19 June 2014

Reliable Identification of Declining Populations in an Uncertain World

Faye d'Eon-Eggertson[1†], Nicholas K. Dulvy[2], and Randall M. Peterman[1]

[1] School of Resource and Environmental Management, Simon Fraser University, Burnaby, BC, Canada

[2] Earth to Ocean Research Group, Department of Biological Sciences, Simon Fraser University, Burnaby, BC, Canada

[†] Corresponding author: fdeonegg@sfu.ca, (778) 686-8216

Co-authors: dulvy@sfu.ca; peterman@sfu.ca

Keywords: Decline indicators; error rates; IUCN; Monte Carlo simulation; Receiver Operating Characteristic; extinction-risk assessment; process variation; observation error

Article type: Letter

Running title: Identifying Declines in an Uncertain World

Abstract: 157 words; Main text: 3,960 words in the Introduction through Discussion

References: 36; Figures: 4; Tables: 2

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/conl.12123.

Abstract:

Assessments of extinction risk based on population declines are widely used, yet scientists have little quantitative understanding of their reliability. Incorrectly classifying whether a population is declining or not can lead to inappropriate conservation actions or management measures, with potentially profound societal costs. Here we evaluate key causes of misclassification of decline status and assess the reliability of 20 decline metrics using a stochastic model to simulate time series of population abundance of sockeye salmon (*Oncorhynchus nerka*). We show that between-year variability in population productivity (process variation) and, to a lesser extent, variability in abundance estimates (observation error) are important causes of unreliable identification of population status. We found that using all available data, rather than just the most recent three generations, consistently improved the reliability of risk assessments. The approach outlined here can improve understanding of the reliability of risk assessments, thereby reducing concerns that may impede their use for exploited taxa such as marine fishes.

Introduction

The IUCN Red List categories and criteria (IUCN 2013) are among the most widely used methods for classifying the extinction risk of various species, with more than 70,000 species assessed to date. However, the validity of such assessments is often called into question when the conservation imperative conflicts with a desire to continue exploiting species, such as marine fishes (Mace and Hudson 1999). Extinction risk criteria were originally developed for terrestrial species and the controversy over their application to marine fishes may have hindered aquatic conservation (e.g., Powles et al. 2000; Reynolds et al. 2005; Vincent et al. 2013). Some simulation analyses have shown that IUCN decline

criteria may raise false alarms by overstating the risk of extinction of widely distributed marine fishes such as bluefin tuna (Matsuda et al. 1998; Rice & Legacé 2007). However, recent empirical analyses of the reliability of various risk criteria show that false alarms (the incorrect classification of a species as threatened when it is not) is quite low for many commercially-exploited marine fishes and Pacific salmon (Dulvy et al. 2005; Porszt et al. 2012).

Now the key challenge is to understand the causes of misclassification and to identify the most reliable indicators of risk, that is, those that have the highest probability of correctly reflecting the actual condition of the assessed species or population. One principal concern is that misclassification can arise from between-year environmentally-driven variability in productivity (process variation) and/or census variability arising from imprecise estimates of abundance (observation error) (Wilson et al. 2011; Connors et al. 2014). Here, we build upon past studies and use simulations that take into account these two sources of variation while evaluating the reliability of various population-decline metrics.

We quantify and rank the relative reliability of various decline indicators based on their ability to correctly classify the overall long-term abundance trajectory as decreasing or not. Such evaluations are particularly critical because incorrectly classifying a population as decreasing enough to be declared a concern (false positive, Table 1) will divert limited conservation funds from other, more at-risk populations. Similarly, incorrectly classifying a population as not declining, when it actually is declining (false negative, Table 1), could lead to eventual loss of the population because remedial action is not taken.

We focus on sockeye salmon (*Oncorhynchus nerka*), because its population dynamics are characterized by substantial variability, yet are understood well enough to develop

realistic simulation models. Furthermore, it is a widely distributed fish species of conservation concern in Canada (Irvine et al. 2005). Major decreases in these exploited sockeye populations can be ecologically, economically, and socially important, even if the probability of extinction is low (DFO 2005). Thus, instead of focusing on extinction risk *per se*, we examine the closely related question of whether an assessment of a population decline in a given period appropriately reflects that population's long-term trend, including the period subsequent to that assessment. Our indicators can thus be viewed as reflecting one symptom of extinction risk (i.e., a continuing decrease in abundance) or other serious conservation situations (Mace et al. 2008).

We extend previous simulation analyses (Punt 2000; Holt et al. 2009; Wilson et al. 2011) by (1) estimating rates of misclassification (false negatives and false positives) for a broad range of thresholds of decline that could be required for classifying a population as being at risk, (2) examining 20 indicators of decreasing abundance and their associated methods of data analysis, and (3) exploring unequal weightings of types of classification errors.

We found that in the presence of even small process variation or observation error, indicators of decline in population abundance that use all of the available data are more reliable than other indicators, including the widely used indicator of recent rate of decline in the last 10 years or 3 generations (IUCN 2013). We also show that the ranking of decline indicators can be affected by how decision makers weight the relative importance of avoiding false positives compared to false negatives. The wider application of our approach can help decision makers to more effectively use the restricted funds available for conservation actions.

