ICES Journal of
Marine Science

ICES Journal of Marine Science (2019), 76(7), 2102–2113. doi:10.1093/icesjms/fsz163

Original Article

A framework for assessing which sampling programmes provide the best trade-off between accuracy and cost of data in stock assessments

Melissa L. Muradian [1*,†], Trevor A. Branch[2], and André E. Punt[2]

[1]*Quantitative Ecology and Resource Management, University of Washington, Ocean Teaching Building, Suite 300, Box 357941, Seattle, WA 98195, USA*

[2]*School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA 98195, USA*

**Corresponding author: tel: + 1 208 652 3567; e-mail: melissa@henrysfork.org.*

[†]Present address: Henry's Fork Foundation, 801 Main St, PO Box 550, Ashton, ID 83420, USA.

Muradian, M. L., Branch, T. A., and Punt, A. E. A framework for assessing which sampling programmes provide the best trade-off between accuracy and cost of data in stock assessments. – ICES Journal of Marine Science, 76: 2102–2113.

Received 7 March 2019; revised 25 July 2019; accepted 27 July 2019; advance access publication 3 September 2019.

Fisheries surveys are required to assess the status of fish populations but are rarely evaluated to determine which data provide most information for least cost. We develop such a method focused on Pacific herring (*Clupea pallasii*) in Prince William Sound, Alaska. This population collapsed in 1992–93 and an intensive monitoring programme has been developed to assess why herring have not yet recovered, including the development of a Bayesian stock assessment model. We conducted a Monte-Carlo simulation study that calculated the deterioration in assessment performance when each survey was excluded, which allowed us to assess the trade-off between cost and improvement in model performance from including each survey data. For $10,000 a year the disease survey reduces bias and imprecision in current biomass by 34% on average, increases model reliability by 22%, and decreases by 31% the probability of a false management conclusion related to regulating the fishery. For $350,000 a year the diver survey reduces bias and imprecision by 12%, increases model reliability by 6%, and decreases the probability of a false management conclusion by 23%. The framework presented here can be used in other fisheries to weigh the costs and benefits of alternative sampling programmes in estimating current biomass.

Keywords: Bayesian methods, Pacific herring, simulation, stock assessment, value of information

Introduction

Stock assessment is the practice of fitting a population dynamics model to data to estimate stock status and provide a basis for regulating harvesting. Development of the foundational theories underlying stock assessment began in the early 20th century. Today, stock assessment is a burgeoning sector of research driven by advancements in sampling technologies and computing, and a steadily growing human population, which adds to the impetus for sustainable fisheries management.

Many managed global fisheries are classified as data limited (having only time-series of total landings available to an assessment model or worse) and numerous other stocks are managed using catch data and at least one index of abundance. However, using multiple data types that include catch information, an index of abundance, and additional time-series (such as age- or length-compositions) can lead to better estimation of stock abundance, productivity, and mortality, which are essential to reliable stock predictions (Deriso *et al.*, 1985; Wetzel and Punt, 2011). There are many commercially important stocks that have multiple associated scientific surveys, and these collect several data types used to inform more biologically realistic and statistically complex assessment models called integrated assessment models (Fournier and Archibald, 1982; Maunder and Punt, 2013). Additional data types can include female fecundity information, population sex ratio, genetic information, the outcomes of stomach content analysis, predator interaction levels and abundance, and growth information. It is expensive to conduct the surveys that collect these data, but often little attention is paid to which surveys yield the best return on the money spent, in accuracy and precision of management reference points.

The consequences of using inaccurate and imprecise assessment models can be severe: overfishing may lead to the collapse of fish populations, whereas underutilization may lead to millions of dollars in economic losses or deficits in food security. It is difficult to determine the accuracy of an assessment model in estimating the true population state in the absence of quantitative analysis. However, simulation studies can be used to fill this gap. Simulation studies involve generating realizations of the true population state, approximating the data sampling process, then applying the assessment model to each simulation to evaluate the difference between the simulated truth and the estimate from the assessment model.

