AN EVALUATION OF CRITERIA FOR ASSESSING CONSERVATION STATUS OF FRASER RIVER SOCKEYE SALMON CONSERVATION UNITS

by

Erin Joanna Porszt
B.Sc., University of Guelph, 2002

RESEARCH PROJECT
SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF RESOURCE MANAGEMENT

In the
School of Resource and Environmental Management

Project No. 473

SIMON FRASER UNIVERSITY

Summer 2009

APPROVAL

Name: **Erin Joanna Porszt**

Degree: **Master of Resource Management**

Title of Thesis: **An Evaluation of Criteria for Assessing Conservation Status of Fraser River Sockeye Salmon Conservation Units**

Project Number: **473**

Examining Committee:

Chair: **Cameron MacKenzie**
Master of Resource Management Candidate

____
Dr. Randall M. Peterman
Senior Supervisor
Professor, School of Resource and Environmental Management

____
Dr. Nicholas K. Dulvy
Supervisor
Associate Professor, Department of Biological Sciences

____
Dr. James Irvine
Supervisor
Research Scientist, Fisheries and Oceans Canada

____
Dr. Andrew Cooper
Supervisor
Associate Professor, School of Resource and Environmental Management

Date Defended/Approved: ______________________________________

ABSTRACT

Fisheries management agencies and conservation organizations use various criteria to determine whether fish populations are a conservation concern. By conducting retrospective analyses using historical data for 18 Fraser River (B.C.) sockeye salmon conservation units (CUs), we evaluated the effectiveness of 20 criteria that measure time trends in spawner abundance to determine how likely they are to correctly categorize conservation status of salmon populations. We used a Receiver Operating Characteristic (ROC) approach to quantify the probability of each criterion correctly distinguishing between a declining and non-declining CU. Those criteria that measured the extent of decline from an estimated historical baseline were most reliable and consistently outperformed the widely used International Union for the Conservation of Nature (IUCN) criterion of percent decline in abundance over the most recent three generations. We therefore urge scientists to evaluate the statistical performance of their criteria for classifying conservation status before applying them in decision making.

Keywords: Fraser River sockeye, Receiver Operating Characteristic curves, Wild Salmon Policy, status assessment

ACKNOWLEDGEMENTS

I thank Randall Peterman for his support, expertise, advice, and enthusiasm. I also thank the other members of my committee: Jim Irvine, Nick Dulvy, and Andy Cooper. I greatly appreciate the time, guidance, and expertise each of them has afforded me throughout this project. Special thanks to Carrie Holt from Fisheries and Oceans Canada, for her thoughtful input, reviews, and encouragement. For assistance in providing data and advice, I thank Tracy Cone from Fisheries and Oceans Canada. I would also like to thank the Fisheries Science and Management Research Group at Simon Fraser University, for their support and encouragement.

This research was funded by an NSERC individual research grant (via Randall Peterman), Fisheries and Oceans Canada (via Jim Irvine), a MITACS Accelerate Internship, and a Faculty of Applied Sciences Graduate Fellowship awarded to Erin Porszt.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Numerous organizations worldwide determine the biological status of species for conservation, management, or socio-economic purposes. These organizations include the International Union for the Conservation of Nature (IUCN), the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), the United States (U.S.) National Marine Fisheries Service (NMFS) and the U.S. Fish and Wildlife Service through the U.S. Endangered Species Act, and the Committee on the Status of Endangered Wildlife in Canada (COSEWIC). The criteria used to assess status may differ among species and may vary among organizations for the same species. There is therefore a need to evaluate the performance of the range of conservation criteria used to assess status. We evaluated a set of criteria for their reliability as indicators of declining populations of sockeye salmon (*Oncorhynchus nerka*) in the Fraser River, British Columbia (B.C.), Canada. This research may provide Canada's Department of Fisheries and Oceans (DFO) and others (e.g., the Pacific Salmon Commission) with information regarding which criteria most reliably distinguish between declining and non-declining populations. The results should contribute to successful management and conservation.

Wild Salmon Policy

Due to the high economic and cultural value of Pacific salmon (*Oncorhynchus* spp.), the conservation of these fishes is of utmost importance to

Canadians. The goal of Canada's policy for conservation of wild Pacific salmon ("Wild Salmon Policy", or WSP), developed by the DFO, is to "restore and maintain healthy and diverse salmon populations and their habitats for the benefit and enjoyment of the people of Canada in perpetuity" (DFO 2005). This overarching goal has three specific objectives: (1) safeguard the genetic diversity of wild Pacific salmon, (2) maintain habitat and ecosystem integrity, and (3) manage fisheries for sustainable benefits (DFO 2005).

The implementation of the WSP will occur through the adoption of six strategies, and the first strategy is standardized monitoring of wild salmon status (DFO 2005). The first action step of this strategy is to identify conservation units (CUs), which are biological groupings described by areal extent of freshwater distribution, for all species of wild Pacific salmon (DFO 2005). These CUs have now been identified (Holtby and Ciruna 2007). A CU is "a group of wild salmon sufficiently isolated from other groups that, if extirpated, is very unlikely to recolonize naturally within an acceptable time frame, such as a human lifetime" (DFO 2005). DFO is in the process of shifting its management and assessment of wild salmon to be based on these CUs in order to maintain the geographic and genetic diversity of the species.

The second action step of the first strategy of the WSP is to develop criteria to assess (i.e., estimate the status of) CUs and identify benchmarks to indicate biological status in terms of whether the CU is a conservation concern (DFO 2005). DFO suggests four classes of possible indicators of biological status: (1) spawner abundance, (2) temporal trends in spawners, (3) spatial distribution (e.g., distribution of spawners across spawning sites and habitat types), and (4) fishing mortality (Holt et al. 2008). The criteria will consist of quantifiable measures of these indicators of biological status, with benchmark levels of those criteria used to describe CU status. Each criterion will have an upper and a lower benchmark. The upper benchmark is where “there would not be a high probability of losing (i.e., extirpation of) the CU”, and the lower benchmark is “at a level of abundance high enough to ensure there is a substantial buffer between it and any level of abundance that could lead to a CU being considered at risk of extinction by COSEWIC" (DFO 2005).

COSEWIC

COSEWIC is an independent committee that uses scientific decision-making criteria and expert opinions to classify Canadian species into categories of threat (COSEWIC 2006). The basis for these criteria is the revised IUCN Red List and Categories, which is a widely recognized system for determining the threat status of populations and species (Mace et al. 2008). The IUCN has three categories of increasing threat into which it classifies species (Vulnerable, Endangered, and Critically Endangered) (IUCN 2001). COSEWIC’s analogous categories of increasing threat are Special Concern, Threatened, and Endangered (COSEWIC 2006).

COSEWIC acts as an advisory organization that undertakes biological assessments to recommend which populations and species are a sufficient conservation concern that they should be considered for protection under Canada’s Species at Risk Act (SARA), which was proclaimed in 2003. On the basis of COSEWIC's recommendations and a consideration of "socioeconomic implications", the Federal Cabinet decides which species to add to the legal list of species at risk. Such a listing subsequently mandates legal protection, and recovery planning and implementation (Irvine et al. 2005). At present, COSEWIC considers four populations of Canadian Pacific salmon to be "at risk of extinction" (COSEWIC 2008), although none of these populations are on the legal list of species at risk under SARA. The Okanagan population of chinook salmon (*O. tshawytscha*) is categorized as threatened, and the Sakinaw Lake population of sockeye salmon, the Cultus Lake population of sockeye salmon, and the Interior Fraser population of coho salmon (*O. kisutch*) are all categorized as endangered by COSEWIC (the highest level of concern).

To determine which category of conservation status should be assigned to a species or population unit, COSEWIC uses five criteria to infer chance of extinction. When the conditions for any one of the five criteria have been met, this can result in the classification of the population or species as a conservation concern. These criteria include criterion A, "Declining Total Population", which measures the change in adult population size over the past 10 years or three generations, whichever is longer (COSEWIC 2006). COSEWIC's other four criteria consider absolute abundance and spatial distribution (COSEWIC 2006). COSEWIC and the IUCN often use criterion A when assessing the threat status of marine and anadromous fish species (e.g., COSEWIC 2003a, COSEWIC 2003b, Rand 2008). This frequency of use may be because criterion A is applicable in the all-too-frequent cases where the absolute abundance and/or distribution of the species are unknown. Absolute abundance is unknown for many CUs of Pacific salmon because some methods used to estimate number of spawners provide only indices of relative abundance (Tracy Cone, DFO, 100 Annacis Parkway, Unit 3, Delta, B.C., V3M 6A2, pers. comm.)

Categorising threat

Despite the frequency of its use, there is debate surrounding the validity of the decline criterion, COSEWIC and IUCN criterion A, for assessing the "extinction-risk" of marine fishes and other species. Two studies maintain that the decline criteria may be overly conservative biologically and incorrectly classify sustainably exploited species as being of conservation concern (Matsuda et al. 1998, Punt 2000). One of these studies used southern bluefin tuna data (Matsuda et al. 1998), while the other used a simulation model for six marine fish species (Punt 2000). In contrast, other analyses using empirical and meta-analytical approaches support the validity of these decline criteria in representing marine fish status (Hutchings 2001, Dulvy et al. 2005). More generally, controversy surrounds the entire process of estimating and classifying organisms into categories based on their chance of extinction, in part due to considerable variation in the precision and defensibility of available risk assessment methods (Dulvy et al. 2004). For instance, there can be slight variations among scientists in the application of the same conservation criterion, and controversy can arise as these scientists defend their case-specific assumptions and methods. To complicate matters, we have limited experience to date with quasi-extinctions (extremely low abundances), let alone extinctions, in the marine environment

(Powles et al. 2000). Therefore, doubt exists as to whether criteria developed for assessing the chance of quasi-extinction for terrestrial species are applicable to marine species (Powles et al. 2000).

There is a common perception, although not supported by empirical evidence, that marine fishes have low vulnerability to quasi-extinction due to their life-history strategies, and thus there is little need to assess their chance of quasi-extinction. The perception is that marine fishes have a high potential for recovery from low abundance due to their high productivity and highly variable populations. However, there is little empirical or theoretical evidence demonstrating that highly fecund species have a chance of extinction any lower than those of low fecundity due in part to the different life-history strategies (e.g., egg-survival rates) that are exhibited by these two groups of species (Sadovy 2001). Thus, fecundity alone is not an appropriate indicator of vulnerability (Sadovy 2001). Empirical evidence also suggests that fish populations do not fluctuate more than populations of other species such as mammals, birds, and butterflies, and that they may exhibit similar vulnerability to quasi-extinction as other organisms (Dulvy et al. 2003). This vulnerability to quasi-extinction thus necessitates the assessment of chance of quasi-extinction for marine fish species.

The controversy surrounding estimation of chance of extinction is especially large for commercially exploited species. A common argument is that commercially exploited species will undergo economic extinction before biological quasi-extinction. However, this may not be the case for non-target species that are caught as by-catch in multispecies fisheries, species with schooling behaviour, or species with high commercial value such as Pacific salmon, especially if this value increases when the species becomes rare (Dulvy et al*.* 2003).

The above debates about methods for estimating the level of conservation concern of a population strongly suggest the need to carefully select indicators of that status and to test the effectiveness/accuracy of these indicators. Therefore, our research objective was to estimate the reliability of various measures of conservation status for Pacific salmon. We begin by reviewing the types of criteria currently being used and then discuss ways of evaluating their reliability.

Possible criteria of spawner abundance trends

Different case examples use or suggest various methods for assessing biological status of fishes based on trends in abundance. For instance, Musick (1999) examined the problem of which quantitative criteria that are related to population decline best reflect the chance of quasi-extinction for marine fishes. That paper resulted in the development of a separate set of criteria, now used by the American Fisheries Society (AFS), with the goal of better reflecting population resilience than the criteria used by IUCN. In the AFS system, classification of populations as vulnerable or otherwise depends on a two-tiered system, which first assigns populations into one of four categories of productivity and then compares the three-generation change in abundance to a corresponding threshold. Another study evaluated the performance of AFS criteria with respect to limit reference points for recruitment overfishing across 76

European exploited stocks (Dulvy et al. 2005). They found that the AFS criteria have a higher probability than IUCN decline criteria of overlooking species that are exploited unsustainably based on limit reference points. Such examples are referred to as being false negatives or type II errors because the null hypothesis of no decrease in abundance is incorrectly *not* rejected.

The DFO is considering using temporal trends in spawner abundances as one possible indicator of the biological status of Pacific salmon populations (Holt et al. 2008). The most frequently used method to examine these trends is to measure the rate of change in spawner abundance over the most recent three generations (COSEWIC 2003a, COSEWIC 2003b, Rand 2008). For instance, the COSEWIC status reports on Cultus Lake and Sakinaw Lake populations of sockeye salmon (COSEWIC 2003a, COSEWIC 2003b) considered time trends in spawner abundances (as well as other information) when determining the populations' status. Pacific salmon populations are highly variable, and some of this apparent variability may be attributed to measurement error (Paulsen et al. 2007). Thus, to remove some of the annual "noise" that occurs on top of an underlying trend in population abundance, spawner abundance data are often smoothed using a running mean over a complete generation of $\log_e$ spawner abundances (4 years for Fraser River sockeye salmon, for instance). The COSEWIC guidelines deem the (negative) slope of a straight line fitted by regression to a smoothed time series across three sockeye salmon generations (12 years) to be the best estimate of a constant rate of decline caused by an underlying threatening process such as fishing, loss of habitat, change in predators, food, etc. (COSEWIC 2003b). COSEWIC uses this method to facilitate comparison with the threshold rates of decline (30%, 50%, and 80%) that trigger designation into one of their three categories of threat -- special concern, threatened, or endangered (COSEWIC 2003b).

The authors (COSEWIC 2003b) also examined an alternative procedure that is based on the reduction in abundance between only the first and last year of the three-generation period, as opposed to the annual rate of decline estimated by slope of the regression line fit to the data over the period. However, the authors concluded that limiting the measure of decline over the period to only the percentage decrease between the single data point abundances at the start and end of the period would cause estimates of decline to be greatly affected by annual fluctuations in individual year class strength, and thus sensitive to the particular two years chosen for comparison.

To estimate the three-generation (12-year) change in abundance at monitoring sites within subpopulations of sockeye salmon, the recent IUCN Red List Assessment for Sockeye Salmon used the same method as the COSEWIC assessments (Rand 2008). That assessment used $\log_e$-transformed number of spawners (escapement) data smoothed with 4-year running means, and estimated the annual rate of change over the most recent three generations using the slope of the best-fit line.

To rank methods for assessing the conservation status of Fraser River sockeye, Pestal and Cass (2009) suggested qualitative risk evaluations that include not only looking at trends in abundance over the last three generations but also comparing the most recent abundance (mean spawner abundance over last 4 years) to various benchmarks. These benchmarks consist of the long-term average abundance (overall geometric mean of spawner abundance), the largest observed abundance (highest geometric mean of any 10-year period), the current capacity (estimate of spawner abundance that maximizes sockeye smolt abundance, i.e., juvenile salmon at the time of migration to sea), the potential capacity (yet to be defined by the authors), and capacity indicated by traditional ecological knowledge (yet to be defined).

Another status criterion used is the historical extent of decline, which has been recommended as the ultimate criterion for triggering concern about the long-term viability of a species, with the historical extent of decline examined over as long of a period as possible to enable a meaningful baseline (Mace et al. 2002). These authors also suggested that recent rates of decrease be used in conjunction with historical extents of decline because either indicator alone does not provide enough information about the status of the population (Mace et al. 2002).

The methods to assess status suggested by Mace et al. (2002), and Pestal and Cass (2009) both consider some form of extent of historical decline. Using this extent of decrease from some historical baseline, as opposed to the rate of decrease over the most recent three generations, is advantageous because it minimizes the influence of natural variability and avoids bias introduced by shifting baselines (Pauly 1995) that occur during the survey period

(Dulvy et al. 2006). However, exactly which criteria are the most reliable indicators of biological status for CUs remains a contentious issue.

Therefore, one of the research objectives of this work was to evaluate the reliability of various quantitative criteria for estimating time trends in spawner abundance as indicators of true declining Pacific salmon CUs. We evaluated criteria that were not only based on the recent rate of decline, but also the extent of decline from an estimated historical baseline, with this historical baseline estimated in five different ways, as described later.

Methods to evaluate criteria of biological status

Several procedures have been used to determine the effectiveness of various status-assessment criteria as indicators of appropriate management actions to be taken. For instance, the performance of limit reference points that are often the basis for management recommendations has been evaluated for the North Sea (Piet and Rice 2004). Limit reference points for spawning stock biomass (SSB) and the corresponding fishing mortality define boundaries between being within or outside “safe biological limits”. An unsustainably exploited population falls outside “safe biological limits”, which can, but not always, lead to a recommendation to reduce total allowable catch (TAC). In contrast, a sustainably exploited population categorised as being “within safe biological limits” usually leads to a recommendation of status quo or increased TAC. In the cases reviewed by Piet and Rice (2004), management advice in a given year depended on estimates of whether the species was within “safe biological limits” at that time, however, the estimates of SSB and fishing mortality in the assessment year were highly uncertain due to the ‘tapering’ effect inherent in Virtual Population Analyses. This effect is where the addition of more years of data leads to retrospective estimates of SSB and fishing mortality becoming more consistent. For each stock, Piet and Rice (2004) compared the management action in a given year with the current (relatively uncertain) estimate of the status of the stock at that time in terms of being within or outside the “safe biological limits”. Ideally, if the stock was outside "safe biological limits" in a given year, the management action would have been a reduction in TAC, whereas if the stock was "inside safe biological limits", the action would have been status quo or an increased TAC. This comparison resulted in a proportion of stocks that had true positives and true negatives (collectively called hits or correct management advice), false positives (false alarms, incorrect management advice to reduce catch), and false negatives (misses, another type of incorrect management advice resulting in catches that were too high) for each year. Piet and Rice’s (2004) analysis then evaluated the currently used reference points used to determine “safe biological limits” based on the proportion of hits, misses and false alarms. Our research also tallies the proportion of hits (both true positives and true negatives), false alarms (false positives), and misses (false negatives) in an attempt to evaluate the reliability of various criteria as indicators of status of Fraser River sockeye CUs, which may affect management actions.