Methods

Stochastic population dynamics model

We used a stochastic model to simulate the population dynamics of our case example, semelparous sockeye salmon populations in the Fraser River that have a 4-year life cycle:

$$S_{t+4} = aS_t e^{-bS_t+\varepsilon_t}, \qquad (1)$$

where S_t is spawner abundance in year t. This spawner-to-spawner model is analogous to a Ricker (1975) spawner-to-recruit model, but here the a parameter of equation 1 encompasses biological productivity from spawner-to-adult recruits prior to the onset of fishing, as well as subsequent removals by fishing and in-river pre-spawning mortality (online Supporting Information (SI), part A). The density-dependent parameter b was set to 1/100,000 (an arbitrary scalar of unfished equilibrium abundance). Process variation (between-year variability in productivity) was represented by assuming that ε_t in equation (1) is a lag-1-year autocorrelated process:

$$\varepsilon_t = \Phi\varepsilon_{t-1} + u_t, \quad (2)$$

where the latter error term u_t is normally distributed, ~$N(0, \sigma_u^2)$ (Hilborn & Walters 1992). Process variation can differ among sockeye salmon populations, so in separate analyses we examined levels of σ_u^2 of 0.01, 0.05, 0.1, 0.3, and 0.5. For temporal autocorrelation in that

process variation, we examined values of Φ (the autocorrelation coefficient) of -0.5, -0.25, 0, 0.25, 0.5, and 0.75. These ranges for these two parameters encompassed most reported estimates for North American sockeye salmon populations (Korman et al. 1995) as well as other animal species (SI part A).

We also simulated observation error:

$$S_{observed} = S_{t+4}e^{v_t} \qquad (3)$$

where $v_t \sim N(0, \sigma_v^2)$, which reflects the interannual variability in observed abundance due to sampling error (Hilborn & Walters 1992). The σ_v^2 values of 0.0, 0.05, 0.1, 0.3, and 0.5 that we explored included the estimated range of observation error for Pacific salmon (Hilborn & Walters 1992) and other animal species (SI part A).

We simulated population dynamics for 76 years, which consisted of a 12-year initialization period, a 52-year (13-generation) "evaluation period", and finally a "subsequent period" of 12 years (3-generations) (Figure 1a). We ran 500 Monte Carlo trials for each scenario; a scenario consisted of one combination of the magnitude of process variation (σ_u^2) and observation error (σ_v^2). Each population (i.e., each Monte Carlo trial) in each scenario was randomly assigned a productivity parameter value (the *a* in equation 1) drawn from $\sim N(1,0.3)$, representing differences in productivity among different populations (some decreasing and some not due to local variables such as fishing rate, habitat quality, and environmental conditions) (SI part A).

Estimation of indicators during the "evaluation period"

We used 20 indicators of time trends in adult abundance to assess whether our populations were declining or non-declining during the "evaluation period" (Table 2 and SI part B). These indicators consisted of different combinations of three components. First was the time frame examined, either (a) the most recent three generations (i.e., 12 years in our case) in the evaluation period prior to the status-assessment year, (b) the entire historical evaluation period prior to the assessment year, or (c) the time series since maximum abundance, regardless of where it occurred in the evaluation period (Table 2). The second component was the data-handling procedure (smoothed vs. unsmoothed data, and raw data vs. $\log_e$-transformed data vs. means of 4-year generations; Table 2). Third, changes in abundance over time were measured using either a linear regression (used to calculate either a rate or extent of change), or a difference in mean abundance between two periods.

For each relevant simulated year of the 13-generation evaluation period, we determined for each of the 20 indicators whether the indicator classified the current status of the population as "declining" or "non-declining". To do this, in a set of separate calculations, we compared the estimated decrease in abundance against various thresholds that delineated "declining"; those thresholds ranged from 0 to 100% in increments of 1% (SI part B). That status classified during the evaluation period (e.g., Figure 1a) was then compared to the overall long-term status, which included the subsequent period (Figure 1b) as described next.

Overall long-term population trend

We focused on how well the 20 risk indicators reflected the actual overall true long-term trend of each simulated population. To determine whether that long-term trajectory of

the population was declining, we estimated the trend in abundance from the end of the initialization period through the subsequent period, as shown in Figure 1b (also SI part C).

For one set of simulations, if the population decreased more than 90%, which was the base-case boundary condition indicating a conservation concern, then the overall long-term status was classified as "declining"; otherwise it was classified as "non-declining". We also examined other commonly used boundary conditions of 50% and 70% decline (COSEWIC 2011; IUCN 2013).

Measuring reliability

We defined the reliability of an indicator as its probability of correctly distinguishing between declining and non-declining populations. Specifically, for a given indicator of decline and a given scenario, we compared the indicator's assessed status of the population (declining or not) in each year of the evaluation period to the estimated long-term status (as described above), resulting in either a true negative (TN), true positive (TP), false negative (FN) or false positive (FP) outcome (Table 1). For each of the 20 indicators, these different categories of outcomes were tallied across all years and 500 trials. We then used a Receiver Operating Characteristic (ROC) analysis to combine into a single metric for each decline indicator the true and false positive rates that were produced across a wide range of thresholds of decline in abundance for classifying population status, from 0 to 100% in 1% increments (Burgman 2005; Porszt et al. 2012; SI part D). The resulting area under the ROC curve (AUC) reflects the ability of a given indicator to correctly distinguish whether a population is declining; higher AUC values mean greater reliability (Hibberd & Cooper 2008). An AUC value of 0.97 reflects an indicator with a near-perfect ability to distinguish a declining from a non-declining population, whereas AUC = 0.51 (falling near the 1:1 line) indicates about a 50/50 chance of correctly making that distinction (Figure 1c).

Different error weightings

One limitation of an ROC analysis is that it inherently attributes equal weighting to false positive and false negative errors; hence, the importance of avoiding each type of error is the same. That equality is unrealistic in common situations where resource managers must make trade-off decisions regarding conservation and resource-use objectives (Peterman 1990;

Mapstone 1995; Field et al. 2004; Dulvy et al. 2006). If one type of error is considered more serious than the other, then managers can specify a desired relative weighting or ratio of the error rates (for instance, that the more heavily weighted false negative rate must be less than half the false positive rate). Therefore, we also investigated the reliability of the 20 indicators of decline across a wide range of unequal error weightings.