Several studies have investigated the impact of data characteristics on assessment results. For example, Yin and Sampson (2004) evaluated the impact of observation error on model output and found that increasing the sample size for age-composition data is most important for accurately estimating ending year biomass. Magnusson and Hilborn (2007) explored the value of trends in catch data and the information contained in age-composition and abundance time-series, and found that integrated assessment models are surprisingly resilient to a lack of contrast in catch series data, but that age-composition data are important for reliably estimating the mortality rate and productivity of a stock. Wetzel and Punt (2011) found that adding length-composition data to data-limited assessments (those using only catch information) improved model performance, and Ono *et al.* (2015) found that the value of data type, quantity, and quality varies according to the life-history characteristics of the stock, but that an infrequent sampling programme run over a longer period is more informative than a shorter, more frequent survey programme.

This study builds upon previous work by quantifying the relative value of various data types of return on investment by asking what subset of sampling programmes will result in the least loss of accuracy, the lowest estimation bias, and the lowest cost of collection. A general method was developed that can be used to assess the present return on money invested in survey programmes. This analysis involved a simulation study that employed a leave-one-survey-out approach followed by summarizing the resulting accuracy and precision of model estimates to investigate how data types collected from an individual survey impact management reference points. It was assumed that omitting the most informative data types from an assessment model would lead to the poorest estimation performance.

This method was applied to Pacific herring (*Clupea pallasii*) in Prince William Sound, Alaska, which were historically harvested and commercially valuable, and have been monitored and managed by the Alaska Department of Fish and Game (ADF&G) for over 50 years (Funk and Sandonne, 1990; Thomas and Thorne, 2003). The Exxon Valdez oil spill occurred inside of the Sound in late March 1989, during the prespawning period of the local herring stock, and herring were adversely affected by the spill in that and the following year (Hose *et al.*, 1996; Norcross *et al.*, 1996; Brown and Baker, 1998; Marty *et al.*, 1999). Despite these adverse effects, the population maintained relatively high biomass throughout these and subsequent years, but eventually collapsed in 1992–1993 and has yet to recover (Quinn *et al.*, 2001; Hulson *et al.*, 2008; Thorne and Thomas, 2008; Pearson *et al.*, 2012; Muradian *et al.*, 2017). The fishery itself has been closed since 1998.

Because the spill, and subsequent collapse of the herring population, monitoring of Prince William Sound herring has been a key research objective in the region, resulting in the implementation of targeted monitoring programmes including diver surveys, hydroacoustic surveys, oceanographic data collection, and surveys to assess marine predator abundance and their interactions with herring. Thus, Prince William Sound herring have a long history of management and a well-funded monitoring programme, making this population an ideal case study to investigate the trade-offs between data collection cost and model accuracy. Using the Bayesian age-structured assessment model developed for this population (Muradian *et al.*, 2017), we conducted a Monte-Carlo simulation study to explore, which data types when omitted lead to lower levels of estimation accuracy, consistent bias, and higher probabilities of incorrect management action, which is defined as closing the fishery when it should be opened, or allowing fishing when the fishery should be closed.

Methods

This simulation study, conducted during 2014–2015, used an age-structured model fitted using Bayesian methods as the operating (or data generation) model. The operating model provided the true state of the population and was used to generate 100 pseudo datasets with realistic uncertainty from the Bayesian posteriors. Multiple estimation models (the test scenarios) were fit to these 100 simulated datasets and model performance was evaluated by comparing estimates to the truth. Equations required are listed in the tables; thus Equation (1.5) is the fifth numbered equation in Table 1.

Operating model

The operating model is based on the Bayesian Prince William Sound herring age-structured assessment model (Muradian *et al.*, 2017). The modelling horizon spanned 34 years (1980–2013), and the modelled population included herring of ages 3–8, with a "plus group" for ages 9 and above. Some proportions of age-3 and -4 herring were assumed to be mature, and cohorts were assumed to be fully mature by age 5. Cohort size was reduced mid-year and at the end of the year by natural mortality and discrete fishing mortality.

