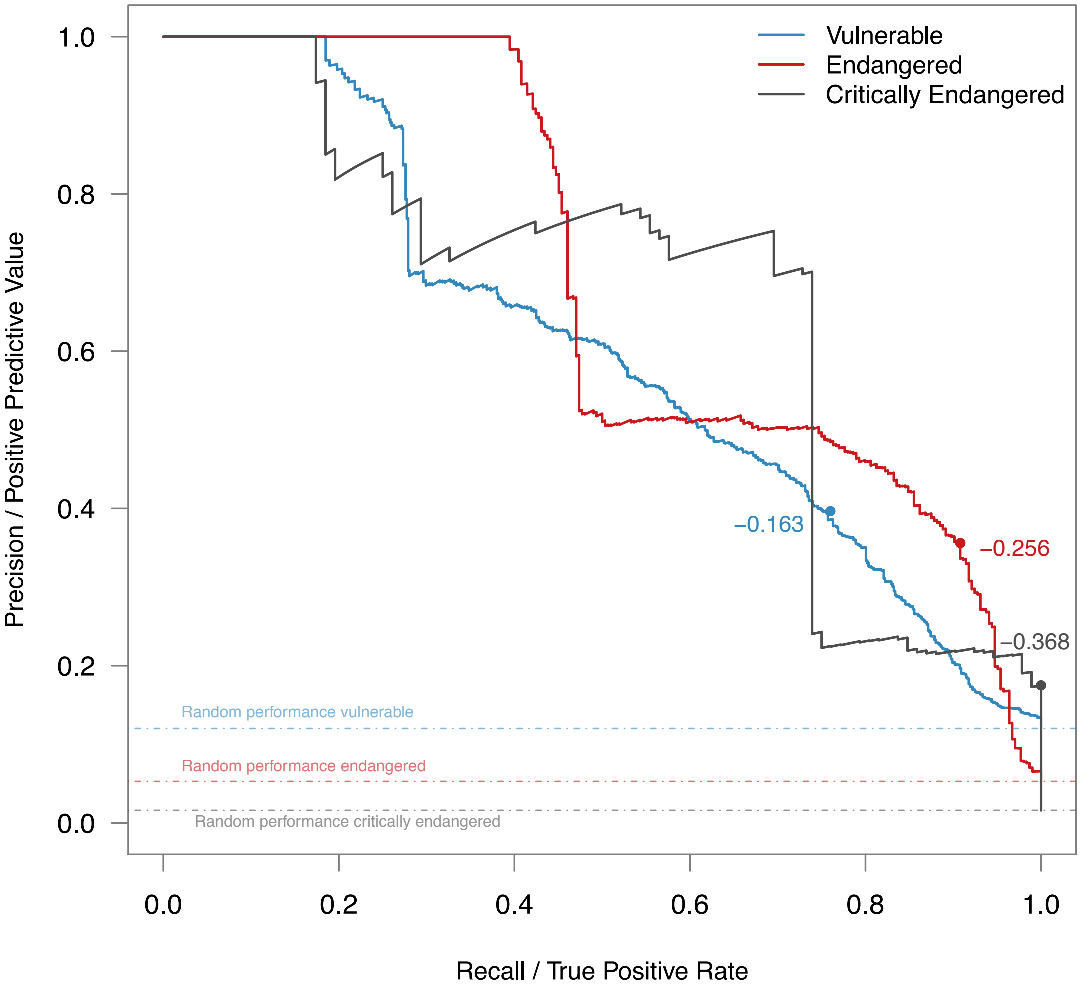

Figure A14 - **The PR curve for SSCOM at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds identified in the ROC analysis, representing percent change over the last ten years, are presented on the figure. As a random performance in PR curves varies with the class skew, the three corresponding random classifier performances are depicted by the dashed lines.**