Results

We first describe results for our base case, in which we assumed that the lag-1 autocorrelation coefficient (Φ) in process variation was zero, before showing how our results are affected by a wide range of positive and negative Φ values. At extremely low levels of both process variation and observation error, all 20 indicators were almost equally reliable at discriminating between declining and non-declining populations (AUC values >0.9) (Figure 2a). However, reliability decreased with greater process variation and observation error (compare AUCs of Figures 2b, c, and d with Figure 2a). Indicators that used a historical baseline starting at the beginning of the time series (in particular indicators #4, 6, 9, and 20) consistently outperformed the other two types of indicators (e.g., Figures 2a-d; also see SI part E, Figures SI-1 and SI-2). Even with fairly low process variation or observation error, indicators calculated from a historical baseline were generally more reliable than either (1) indicators with a baseline based on the maximum abundance anywhere in the time series, or (2) indicators based on decline over the most recent 3 generations (Figure 2). The latter group included indicator #2, which is analogous to the commonly used IUCN criterion A measured over only the past three generations. However, the IUCN criterion A metric performs far

better when the full time series is used rather than just the past three generations (indicators #19 and #20).

Indicators varied in how robust they were to increases in process variation, σ_u^2, and observation error, σ_v^2. The AUC of the majority of indicators decreased by ~10% when process variation increased from 0.01 to 0.5, whereas the AUC of indicators #1, 2, 5, 17, and 18 decreased at least twice as much as that (Figures 2a and b, also SI part E, Figures SI-1 and SI-2). The latter group of indicators included ones that either evaluated the rate of decline over the most recent 3 generations (#1 and 2), used raw abundance (#17 and 18), or estimated decline based on the first corresponding population-cycle year (#5). By contrast, most indicators were less sensitive to increases in observation error, exhibiting only slight decreases in reliability (drop in AUC of ~5%) when σ_v^2 was increased from 0 to 0.5 (Figures 2a and c, also SI part E, Figures SI-1 and SI-2). Exceptions to this trend included indicators #1 and 2 (recent rate of decline), which were much more sensitive to observation error, with a drop in AUC of ~20% when σ_v^2 increased from 0 to 0.5 (Figures 2a and c, also SI part E, Figures SI-1 and SI-2).

The AUCs of all indicators were relatively insensitive to changes in the boundary used to classify a population as declining over the long term, i.e., 50%, 70%, or 90% (AUCs changed by <4%), but again, indicators #1 and 2 changed the most (SI part E, Figure SI-3). The AUCs of all indicators declined with increases in the mean *a* parameter, but the rank order of the indicators was relatively insensitive to changes in mean *a* and *b* parameters (SI part E, Figure SI-4).

Conservation objectives, error weightings, and error tolerance

There is an inherent trade-off between false positive rate (FPR) and false negative rate (FNR); both errors cannot be completely avoided (SI part E, Figure SI-5). If managers explicitly specify a desired relative error weighting, then the indicators can be optimized to target this weighting by adjusting the threshold that signals a declining population. The optimal threshold depends on the manager's relative preferences and the indicator being used (SI part E, Figure SI-5). The top-ranked indicators #4, 6, 9, and 20, which used some historical baseline, were best across the majority of error weightings for different magnitudes of process variation (SI part E, Table SI-1a). Indicators 4 and 20 were best for almost all weightings and levels of observation error (SI part E, Table SI-1b). These results still held at high levels of temporal autocorrelation in process variation (SI part E, Tables SI-1c and d).

For cases in which managers have an upper limit to the acceptable false positive error rate, we estimated the false negative rate that would result from the trade-off noted in SI part E, Figure SI-5. Figure 3 shows the effect of variance in observation error (X-axis of panels a and b) and process variation (panels c and d) on this nonlinear trade-off between the acceptable false positive rate (Y-axis) and the resulting false negative rate (contours). The high-ranking indicator #4 reflects the change in abundance from the first generation in the time series (Figures 3a, c), whereas the commonly used indicator #2 measures decline over only the last 3 generations (Figures 3b, d). For example, if managers using that decline indicator #2 decided that it would be unacceptable to have greater than a 0.1 probability of a false positive error (i.e., probability of incorrectly concluding that a population was declining), and if the population had a high level of observation-error variance (say σ_v^2=0.3), then they could expect a false negative rate of about 0.4, as shown on the contour plot (white dot in Figure 3b). This means that if the population is actually not declining, there will be a

10% chance of the indicator reporting that it is declining, but if the population is actually declining, there will be about a 40% chance of the indicator erroneously reporting that it is not declining. If managers want a lower false negative rate, then they have three choices: increase their tolerance for false positives, decrease observation error, or pick a different indicator. As an example of the latter, given the same situation, indicator #4, which reflects change in abundance since a historical baseline, would provide a much more desirable false negative rate of slightly less than 5% (white dot in Figure 3a). Regardless of the magnitude of observation error or process variation, indicator #4 has lower rates of false negative errors than analogous situations for indicator #2 (compare contour values in Figure 3a with those in 3b; also compare 3c with 3d). See SI part E, Figure SI-6 for analogous contour plots for all 20 indicators.