Uncertainty in the population process was accounted for by drawing sets of parameter values from the joint posterior distribution constructed using the Markov-chain Monte-Carlo algorithm included in AD Model Builder (Fournier *et al.*, 2012). Using the operating model, each set of parameter values produced a realistic simulation of "true" dynamics given uncertainty in time-series of recruitment, mortality, and spawning biomass (Figure 1). This step also produced simulations of the "true" indices of biomass and age-compositions from the fishery and survey [Equations (1.1)–(1.6)].

Data generation

The following data were assumed to be available to all estimating models without error: catches (from each of four gear-specific fisheries: gillnet, sac-roe, food and bait, and spawn-on-kelp),

Table 1. Key estimated values used in the operating model, with equations, and description of how observation error was added during the data generation step.

Quantity	Expected value	Equation	Sampling distribution	Equation
Estimated ADF&G acoustic biomass, mt	$\hat{H}_{1,y} = B_y\, e^{q_1}$	1.1	$\hat{H}_{1,y} = B_y\, e^{q_1+\varepsilon_{H_1}-c}$ $\epsilon_{H_1} \sim \mathbf{N}(0,\ \sigma_{H_1} = 0.29)$	1.7
Estimated PWSSC acoustic biomass, mt	$\hat{H}_{2,y} = B_y\, e^{q_2}$	1.2	$\hat{H}_{2,y} = B_y\, e^{q_2+\varepsilon_{H_2}-c}$ $\epsilon_{H_2} \sim \mathbf{N}(0,\ \sigma_{H_2} = 0.35)$	1.8
Estimated sac-roe fishery age-composition	$\hat{\Theta}_{1,y,a} = \frac{V_a\, N_{y,a}}{\sum_{a\in A}(V_a N_{y,a})}$	1.3	$\hat{\Theta}_{1,y,a} \sim \mathbf{D}\left[(120-1)\left(\frac{V_a N_{y,a}}{\sum_{a\in A}(V_a N_{y,a})}\right)\right]$	1.9
Estimated spawning age-composition	$\hat{\Theta}_{Sp,y,a} = \frac{\rho_{M,a,.} N_{y,a}}{\sum_{a\in A}\left(\rho_{M,a,.} N_{y,a}\right)}$	1.4	$\hat{\Theta}_{Sp,y,a} \sim \mathbf{D}\left[(40-1)\left(\frac{\rho_{M,a,.} N_{y,a}}{\sum_{a\in A}\left(\rho_{M,a} N_{y,a}\right)}\right)\right]$	1.10
Estimated naturally spawned eggs (diver survey), trillions	$\hat{E}_y = 10^{-6}\rho_{f,y}\sum_{a\in A}\left(\sim N_{y,a} f_{y,a}\right)$	1.5	$\hat{E}_y = 10^{-6}\rho_{f,y}\sum_{a\in A}\left(\tilde{N}_{y,a} f_{y,a}\right)e^{\epsilon_E - c}$ $\epsilon_E \sim \mathbf{N}(0, \sigma_E = 0.35)$	1.11
Estimated milt (aerial survey), mile-days	$\hat{T}_y = \frac{(1-\rho_{f,y})\tilde{B}_{\text{post},y}}{e^{q_T}}$	1.6	$\hat{T}_y = \frac{(1-\rho_{f,y})\tilde{B}_{\text{post},y}}{e^{q_T}} e^{\epsilon_T - c}$ $\epsilon_T \sim \mathbf{N}(0, \sigma_T = 0.32)$	1.12

N() denotes normally-distributed random variables, **D**[] denotes Dirichlet-distributed random variables, and c denotes the lognormal correction term $\boldsymbol{\sigma}^2/2$. Only the diver survey is treated as an absolute index of abundance.

empirical weight-at-age for the population in each year, annual sex ratios, and female fecundity-at-age. For the estimation models that assumed disease was impacting the population, the reported disease data (per cent infection) were provided without observation error, consistent with how these data were included in the operating model. For biomass indices, observation error was added to the "true" simulation indices using bias-corrected, lognormal errors with log standard deviations estimated by fit of the Bayesian assessment model to the actual data [Equations (1.7)–(1.8), (1.11)–(1.12)].