Another study tested for consistency between the population status contained in fisheries stock-status assessments and threat-status assessments

(Dulvy et al. 2005). For North-East Atlantic marine populations, the authors evaluated threat status by comparing estimated rates of decline in abundance (IUCN and AFS criteria) with exploitation status, which is based on whether fished species are within or outside “safe biological limits”. Their results showed that the conservation status assigned by the decline criteria was consistent with the exploitation status when applied to exploited marine stocks, based on the resulting proportion of hits, misses and false alarms. Most importantly, the study found no evidence of false alarms (false positives), which counters concerns of a “collision of control rules” whereby conservation-oriented decline criteria may be triggered while stocks are within safe biological limits. This finding differs from previous analyses that suggested decline criteria may be prone to false alarms when applied to exploited marine species (Matsuda et al. 1998 and Punt 2000). The difference in outcomes between these studies may have resulted from IUCN’s increase of the decline thresholds of criterion A. For instance, in 2001, IUCN changed the threshold of decline over the greater of 10 years or three-generation spans that would trigger a classification of "vulnerable" from 20% to 50%, assuming the causes of the decline are reversible and understood and have ceased (Dulvy et al. 2005). In other words, a population needed to have a greater rate of decline than previously to be considered "vulnerable", "threatened", or "endangered".

Another comparative analysis evaluated decline criteria by asking not whether decline criteria are consistent with fisheries stock status assessments, but instead whether threat assessments are a reasonable predictor of future population trajectory (Rice and Legace 2007). Using historical data, the analysis assessed the probability that a stock that was flagged as a conservation concern by the triggering of a decline criterion was on a trajectory to extinction (Rice and Legace 2007). For years subsequent to the year in which the decline criterion was triggered, Rice and Legace (2007) calculated the probability of observing estimates of SSB at least as large as those estimated in the assessments if the stock was truly on a trajectory to extinction. A true positive was a triggering of the decline criterion when subsequent SSB estimates indicated that there was a high probability the stock was on or below the trajectory to extinction. A "false alarm" (i.e., false positive) was a triggering of the decline criterion when subsequent SSB estimates indicated that the stock had a low probability of declining to extinction. The ability of the decline criterion to predict the future population trajectory (i.e., proportion of true and false positives) was highly dependent on the chosen starting point of this future trajectory period (Rice and Legace 2007).

The above studies demonstrate the utility of estimating rates (probabilities) of occurrence of true and false positives as well as true and false negatives as a method of evaluating appropriateness of the various criteria used for assigning conservation status. Receiver Operating Characteristic (ROC) analysis, described below, is a method that can take these studies a step further by combining these four probability components into a single measure, as we explain later. ROC analysis is also useful in another way. Previous evaluations of conservation criteria have been limited to only a few thresholds at a time, where the threshold is the level that results in a change of classification. For instance, when evaluating IUCN's criterion A, which measures the rate of decline over the most recent three generations, the threshold would be the percent decline that results in a classification of vulnerable, endangered, or critically endangered (i.e., a decline greater than 50%, 70%, or 90%, respectively, over the three generations, assuming the causes of the decline are reversible, understood and have ceased). In contrast, ROC analysis estimates the true and false positive rates of a conservation criterion over a wide range of thresholds, thus providing more insight into the overall reliability of the criterion.

Determination of the reliability of tests, criteria, or models is an important aspect of many scientific fields. ROC analysis is a method used regularly in medical science and less frequently in ecology to assess the reliability of various tests and models. ROC analysis originated in the signal detection field in World War II to describe the effectiveness of radar receivers at distinguishing “noise” from “signal plus noise” (Vida 1993). In the medical sciences, ROC methods compare the reliability of diagnostic tests at detecting a given disorder (i.e., ability of the test to distinguish correctly between a patient with a given disorder and one without the disorder) (Vida 1993, Hanley and McNeil 1982, and Hibberd and Cooper 2008). In ecology, some authors have used ROC methods to compare the effectiveness of models at predicting species’ spatial distribution from occurrence data (i.e., ability of models to predict correctly whether unmonitored areas are inhabited by a species) (Elith et al. 2006, Fielding and Bell 1997, Pearson 2007). This ROC concept also has potential for application in fish conservation and management. For instance, it can be used to evaluate the reliability of various criteria, such as measures of time trends in salmon spawner abundance, at correctly distinguishing CUs that are of conservation concern from those CUs that are not. We therefore apply the ROC method here, the first time (to our knowledge) that it has been used in fish conservation and management.

ROC analysis evaluates the performance of tests that have four possible outcomes, depending on the true state of nature: (1) true positive, (2) false positive, (3) true negative, and (4) false negative. These outcomes correspond, respectively, to the four possible outcomes for a statistical test of some null hypothesis, depending on the true state of nature: (1) correctly rejecting the null hypothesis of no effect, (2) type I error (rejecting the null hypothesis when it should not have been), (3) correctly not rejecting the null hypothesis, and (4) type II error (not rejecting the null hypothesis when it should have been) (Peterman 1990). Power ($1-\beta$) is the probability of occurrence of a true positive (correctly rejecting the null hypothesis), alpha (α) is the probability of occurrence of a false positive (type I error), ($1-\alpha$) is the probability of occurrence of a true negative (correctly not rejecting the null hypothesis), and β is the probability of occurrence of a false negative (type II error), (Dixon and Massey 1969).

In the context of evaluating criteria as indicators of CU status, Table 1 illustrates the definition of a true positive, false positive, true negative, and false negative. Conservation criteria estimate the status of a CU as either declining or not declining, whereas the true status of a CU depends on its spawner abundance in years subsequent to the analysis. A true positive arises in a year in which the criterion correctly signals a decline, whereas a false positive (type I error) arises in a year in which the criterion incorrectly signals a decline. A true negative arises when the criterion correctly does not signal a decline, whereas a false negative (type II error) arises when the criterion incorrectly fails to signal a decline. The true positive rate (power, 1-β) is the probability of the conservation criterion signalling a decline out of all years where the population is actually declining. The false positive rate (α) is the probability of the conservation criterion signalling a decline out of all years where the population is not declining. The true positive rate and the false positive rate account for all four elements of Table 1 because their rates are the complement of their corresponding false negative and true negative rates, respectively. Clearly, based on the Wild Salmon Policy objective to avoid having CUs that are considered “at risk of extinction” by COSEWIC, we want criteria for assessing status of salmon populations to have high true positive (low false negative) rates and low false positive (high true negative) rates.

ROC analysis provides a way to combine these true positive rates and false positive rates into a concise measure of reliability for a given criterion. ROC analysis generally uses two key descriptors of test behaviour that are called “sensitivity”, which is equivalent to the true positive rate, and “specificity”, which is the true negative rate (1-false positive rate or 1-α) (Vida 1993). For example, in medical sciences, “sensitivity” is defined as how good a diagnostic test is at correctly identifying subjects who have a given disorder (true positive rate or power), whereas “specificity” is defined as how good the test is at correctly identifying subjects who do not have the disorder (true negative rate or 1-false positive rate) (Hibberd and Cooper 2008). An ROC curve is composed of points that are the true positive rate (power, 1-β; Y-axis) and the false positive rate (α; X-axis) values for a given criterion at each possible threshold, i.e., level of the given criterion that results in a change of classification category. For instance, a threshold of 20% could be evaluated as the percentage change in abundance over a period that results in the classification of a population as a conservation concern. Other thresholds of 5%, 10%, 30%, 40%, 50%, etc. would also be used, with each case generating a point on the ROC curve.

The area under the resulting ROC curve (AUC) provides a single measure that summarizes performance across the full range of thresholds (Pearson 2007). If the evaluation of a criterion is limited to only one threshold that results in a change of classification, this assessment does not take into account all of the information provided by that criterion because the true positive rate and false positive rate of the criterion will vary among thresholds. For instance, ecologists know very well that if a threshold for an estimated effect size of concern is set too low, the false positive rate will be too high. In contrast, the ROC curve approach, and the AUC measure in particular, reflect the performance of a criterion of conservation concern across all thresholds that might be used by someone. The AUC can vary between 0 and 1. A criterion that perfectly indicates the true status of a population generates an ROC curve that falls along the left axis and across the top of the plot (AUC=1), whereas a criterion that is no better than a random coin toss at correctly categorizing the status of a population generates an ROC curve that closely falls on the 1:1, or 45° line (AUC=0.5) (Pearson 2007). Figure 1 shows two hypothetical ROC curves. One curve has a high AUC (0.94), with points in the upper-left corner of the plot. The other curve has an AUC of 0.5 and points that roughly follow the line of equality where the false positive rate equals the true positive rate.

The AUC measures the probability that a criterion will distinguish correctly between cases that differ in their true status. In medical science, AUC is the probability of a test correctly distinguishing between a randomly chosen diseased and non-diseased patient. Medical science and fisheries management may have similar objectives, i.e., to maintain a low chance of death or extinction. In fisheries management applications, AUC gives the probability that the criterion will correctly distinguish between a year when a population has a true status of declining and a year when it has a true status of not declining. Therefore, when comparing two criteria of conservation concern such as one measuring the rate of decline over three generations and another measuring rate of decline over all years, the one with a larger AUC is a better indicator of CU status because it has a higher probability of correctly distinguishing between declining populations and those that are not actually declining. This ROC approach holds substantial promise for evaluating criteria that assess status in Pacific salmon and other fishes.

Case study: Fraser River sockeye salmon

To demonstrate the applicability of the ROC method, we used data on 18 conservation units identified by DFO for Fraser River, B.C. sockeye salmon. The Fraser River watershed includes both a highly urbanized region in southwestern B.C., and a less disturbed section in central and eastern B.C. The Fraser River has generally been the largest and most productive salmon-producing watershed in B.C. and supports the largest salmon fishery in Canada (Northcote and Larkin 1989). Fraser River sockeye pass through marine, estuarine, and freshwater environments during their return migration to spawning grounds, and fisheries take place on these migratory adults as they return (English et al. 2005). Sockeye salmon enter the Fraser River to spawn from June through October and each stock has adapted to normally enter the river at a particular time (English et al. 2005). For management purposes, Fraser sockeye are categorized into four groups based on their adult migratory timing patterns or runs (1) Early Stuart River (late June to late July), (2) Early Summer (mid-July to mid-August), (3) Summer (mid-July to early September), and (4) Late Summer (early September to mid-October) (English et al. 2005).

Sockeye salmon exhibit substantial life history variations. Lake-type sockeye typically spawn in lakes, or in tributaries associated with lakes, with the offspring rearing in these nursery lakes for at least 1 year before migrating to the ocean (Burgner 1991). This is generally the most widespread and abundant life-history type in B.C. (Beacham et al. 2004), and there are approximately 30 CUs of lake-type sockeye salmon found in the Fraser River watershed (DFO 2008).

Sockeye salmon in the Fraser River also exhibit a river-type life history, where they spawn in tributaries and mainstem side channels, and juveniles rear in rivers rather than lakes before migrating to the ocean (Beacham et al. 2004). There are approximately seven CUs of river-type sockeye salmon found in the Fraser River watershed (DFO 2008).

The vast majority of Fraser River sockeye (~ 94%) mature and die at age 4 years (Schnute et al. 2000). This 4-year life cycle results in the stocks having four distinct cycle lines (brood lines) associated with particular years of return, with little gene exchange among them (Schnute et al. 2000, Ricker 1997). Many Fraser sockeye stocks exhibit cyclic dominance, in which the stock abundance in one cycle often substantially dominates the abundance in other cycles, with the persistently high-abundance cycle line referred to as the "dominant" cycle for the stock (Schnute et al. 2000). In such populations, a "sub-dominant" line and two weak lines ("off-cycle") are often present, with the sub-dominant line being 10-25% as large as the dominant one, and the weak lines less than 1% as large as the dominant line (Ricker 1997).

The methods used to estimate sockeye salmon spawner abundance within the Fraser River watershed vary depending on the anticipated size of the run (Eggers and Irvine 2007). For most of the Fraser River sockeye populations, visual surveys usually generate estimates when anticipated spawner abundances are less than 25,000, and counting fences (weirs) and mark-recapture generate estimates when anticipated spawner abundances exceed 25,000 (Eggers and Irvine 2007). In general, fence counts or mark-recapture methods provide absolute abundance estimates with higher accuracy and precision than other methods (Tracy Cone, DFO, pers. comm.).

Fraser River sockeye CUs also exhibit a wide range of abundance trends over time and qualities of spawner abundance estimates. This diversity of cases within the Fraser watershed makes the region ideal for evaluating the reliability of various criteria for measuring time trends in abundance and the robustness of the criteria across a wide range of conditions.

Research objective

Therefore, we used sockeye salmon CUs in the Fraser River watershed to evaluate the reliability of criteria that estimate time trends in spawner abundance at classifying the conservation status of those CUs. Our research objective was to determine which quantitative criteria that describe time trends in abundance are the most reliable indicators of biological status for Pacific salmon CUs. Specifically, we aimed to determine which decline criteria most frequently correctly indicate whether CUs are truly declining or not (i.e., have a downward future trajectory or not).

METHODS

We conducted retrospective analyses to evaluate 20 criteria that estimate time trends in spawner abundance to assess the conservation status of Fraser River sockeye. We did this by using the criteria to estimate the decline status of 18 CUs in each year of historical data and comparing that estimated status to the true status of the CUs, which was based on the trajectory of spawner abundance that was observed subsequent to when the status was determined. These retrospective analyses used conservation criteria to determine the status that would have been estimated for a given CU in each year included in the historical dataset and compared that status to the true status in the same year. The 20 conservation criteria evaluated here resulted from a combination of possible spawner abundance estimates, periods of decline, and historical baselines, as elaborated below. Receiver Operating Characteristic (ROC) curves combined the true positive rate and the false positive rate of each conservation criterion across a range of thresholds, and the area under the ROC curve is the probability that a criterion will correctly distinguish between CUs in which abundance is declining from those that are not declining.

We empirically compared the performance of various decline criteria by analyzing spawner abundance data (starting as early as 1938 for some CUs through to 2007) for 18 CUs of sockeye salmon, which cover all four run-timing groups that occur in the Fraser River watershed (Table 2). All spawner abundances (S) were transformed with $\log_e(S+1.01)$ prior to analysis, except for calculating the geometric mean of a generation or the few cases where we compared raw spawner abundances to a geometric mean (Gotelli and Ellison 2004).

Conservation criteria

The decline criteria (variously known as "threat criteria", "conservation-status criteria", or "conservation criteria") evaluated here were variants of COSEWIC's criterion A, "Declining Total Population", which measures the change in adult population size over the past 10 years or three generations, whichever is longer. The simplicity of COSEWIC's definition belies a range of ways to interpret and apply criterion A. Specifically, the conservation-status criteria evaluated here consist of various combinations of type of spawner abundance estimates, periods over which the decline was estimated, and historical baselines used for comparison, if any.

We considered five spawner abundance estimates: (1) untransformed (raw) spawner abundance, (2) spawner abundance estimated by the linear time-trend model fit to unsmoothed data, (3) spawner abundance estimated by the linear time-trend model fit to smoothed (running generational (4-year) mean) data, (4) geometric mean of observed abundance of a generation using unsmoothed data, or (5) geometric mean of observed abundance of a generation using smoothed data. The linear-trend model used robust regression to downweight the influence of outliers on estimates of rate or extent of change

(Venables and Ripley 2002). Least squares regression assumes that errors follow a normal distribution and minimizes the squared residuals. In contrast, robust regression is useful when the assumption of normality of residuals is not met, say for cases with highly variable data and outliers, because it minimizes the absolute value of residuals (Gotelli and Ellison 2004). Regressions were performed over periods with at least ten data points.

Several CUs have years with missing data because it is difficult to survey every site every year. This is most problematic when smoothing the data across each non-overlapping 4-year period, because a missing data point for a dominant cycle year could dramatically reduce the generational mean (i.e., smoothed estimate) that includes that year, and a missing data point for a small-run, off-cycle year would have the opposite effect. Therefore, for all criteria that used data smoothed with a running mean, we inserted abundance estimates for any missing years by interpolating a value from the mean of the same cycle year from the immediately previous and subsequent generations (i.e., only the single generation before and the single generation after the missing point). For example, if abundance was missing for a CU in 1990, we interpolated the missing value with the mean of abundances in 1986 and 1994. We inserted an interpolated value for no more than one year per generation cycle (i.e., specific 4-year period), and if the corresponding cycle year of either of the closest two generations was missing (e.g., 4 years previous or subsequent to missing point), we used the corresponding cycle year no more than two generations away to calculate the mean used in the interpolation. The analysis excluded years where a CU had insufficient data to allow for such interpolations.

The change in spawner abundance was estimated over two time periods. Over the short term, it was estimated as the recent rate of decline over the last three generations. Over the long term, the change in abundance was estimated as either the annual rate of change over all available years, or as the extent of decline from an estimated historical baseline. Both rates of decline and extent of decline were expressed as annual rate (%) of change per year. Historical baselines were identified for each CU. However, for cases where the current abundance was not compared to a historical baseline, we calculated the change in abundance between spawner abundance estimates in the same category of brood year (i.e., dominant, sub-dominant, or off cycle years), to account for the CUs' four distinct cycle lines.

There are several ways to define a historic baseline of spawner abundance. We considered five definitions, and specific values of these baselines varied among CUs (depending on historical spawner abundance levels): (1) the first abundance data point in the time series, (2) the maximum abundance of the first five points in the time series (regardless of cycle year) (3) the geometric mean abundance of the first generation, (4) the maximum abundance of all points in the time series (regardless of cycle year), and (5) the maximum geometric mean abundance of any three-generation (12-year) period. The extent of decline was measured between each of these historical baseline abundances and abundance in subsequent years. For baselines based on maximum abundance (maximum of all points and maximum three-generation period), the extent of decline was calculated fewer times because for some CUs, the maximum abundance fell in the mid-to-late years of the time series.