The AUCs of all indicators declined at high levels of temporal autocorrelation in process variation, Φ, with indicators #1 and 2 being the most affected (Figure 4, SI part E, Figure SI-7). The higher reliability of indicators that were based on some historical baseline also held across a wide range of autocorrelation (Figure 4, SI part E, Figure SI-7). The stronger the positive autocorrelation, the larger the advantage of indicators based on historical baselines compared to indicators based on just the last three generations (#1 and 2) (Figure 4, SI part E, Figure SI-7). The addition of observation error, even at a very high level (σ_v^2=0.5), did not have an interaction effect with increasing levels of temporal autocorrelation in process variation. Specifically, there was almost no difference in either the absolute or relative performance of the different indicators beyond what was found in the base case (i.e., without temporal autocorrelation in process variation) and in the case with autocorrelation with no observation error (SI part E, Figure SI-8).

Discussion

A key problem in conservation biology is that the reliability of most extinction-risk metrics or indicators is at present largely uncertain, i.e., their ability to correctly categorize a population as being at risk is unclear. However, this uncertainty could be reduced, and outcomes from conservation decision-making could be improved, if more quantitative analyses were done like those presented here, which explicitly take into account uncertainties resulting from errors in abundance estimates and natural variability in ecological processes. Other scientists have also documented differences in reliability among indicators (Wilson et al. 2011; Porszt et al. 2012; Regan et al. 2013). We found (as did the empirical analysis of Porszt et al. 2012) that indicators that measure the extent of decline from a historical baseline tend to better reflect a population's long-term status than either (1) a decline from some maximum abundance, or (2) the widely-used IUCN rate of decline over the previous 3 generations. As well, some indicators are more predisposed to making certain types of errors, but generally process variation reduced reliability more than observation error, as was also shown by Wilson et al. (2011).

For commonly used IUCN-like indicators #1 and 2 (declines in abundance over the most recent 3 generations), we found that reliability decreased substantially more than the reliability of other indicators when process variation, observation error, or temporal autocorrelation in process variation increased. This poor performance is likely due to the assessment of status over a short period (the most recent 12 years vs. up to 52 years), which makes long-term trends in abundance more difficult to identify amid the "noise" of process variation and observation error. In contrast, indicators based on some early historical baseline

were the least sensitive to increases in either process variation or observation error. Although comparison with early portions of time series is recommended in some technical guidelines (Mace et al. 2002), many schemes emphasize recent rates of decline. There are two ways to calculate the IUCN criterion A: using only the most recent three generations of data or using all available data. IUCN recommends the use of all data only when "populations fluctuate wildly or oscillate with periods longer than the generation time" (IUCN 2013, page 26). However, here we show that the full span of data should be used not just for these special cases, but whenever there is process or observation error, i.e., for any real time series. We found that indicators that reflect the rate of decline over a three-generation time span can be as reliable as the best-performing indicators, but only when that rate is calculated using all available data and applied to those three most recent generation spans. Our findings could easily be incorporated into technical guidance for use in IUCN and other related conservation assessments.

For two other reasons, we recommend using all data from a time series, not just the most recent data. First, indicators of the recent rate of decline ignore the elevated extinction risks when recent abundances are relatively stable but are much lower than in an earlier period (the shifting baseline syndrome of Pauly 1995). Second, emerging evidence suggests that for heavily-exploited marine fishes, process variation may increase at low population densities owing to truncated age structures (Minto et al. 2008; Shelton and Mangel 2011), reducing recovery rates and elevating risk (Keith and Hutchings 2012; Neubauer et al. 2013).

Our model simulated semelparous sockeye salmon populations. However, we encourage other researchers to challenge assumptions that their current indicators reliably reflect actual population trends, and estimate that reliability for their own populations,

especially those with different life-history traits or population dynamics (Dulvy et al. 2004). Those future analyses (either empirical such as Porszt et al. 2012 or simulations like Wilson et al. 2011 and here) should explicitly take into account interannual variation in biological processes (and associated temporal autocorrelation), as well as errors in estimates of abundance, because these sources of variation can differentially affect the reliability of various indicators. People making management policies and conservation decisions should request that such tests of reliability be conducted before any indicators are used as input to risk assessments (Porszt et al. 2012).

Decision makers must also consider the relative importance, or severity of consequences, of false positives and false negatives when ranking indicators. For instance, in fisheries management, the economic and social costs associated with false negatives can often be as serious as false positives (Peterman 1990; Mapstone 1995; Field et al. 2004; Dulvy et al. 2006). Different stakeholders will often have different error weighting preferences, in part because the costs of these errors might be borne by different groups (Peterman 1990), which can present challenges in selecting error weightings. Despite the consistently high rank of certain indicators, we also showed that in some cases, the ranking of indicators can be influenced by the relative magnitude of weightings placed on the two types of errors, as well as a stated maximum acceptable error rate. Decision makers should therefore be explicit and transparent about those weightings and acceptable error rates, and mindful of which stakeholders will be affected by these (sometimes implicit) decisions. Standard methods of risk assessment and decision analysis can further help managers choose appropriate indicators of population risk (Burgman 2005).

Given the prominent role that IUCN and other indicators of conservation concern play in species status assessments worldwide, our results should prompt scientists who develop and evaluate extinction-risk criteria to use simulation and empirical analyses to understand the performance and reliability of any proposed criteria. We found that use of all available data, rather than the common practice of using just the most recent three generations, consistently increased reliability of risk assessments. Quantitative analyses such as ours, which can estimate reliability of indicators, are a relatively inexpensive way to ensure that management agencies make the best decisions about their use of limited conservation funds.

Acknowledgments

We are grateful for support for this research provided by Simon Fraser University, the Natural Sciences and Engineering Research Council of Canada, and the Canada Research Chairs Program, Ottawa, Canada. We also thank Brendan Connors and John Reynolds for their discussions and suggestions on a draft manuscript. N.K.D. acknowledges the National Center for Ecological Analysis and Synthesis (NCEAS), Santa Barbara, CA, USA Working Group on ‘Red flags and species endangerment’ for helpful discussion.