Observation error was added to simulated catch-at-age data from the sac-roe fishery and simulated age-composition data collected by the herring-spawn survey. These data were assumed to be Dirichlet-distributed about the "true" operating model age-compositions. A Dirichlet distribution was used to account for overdispersion in the observed age-composition data because of non-random age- or size-specific spatial variation in prespawning and spawning herring aggregations (Hulson *et al.*, 2012). An overdispersion sample size multiplier of 5 is appropriate for this population (Muradian *et al.*, 2017) and was used in the Dirichlet distribution, which brought the effective sample sizes for multinomial-distributed age-compositions to 120 for the sac-roe fishery and 40 for the herring-spawn survey [Equations (1.9)–(1.10)]. These fishery and survey effective sample sizes were supplied as known to the estimation models.

Estimation models

Five Bayesian estimation models were created, in addition to a base estimation model to examine the relative value of information provided by each survey to the age-structured assessment model. Comparing model performance of a case lacking survey x to the base case (which contains survey x) reveals the ways in which model performance is altered by the addition of survey x.

(i) *Base case:* this was the reference model that used all collected data types and matches the structure of the operating model (Table 2). This case included three relative time-series of biomass, one absolute index of biomass, and two time-series of interannual variation in disease infection used as indices of disease mortality for age-specific disease mortality for the protozoan parasite *Ichthyophonus hoferi* and for the North American strain of viral haemorrhagic septicaemia virus (VHSV; Muradian *et al.*, 2017).

(ii) *Omit disease survey*: examined the influence of removing the two time-series of disease infection. This case reflects the scenario in which disease-infection data had never been collected nor analysed. This case assumed that the additional mortality events that contributed to the collapse in 1992–1993 were temporally isolated, and assumes no additional mortality after 1993 as this reflects a realistic estimation model that would result from modelling efforts had a disease survey never been initiated. This estimation model also differed from the other cases by holding plus-group mortality at 1.14 year^{-1} (the midpoint of the prior) to address convergence issues associated with removing the additional mortality provided by the disease index (Table 2). This assumption did not impact the relative ranking of the performance as measured by the metrics listed below of the estimation model for this case (results not shown).

(iii) *Omit diver survey*: tested the value of the diver survey by omitting the egg deposition data (an absolute index of female spawning biomass) and the fecundity-at-age data that were collected in 1984, 1989–1993, and 1994–1997. A key consequence of this removal is that this estimation model has no absolute index of biomass, because the egg deposition data are assumed to be equal (in expectation) to female spawning abundance multiplied by fecundity (Table 2).

(iv) *Omit ADF&G hydroacoustic survey*: tested the value of the ADF&G hydroacoustic survey, which collected an independent relative index of total biomass of age-3 and older fish each spring from 2005 to 2009.

(v) *Omit Prince William Sound Science Center (PWSSC) hydroacoustic survey*: tested the value of the PWSSC hydroacoustic survey, which also collected an independent relative index of total biomass of age-3 and older fish each spring, but ran every year from 1995 to 2012.