All combinations of these types of spawner abundance estimates, periods of decline, historical baselines, and measures of decline created 20 conservation-status criteria for our initial evaluation of their effectiveness as indicators of declining CUs (i.e., CUs with decreasing abundance of sockeye salmon spawners) (Table 3). The Appendix provides a detailed description of each of these 20 criteria evaluated in the analysis.

Some of the 20 conservation-status criteria are only slight variations of other criteria. For instance, some criteria (criteria 10, 11, 15, and 16) correspond closely with other criteria (criteria 12, 13, 17, and 18, respectively); the only difference is that the former criteria compare a sliding window of generations to the historical baseline, whereas the latter criteria compare non-overlapping 4-year generational blocks to the historical baseline. The former method allows for an assessment of status every year, whereas the latter method only does so every 4 years.

Estimated and true conservation status of CUs

For each year and CU, we compared the status that was estimated by each conservation criterion to the true status; the latter was based on the future trajectory of spawner abundance in that CU using the data subsequent to each year when status was estimated. For each of the 20 decline criteria, we estimated the number of triggering events (criterion conditions met, as defined below) that would have occurred in the past for each of the 18 CUs. A triggering event occurred when the particular criterion estimated the status of the CU to be declining, and a non-triggering event occurred when the criterion estimated the status of the CU to be not declining.

To determine how well a particular criterion works at identifying the actual subsequent time trend in abundance of a conservation unit, we examined the data to determine the true status in each year, which was based on abundance in years subsequent to the criterion's categorization. Specifically, for each year we defined the true status of CUs by using one of two metrics that describe the change in spawner abundance over a future period; "future" was either the subsequent three generations or all remaining years in the time series (with a minimum of 10 years remaining). For both metrics, the decline status was based on the spawner-to-spawner (SS) ratio, which is the ratio of the estimated number of spawners at the end of the future period to the estimated number of spawners in the year of analysis. We calculated this SS ratio as the change in abundance between best-fit estimates of spawner abundance from the robust regression of abundance on years. In a given year, we categorised the true status of a CU as "declining" if the SS ratio was less than or equal to the ratio that corresponds to a given threshold of percent decline (e.g., a 30% decline would be equivalent to a SS ratio of 0.7, or (100-threshold) / 100). Such cases would indicate a percent decrease in abundance greater than or equal to the threshold level over the future period. The true observed amount of decrease in the future period was compared to thresholds of 30%, 50%, 70% or 90% decline. The percent decline over the future period is equivalent to an effect size in statistical power analysis, where the effect size is the magnitude of the true effect that you are trying to detect (Peterman 1990). Here, we are evaluating the ability of various conservation criteria to detect a future decline in abundance of a given percentage (i.e., effect size). Table 4 lists the eight scenarios or different definitions of a CU with a true status of declining that resulted from variations in future period and percent decrease over that period.

We evaluated the performance of each criterion across all CUs by comparing the criterion's estimated status in a given year with the true status in the same year as defined above. That comparison was based on the proportion of years that could be categorised as "true positives", "true negatives', "false positives (type 1 error)", or "false negatives (type 2 error)" (Table 1). A true positive was a triggering event (i.e., conditions of conservation-status criterion were met, such as the population decreased by more than 30% over three generations) in a year where the CU had a true status of declining, and a true negative was a non-triggering event in a year where the population had a true status of not declining. A false positive (type I error) was a triggering event in a year where the CU had a true status of not declining, and a false negative (type 2 error) was a non-triggering event in a year where the CU had a true status of declining.

Evaluation of conservation criteria

We used two quantities, the true positive rate (power, 1-β, 1-false negative rate) and the false positive rate (α, 1-true negative rate), to evaluate the conservation-status criteria. The true positive rate of a criterion is its effectiveness at (i.e., probability of) correctly identifying years when a CU is truly going to decline subsequently. We calculated the true positive rate as the number of true positives out of all years in which the CUs have a true status of declining (Vida 1993):

(1) True Positive Rate = True Positives / (True Positives + False Negatives).

We calculated the false positive rate or the probability of a type 1 error, as the number of false positives out of all years in which the CUs have a true status of not declining (Vida 1993):

(2) False Positive Rate = False Positives / (False Positives + True Negatives).

At each threshold (e.g., the percent change in abundance over a period that results in the classification of a population as declining), we calculated one true positive rate and one false positive rate for each criterion across all 18 CUs and all years based on the number of true positives, true negatives, false positives, and false negatives.

Recall that the area under ROC curves (AUC) concisely describes the combination of true positive rates and false positive rates for each criterion of conservation status. The ROC curves were derived by calculating the true and false positive rates for a given criterion over a wide range of quantitative thresholds of change in abundance that trigger a change in classification status from not declining to declining. For instance, the thresholds ranged from a 100%

reduction in abundance required before resulting in a triggering event, all the way to the maximum increase in abundance that occurred over any period for any CU resulting in a triggering event. When the threshold resulting in a triggering event was a 100% decrease in abundance, both the false positive rate (α) and the true positive rate (power) will equal zero because there are no triggering events at all (no true or false positives). When the threshold resulting in a triggering event was the maximum increase in abundance over any given period, both the false positive rate and the true positive rate will equal one because every single year is a triggering event (no false or true negatives).

We estimated the AUCs of the ROC curves using the trapezoidal method, which involves connecting the points on the ROC curve with straight lines and summing the areas of the resulting triangles and trapezoids (Vida 1993). The AUC of each criterion can range from zero to one.

Higher ranks were given to conservation criteria with higher AUC values. A higher AUC value indicates that a criterion has a greater chance of correctly identifying a declining CU than a criterion with a lower AUC value.

Sensitivity analyses

We also performed several sensitivity analyses, including one that examined the effect of the way in which true status was defined on rankings of criteria as indicators of declining CUs. As we show later in the Results section, scenarios A-D (Table 4), which define true status based only on the subsequent three generations of data, led to poor results and so those cases were eliminated from further discussion. Within each of the four remaining scenarios or definitions of a declining CU (cases E-H in Table 4), we calculated AUC values for the 20 decline criteria and ranked them accordingly, as well as calculating the median AUC value of each criterion across all four scenarios. We used median instead of mean AUC to reduce the chance of having situations in which an outlier AUC value for a criterion in a given scenario strongly affects the criterion's ranking.

We also examined the effect that the quality of data had on the ranking of the criteria. We explored two categories of spawner abundance data, either all 18 CUs or only the best-quality data (sites within the various CUs in which spawner abundance was estimated by either fence counts or mark-recapture methods, Table 5, Tracy Cone, DFO, pers. comm.).

We also did a sensitivity analysis on a major change in harvest rates, which could confound our results and interpretation. In 1995 there was a dramatic change in salmon management within the Fraser River, resulting from increased concern about issues of biological conservation. As a result, the exploitation rate across all Fraser River sockeye run-timing groups significantly decreased from previous years (Figure 2) (Mike Lapointe, Pacific Salmon Commission, 600 - 1155 Robson Street, Vancouver, B.C., V6E 1B5, pers. comm.). To examine whether this change in exploitation rate had an effect on the ranking of conservation criteria, we compared the results of the baseline analysis for all years (up through 2007) to one limited to only pre-1995 years.

CU aggregations

Two final sensitivity analyses examined the effect of the degree of spatial aggregation on results. Despite the goals of the Wild Salmon Policy, Fraser River sockeye CUs may remain aggregated into the four major run-timing groups for management purposes (Irvine and Fraser 2008). It is unknown how this will affect the status of individual CUs within the management aggregations because past management actions have not consistently reflected conservation concerns of individual components of run-timing groups (COSEWIC 2003a, COSEWIC 2003b). In addition, the recent IUCN status assessment of sockeye salmon did not evaluate status at the CU level, but rather at what they referred to as the subpopulation level, defined by freshwater and marine eco-regional groupings and genetic differentiation (Rand 2008). The subpopulations represent coarse units defined by extremely low rates of gene flow, and may contain numerous spawning sites (Rand 2008). One of their subpopulations (their #68) covers a geographic area that is composed of several Fraser sockeye CUs.

We grouped the 18 CUs from our analysis into their corresponding run-timing groups, as well as a population aggregate similar to that used in the recent IUCN Red List Assessment for sockeye salmon (Rand 2008), to evaluate the effect that aggregation of Fraser sockeye CUs for purposes of assessment and management decisions had on the status of individual CUs within the management groups. The IUCN used data from 33 individual spawning sites from 10 CUs for subpopulation #68, whereas our aggregation was not limited to just individual spawning sites but instead used the data from the 11 CUs that fell within the geographic range defined by IUCN subpopulation #68 (Table 5).

After the baseline analyses ranked the criteria based on their AUC values, we then assessed how often across all past years one of the top-ranked criteria would have identified a rate of decline in each run-timing group, IUCN subpopulation #68, and individual CUs within the aggregates that was large enough to result in a classification of threatened or endangered based on COSEWIC's thresholds of decline. A rating of threatened or endangered occurs if the decrease in abundance over the period measured by the criterion is greater than or equal to 30% or 50%, respectively, assuming that the causes of the decrease in abundance may not have ceased, may not be understood, or may not be reversible.

RESULTS

When the methods were applied to historical data for the 18 Fraser River sockeye conservation units (CUs), the performances varied considerably across the 20 criteria for identifying decreases in salmon abundance, as illustrated by the median AUCs across the eight definitions of a true declining CU (Table 4). These medians ranged from 0.33 to 0.67 (i.e., probabilities of correctly distinguishing between CUs that were declining and those that were not). Under scenarios that were based on defining true status using only the subsequent three generations of data (i.e., scenarios A-D), the conservation criteria had consistently lower AUCs (i.e., the majority of them were between 0.5 and 0.6) than under those scenarios that defined true status based on all subsequent years (which had higher proportions of criteria with AUC values $\geq$ 0.6) (Figure 3).

It is not worthwhile to continue ranking criteria based on AUC values for scenarios in which all criteria have low AUC values (probabilities of correctly categorizing decline status near that of a coin toss) because no criterion performs well. We therefore eliminated scenarios A-D in all subsequent analyses, and subsequently calculated median AUC values for each criterion only across scenarios E-H, for which a true status of declining was defined by a decrease in abundance of a given amount over all remaining years in the time series. Note that this removal from further consideration of the four cases for determining the *true* status does *not* apply to the criteria for estimating the *current* status; the latter criteria still include cases based on calculations over the most recent three generations.

Among the 20 criteria of conservation status, there is a range of median probabilities of correctly identifying whether CUs are actually declining or not, as measured by median AUC values across scenarios E-H (Figure 4a). This result indicates that there is a difference in reliability of conservation criteria at distinguishing between declining and non-declining CUs. Criterion 13 has the highest median probability of correctly distinguishing between declining and non-declining CUs. This criterion measures the extent of decline between the geometric mean spawner abundance of a generation and the geometric mean abundance of the first generation in the time series, using natural log abundance data smoothed via a moving 4-year average (Table 3). In this criterion 13, generational means move in non-overlapping blocks and thus we can only assess whether the conditions for a given criterion are met every four years (1 generation). In general, criteria that measure the extent of decline from a historical baseline anchored at the beginning of the time series perform better than those criteria that measure either rate of decline over three generations or extent of decline from a historical baseline anchored at the maximum abundance that occurred during the time series (Figure 4a). This advantage of the first group of criteria was between 1 and 49% greater probability of correctly classifying decline status, as measured by the difference between median probabilities.

We compared the performance of all criteria to a frequently used version of COSEWIC's criterion A, which measures the change in abundance over three generations of smoothed data (criterion 2 in Figure 4). The performance of this standard COSEWIC decline criterion was moderately good, but nine criteria were better and 10 were worse at distinguishing between declining and non-declining CUs. The conservation criteria with the highest five median probabilities had an 8-10% greater chance of correctly categorizing a CU as declining or not than the standard COSEWIC decline criterion (criterion 2 here). However, this criterion 2 had an 8-38% greater chance of correctly classifying decline status than criteria that measured extent of decline from the maximum abundance that occurred anywhere in the time series.

The most reliable criteria (highest probabilities of correctly classifying decline status, measured by AUC values) depended on the definition of a true declining CU, as shown by the proportion of the four scenarios in which criteria ranked either in the top three or top five criteria based on the largest AUC value (Figure 4b). Nevertheless, all criteria that were ever ranked in the top three or top five are criteria that measure the extent of decline from a historical baseline anchored at the beginning of the time series. Once again, criterion 13 ranked highly, with 100% of the scenarios ranking it in the top five, and 75% of the scenarios ranking it in the top three (Figure 4b). Criteria that measure the rate of decline over three generations (criteria 1 and 2) or the extent of decline from the maximum abundance in the time series (criteria 5, 6, and 14 to 18) never ranked in the top three or five of any scenario.

Criteria that measure change in abundance using smoothed data generally performed better than those criteria that measure change in abundance over the same period using unsmoothed data. For all pairs of criteria (only differing in their use of either smoothed or unsmoothed data), those based on smoothed data (criteria 2, 4, 6, 8, 11, 13, 16, 18, and 20) have a greater median probability of correctly classifying decline status than those using unsmoothed data (1, 3, 5, 7, 10, 12, 15, 17, and 19), except in the case of measuring the extent of decline from the maximum year in the time series (criteria pair 5 and 6). Excluding that exception, the smoothed criteria had a 4 to 29% greater chance of correctly distinguishing between declining and non-declining CUs when compared to using unsmoothed data.

Sensitivity analyses

We first compared results of the baseline analysis that used all years of data with those of an analysis based only on pre-1995 data to determine whether reduced exploitation rates beginning in 1995 confounded our interpretation of the reliability of the conservation-status criteria. Most criteria were robust to the potential confounding effect of changes in harvest rate because they only experienced a small change (ranging from 3-12%) in their chance of correctly identifying a declining CU when the analysis was limited to pre-1995 years (data not shown). However, several criteria that measure decline from the maximum abundance (criteria 6, 16, 17, and 18) were not robust to the change in harvest rate and experienced a large increase (ranging from 21-41%) in their chance of correctly identifying a declining CU when the analysis was limited to pre-1995 years.

The rank order of the criteria based on their median probabilities of correctly classifying decline status also generally stayed the same whether the analysis used all years of data or was limited to pre-1995 data (Figure 5a). The two best-performing criteria (11 and 13) were still ones that measure the extent of decline from a historical baseline anchored at the beginning of the time series, however, this change in data period caused an improved performance of criteria 16 and 18 compared to the baseline case; they became ranked 7th and 3rd, respectively (Figure 5a). All criteria that were ranked in the top three or top five, using either all years of data or only pre-1995 years, measure the extent of decline from a historical baseline anchored at the beginning of the time series, except for criteria 16 and 18, which appeared in the top five criteria when the analysis was limited to pre-1995 years (Figure 5b). These two criteria measure the extent of decline between the geometric mean of each generation and the geometric mean of the maximum three-generation period in the time series, using smoothed data. Generations move in a sliding window one year at a time for criterion 16 and in non-overlapping 4-year blocks for criterion 18.

Another indication of reliability of criteria for conservation concern is that their probability of correctly categorizing status should increase for analyses that are limited to only the best-quality data, as opposed to all available data. Our analysis of the best-quality data resulted in most criteria exhibiting an increase in the median probability of correctly classifying decline status across scenarios E-

H. All exceptions are criteria that measure extent of decline from the maximum abundance in the time series (criteria 5, 6, 14, 16, 17, and 18). However, the median probabilities of correctly classifying decline status for these particular criteria are generally low regardless of the quality of data used (i.e., median probabilities ranged from 0.07 to 0.59 across both sets of data), so the difference between using all data or best-quality data is relatively small and these criteria still perform poorly.

Analyses with either all data or only best-quality data resulted in slight differences in rank order of criteria based on their median probabilities of correctly distinguishing between declining and non-declining CUs. However, the general pattern remained the same regardless of data quality. The median probabilities of correctly categorizing decline status were once again consistently highest for criteria that measure extent of decline from the beginning of the time series, with criteria measuring rate of decline over three generations ranking in the middle, and criteria measuring extent of decline from the maximum abundance consistently ranking lowest (Figure 6a). Limiting the analysis to best-quality data resulted in criterion 13 having the second highest median probability, just below criterion 11 (Figure 6a). Criterion 11 only differs from criterion 13 in that the generations compared to the historical baseline (geometric mean abundance of the first generation) move in a sliding window for criterion 11 as opposed to non-overlapping blocks; this allows an assessment of criterion conditions every year. The top-five-ranked criteria differ when the quality of data used is changed, but regardless, all criteria ranked in the top three or five are those that measure the extent of decline from a historical baseline anchored at the beginning of the time series (Figure 6b).

Collectively, these results suggest that criteria measuring extent of decline from a historical baseline anchored at the beginning of the time series consistently outperform other criteria in terms of having an increased chance of correctly distinguishing between a declining and non-declining CU, regardless of data quality.

True and false positive rates

Although median AUC as reported above provides a concise measure of performance of criteria of conservation status, some scientists or managers will also be interested in the components of that statistic, i.e., the rates of occurrence of true positives (correct detection of declines) and false positives (incorrect declaration of decline). For a given conservation criterion, the points on the ROC curve relate to different thresholds of percentage change in abundance that cause a triggering event. For each criterion, in each scenario, different true and false positive rates are produced at each threshold. The ROC curve for criterion 11 in scenario E, using the best-quality data only, has an AUC of 0.9 (Figure 7). This curve demonstrates the possible trade-offs between true and false positive rates. For instance, a manager may choose one of COSEWIC’s conventional threshold levels for the decrease in abundance (30%, 50%, 70%, or 90%) that results in a triggering event; circled points on the left-hand-side of the ROC curve indicate these thresholds for criterion 11 (Figure 7). Figure 8 provides the true positive rate and false positive rate corresponding to each threshold. Reducing the percent decline that would result in a triggering event from 90% to 30% increases the true positive rate from 0.11 to 0.73 (Figure 8). The probability of false negatives (type II errors) is 1- the true positive rate (i.e., 1- power); therefore the true positive rate of 0.73 is equivalent to a 27% chance of a false negative (type II error) occurring when the threshold level of decline is 30%. For criterion 11, with this reduction of the threshold of decline from 90% to 30%, the false positive rate only increases from 0.00 to 0.02 (Figure 8), indicating that it is beneficial to reduce the threshold of decline in this case because we get a large increase in the true positive rate (power) but a very small increase in the chance of a false positive (type I error).