Long Time Series

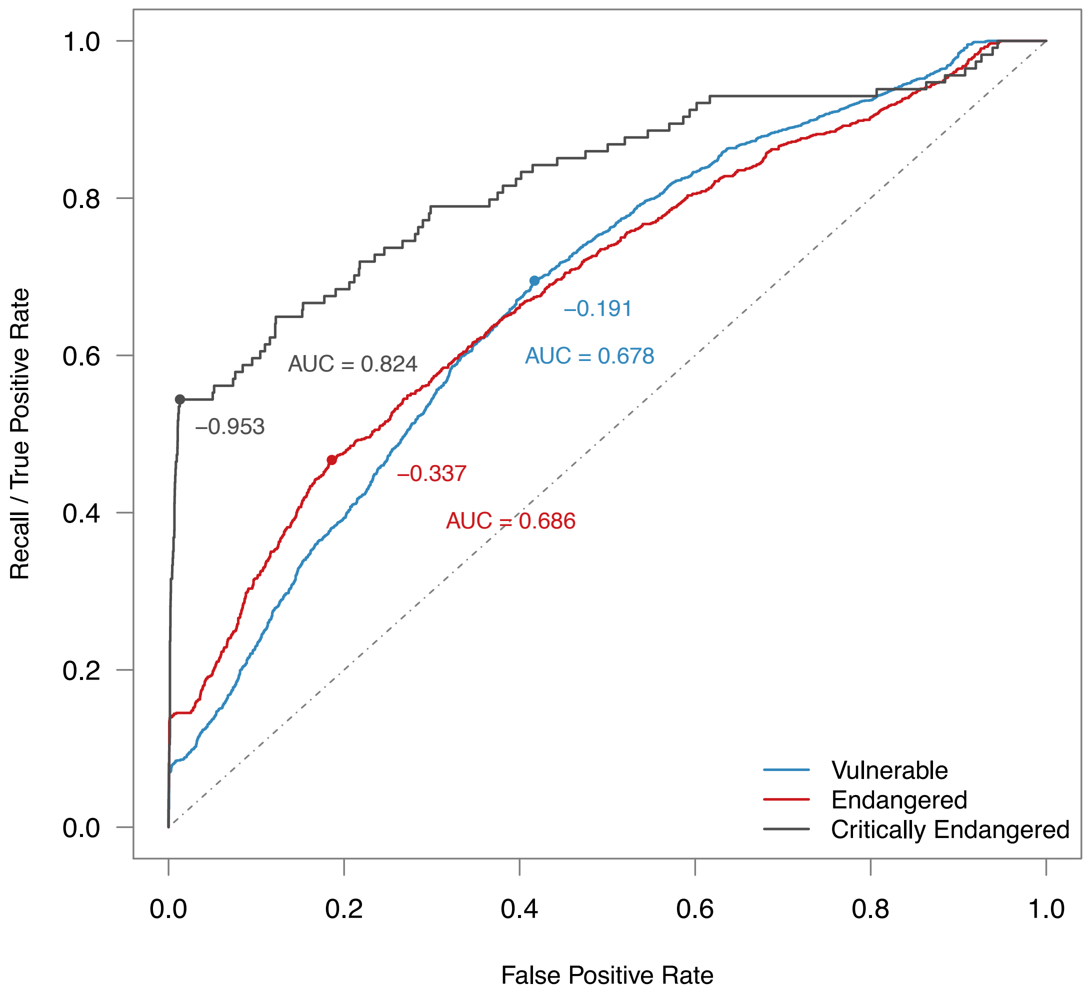

Figure A15 - **The ROC curve for SSCOM at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds, representing percent change over the whole time series, and the AUC values are presented on the figure, with the performance of a random classifier represented by the dashed grey line.**

Table A8 - **The confusion matrices for SSCOM at the three different optimal thresholds for a) Vulnerable, b) Endangered and c) Critically Endangered over the whole time series. The class imbalance indicated in the tables is the ratio of threatned to non-threatened stocks.**

a)

At the optimal threshold -19.1 %		True Status	
		Vulnerable	Not Vulnerable
Predicted Status	Vulnerable	1392	1570
	Not Vulnerable	610	2188
Class Imbalance = 1:1.9		2002	3758

b)

At the optimal threshold -33.7 %		True Status	
		Endangered	Not Endangered
Predicted Status	Endangered	437	896
	Not Endangered	499	3928
Class Imbalance = 1:5.2		936	4824

c)