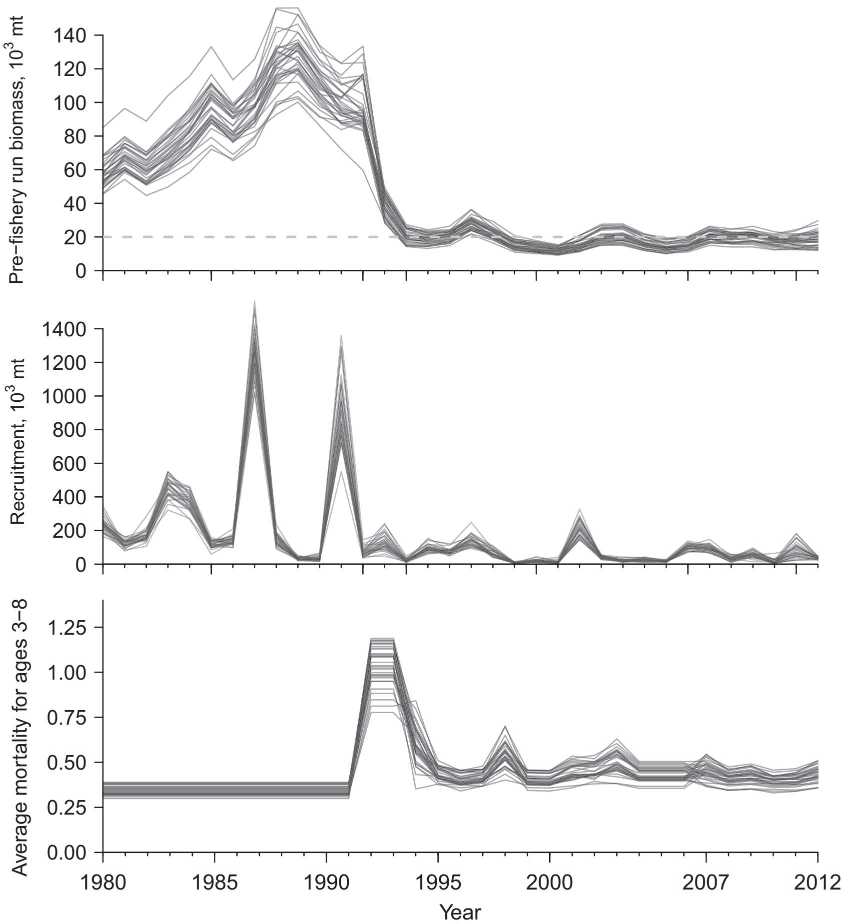

Figure 1. The 100 simulations of true biomass, recruitment, and mortality generated by the operating model. After observation error was added to each simulation, these became the basis for the 100 datasets.

(vi) *Omit aerial survey*: tested the value of the aerial survey that ran every year from 1980 to 2012 by removing the mile-days of milt time-series. During the herring spawning events each spring, aerial surveys fly over the spawning sites and measure the linear extent of milt clouds in miles of corresponding coastline per day. Mile-days of milt data were collected every year since 1980 and provide a relative index of male spawning biomass.

The impact of omitting the age-/sex-/size-survey that collects spawning age-composition data and size-at-age was not investigated because the goal of this study is to identify the value of an individual survey, and data from multiple surveys rely on the information collected by the age-/sex-/size-survey. For example, preparation of the hydroacoustics data relies on the length and weight data collected by the age-/sex-/size-survey to convert acoustics data into biomass estimates (Csepp *et al.*, 2011). Therefore, the situation reflected by an estimation model that removed the age-/sex-/size-survey would defy the premise of this investigation by affecting data from the several other surveys as well.

Survey collection costs

Annual survey costs summed over the years in the time-series are reported in Table 3. Best estimates of recent annual survey costs were obtained from the scientists leading the surveys (P. Hershberger, S. Moffitt, and W.S. Pegau, pers. comm.) because published budgets include sums of money for programmes not directly tied to the cost of running a survey. The most recent year of each survey differs because some surveys are no longer conducted (Muradian *et al.*, 2017). Adjustment for inflation was necessary to bring past costs to US dollars in the year 2015 so that costs could be compared (Bureau of Labor Statistics, 2015) because the time-series are patchy, of differing length, and start in different decades.

Development of posterior distributions

As explained under "Data generation", 100 simulated datasets were retained for simulations. A larger number of simulated datasets were created (293) and we retained the first 100 for which all 6 estimation models obtained a positive-definite Hessian matrix (without which the MCMC algorithm did not run automatically), and for which the MCMC algorithm reached convergence. The result was 600 converged models, each of which had been run for 1 100 000 cycles, with a burn-in of 10% and thinned every 1000 draws, for all 1000 posterior samples. Convergence was assessed using the Geweke statistic (Geweke, 1992) to judge stationarity of the chain, low autocorrelation for all parameters, and visual assessment of the trace plots.

Table 2. Description of the base and each alternative estimation model by difference in estimated parameters and constant values.