In our analysis, the point corresponding to the upper-left-most corner on the ROC curve indicates an “ideal” threshold of change in abundance for that criterion, because it has the lowest chance of an error (i.e., lowest (chance of false positive + chance of false negative)), which is indicated by a high true positive rate and a low false positive rate. In our example (Figure 7), a square indicates this “ideal” point for criterion 11 and it corresponds to a threshold of a 10% decrease in abundance over the period measured by the criterion that causes a triggering event. Figure 8 shows that this further reduction in the threshold of decline (from 90% to a 10% decrease) once again increases the true positive rate with only a small increase in the false positive rate.

CU aggregations

We evaluated the effect on the status of individual CUs within the management groups of the level of spatial aggregation of Fraser sockeye CUs that is used for assessment and management purposes. This evaluation assessed how often criterion 13, one of the top-ranked criteria in our other analyses, classified both the CU aggregates (run-timing groups and IUCN subpopulation #68) and the individual CUs within the aggregates as threatened or endangered based on the thresholds of decline used by COSEWIC (30% and 50% decline, respectively, assuming the causes of the decline may not have ceased, or may not be understood or reversible). Criterion 13 never triggered a threatened or endangered rating for IUCN subpopulation #68 at any time, and the Late Summer run-timing group was the only run-timing aggregate to experience a decline (measured by the conditions of criterion 13) that was large enough to trigger a COSEWIC threat rating, threatened in one year, 2003 (Table 6) . However, there are numerous years where individual CUs within the aggregates experienced declines large enough to classify them as threatened or endangered based on the conditions of criterion 13 and the thresholds of decline used by COSEWIC (Table 6). These results most likely underestimate the number of years where both the individual CUs and the larger CU aggregates would have been classified as threatened or endangered because criterion 13 only assesses status every four years, using non-overlapping 4-year generational blocks of spawner abundances.

DISCUSSION

There are many different methods to assess the biological status of CUs. COSEWIC criteria are not limited to examining time trends in abundance; they also consider absolute abundance and spatial distribution (COSEWIC 2006). Diversity and population growth rate are also considered as key parameters for evaluating salmonid population viability (McElhany et al. 2000). The DFO is considering a combination of several criteria that account for all of the above-suggested factors when assigning biological status to salmon under the WSP. Our analysis is limited to evaluating possible criteria of one component of biological status, time trends in abundance, but such analyses could be done on other criteria in the future.

This analysis estimated the probability that various criteria will generate four different types of outcomes (i.e., true positives, false positives, true negatives, and false negatives) but took further steps by showing the relationship among these outcomes through the development of ROC curves. The limitation of other analyses to one threshold is equivalent to estimating only one point on the ROC curve, because each point (Figure 1) results from estimating the probability of the four outcomes at a specific threshold. Our analysis examines the performance of decline criteria across a range of thresholds and thus provides much more information about the overall reliability of the criteria.

This evaluation of criteria for assessing whether a CU is truly declining used two main methods to rank reliability of criteria: (1) median area-under-the-curve (AUC) across scenarios that only differ in their definition of true status (various future periods and percent declines over that period), and (2) proportion of those scenarios when a given conservation criterion ranks in the top three or five criteria; a high ranking results from a large AUC. It is difficult to use rank alone to evaluate the performance of criteria because, for instance, the top two criteria may have very small differences between their AUC values. If we only consider rank and not the actual AUC values, this can lead to a criterion appearing to have better performance than it actually does, because there is no quantification of the difference in performance between criteria. Regardless, we show that criteria measuring extent of decrease from a historical baseline anchored at the beginning of the time series consistently perform better (based on their higher AUCs and rank) than criteria that measure either rate of decline over three generations or extent of decline from the maximum abundance in the time series. We also show that criteria that smooth abundance estimates generally perform better for Fraser sockeye than criteria using unsmoothed abundances.

This analysis based estimated status on past spawner abundances and true status on future spawner abundances, which implicitly assumed that past management actions (harvest rates), environmental conditions, and productivity will also apply in the future. However, estimated status and true status may be decoupled by changes in these factors between past and future years. For instance, successful management intervention may increase spawner abundance, erroneously leading to the conclusion that an estimated status of declining was unwarranted (false positive). Ideally, one should account for year-by-year management intervention to rule out the confounding effect of management actions on the evaluations of conservation criteria. Such a detailed analysis was not possible because the Fraser sockeye CUs have historically been aggregated into their run-timing groups for management purposes, with CUs that migrate upstream during the same time period being harvested simultaneously. Therefore, data on yearly harvest rates for each specific CU are currently not available. Nonetheless, we tested the sensitivity of our findings to a single large-scale region-wide management intervention, the 1995 reduction of sockeye exploitation rates across all run-timing groups that has continued since that time. The reliability of most criteria was robust to this change in harvest rate.

The number of spawners in any given year is not only affected by harvest rate but also by past spawner abundance in the CU due to the possible density-dependent stock-recruitment relationship (i.e., spawning stock size affects subsequent number of returning adults of the year class produced by that spawning), abundance of predators, marine survival conditions, and enroute mortality (Nelitz et al. 2006). Spawner abundance is also affected by outcome uncertainty (Holt and Peterman 2006), the difference between realized and target mortality rates. We evaluated the reliability of various decline criteria as indicators of the true status of CUs as defined by the net response of abundance to all of these confounded factors in the years subsequent to the year of analysis.

There are only slight differences in conclusions about which criteria perform best, regardless of whether we evaluated criteria based on median AUC across scenarios or the proportion of scenarios in which a given criterion ranks highly, whether data are all available data or only best-quality data, and whether analyses use all available years of data or only pre-1995 years. Criteria that measure the extent of decline from a historical baseline anchored at the beginning of the time series have greater reliability (i.e., increased chance of correctly distinguishing between a declining and non-declining CU) than all others and appear robust to the method of evaluation. The better performance of this group of conservation criteria indicates that use of the standard COSEWIC and IUCN decline criterion of the change over the most recent three generations may reduce the chance of correctly distinguishing between declining and non-declining CUs. However, this standard COSEWIC criterion has a greater chance of correctly identifying a declining CU than criteria that measure extent of decline from the maximum abundance in the time series. Not only do the latter criteria with historical baselines based on maximum abundance perform poorly, but it is also dangerous to compare future abundances to such a baseline because the maximum abundance may have arisen under the most productive climate regime in recent history (Beamish et al. 1999). Hence, any method assuming maximum historical abundance as the baseline will be vulnerable to errors due to uncertainty in future climate regimes and salmon productivity.

Some conservation-status criteria are only slight variations on others and therefore may be highly correlated, especially within the three main groups of criteria (i.e., extent of decline from a historical baseline anchored at the beginning of the time series, rate of decline over three generations, and extent of decline from a historical baseline that is the maximum abundance in time series). Analysts can choose a method for measuring time trends in spawners on a case-by-case basis, depending on the data available and specifics of the CU (e.g., life-history strategy, productivity, fishing mortality). For instance, the choice of conservation criterion may depend on the age-structure of the CU, whether the CU exhibits cyclic patterns in abundance, the quality of abundance estimates, or fishing mortality in given years. Some criteria only allow for assessment of status every four years because they compare 4-year non-overlapping generational blocks to the historical baseline. Managers may have preferences regarding the appropriateness of these criteria, depending on whether or not they want to assess status every year. However, some criteria that only assess status every four years were found very reliable (e.g., criterion 13), indicating that their use may be advantageous because they give a good indication of a declining CU, without having to be assessed every year.

An entire subset of conservation-status criteria, those estimating the reduction from the maximum spawner abundance, had consistently poor reliability as indicators of declining CUs. Not only were the median AUCs of this group of criteria consistently the lowest, but it was the only group with any criteria that exhibited large differences in median AUC when the analysis was limited to pre-1995 data as opposed to all years, suggesting that they are not robust to the potential confounding effect of changes in harvest rates. However, keep in mind that the design of these criteria means they were measured less often than other criteria. If the maximum abundance of a CU occurred in the mid-to-late years of the time series, there were fewer years subsequent to this period where the conditions of the criterion could be measured than criteria using other periods for comparison. This resulted in criteria that estimated the extent of decline from maximum spawner abundance using fewer true positives, true negatives, false positives, and false negatives to calculate the true and false positive rates for ROC curves.

Although measurement of the extent of decline from early years is the most robust method of estimating time trends in abundance, there may be cases (e.g., CUs with extremely short time series or large amounts of missing data) where the DFO may need to use other indicators such as absolute spawner abundances, spatial distribution of spawners, and fishing mortality to assess the biological status of salmon populations. The most applicable indicator of biological status may vary not only among CUs, but also within CUs. For instance, time trends in spawner abundance may be most applicable to dominant cycle years of Fraser sockeye, whereas absolute spawner abundance may be most applicable to weak cycle lines, which have much lower abundances.

One lesson from this research is that it is important to evaluate the statistical properties of various criteria of conservation status instead of assuming that the most frequently used criterion is the best indicator. We demonstrate that there can be considerable differences among possible criteria in the probability of drawing correct conclusions about whether a CU is declining. Despite its frequency of use, the commonly used criterion for determining status of Pacific salmon populations (the criterion measuring the rate of decline over three generations of smoothed data) does not actually have the best performance and can be improved upon.

As noted in the Introduction, other studies have evaluated criteria that are used to assess biological status for conservation or management purposes. The reliability of management advice based on various limit reference points was evaluated (Piet and Rice 2004), and another study evaluated IUCN decline criterion A by comparing its assessment of status with one derived from limit reference points (Dulvy et al. 2005). The latter analysis, which used empirical meta-analytical approaches, found that the IUCN decline criterion A1 (using thresholds of decline of 50%, 70%, and 90%) provided warnings of population collapse consistent with those provided by fisheries stock assessments, with no evidence of false positives (Dulvy et al. 2005). This contrasts with previous analyses using simulation approaches that evaluated IUCN decline criterion A and that found a high chance of incorrectly classifying species as threatened (false positives) (Punt 2000 and Matsuda et al. 1998). These results are not directly comparable to ours because we evaluated the reliability of different interpretations of the commonly used IUCN decline criterion A, as opposed to comparing a single interpretation against other methods of assessment.

Rice and Legace (2007) found that, when estimating the probability that a stock that met the IUCN decline criterion A was actually on a trajectory to extinction, the number of true and false positives was highly dependent on the assumed starting point of the future trajectory. If the starting point was when the threshold of the decline criterion was first exceeded, then there was a high probability that the subsequently observed abundance estimates would be less than or equal to those expected if the population was on a true trajectory to extinction (i.e., true positive, criterion correctly indicated a conservation issue) (Rice and Legace 2007). However, if the starting point was the year in which the decline reached its trough as opposed to when the threshold was first exceeded, there was a increased chance that the subsequently observed abundance estimates would be greater than those expected if the population was on a true trajectory to extinction (i.e., false positive, criterion incorrectly indicated a conservation issue). Our analysis used the year in which the threshold of decline was first exceeded as the starting point of the future trajectory that defines true status, as opposed to the year when the decline reached its trough because the latter starting point may be biased towards recovery, i.e., selecting an especially low starting point increases the chance of having larger abundances in subsequent years, and therefore more false positives (Rice and Legace 2007).

Our research is not limited to assessing the chance of a false positive or false negative at only a few thresholds of decline that trigger a classification of conservation concern, as are the above studies (Piet and Rice 2004, Dulvy et al. 2005, and Rice and Legace 2007), but it instead determines these probabilities over a range of thresholds. Therefore, the conclusions from the above studies regarding the tendency for the IUCN decline criterion to produce numerous false positives or not is only one small part of the discussion. Different thresholds of decline can result in drastic differences in the chance of false positives and negatives for a given conservation-status criterion. At a given threshold for percentage change in abundance, a given criterion will tend to produce more false positives than false negatives, whereas the same criterion will tend to produce more false negatives at a different threshold. AUC values are indicators of the overall performance of criteria across all possible thresholds of change in abundance. However, it is useful to examine the true positive rate (probability of avoiding a type II error) and the false positive rate (α, probability of making a type I error) values at specific thresholds of change in abundance. Managers need to take all of this information into account when choosing a criterion and threshold of decline to assess conservation status, and this choice will in part depend on their risk-tolerances for both false positives and false negatives, as we discuss below.

Trade-offs

ROC curves convey information about the chances of both false positives and negatives in a way that allows managers to visualize trade-offs they need to make between the two errors. Managers of Fraser River sockeye salmon can select the appropriate threshold of change in spawner abundance that results in the true and false positive rates that they find acceptable. Various managers will have different risk-tolerances for false positives and false negatives that are specific to their situation and priorities. A high true positive rate indicates a low chance of a false negative (type II error), and a low false positive rate indicates a low chance of a false positive (type I error). Ideally, managers would like a low chance of both errors, but this is difficult to achieve. Therefore, managers often have to make trade-offs between levels of each error that reflect their personal perception of the detrimental effects of each type of error. A manager's risk-tolerance will depend on the perceived cost of both errors as well as their respective probabilities of occurrence. When managers think explicitly about these risk tolerances, our results and use of the ROC method will help by providing estimates of the probability of different types of errors occurring for a given criterion for determining conservation status.

Estimation of perceived costs is a different matter, though. The cost of false positives and false negatives are determined by the relative importance of their consequences (economic, political, environmental, and social costs) (Mapstone 1995). When assessing whether a CU is declining, a false positive may result in unnecessarily restricting fishing (incorrectly classifying a CU as declining when it actually is not), which could have social and economic costs. In the case of Fraser River sockeye salmon, the cost of a false positive may be a decrease in revenue from reduced harvest on a run-timing group that contains a CU that was mistakenly classified as declining. An additional cost of false positives is that they can undermine truly needy conservation situations by diverting scarce resources unnecessarily. A false negative (incorrectly classifying the CU as non-declining when it really is) could result in irreversible depletion of salmon or loss of a CU. The cost of a false negative (not reducing harvest on a declining CU) may be the loss of all future revenue from the harvest of that CU, the loss of catch for First Nations (possibly for important food, social, or ceremonial purposes), the public's loss of a culturally important population, and ecosystem effects such as loss of nutrients /biodiversity in an area (Kappel 2005).

The cost of these different types of errors may be incurred over different time scales. Costs of a false positive (e.g., loss of revenue from catch) would be incurred in the short-term and the costs of a false negative (e.g., loss of biodiversity) over the longer term. In addition, the cost of false negatives may be incurred by different groups than the cost of false positives (Peterman 1990). For instance, those involved in the fishery may experience a decrease in profits due to a false positive, whereas the costs of a false negative are suffered by not only the fishing industry but also First Nations, the public, and the ecosystem as a whole. Determination of not only the costs but also who will suffer the consequences of both types of errors is an important part of the trade-off process for managers, which they can address by carefully articulating their management objectives.

Mapstone (1995) suggested setting the acceptable ratio of false positives and false negatives according to the relative costs of committing those errors (i.e., α/β = cost of false negative / cost of false positive). In our case, this means choosing a criterion and a threshold of decline that has a ratio of the false positive rate to false negative rate equal to the relative cost of incorrectly classifying a CU as declining (false positive) versus incorrectly classifying a CU as non-declining (false negative). Some argue that in the absence of sufficient information to determine the costs of errors, the potential costs of both errors should be considered equal (Mapstone 1995). If managers assume both types of errors have equal costs, they may want an equal probability of occurrence of both false positives and negatives (i.e., $\alpha/\beta = 1$). However, the cost of false negatives often exceeds the cost of false positives, in which case managers may want equal *expected* costs for the two types of error (probability of the error occurring multiplied by the cost of that error) (Peterman 1990). Mapstone (1995) suggested a method for determining which criterion and threshold of decline managers should choose based on their risk-tolerances for both false positives and false negatives. The first step was to determine both the maximum chance of false positives occurring and the maximum chance of false negatives occurring that would be acceptable to the manager. Then these managers could choose a conservation criterion with a threshold of decline that would produce a chance of both errors below those maximum acceptable levels and that also corresponds to the acceptable ratio of the probability of both types of errors occurring (i.e., α/β) that was based on their relative costs.

ROC curves suggest that the ideal threshold of decline is the one resulting in the point in the most-upper-left corner of the plot, because it has the lowest chance of an error. However, a manager may not find acceptable a true positive rate less than some constraint level. This situation requires the manager to choose a conservation threshold with the acceptable true positive rate and to make a trade-off by having to accept the false positive rate that corresponds to that true positive rate. A manager may only be willing to accept a very low chance of a false negative, for example, a chance of a false negative less than or equal to 16%, corresponding to an 84% chance of a true positive (triangle for criterion 11 in Figure 7). This means she/he will have to accept a greater chance of a false positive, in this case a 22% chance of a false positive (Figure 7). This requires the manager to use a conservation threshold that considers any *increase* in abundance less than 80% over the period as a triggering event signalling a declining CU, which is of course inappropriate. For criterion 11, this change in the threshold from the “ideal” threshold of a 10% decline in abundance to an 80% *increase* in abundance resulted in a small increase in the true positive rate of the criterion and a large increase in the false positive rate (Figure 8).

Managers must decide which threshold to choose based on what they (and society) regard as a decline rate that reflects a concern for conservation status (which will imply what true and false positive rate they are willing to accept even if they do not explicitly consider those probabilities). ROC curves allow managers to see that if they are only willing to accept some given level of false positives or false negatives, this may result in a high chance of the other type of error.

Some conservation criteria may tend to make more false positives than false negatives (or vice versa), as measured by whether the majority of thresholds of change in abundance have a greater chance of making one type of error or the other. Most criteria evaluated here tended to produce more false negatives, i.e., failing to trigger a conservation concern when the population subsequently declines. For instance, the majority of thresholds for criterion 11 have a greater chance of occurrence of false negatives than false positives

(Figure 7). However, this criterion has an AUC of 0.9, indicating an overall 90% chance of correctly distinguishing between a declining and non-declining CU. Although there is a greater chance of false negatives as opposed to a false positives at the majority of thresholds, the proximity of points on the ROC curve to the upper-left corner of the plot indicates that there is a small chance of either type of error at those thresholds. If the sole objective is to minimize the probability of both types of error, then a manager only needs to choose a threshold of percentage decline near the top left corner of the ROC curve to satisfy that objective. If consideration of the relative costs of the two types of errors is included in the manager's objectives, then this could move the choice of a threshold higher or lower.