References

Burgman, M. (2005). Risks and decisions for conservation and environmental management. Cambridge University Press, U.K., 488 pp.

Connors, B.M., Cooper, A.B., Peterman, R.M., & Dulvy, N.K. (2014). The false classification of extinction risk in noisy environments. *Proceedings of the Royal Society B,* **281**, 20132935. DOI:10.1098/rspb.2013.2935.

COSEWIC (Committee on the Status of Endangered Wildlife in Canada) (2011). COSEWIC's assessment process and criteria. Available from http://www.cosewic.gc.ca/pdf/Assessment_process_and_criteria_e.pdf (accessed October 2012).

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

Dulvy, N.K., Jennings, S., Goodwin, N.B., Grant, A. & Reynolds, J.D. (2005). Comparison of threat and exploitation status in north-east Atlantic marine populations. *Journal of Applied Ecology*, **42**, 883-891.

Dulvy, N.K., Jennings, S., Rogers, S.I. & Maxwell, D.L. (2006). Threat and decline in fishes: an indicator of marine biodiversity. *Canadian Journal of Fisheries and Aquatic Sciences*, **63**, 1267-1275.

Field, S.A., Tyre, A.J., Jonzen, N., Rhodes, J.R. & Possingham, H.P. (2004). Minimizing the cost of environmental management decisions by optimizing statistical thresholds. *Ecology Letters*, **7**, 669–675.

Fisheries and Oceans Canada (DFO) (2005). Canada's policy for conservation of wild Pacific salmon. DFO, Vancouver, B.C., Canada. Available from http://www.pac.dfo-mpo.gc.ca/publications/pdfs/wsp-eng.pdf (accessed January 2012).

Hibberd, P.L. & Cooper, A.B. (2008). Methodology: statistical analysis, test interpretation, basic principles of screening with application for clinical study. In: *Walker's pediatric gastrointestinal disease: pathophysiology, diagnosis, management* (eds. Kleinman, R.E., Goulet, O., Mieli-Vergani, G., Sanderson, I.R., Sherman, P.M. & Shneider, B.L.). 5th edition. B. C. Decker, Hamilton, Ontario.

Hilborn, R. & Walters, C.J. (1992). Chapter 7: Stock and Recruitment. In *Quantitative Fisheries Stock Assessment: Choice, Dynamics, and Uncertainty*. Chapman and Hall, New York.

Holt, C.A., Cass, A., Holtby, B. & Riddell, B. (2009). Indicators of status and benchmarks for conservation units in Canada's Wild Salmon Policy. DFO Canadian Science Advisory Secretariat Research Document 2009/058. viii + 74 p. Ottawa, Ontario, Canada. Available from http://www.dfo-mpo.gc.ca/CSAS/Csas/Publications/ResDocs-DocRech/2009/2009_058_e.pdf (accessed August 2012).

Irvine, J.R., Gross, M.R., Wood, C.C., Holtby, L.B., Schubert, N.D. & 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) Standards and Petitions Subcommittee (2013). Guidelines for using the IUCN Red List Categories and Criteria. Version 10.1 (September 2013). IUCN, Gland, Switzerland. Available from

http://www.iucnredlist.org/documents/RedListGuidelines.pdf (accessed November 2013).

Keith, D. M. and Hutchings, J. A. (2012). Population dynamics of marine fishes at low abundance. *Canadian Journal of Fisheries and Aquatic Sciences*, **69**, 1150-1163.

Korman, J., Peterman, R.M., & Walters, C.J. (1995). Empirical and theoretical analyses of correction of time series bias in stock-recruitment relationships of sockeye salmon. *Canadian Journal of Fisheries and Aquatic Sciences*, **52**, 2174-2189.

Mace, G.M., & Hudson, E.J. (1999). Attitudes toward sustainability and extinction. *Conservation Biology*, **13**, 242-246.

Mace, P.M., et al. (2002). National Marine Fisheries Service (NMFS) / Interagency Working Group evaluation of CITES criteria and guidelines. Technical memorandum NMFS-F/SPO-58. National Oceanic and Atmospheric Administration, Rockville, Maryland.

Mace, G.M., Collar, N.J., Gaston, K.J., Hilton-Taylor, C., Akçakaya, H.R., Leader-Williams, N., Milner-Gulland, E.J., & Stuart, S.N. (2008). Quantification of extinction risk: International Union for the Conservation of Nature's (IUCN) system for classifying threatened species. *Conservation Biology*, **22**, 1424-1442.

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

Matsuda, H., Takenaka, Y., Yaharat, T. & Uozumi, Y. (1998). Extinction risk assessment of declining wild populations: The case of the southern bluefin tuna. *Researches on Population Ecology*, **40**, 271-278.

Minto, C., Myers, R. A. and Blanchard, W. (2008) Survival variability and population density in fish populations. *Nature*, **452**, 344-347.

Neubauer, P., Jensen, O. P., Hutchings, J. A. and Baum, J. K. (2013) Resilience and Recovery of Overexploited Marine Populations. *Science*, **340**, 347-349.

Pauly, D. (1995). Anecdotes and the shifting base-line syndrome of fisheries. *Trends in Ecology and Evolution*, **10**, 430.

Peterman, R.M. (1990). Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 2-15.

Porszt, E.J., Peterman, R.M., Dulvy, N.K., Cooper, A.B. & Irvine, J.R. (2012). Reliability of indicators of decline in abundance. *Conservation Biology*, **26**, 894-904.