At the optimal threshold -95.3 %		True Status	
		Critically Endangered	Not Critically Endangered
Predicted Status	Critically Endangered	62	72
	Not Critically Endangered	52	5572
Class Imbalance = 1:49.5		114	5646

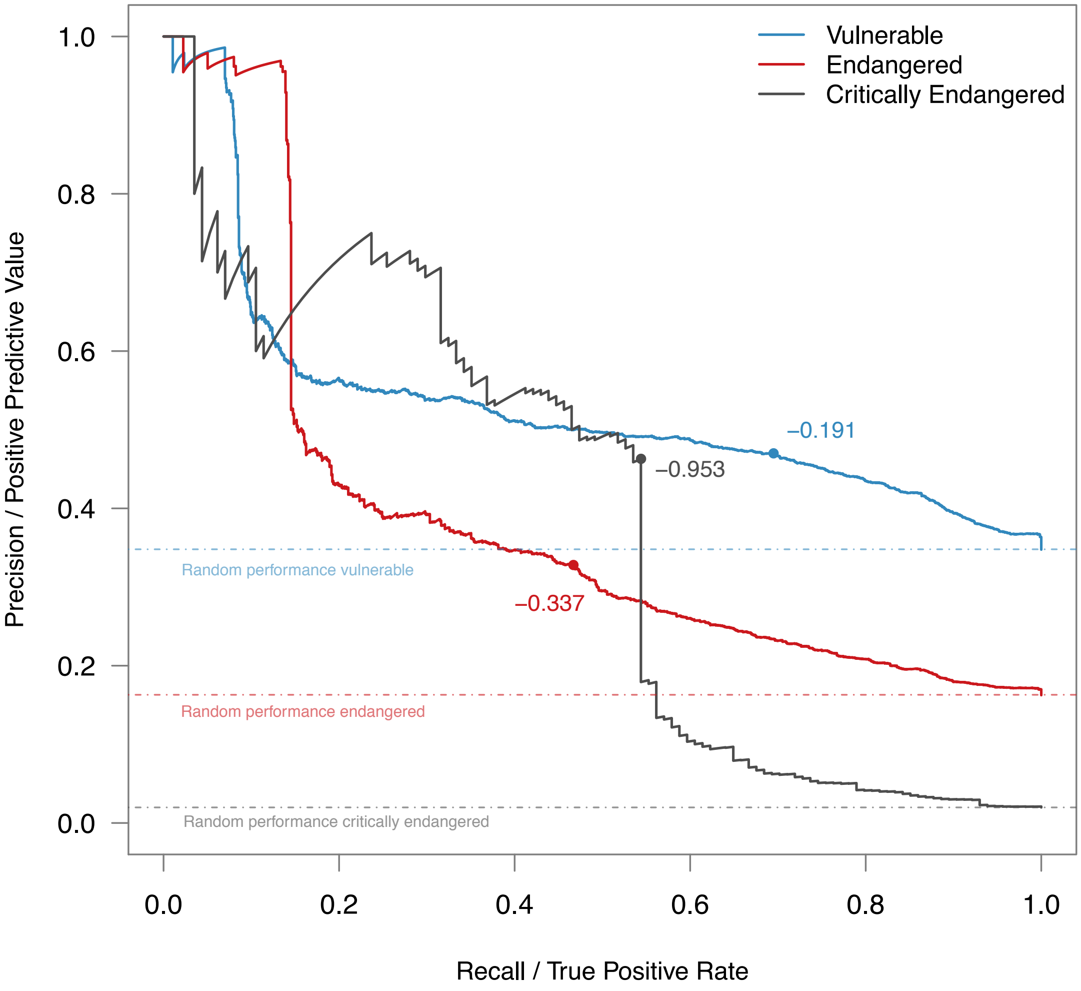

Figure A16 - ***The PR curve for SSCOM at three different threat thresholds: vulnerable, endangered and critically endangered. The optimal thresholds identified in the ROC analysis, representing percent change over the whole time series, are presented on the figure. As a random performance in PR curves varies with the class skew, the three corresponding random classifier performances are depicted by the dashed lines.***