Most criteria tended to produce more false negatives than false positives, whereas criteria 5, 14, 15, and 17 (which all measure the extent of decline, using unsmoothed data, from a historical baseline anchored at the maximum abundance in the time series) produced the reverse. For most thresholds of change in abundance considered, criteria 5, 14, 15, and 17 had a greater chance of false positives than false negatives. For instance, if a manager prefers making false positives, criterion 5 and the thresholds to the right of the dashed line in Figure 7 may seem like a good option. However, these criteria (5, 14, 15, and 17) all have relatively low AUCs, with median AUCs across scenarios E-H of 0.61, 0.37, 0.45, and 0.33 respectively (Figure 4a). Therefore, under most scenarios, these criteria are not able to distinguish between declining and non-declining CUs much better than at random and some criteria are consistently worse than that. These low AUC values are a result of a high chance of both errors at the majority of thresholds, even though the chance of a false positive is greater than a false negative (Figure 7). These results suggest that the tendency of certain criteria to produce more false positives than false negatives may result from those criteria having a greater chance of false positives at a given threshold than other criteria, as opposed to a lower chance of making false negatives (Figure 8). Depending on the managers' priorities, they may wish to use a criterion that has a lower AUC, but a higher true positive rate and lower false positive rate at a given threshold of change in abundance.

The results of this analysis may be useful to the DFO and other organizations in their process of selecting criteria for assessing biological status of spatial units of fish populations that are based on quantifying time trends in spawner abundance. The best-performing conservation-status criteria may have other applications, such as biodiversity indicators. For example, over a range of historical years, the rate and extent of decline that numerous fish species had experienced up to a given year was measured and each species was assigned a threat categorization based on the thresholds of decline used to assess status by the IUCN A1 decline criterion (Dulvy et al. 2006). The threat categorization for each species in each year was weighted as vulnerable = 1, endangered = 2, and critically endangered = 3. The authors then reported trends in biodiversity via a composite indicator based on the weighted average of threat scores of individual species in each year. The proportion of threatened fishes and their degree of threat determined the composite indicator value, which acts as an indicator of biodiversity trends and can be used to judge progress in relation to management objectives.

CU aggregations

The maintenance of species diversity and, by implication, resilience to environmental change requires the assessment and maintenance of individual locally-adapted populations (Hilborn et al. 2003). Spatial distribution of spawners across spawning sites and habitat types is a possible surrogate for biological diversity because Fraser sockeye that spawn in different environments may be adapted to the local conditions and may respond differently to human and natural stressors (Irvine and Fraser 2008). Local adaptations improve survival in specific habitats and consequently increase the population’s productivity (Irvine and Fraser 2008). The identification of CUs within the Wild Salmon Policy (WSP) attempts to protect genetic diversity, demonstrated by local adaptations, within each salmon species (Irvine and Fraser 2008).

Aggregating several CUs into one large unit before assessing status may give the incorrect impression that the salmon are better or worse off than they actually are. For instance, the lack of a threatened/endangered classification at the aggregate level may hide the fact that individual CUs within that aggregate may be classified as threatened or endangered, as illustrated in Table 6. Management and assessment of status based on an aggregate of several CUs may not allow for the maintenance of some CUs, because their decline goes unnoticed within a CU aggregate, resulting in no initiation of actions to halt the decrease. Results in Table 6 show that while the CU aggregates may not be classified as conservation concerns, individual CUs within the aggregates can frequently exhibit declines in abundance large enough to warrant conservation concern based on COSEWIC's thresholds of decline. This situation contradicts the WSP, which aims to ensure CUs “are at a level of abundance high enough to ensure there is a substantial buffer between it and any level of abundance that could lead to a CU being considered at risk of extinction by COSEWIC” (DFO 2005). It is therefore important to assess the status of Pacific salmon at the CU level.

Even if managers of Fraser sockeye continue to apply exploitation rates at the level of run-timing groups, assessments of status still need to be done at the individual CU level. If the biological status of the individual CUs is known, then the exploitation rate on the aggregate containing a CU of conservation concern can be adjusted to account for the trade-offs that managers deem appropriate between profits from catch and diversity of Fraser sockeye.

Limitations

A limitation of ROC analysis is that selecting discrete thresholds of change in abundance that result in a triggering event means the trapezoidal rule will tend to underestimate the area under what is in reality a smooth ROC curve (Hanley and McNeil 1982). The smaller the intervals between thresholds (i.e., more continuous as opposed to discrete thresholds), the closer the trapezoidal estimate of area under the curve will come to the true smoothed AUC (Hanley and McNeil 1982). This may not be detrimental to our analysis, because (a) such biases are likely very small relative to differences in AUCs among criteria, and (b) managers are likely to be more interested in the relative performance of criteria and trade-offs between true and false positive rates, as opposed to the accuracy of particular AUC values.

Some CUs had years with missing spawner abundance estimates and this may lead to biased estimates from the linear time-trend model (i.e., biased rates of change over a period), and thus possibly incorrectly estimated CU status. Interpolation of these missing points using the mean of the same cycle year from the immediately previous and subsequent generations may still be problematic because it ignores the uncertainty of the interpolated values (Nakagawa and Freckleton 2008). Data augmentation and multiple imputation are two methods recommended by statisticians to deal with missing data because they not only provide unbiased parameter estimates but they also provide information on the impact of missing data on parameter estimation (Nakagawa and Freckleton 2008). Although there was not a need to interpolate numerous years of missing data in this case, suggesting that our results are likely robust to changes in those few interpolated estimates of spawner abundance, it may be worthwhile to explore in the future the effect that missing data has on the ranking of criteria.

The main limitation of our analysis is the inevitable confounding of the change in a CU's spawner abundance over any period with biological factors and management actions. We tried to overcome this confounding by evaluating the reliability of criteria at identifying CUs that are exhibiting a net decline in spawner abundance. That is, we defined the true status of a CU as declining or not based on the overall change in spawner abundance over the period, regardless of the management action or environmental conditions of the time. We also compared results for a period with persistent high harvest rates with results that included that period plus one with substantially lower harvest rates; we found few changes in conclusions.

The change in CU spawner abundance to a given threat in a given year may be highly dependent on the specific CU and environmental conditions of the time (i.e., CU productivity, marine and freshwater survival, and enroute mortality may depend on environmental conditions). However, those influencing conditions are not usually known at the individual CU level in each year, and scientists will need to judge which conservation criteria to use based on the general performance of criteria across many situations. Therefore, to determine the effectiveness of a given criterion at correctly distinguishing between declining and non-declining CUs, where effectiveness is not specific to any given CU or period, we tallied the number of true positives, false positives, true negatives, and false negatives for each criterion across a wide range of CUs and years. We chose a highly varied group of 18 CUs for the analysis, which cover a range of geographic spawning sites, management actions, environmental conditions, productivity, quality of spawner abundance estimates, and population trajectories. These varied CUs allowed us to evaluate the overall effectiveness of each criterion across a wide range of conditions.

Our analysis is also limited in that the performance of criteria as indicators of declining CUs is contingent on our definition of a declining CU. We defined the true status of a CU as declining based on its future trajectory subsequent to the year in which the classification of status was determined by a criterion. To overcome the dependence of the ranking of the criteria on a single definition of a declining CU, we plotted ROC curves for each criterion across a range of definitions of declining CUs, varying in their percent declines and future periods. We ranked the criteria based on their performance across various definitions of a declining CU to determine the criterion with performance that is robust to changes in this definition. However, by basing the true status on the future trajectory of CUs and the estimated status on the past trajectory, we are assuming that management actions and environmental conditions have not changed over the entire period, which is obviously not the case.

Future research

Because management actions, population dynamics, and environmental conditions all confound time trends in spawner abundance, there is a need to explore other ways to define the true status of CUs as declining that are not dependent on future time trends in abundance. Methods that determine what the exploitation rate was on individual CUs over past years can be used to assess whether a decline over a given period was a true decline in productivity or only a result of a high exploitation rate, and may be used to evaluate the effectiveness of the criteria at signalling this decline.

The DFO, the IUCN, and COSEWIC are not limiting themselves to criteria measuring trends in abundance to assess the biological status of CUs, therefore it may be worthwhile to perform a similar retrospective analysis for the other classes of indicators, such as current spawner abundance (rather than time trends in it) and spatial distribution of spawners. In addition, retrospective analyses of the effectiveness of the criteria for different species of salmon inhabiting different regions would be useful.

Furthermore, all criteria evaluated here should be evaluated using simulation modeling in the future. Such an approach has been used to evaluate the performance of a subset of criteria for measuring current spawner abundance, which is one of the proposed indicators of biological status under the WSP (Holt et al. 2008). A similar approach applied to criteria measuring time trends in abundance would allow researchers to see how the criteria perform when the true status of the CU, as well as the population dynamics, management actions, and environmental conditions, are under the control of the modeler and are therefore known with certainty. This method would differ from our analysis, in which there is considerable uncertainty about these other factors, and in which a CU's change in spawner abundance in any given period may be confounded by many variables. In a simulation model, criteria could be evaluated across a wide range of conditions to see which ones are most robust to changes in population dynamics, management actions, and environmental conditions.

Conclusion

This research not only explores the reliability of various conservation-status criteria at detecting true declining CUs, but also demonstrates the utility of Receiver Operating Characteristic analysis for evaluating the performance of these criteria. ROC analysis may have many useful applications in fisheries management (and other fields) due to its ability to measure true and false positive rates over a range of thresholds used to detect a given effect (e.g., percent decline in abundance), and combine this information into one measure of overall reliability, i.e., area-under-the curve. In this case, ROC analysis found conservation-status criteria that measured decline from the earliest available period to be the most reliable indicators of declining Fraser sockeye conservation units.

REFERENCES

Beacham, T.D., McIntosh, B., and MacConnachie, C. 2004. Population structure of lake-type and river-type sockeye salmon in transboundary rivers of northern British Columbia. J. Fish Biol. **65**: 389-402.

Beamish, R.J., Noakes, D., McFarlane, G.A., Klyashtorin, L., Ivanov, V.V., and Kurashov, V.1999. The regime concept and natural trends in the production of Pacific salmon. Can. J. Fish. Aquat. Sci. **56**:516-526.

Burgner, R. L.1991. Life history of sockeye salmon *Oncorhynchus nerka*. In Pacific salmon life histories. Groot, C. & Margolis, L., eds. Vancouver: University of British Columbia Press. pp. 3–117.

COSEWIC (Committee on the Status of Endangered Wildlife in Canada). 2003a. COSEWIC assessment and status report on the sockeye salmon *Oncorhynchus nerka* (Cultus population) in Canada. Committee on the Status of Endangered Wildlife in Canada. Ottawa. ix + 57 pp.

COSEWIC. 2003b. COSEWIC assessment and status report on the sockeye salmon *Oncorhynchus nerka* Sakinaw population in Canada. Committee on the Status of Endangered Wildlife in Canada. Ottawa. ix + 35 pp.

COSEWIC. 2006. COSEWIC's assessment process and criteria, Canada. Available from http://www.cosewic.gc.ca/eng/sct0/assessment_process_e.cfm

COSEWIC. 2008. Canadian wildlife species at risk. Committee on the Status of Endangered Wildlife in Canada. Available from http://www.cosewic.gc.ca/eng/sct0/rpt/rpt_csar_e.cfm [accessed 9 December 2008]

DFO (Fisheries and Oceans Canada). 2005. Canada's policy for conservation of wild Pacific salmon. Available from http://www-comm.pac.dfo-mpo.gc.ca/publications/wsp/wsptoc_e.htm. Vancouver, British Columbia.

DFO. NuSEDS V2.0 Regional Adult Salmon Escapement Database. Georgia Strait Pink and Sockeye: 1965-2005. Retrieved April 23, 2006.

DFO. 2008. Summary list and maps of conservation units in BC. Available from http://www-comm.pac.dfo-mpo.gc.ca/pages/consultations/wsp/Presentations/PublicForum_CU list handout_v1c_24Mar08.pdf

Dixon, W.J., and Massey, Jr., F.J. 1969. Introduction to statistical analysis, 3rd edition. McGraw Hill Book Co., New York. 638 p.

Dulvy, N.K., Ellis, J.R., Goodwin, N.B., Grant, A., Reynolds, J.D., and Jennings, S. 2004. Methods of assessing extinction risk in marine fishes. Fish Fish. **5**: 255-276.

Dulvy, N.K., Jennings, S., Goodwin, N.B., Grant, A., and Reynolds, J.D. 2005. Comparison of threat and exploitation status in North-East Atlantic marine populations. J. Appl. Ecol. **42**: 883-891.

Dulvy, N.K., Jennings, S., Rogers, S.I., and Maxwell, D.L. 2006. Threat and decline in fishes: an indicator of marine biodiversity. Can. J. Fish. Aquat. Sci. **63**: 1267-1275.

Dulvy, N.K., Sadovy, Y., and Reynolds, J.D. 2003. Extinction vulnerability in marine populations. Fish Fish. **4**: 25-64.

Eggers, D.M., and Irvine, J.R. 2007. Trends in abundance and biological characteristics for north Pacific sockeye salmon. N. Pac. Anadr. Fish Comm. Bull. **4**: 53-75.

Elith, J., Graham, C.H., Anderson, R.P., et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography, **29**: 129-151.

English, K.K., Koski, W.R., Sliwinski, C., Blakley, A., Cass, A., and Woodey, J.C. 2005. Migration timing and river survival of late-run Fraser River sockeye salmon estimated using radiotelemetry techniques. T. Am. Fish. Soc. **134**:1342-1365.

Fielding, A.H., and Bell, J.F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. **24**: 38-49.

Gotelli, N.J., and Ellison, A.M. 2004. A primer of ecological statistics. Sinauer Associates, Inc., Sunderland. 510 p.

Hanley, J.A., and McNeil, B.J. 1982. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. Radiology, **143**: 29-36.

Hibberd, P.L. and Cooper, A.B. 2008. Methodology: Statistical analysis, test interpretation, basic principles of screening with application for clinical study. Chapter 58 in Kleinman, R.E., O. Goulet, G. Mieli-Vergani, I.R. Sanderson, P.M. Sherman, B.L. Shneider (eds.), Walker's Pediatric Gastrointestinal Disease: Pathophysiology, Diagnosis, Management, Fifth Edition. BC Decker, Ontario, Canada.

Hilborn, R., Quinn, T.P., Schindler, D.E., and Rogers, D.E. 2003. Biocomplexity and fisheries sustainability. PNAS. **100**: 6564-6568.

Holt, C., Cass, A., Holtby, B., and Riddell, B. 2008. Indicators of status and benchmarks for conservation units in Canada's wild salmon policy. DFO Working Paper for PSARC Workshop January 5-6, 2009, Vancouver, B.C., available from the authors at Pacific Biological Station, Nanaimo, B.C.

Holt, C.A., and Peterman, R.M. 2006. Missing the target: uncertainties in achieving management goals in fisheries on Fraser River, British Columbia, sockeye salmon (*Oncorhynchus nerka*). Can. J. Fish. Aquat. Sci. **63**: 2722-2733.

Holtby, L.B., and Ciruna, K.A. 2007. Conservation Units for Pacific salmon under the Wild Salmon Policy. Canadian Science Advice Secretariat, Research Document 2007/070, 350 pp.

Hutchings, J.A. 2001. Conservation biology of marine fishes: perceptions and caveats regarding assignment of extinction risk. Can. J. Fish. Aquat. Sci. **58**: 108–121.

Irvine, J.R., and Fraser, G.A. 2008. Canada's wild Pacific salmon policy and the maintenance of diversity. Am. Fish. Soc. Symp. **49**: 391-398.

Irvine, J.R., Gross, M.R., Wood, C.C., Holtby, L.B., Schubert, N.D., and Amiro, P.G. 2005. Canada's Species at Risk Act: An opportunity to protect "endangered" salmon. Fisheries, **30**: 11-19.

IUCN (International Union for Conservation of Nature). 2001. IUCN red list categories and criteria: Version 3.1, Gland Switzerland and Cambridge, UK. Available from http://www.iucnredlist.org/static/categories_criteria_3_1

Kappel, C.V. 2005. Losing pieces of the puzzle: threats to marine, estuarine, and diadromous species. Front. Ecol. Environ. **3**: 275-282.

Mace, G.M., Collar, N.J., Gaston, K.J., Hilton-Taylor, C., Akcakaya, H.R., Leader-Williams, N., Milner-Gulland, E.J., and Stuart, S.N. 2008. Quantification of extinction risk: IUCN's system for classifying threatened species. Conserv. Biol. **22**: 1424-1442.

Mace, P.M. et al. 2002. NMFS/Interagency working group evaluation of CITES criteria and guidelines. NOAA Technical Memorandum NMFS-F/SPO-58.

Mapstone, B.D. 1995. Scalable decision rules for environmental impact studies: effect size, type I, and type II errors. Ecol. Appl. **5**: 401-410.

Matsuda, H., Takenaka, Y., Yahara, T., and Uozumi, Y. 1998. Extinction risk assessment of declining wild populations: the case of the southern bluefin tuna. Res. Popul. Ecol. **40**: 271–278.

McElhany, P., Rucklelshaus, M.H., Ford, M.J, Wainwright, T.C., and Bjorkstedt, E.P. 2000. Viable salmonid populations and the recovery of evolutionarily significant units. U.S. Dept. Commer., NOAA Technical Memorandum NMFS-NWFSC-42.

Musick, J.A. 1999. Criteria to define extinction risk in marine fishes - the American Fisheries Society initiative. Fisheries, **24**: 6-14.

Nakagawa, S., and Freckleton, R.P. 2008. Missing inaction: the dangers of ignoring missing data. Trends Ecol. Evol. **23:** 592-596.