Powles, H., Bradford, M.J., Bradford, R.G., Doubleday, W.G., Innes, S. & Levings, C.D. (2000). Assessing and protecting endangered marine species. *ICES Journal of Marine Science*, **57**, 669-676.

Powles, H. (2011). Assessing risk of extinction of marine fishes in Canada - the COSEWIC experience. *Fisheries*, **36**, 231-246.

Punt, A.E. (2000). Extinction of marine renewable resources: a demographic analysis. *Population Ecology*, **42**, 19-27.

Regan, T.J., Taylor, B.L., Thompson, G.G., Cochrane, J.F., Ralls, K., Runge, M.C., & Merrick, R. (2013). Testing Decision Rules for Categorizing Species' Extinction Risk to Help Develop Quantitative Listing Criteria for the U.S. Endangered Species Act. *Conservation Biology*, **27**, 821–831.

Reynolds, J.D., Dulvy, N.K., Goodwin, N.B. & Hutchings, J.A. (2005). Biology of extinction risk in marine fishes. *Proceedings of the Royal Society of London B*, **272**, 2337–2344.

Rice, J. C. & Legacé, E. (2007). When control rules collide: a comparison of fisheries management reference points and International Union for Conservation of Nature

(IUCN) criteria for assessing risk of extinction. *ICES Journal of Marine Science*, **64**, 718-722.

Ricker, W.E. (1975). Computation and interpretation of biological statistics of fish populations. Fish. Res. Board. Can. Bull. No. 191.

Shelton, A. O. and Mangel, M. (2011). Fluctuations of fish populations and the magnifying effects of fishing. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 7075-7080.

Venables, W.N. & Ripley, B.D. (2002). Modern applied statistics with S. 4th edition. Springer Science, New York.

Vincent, A.C. J., Sadovy de Mitcheson, Y., Fowler, S. L. and Lieberman, S. (2013). The role of CITES in the conservation of marine fishes subject to international trade. *Fish and Fisheries*, doi: 10.111/faf.12035

Wilson, H.B., Kendall, B.E. & Possingham, H.P. (2011). Variability in population abundance and the classification of extinction risk. *Conservation Biology*, **25**, 747-757.

Table 1. Categories of possible outcomes when an indicator's assessed current status of a population in a given year is compared to the overall long-term status of the same population. Note that the probabilities for two cases, false positives (α) and false negatives (β), are sufficient to determine the probabilities for the other two cases; the true negative rate is 1-α, and the true positive rate is 1-β.

		Indicator's assessed current status:	
		Non-declining	**Declining**
Overall long-term status:	**Non-declining**	True negative (TN) 1-α	False positive (FP) α (*Type I error*)
	Declining	False negative (FN) β (*Type II error*)	True positive (TP) 1-β

Table 2. A summary of the 20 indicators of population decline evaluated here for assessing status during the simulated evaluation period (based on indicators suggested by IUCN 2013, Mace et al. 2002, and Holt et al. 2009, among others). Further details are in online Supporting Information part B.

Indicator	**Periods***	**Smoothed**[†]	**Transformation**[‡]	**Change in abundance**[§]
1	Recent 3 gen.	No	Log_e	Regression: rate
2	Recent 3 gen.	Yes	Log_e	Regression: rate
3	Hist.: first year	No	Log_e	Regression: extent
4	Hist.: first year	Yes	Log_e	Regression: extent
5	Hist.: cycle year	No	Log_e	Regression: extent
6	Hist.: cycle year	Yes	Log_e	Regression: extent
7	Max.: single year	No	Log_e	Regression: extent
8	Max.: single year	Yes	Log_e	Regression: extent
9	Hist.: first gen.	No	Mean: Sliding window	Two means
10	Hist.: first gen.	Yes	Mean: Sliding window	Two means
11	Hist.: first gen.	No	Mean: Generation block	Two means
12	Hist.: first gen.	Yes	Mean: Generation block	Two means
13	Max.: 3 gen.	No	Mean: Sliding window	Two means
14	Max.: 3 gen.	Yes	Mean: Sliding window	Two means
15	Max.: 3 gen.	No	Mean: Generation block	Two means
16	Max.: 3 gen.	Yes	Mean: Generation block	Two means

17	Hist.: first gen.	No	Raw	Two means
18	Max.: 3 gen.	No	Raw	Two means
19	Hist.: first year	No	Log_e	Regression: rate
20	Hist.: first year	Yes	Log_e	Regression: rate

*Periods -- "Recent 3 gen." is the percent decline over the most recent 3 generations (12 years) analogous to criterion A of IUCN (2001) and COSEWIC (2011); "Hist." is the percent decline from some historical baseline since the beginning of the time series (either the first year of the evaluation period, the geometric mean abundance of the first 4-year generation, or the first corresponding cycle year in the time series). See Supporting Information part A for explanation of "cycle" year; "Max." is the percent decline from the maximum geometric mean abundance in the time series (either single year or 3-generation period) anywhere in the evaluation period of the time series.

†Smoothed -- Abundance estimates were either smoothed with a 4-year (1-generation) running mean or were left unsmoothed.

‡Transformation -- Abundance estimates were either raw values, log_e-transformed values, or the geometric mean abundance of 4-year generations where the generations either moved one year at a time in sliding windows or in 4-year-generation blocks with no overlap of years (i.e., status only assessed every four years). "Raw" refers to the original, untransformed data.

§Change in abundance -- "Regression" means that changes in abundance over the designated period were measured using robust linear regression to minimize the influence of outliers (Venables & Ripley 2002) and then calculated as either a rate or extent of change; "Two means" refers to changes in abundance that were measured as a percent decline from a mean abundance during some baseline period to either the current year's abundance or the mean abundance in the current generation being evaluated.