Nelitz, M., Murray, C., Porter, M. and Marmorek, D.R. 2006. Managing Pacific salmon for ecosystem values: Ecosystem indicators and the wild salmon policy. Prepared for the Pacific Fisheries Resource Conservation Council by ESSA Technologies Ltd. Available from www.fish.bc.ca.

Northcote, T.G., and Larkin, P.A. 1989. The Fraser River: a major salmonine production system. *In* Proceedings of the International Large River Symposium, Honey Harbor, Ont., Canada, September 14–21, 1986. *Edited by* D.P. Dodge. Can. Spec. Publ. Fish. Aquat. Sci. No. 106. pp. 172–204.

Paulsen, C.M., Hinrichsen, R.A., and Fisher, T.R. 2007. Measure twice, estimate once: Pacific salmon population viability analysis for highly variable populations. T. Am. Fish. Soc. **136**: 346-364.

Pauly, D. 1995. Anecdotes and the shifting baseline syndrome of fisheries. Trends Ecol. Evol. **10**: 430.

Pearson, R.G. 2007. Species' distribution modeling for conservation educators and practitioners. Synthesis. American Museum of Natural History. Available from http://ncep.amnh.org.

Pestal, G.P., and Cass, A. 2009. Using qualitative risk evaluations to prioritize resource assessment activities for Fraser River sockeye. CSAS Research Document 2009/###. To be released on www.dfo-mpo.gc.ca/csas/index.htm.

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

Piet, G.J., and Rice, J.C. 2004. Performance of precautionary reference points in providing management advice on North Sea fish stocks. ICES J. Mar. Sci. **61**: 1305-1312.

Powles, H., Bradford, M.J., Bradford, M.R.G., Doubleday, M.W.G., Innes, M.S., and Levings, C.D. 2000. Assessing and protecting endangered marine species. ICES J. Mar. Sci. **57**: 669-676.

Punt, A.E. 2000. Extinction of marine renewable resources: a demographic analysis. Popul. Ecol. **42**: 19-27.

Rand, P.S. 2008. *Oncorhynchus nerka*. In: IUCN 2008. 2008 IUCN Red List of Threatened Species. Available from http://www.iucnredlist.org.

Rice, J.C., and Legace, E. 2007. When control rules collide: a comparison of fisheries management reference points and IUCN criteria for assessing risk of extinction. ICES J. Mar. Sci. **64**: 718-722.

Ricker, W.E. 1997. Cycles of abundance among Fraser River sockeye salmon (*Oncorhynchus nerka*). Can. J. Fish. Aquat. Sci. **54**: 950-968.

Sadovy, Y. 2001. The threat of fishing to highly fecund fishes. J. Fish Biol. **59**: 90-108.

Schnute, J.T., Cass, A., and Richards, L.J. 2000. A Bayesian decision analysis to set escapement goals for Fraser River sockeye salmon (*Oncorhynchus nerka*). Can. J. Fish. Aquat. Sci. **57**: 962-979.

Venables, W.N., and Ripley, B.D. 2002. Modern Applied Statistics with S, 4th edition. Springer Science +Business Media, LLC., New York. pp. 156-163.

Vida, S. 1993. A computer program for non-parametric receiver operating characteristic analysis. Comput. Meth. Prog. Bio. **40**: 95-101.

APPENDIX

Descriptions of decline criteria

Below are descriptions of all combinations of types of spawner abundance estimates, periods over which the decline was estimated, historical baselines used for comparison, if any, and measures of decline that resulted in the 20 criteria evaluated here as indicators of declining CUs. The types of spawner abundance measures were untransformed (raw) spawner abundances, spawner abundance estimated by the linear time-trend model fit to unsmoothed and smoothed (running generational (4-year) mean) $\log_e$ (spawner abundance) data, or geometric mean of observed abundance of a generation using unsmoothed and smoothed $\log_e$ (spawner abundance) data. The change in spawner abundance was estimated over two time periods; over the short term, it was estimated as the recent rate of decline over spans of three generations. Over the long term, the change in abundance was estimated as either the annual rate of change over all available years, or as the extent of decline from one of five estimated historical baselines. These definitions further elaborate on the descriptions of criteria found in Table 3.

1. Annual rate of decline between spawner abundance estimates three generations apart, estimated by exponentiation of best-fit values from the robust regression of $\log_e$ (abundance) on years.

2. Annual rate of decline between spawner abundance estimates three generations apart, estimated by exponentiation of best-fit values from the robust regression of the running mean of $\log_e$ (abundance) on years.
3. Percent decline between abundance in first year of data series and abundance in all subsequent years (at least 9 years later), using values estimated by exponentiation of best-fit values from the robust regression of $\log_e$ (abundance) on years.
4. Percent decline between abundance in first year of data series and abundance in all subsequent years (at least 9 years later), using values estimated by exponentiation of best-fit values from the robust regression of the running mean of $\log_e$ (abundance) on years.
5. Percent decline between maximum abundance in time series and abundance in all subsequent years (at least 9 years later), using values estimated by exponentiation of best-fit values from the robust regression of $\log_e$ (abundance) on years.
6. Percent decline between maximum abundance in time series and all subsequent years (at least 9 years later), using values estimated by exponentiation of best-fit values from the robust regression of the running mean of $\log_e$ (abundance) on years.
7. Percent decline between maximum abundance of first five data points (not first five years) and all subsequent years (at least 9 years later), using values estimated by exponentiation of best-fit values from the robust regression of $\log_e$ (abundance) on years.
8. Percent decline between the maximum abundance of the first five data points (not first five years) and all subsequent years (at least 9 years later), using values estimated by exponentiation of the best-fit values from the robust regression of the running mean of $\log_e$ (abundance) on years.
9. Percent decline between the geometric mean abundance of the first generation and all subsequent raw spawner abundance values (at least 9 years later).

10. Percent decline between geometric mean abundance of the first generation and the geometric mean abundance of all subsequent generations, where generations move in sliding windows.
11. Percent decline between geometric mean abundance (using running mean of $\log_e$ (abundance)) of the first generation and the geometric mean abundance of all subsequent generations, where generations move in sliding windows.
12. Percent decline between geometric mean abundance of the first generation and the geometric mean abundance of all subsequent generations, where generations move in blocks with no overlap of years.
13. Percent decline between geometric mean abundance (using running mean of $\log_e$ (abundance)) of first generation and the geometric mean abundance of all subsequent generations, where generations move in blocks with no overlap of years.
14. Percent decline between maximum geometric mean abundance of any three-generation period and all subsequent raw spawner abundance values (at least 9 years later).
15. Percent decline between maximum geometric mean abundance of any three-generation period and geometric mean abundance of all subsequent generations, where generations move in sliding windows.
16. Percent decline between maximum geometric mean abundance of any three-generation period (using running mean of $\log_e$ (abundance)) and geometric mean abundance of all subsequent generations, where generations move in sliding windows.
17. Percent decline between maximum geometric mean abundance of any three-generation period and geometric mean abundance of all subsequent generations, where generations move in blocks with no overlap of years.
18. Percent decline between maximum geometric mean abundance of any three-generation period (using running mean of $\log_e$ (abundance)) and geometric mean abundance of all following generations, where generations move in blocks with no overlap of years.

19. Annual rate of decline, estimated by exponentiation of best-fit values from the robust regression of $\log_e$ (abundance) on years, over all years of data (from first corresponding cycle year) up to year of analysis.
20. Annual rate of decline, estimated by exponentiation of best-fit values from the robust regression of the running mean of $\log_e$ (abundance) on years, over all years of data (from first corresponding cycle year) up to year of analysis.

TABLES

Table 1: Four possible outcomes and their probabilities of occurrence, depending on the status estimated by a conservation criterion (such as a criterion measuring rate of change in spawner abundance over the most recent three generations) and the true status of a conservation unit (CU) of Fraser River sockeye salmon in any given year. The null hypothesis of the status estimated by the conservation criterion is that spawner abundance in the CU is not declining (i.e., H_o = no decline). A "triggering event" would occur when the conditions of the conservation criterion were met (e.g., a decline greater than 50% over the most recent three generations) and resulted in an estimated status of declining (top right of table). The true status of a CU depends on its spawner abundance in the years subsequent to the year when the conservation criterion was evaluated, with a true declining CU having a downward future trajectory in spawner abundance and a non-declining CU having a constant or increasing future trajectory in spawner abundance. Probabilities of occurrence of the four types of outcomes are indicated in parentheses.

		Status estimated by applying the conservation criterion	
		H_o = true **Not declining** (no triggering event)	**H_o = false** **Declining** (triggering event)
True status	**Not declining** (no downward future trajectory)	3. True negative (1-α)	2. False positive (type I error, α)
	Declining (downward future trajectory)	4. False negative (type II error, β)	1. True positive (Power = 1-β)

Table 2: Groupings of spawner abundance data for 18 conservation units (CUs) of sockeye salmon within the Fraser River watershed in British Columbia. This subset of 18 Fraser River sockeye CUs covers all four run-timing groups that occur in the Fraser River, referred to as Early Stuart (EStu), Early Summer (ES), Summer (S), and Late Summer (L). Each of the 18 CUs is composed of a combination of a geographic location and run-timing group. An aggregate CU consisting of Chilko-ES and Chilko-S was used because annual spawner abundance estimates for all sites within both CUs have been collected in one mark-recapture study since 1990. We obtained the escapement data for all CUs from Tracy Cone (DFO, pers. comm.), with the exception of Fraser-ES, Quesnel-S, and Mckinley-S, which came from the Salmon Escapement Database System (nuSEDS) (DFO 2006), maintained by DFO (Erik Grundmann, DFO, Pacific Biological Station, 3190 Hammond Bay Road, Nanaimo, B.C., V9T 6N7, pers. comm.).

Run-timing group	**Conservation unit**
Early Stuart (EStu)	• Stuart-EStu • Takla-Trembleur-EStu
Early Summer (ES)	• Chilliwack-ES • Taseko-ES • Nahatlatch-ES • Fraser-ES • Kamloops-ES • Pitt-ES • Shuswap Complex-ES
Summer (S)	• Chilko aggregate (Chilko-ES and Chilko-S) • Takla-Trembleur-S • Fraser-S • Stuart-S • Quesnel-S • Mckinley-S
Late Summer (L)	• Lillooet-L • Cultus-L • Seton-L

Table 3: A summary of all combinations of types of spawner abundance estimates, periods of decline, historical baselines, and measures of decline that resulted in the 20 criteria evaluated here as indicators of declining abundances in salmon conservation units (CUs). Numbers are labels for the criteria. Criteria 10, 11, 15 and 16 are those that compare an annually sliding window of generations to the historical baseline, while criteria 12, 13, 17 and 18 compare the historical baseline to non-overlapping 4-year generational blocks. Criteria 19 and 20 measure the extent of decline in spawner abundance from the first corresponding cycle year in the time series, as opposed to the first year of the time series (i.e., criteria 3 and 4). The Appendix provides a more detailed description of all criteria.

Type of spawner abundance		Measures of decline	Periods over which decline was calculated					
			Rate of decline over the most recent 3 generations	Definitions of historical baseline				
				1st year in data series	Maximum abundance of first five data points	Geometric mean abundance of first generation	Maximum abundance of all years	Geometric mean abundance of maximum 3-generation period
Natural log abundances	Unsmoothed	% change in spawner abundance estimated by linear model using robust regression	**1.**	**3.** **19.** Historical baseline is 1st corresponding cycle year	**7.**		**5.**	
		% change between geometric means of generations				**10.** Sliding window **12.** Generation blocks		**15.** Sliding window **17.** Generation blocks
	Smoothed via a 4-year running mean	% change in spawner abundance estimated by linear model using robust regression	**2.**	**4.** **20.** Historical baseline is 1st corresponding cycle year	**8.**		**6.**	
		% change between geometric means of generations				**11.** Sliding window **13.** Generation blocks		**16.** Sliding window **18.** Generation blocks
Raw		Decline in abundances from historical baseline				**9.**		**14.**

Table 4: The eight scenarios resulting from variations in the definition of the true status of Fraser River sockeye salmon conservation units (CUs) as declining. For any given year, designation of the true status of a CU as declining occurred when a decline greater than or equal to the stated percentage occurred over the stated future period. We determined the area under the Receiver Operating Characteristic curve for each conservation threat criterion within each scenario.

Scenario	Future period that defines true status	Percent decline over future period that defines true status
A	3 generations	30%
B	3 generations	50%
C	3 generations	70%
D	3 generations	90%
E	All remaining years in time series	30%
F	All remaining years in time series	50%
G	All remaining years in time series	70%
H	All remaining years in time series	90%

Table 5: Best-quality data sites (spawner abundance estimated from fence counts or mark-recapture) (left column) and the Fraser River sockeye salmon conservation units (CUs) included in an aggregate unit similar to IUCN subpopulation #68 as defined by Rand (2008), (right column).

Best-quality data sites	CUs included in aggregate similar to IUCN subpopulation #68
• Pitt-ES • Chilko aggregate (Chilko-ES and Chilko-S) • Fraser-S • Horsefly River (from Quesnel-S) • Cultus-L • Birkenhead River (from Lillooet-L)	• Takla/Trembleur-EStu • Stuart-EStu • Taseko-ES • Nahatlatch-ES • Fraser-ES • Chilko aggregate (Chilko-ES and Chilko-S) • Takla/Trembleur-S • Stuart-S • Fraser-S • Quesnel-S • Seton-L

Table 6: Five aggregate populations consisting of the 18 Fraser River sockeye salmon conservation units (CUs) from our analysis grouped into their corresponding run-timing groups, as well as a population aggregate similar to that used in the recent IUCN Red List Assessment for sockeye salmon (the IUCN subpopulation #68, Rand 2008). For each aggregate, the table shows the number of years in which criterion 13, which is one of the top-ranked criteria in our retrospective analyses, would have measured a decline in spawner abundance large enough to result in a classification of threatened or endangered based on the thresholds of decline used by COSEWIC. A rating of threatened or endangered occurred if the decrease in abundance over the period measured by the criterion was greater than or equal to 30% or 50%, respectively. The number of years in which each population aggregate was classified as threatened or endangered (column 3) is compared to the number of years when individual CUs within the population aggregates would have been classified as threatened (column 4) or endangered (column 5) according to the thresholds of decline used by COSEWIC.

Population aggregate	**CUs in aggregate**	**Number of years *aggregate* was classified as threatened or endangered**	**Number of years any *individual* CUs within aggregate were classified as threatened**	**Number of years any *individual* CUs within aggregate were classified as endangered**
IUCN subpopulation #68	11 CUs (Table 5)	0	1 (1 CU threatened)	7 (1 CU endangered)
Early Stuart run-timing group	Stuart-EStu Takla-Trembleur-EStu	0	0	0
Early Summer run-timing group	Chilliwack-ES Taseko-ES Nahatlatch-ES Fraser-ES Kamloops-ES Pitt-ES Shuswap Complex-ES	0	2 (1 CU threatened)	6 (1 CU endangered)
Summer run-timing group	Chilko aggregate Takla-Trembleur-S Fraser-S Stuart-S Quesnel-S Mckinley-S	0	0	0
Late Summer run-timing group	Lillooet-L Cultus-L Seton-L	1 (aggregate threatened)	3 (1 CU threatened)	10 (1 CU endangered) 2 (2 CUs endangered)

FIGURES

Figure 1: Two hypothetical Receiver Operating Characteristic (ROC) curves. Each point on an ROC curve is the true positive rate (power, probability of avoiding a type II error) and the false positive rate (α, probability of a type I error) for that criterion at a given threshold (e.g., percent change in abundance that results in the classification of a conservation unit as declining). The area under the ROC curve (AUC) is the probability of that criterion correctly distinguishing between cases that differ in their true status (e.g., true status of declining or not declining). The dashed line is the line of equality where the true positive rate equals the false positive rate. One curve (line with solid diamonds) has a large AUC (0.94), indicated by the curve passing near the upper-left corner of the plot. The second curve (line with squares) has an AUC of 0.5 and roughly follows the line of equality, which indicates that the criterion is no better than a coin toss at correctly categorizing the status of CUs.

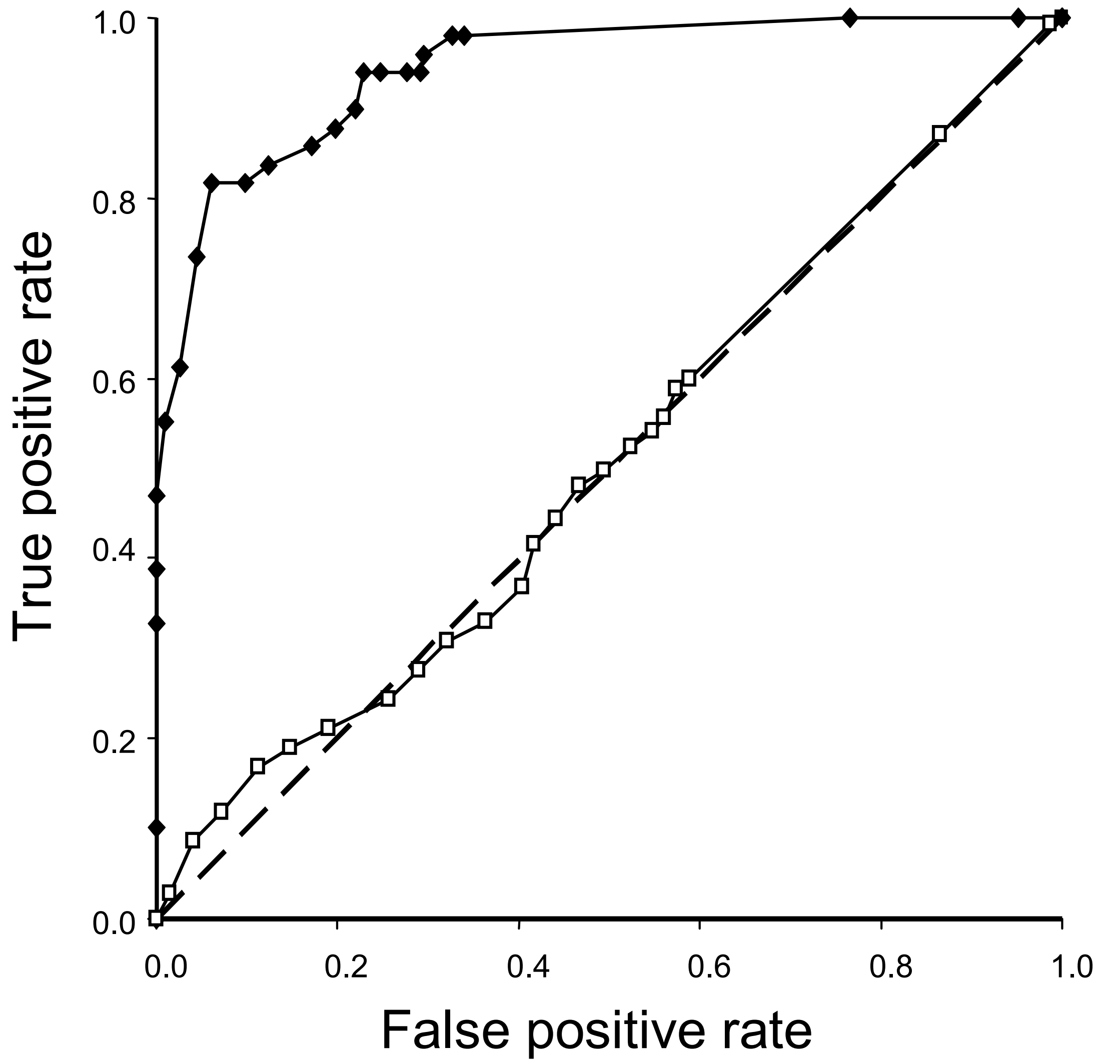

1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

Figure 2: Exploitation rates on adult Fraser River sockeye salmon from 1952-2006 (Mike Lapointe, Pacific Salmon Commission, pers. comm.). In 1995 (dotted line), there was a change in salmon management within the Fraser River, when the sockeye exploitation rate across all run-timing groups substantially decreased from previous years. (a) Exploitation rate on total number of adult Fraser River sockeye salmon across all four run-timing groups. (b) Exploitation rate on each run-timing group of Fraser River sockeye salmon: Early Stuart (solid diamonds), Early Summer (squares), Summer (triangles), and Late Summer (circles).

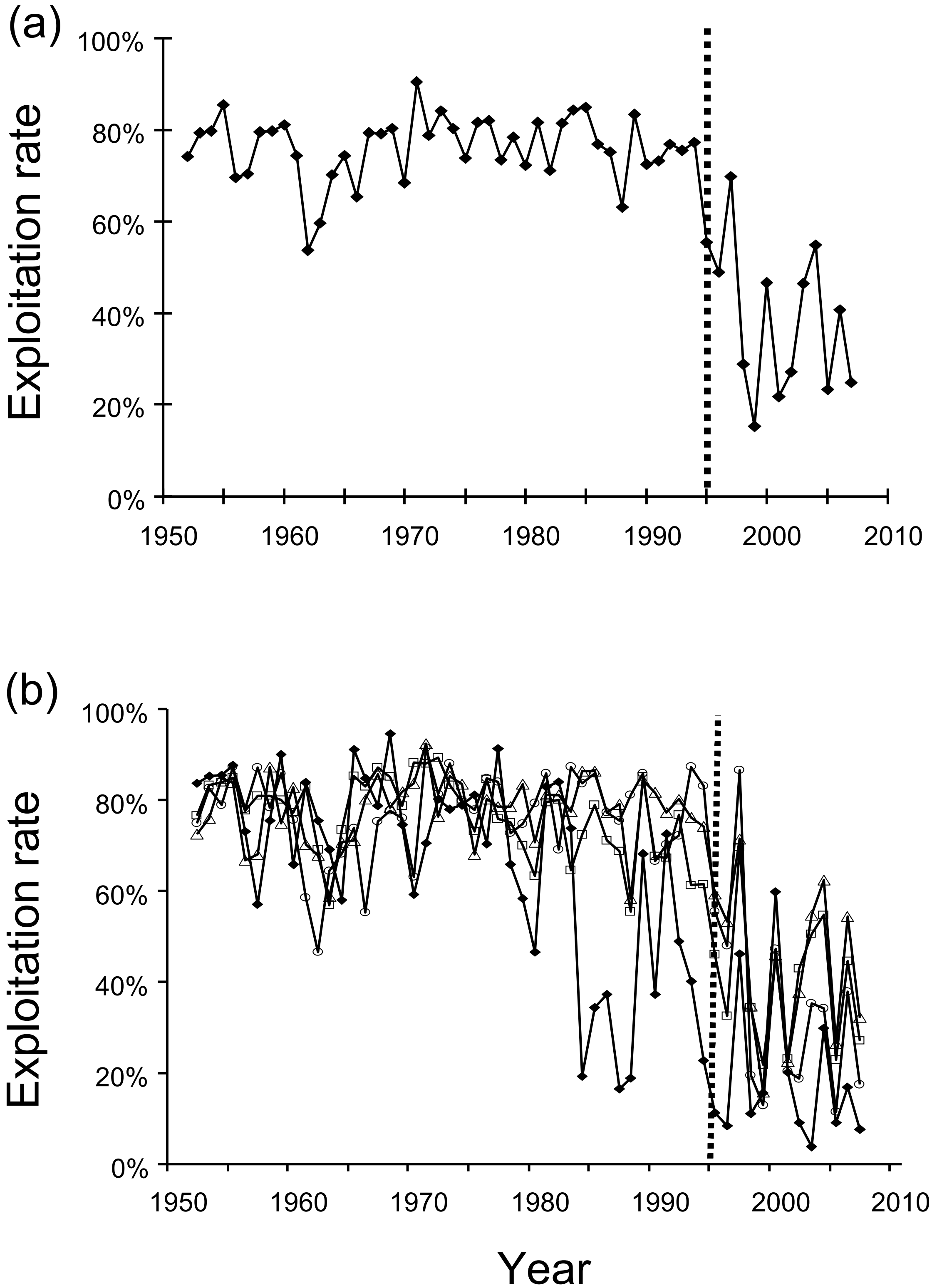
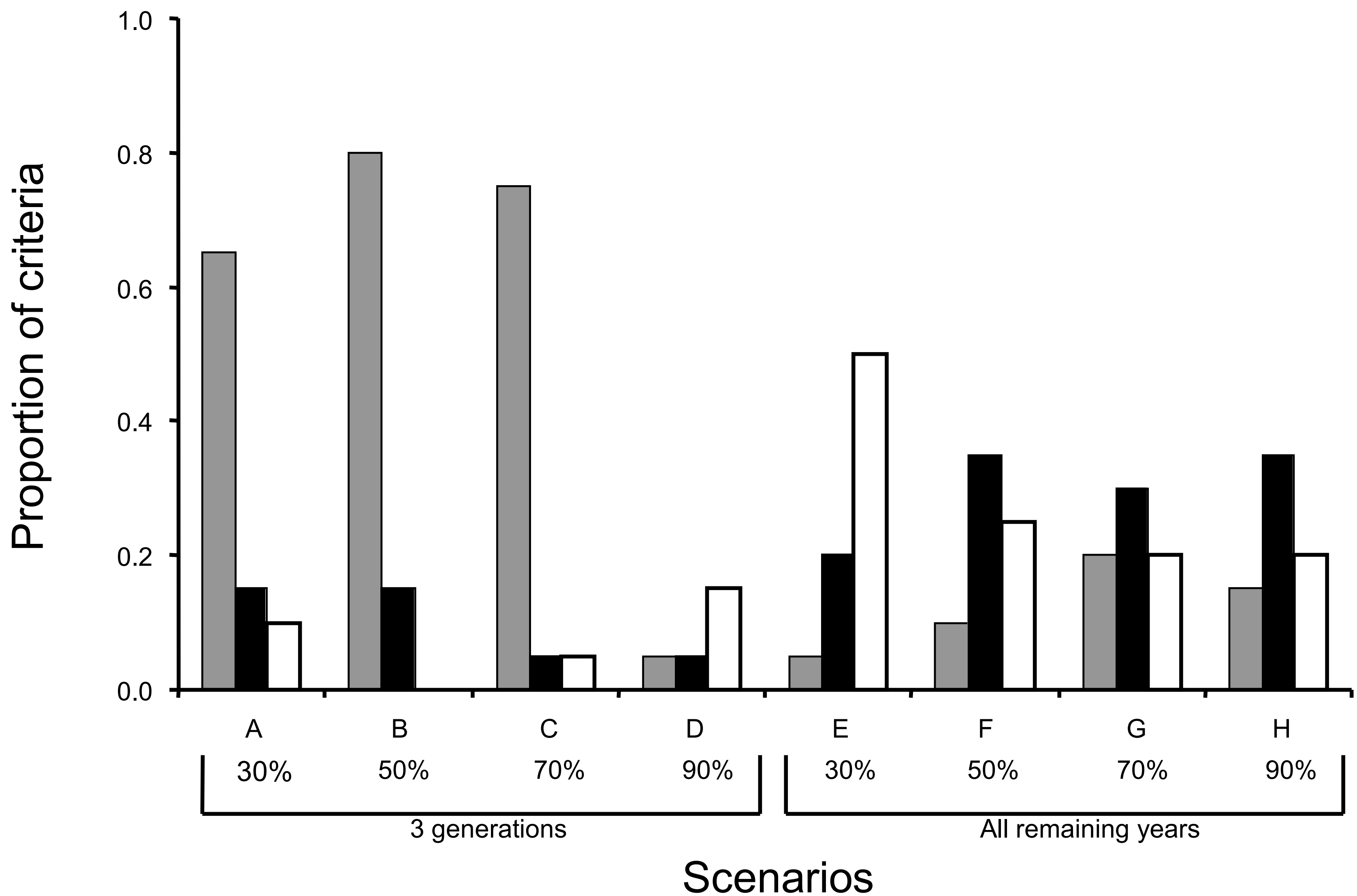
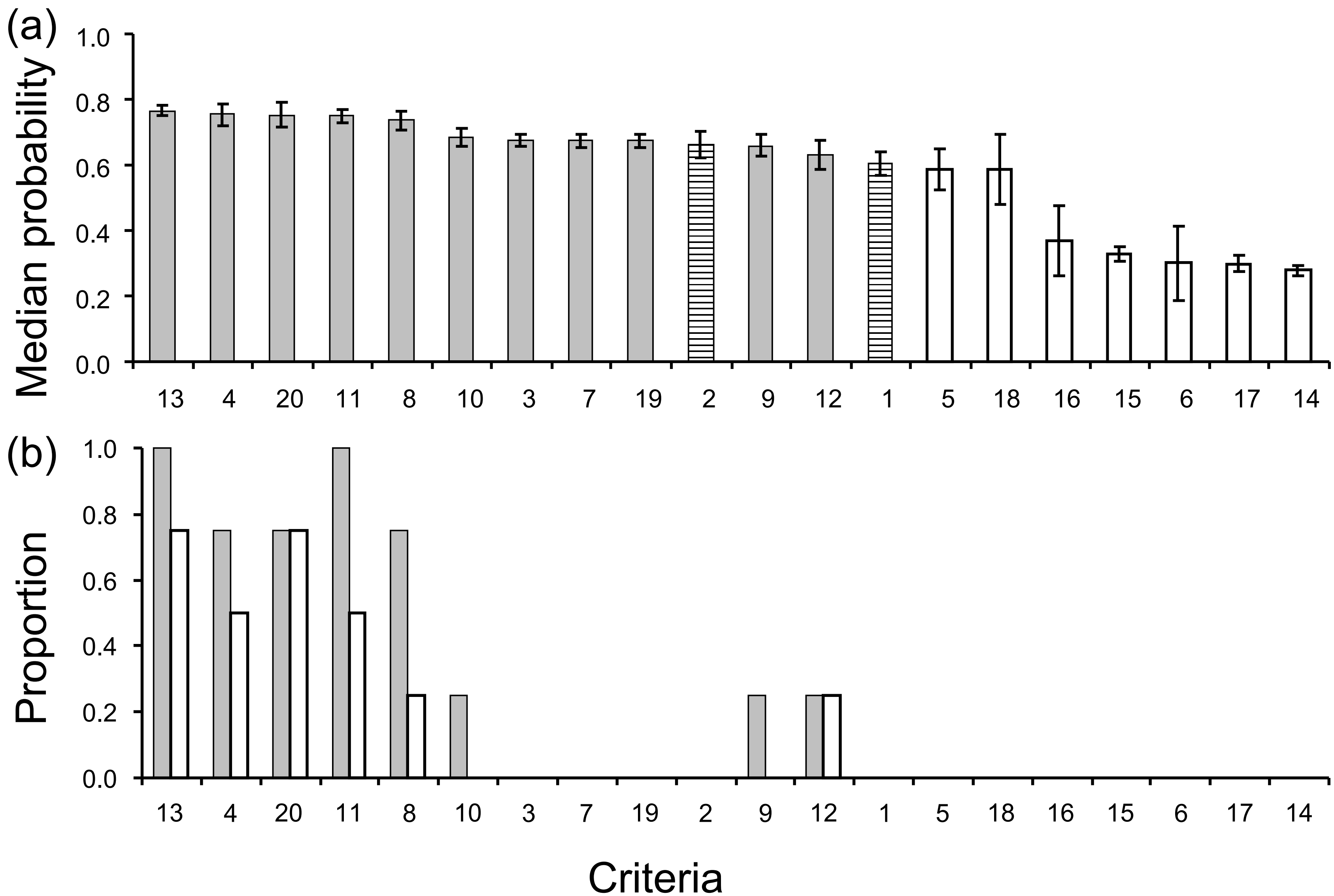

(a)
Exploitation rate
0%
20%
40%
60%
80%
100%
1950
1960
1970
1980
1990
2000
2010
(b)
Exploitation rate
0%
20%
40%
60%
80%
100%
1950
1960
1970
1980
1990
2000
2010
Year

Figure 3: Proportion of the 20 conservation-status criteria within each of the eight scenarios that produced AUC values between 0.5 and 0.6 (grey bars), between 0.6 and 0.7 (black bars), or greater than or equal to 0.7 (white bars). The eight scenarios that differ in their definition of a true declining CU are depicted on the X-axis (A-H) and reflect different future periods over which the decline is measured (subsequent 3 generations or all subsequent years of data) and the percent decline over that future period (30%, 50%, 70%, and 90%) that is required to trigger a true status of “declining”.



Proportion of criteria
1.0
0.8
0.6
0.4
0.2
0.0
A
B
C
D
E
F
G
H
30%
50%
70%
90%
30%
50%
70%
90%
3 generations
All remaining years
Scenarios

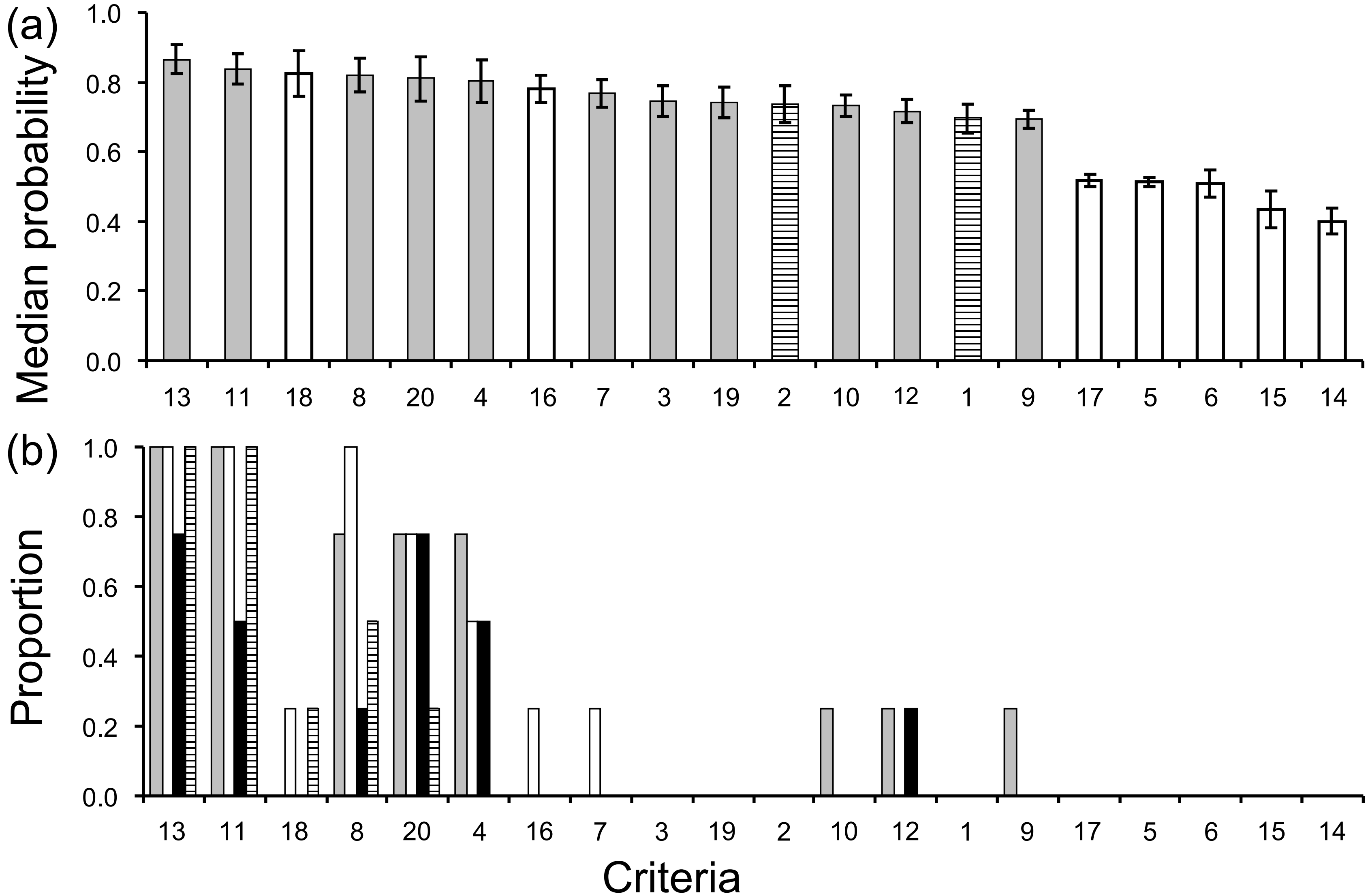
Figure 4: (a) Median probability (and its standard error) that a criterion of population decline can correctly identify whether a Fraser River sockeye conservation unit (CU) is actually declining or not. This probability was measured by the median area under the Receiver Operating Characteristic curve (AUC) of each conservation-status criterion across scenarios E-H. Those scenarios used all remaining years to determine true status of the CUs. Grey bars are for criteria that measure the extent of decline from a historical baseline anchored at the beginning of the time series, bars with horizontal lines are for criteria that measure the rate of decline over three generations, and white bars are for criteria that measure the extent of decline from the maximum abundance in the time series. Criteria are defined by number in Table 3 and the Appendix. (b) Proportion of scenarios E-H in which each conservation-status criterion ranked in either the top three (white bars) or top five (grey bars) out of the 20 criteria, based on AUC values (criterion with higher AUC has a higher rank).