Figure 1. An example of (a) how a decline in population abundance would be assessed using Indicator 17 (percent decline from the geometric mean abundance of the first 4-year generation in the evaluation period (years 13-16) to abundance in the current year of assessment -- year 31 here), for one example Monte Carlo trial output from the simulation model, and (b) how the overall long-term status would be assessed for the same Monte Carlo trial -- the decline in generational mean abundance, as estimated from regression, between the last generation of the initialization period (years 9-12) and the last generation of the subsequent period, also estimated from regression (years 73-76). (c) Example Receiver Operating Characteristic (ROC) curves for two hypothetical indicators created by plotting true positive and false positive rates (Table 1) for each of 101 threshold values, ranging from 0 to 100%, for classifying a population as declining in abundance. The area under the curve (AUC) is also shown, including for the straight-line case for which the indicator would be no better than doing a coin toss.

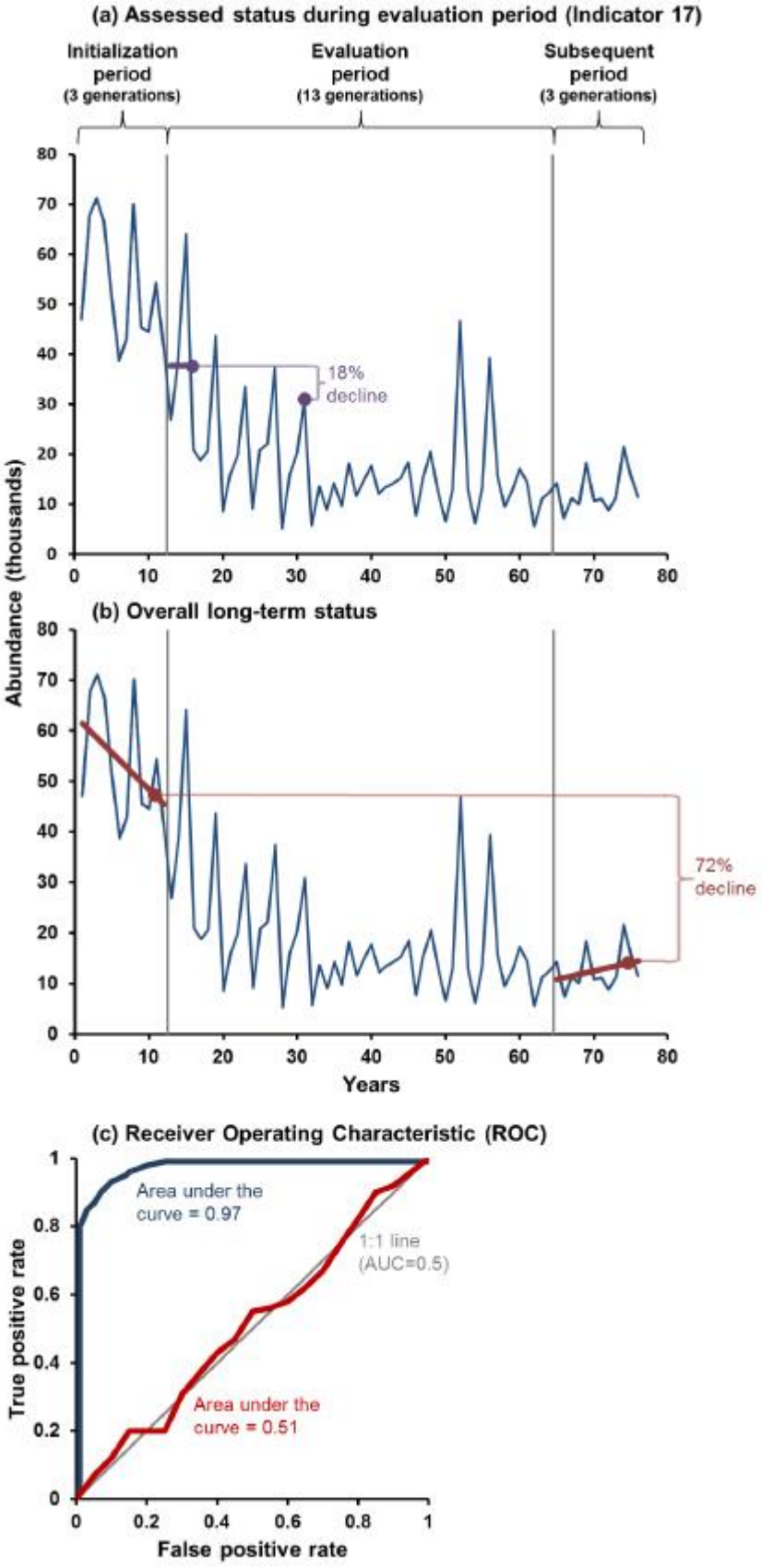

Figure 2. Ranking of indicators based on their reliability (probability of correctly reflecting the long-term status of the population), as measured by the area under the ROC curve (AUC), for (a) low process variation, i.e., low environmentally-driven variability in productivity (σ_u^2 = 0.01) and no observation error (σ_v^2 = 0); (b) high process variation (σ_u^2 = 0.5) and no observation error (σ_v^2 = 0); (c) low process variation (σ_u^2=0.01) and high observation error (σ_v^2 = 0.5); and (d) high process variation (σ_u^2 = 0.5) and high observation error (σ_v^2 = 0.5). Indicators based on the decline in the last three generations are black, indicators based on decline from a historical baseline are white, and indicators based on decline from maximum abundance are gray.