(a)
Median probability
1.0
0.8
0.6
0.4
0.2
0.0
13 4 20 11 8 10 3 7 19 2 9 12 1 5 18 16 15 6 17 14
(b)
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
13 4 20 11 8 10 3 7 19 2 9 12 1 5 18 16 15 6 17 14
Criteria

Figure 5: (a) Median probability (and its standard error) that a criterion of population decline can correctly identify whether a Fraser River sockeye conservation unit (CU) is actually declining or not, when the analysis was limited to pre-1995 years. This probability was measured by the median area under the Receiver Operating Characteristic curve (AUC) of each conservation-status criterion across scenarios E-H, using only pre-1995 data. Those scenarios used all remaining years to determine true status of the CUs. Grey bars are for criteria that measure the extent of decline from a historical baseline anchored at the beginning of the time series, bars with horizontal lines are for criteria that measure the rate of decline over three generations, and white bars are for criteria that measure the extent of decline from the maximum abundance in the time series. Criteria are defined by number in Table 3 and the Appendix. (b) Proportion of scenarios E-H in which each conservation-status criterion ranked in either the top three or top five out of the 20 criteria, based on AUC values (criterion with higher AUC has a higher rank), when the analysis used either all years of data or only pre-1995 years. Grey bars are for criteria ranked in the top five using all years of data and white bars are for criteria ranked in the top five using pre-1995 years. Black bars are for criteria ranked in the top three using all

years of data and bars with horizontal stripes are for criteria ranked in the top three using pre-1995 years.



(a)
Median probability
1.0
0.8
0.6
0.4
0.2
0.0
13 11 18 8 20 4 16 7 3 19 2 10 12 1 9 17 5 6 15 14
(b)
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
13 11 18 8 20 4 16 7 3 19 2 10 12 1 9 17 5 6 15 14
Criteria

Figure 6: (a) Median probability (and its standard error) that a criterion of population decline can correctly identify whether a Fraser River sockeye conservation unit (CU) is actually declining or not, when the analysis was limited to the best-quality data only (Table 5). This probability was measured by the median area under the Receiver Operating Characteristic curve (AUC) of each conservation-status criterion across scenarios E-H, using only best-quality data. Those scenarios used all remaining years to determine true status of the CUs. Grey bars are for criteria that measure the extent of decline from a historical baseline anchored at the beginning of the time series, bars with horizontal lines are for criteria that measure the rate of decline over three generations, and white bars are for criteria that measure the extent of decline from the maximum abundance in the time series. Criteria are defined by number in Table 3 and the Appendix. (b) Proportion of scenarios E-H in which each conservation-status criterion ranked in either the top three or top five out of the 20 criteria, based on AUC values (criterion with higher AUC has a higher rank), when the analysis used either all data or only best-quality data. Grey bars are for criteria ranked in the top five using all data and white bars are for criteria ranked in the top five using best-quality data. Black bars are for criteria ranked in the top three using

all data and bars with horizontal stripes are for criteria ranked in the top three using best-quality data.

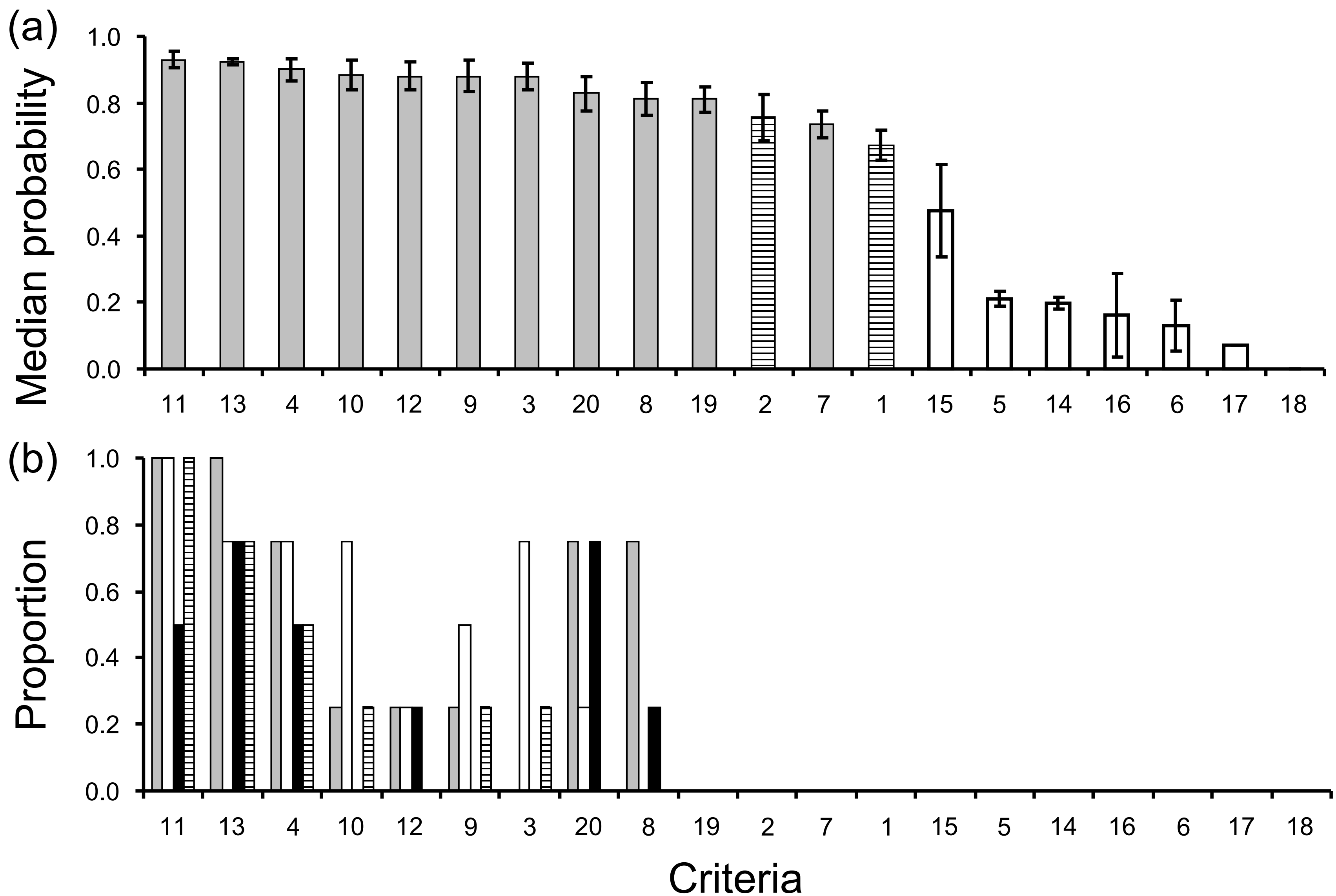
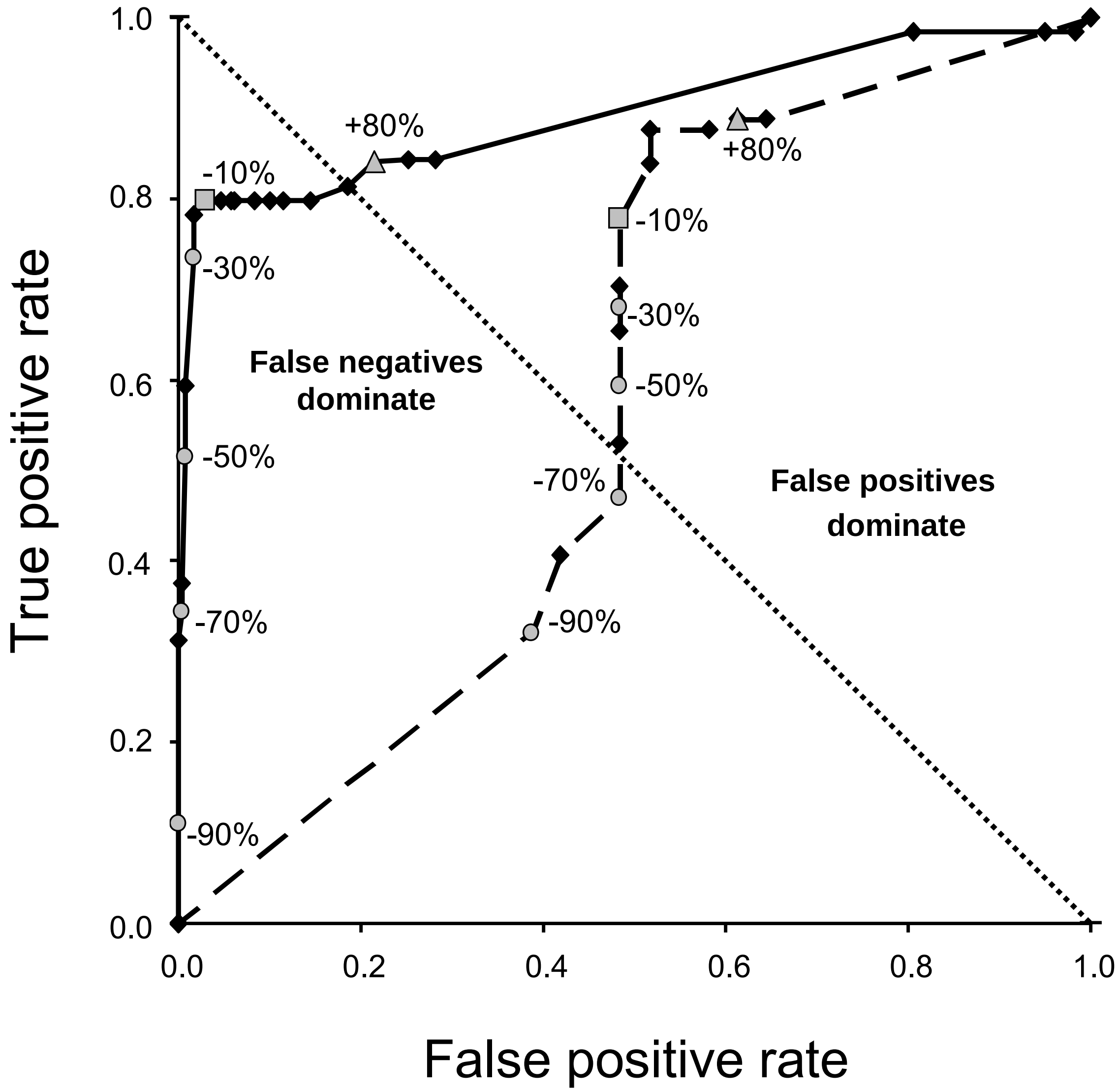
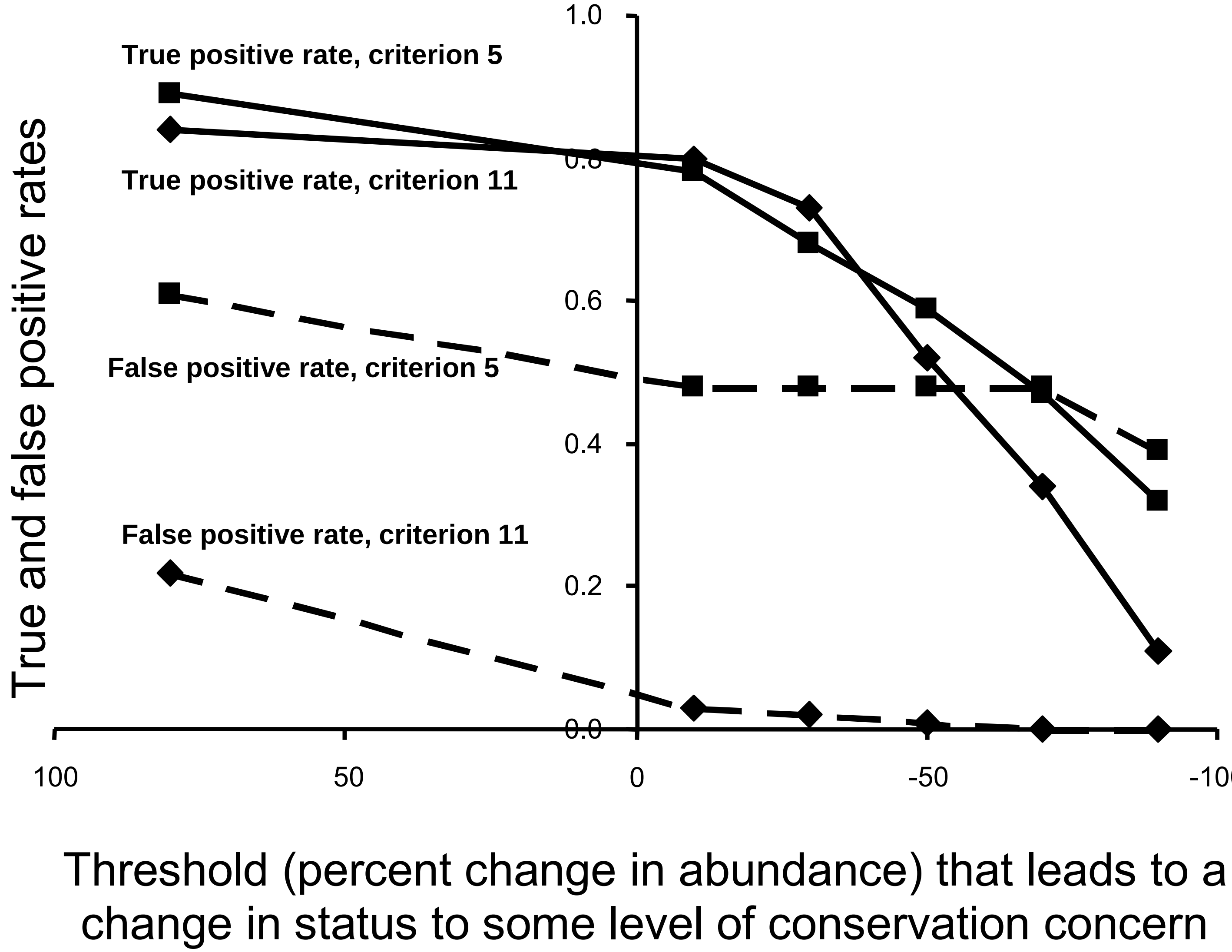

(a)
Median probability
1.0
0.8
0.6
0.4
0.2
0.0
11 13 4 10 12 9 3 20 8 19 2 7 1 15 5 14 16 6 17 18
(b)
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
11 13 4 10 12 9 3 20 8 19 2 7 1 15 5 14 16 6 17 18
Criteria

Figure 7: Receiver Operating Characteristic (ROC) curves for criterion 11 (solid line), which used only best-quality data, and criterion 5 (dotted line), which used all data, in scenario E where a true declining CU would experience a decline greater than or equal to 30% over all subsequent years; area under the curves (AUCs) are 0.9 and 0.58, respectively. For both curves, the four circles, moving up from the bottom-left corner are, respectively, for a 90%, 70%, 50% and 30% decrease in abundance over the period that causes a triggering event for a conservation concern. The square is for using a 10% decrease in abundance and the triangle an 80% *increase* in abundance over the period to trigger a conservation concern. The circle, square, and triangle points are labelled with their corresponding thresholds of change in abundance. The dashed diagonal line indicates where the proportion of false positives (type I errors) equals the proportion of false negatives (type II errors); the latter occur with a probability of β (=1-true positive rate or power). Points to the right of the line are thresholds of change in abundance where there are greater proportions of false positives than false negatives. Points to the left of the line are thresholds of change in abundance where there are greater proportions of false negatives.



1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
True positive rate
False positive rate
+80%
-10%
-30%
-50%
-70%
-90%
+80%
-10%
-30%
-50%
-70%
-90%
False negatives dominate
False positives dominate

Figure 8: The true positive rate (power, probability of avoiding a type II error) and the false positive rate (α, probability of a type I error) corresponding to various thresholds of change in abundance that result in a change of classification of conservation status, for both criterion 11 (area under the Receiver Operating Characteristic Curve (AUC) of 0.9) using best-quality data and criterion 5 (AUC of 0.58) using all data in scenario E, which defines a declining CU as one in which spawner abundance decreases by greater than or equal to 30% over all subsequent years. The solid line with diamonds is for the true positive rate of criterion 11 and the dashed line with diamonds is for the false positive rate of criterion 11. The solid line with squares is the true positive rate of criterion 5 and the dashed line with squares is for the false positive rate of criterion 5. The threshold of change in abundance ranges from a 100% increase in abundance (i.e., 100) to a 100% decrease in abundance (i.e., -100) over the period measured by the conservation criterion.


True positive rate, criterion 5
True positive rate, criterion 11
False positive rate, criterion 5
False positive rate, criterion 11
True and false positive rates
1.0
0.8
0.6
0.4
0.2
0.0
100
50
0
-50
-100
Threshold (percent change in abundance) that leads to a change in status to some level of conservation concern