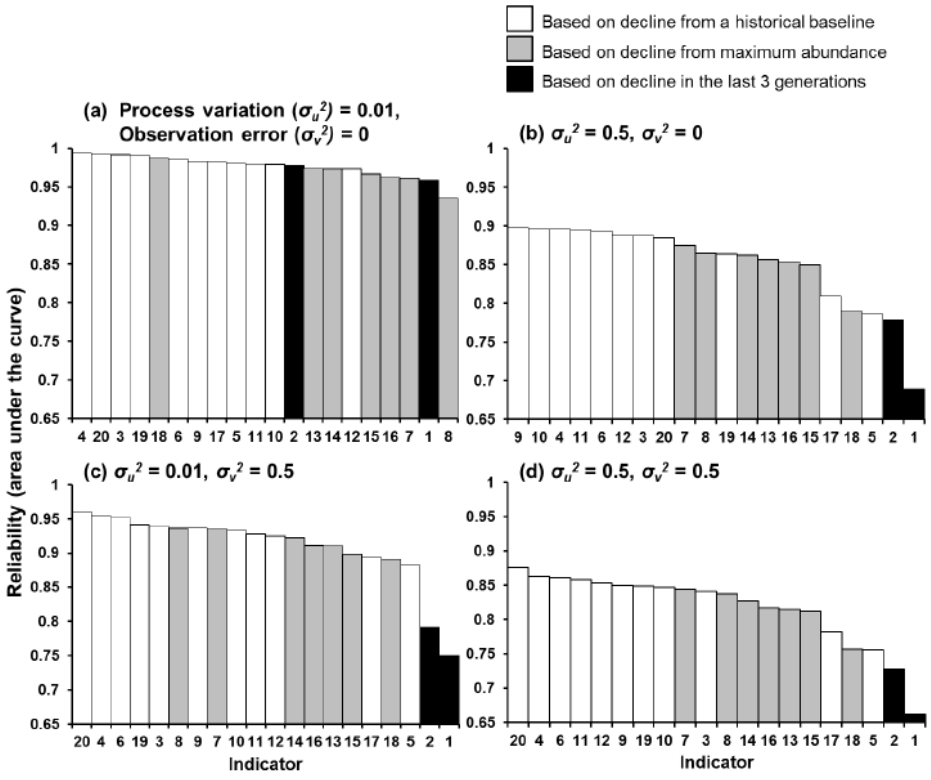

Figure 3. Plots of the minimum rate of false negative (FN) errors (probability values on contours) that can be obtained if the false positive (FP) rate is constrained to be below the maximum acceptable value specified on the Y-axis. False negative rates are shown as a function of observation error, σ_v^2, (but with process variation set at $\sigma_u^2 = 0.01$) for the top-performing indicator in this study #4 (a) and commonly used decline indicator #2 (b). False negative rates are also shown for different levels of process variation, σ_u^2, (but with no observation error, $\sigma_v^2 = 0$) for indicator #4 (c) and indicator #2 (d). The white data point in parts (a) and (b) indicates the false negative rate that will result from a maximum acceptable false-positive rate of 0.1 and an observation error variance of 0.3.

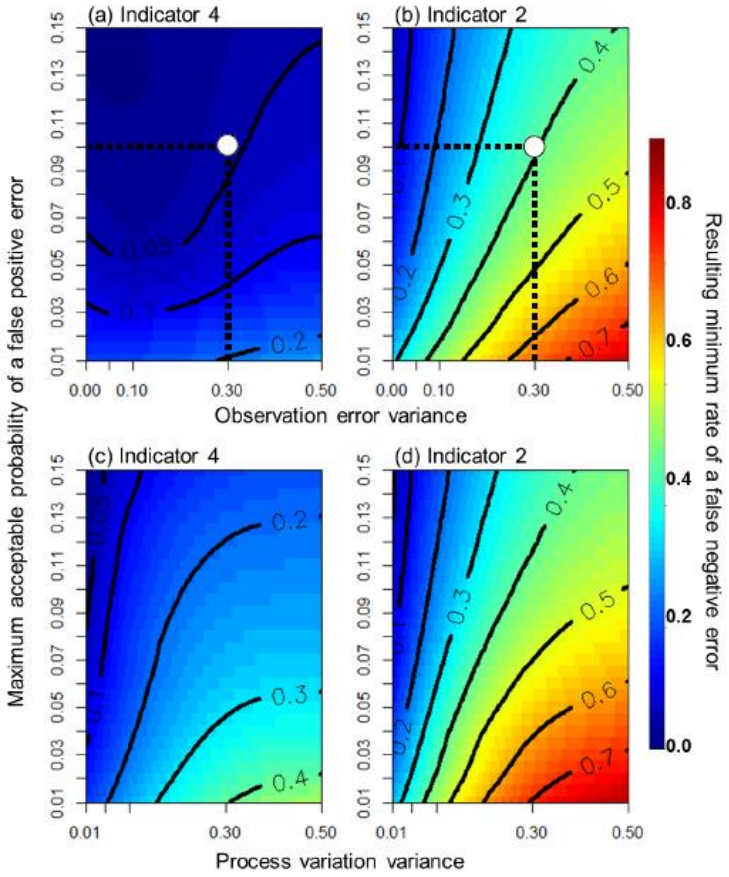

Figure 4. The reliability, or area under the curve (AUC), of the indicators across levels of temporal autocorrelation in process variation (Φ) ranging from -0.5 to 0.75 for (a) low process variation (σ_u^2=0.01) and (b) high process variation (σ_u^2=0.5), both with no observation error. Dark blue represents the highest probability of correctly classifying a population's abundance as decreasing or not, whereas dark red is the lowest probability. Results for other levels of process variation and observation error are found in the SI part E, Figures SI-7 and SI-8.