Index	Parameter	Base	Omit disease survey	Omit diver survey	Omit ADF&G hyd. survey	Omit PWSSC hyd. survey	Omit aerial survey
2.1	Background mortality, ages 3–8	Fixed at 0.25	–	–	–	–	–
2.2	Non-fishery mortality, ages 9^+	$m_{9^+} \sim U(0.3, 2.0)$	**Fixed at 1.14**	–	–	–	–
2.3	VHSV disease scalar, ages 3–4	$\beta_1 \sim U(0.0, 1000)$	**NA**	–	–	–	–
2.4	*I. hoferi* scalar, ages 5–8, 1994–2006	$\beta_{2,1} \sim U(0.0, 25.0)$	**NA**	–	–	–	–
2.5	*I. hoferi* scalar, ages 5–8, 2007–2012	$\beta_{2,2} \sim U(0.0, 25.0)$	**NA**	–	–	–	–
2.6	Disease mortality in 1993, VHSV	$m_{1993,a,1} \sim U(0.0, 5.0)$	–[a]	–	–	–	–
2.7	Disease mortality in 1993, *I. hoferi*	$m_{1993,a,2} \sim U(0.0, 5.0)$	**NA**	–	–	–	–
2.8	Purse-seine fishery gear selectivity	$\alpha_v \sim U(3.0, 5.0)$	–	–	–	–	–
2.9	Purse-seine fishery gear selectivity	$\beta_v \sim U(1.0, 7.0)$	–	–	–	–	–
2.10	ADF&G acoustic scalar, log-link	$q_1 \sim U(-5.0, 5.0)$	–	–	**NA**	–	–
2.11	ADF&G acoustic biomass CV	$\sigma_{H_1} \sim U(0.0, 0.6)$	–	–	**NA**	–	–
2.12	PWSSC acoustic scalar, log-link	$q_2 \sim U(-5.0, 5.0)$	–	–	–	**NA**	–
2.13	PWSSC acoustic biomass additional error	$\sigma_{H_2,B} \sim U(0.0, 0.6)$	–	–	–	**NA**	–
2.14	Egg deposition additional error	$\sigma_{E,B}$ =Fixed at 0.3	–	**NA**	–	–	–
2.15	Milt scalar, log-link	$q_T \sim U(2.3, 7.0)$	–	–	–	–	**NA**
2.16	Milt CV	$\sigma_T \sim U(0.0, 0.6)$	–	–	–	–	**NA**
2.17	Proportion mature at age 3, 1980–1996	$\nu_3 \rho_{M,4,1} \sim U(0.0, 0.75)$	–	–	–	–	–
2.18	Proportion mature at age 4, 1980–1996	$\rho_{M,4,1} \sim U(0.0, 1.0)$	–	–	–	–	–
2.19	Proportion mature at age 3, 1997–2012	$\rho_{M,3,2} \sim U(0.0, 1.0)$	–	–	–	–	–
2.20	Proportion mature at age 4, 1997–2012	$\rho_{M,4,2}$ = Fixed at 0.9	–	–	–	–	–
2.21	Recruitment by year (million), log-link	$\eta_{y,3} = \ln(N_{y,3}) \sim U(0.0, 8.0)$	–	–	–	–	–
2.22	Age-4 abundance in 1980, log-link	$\eta_{1980,4} = \ln(N_{1980,4}) \sim U(0.0, 8.0)$	–	–	–	–	–
2.23	Age-5 abundance in 1980, log-link	$\eta_{1980,5} = \ln(N_{1980,5}) \sim U(0.0, 8.0)$	–	–	–	–	–

Labels for the last five columns refer to the survey being tested by—and therefore omitted from—that case. NA stands for not applicable and denotes parameters that are not present in that case; dashes denote parameter as defined in base case. Use of bold for legibility.

[a]Parameter 2.6 is defined differently for the model in the case omitting the disease survey; a single additional mortality is estimated for ages 3–8 in 1992 and 1993 and not assumed to be explicitly associated with disease.

Table 3. Information on the survey programme being tested by each case: survey being omitted by that case, the data types collected by that survey, the number of years that survey was run (number of years in each time-series), the total cost of running the tested survey, and the annual cost of running the most recent survey in the series are shown.

Survey tested (omitted)	Data types collected by survey	*n*	Total survey cost, ψ	Annual survey cost
Disease	Index of VHSV/ulcer infection Index of *I. hoferi* infection	19	\$152 898	\$10 000
Diver	Eggs deposited (trillions) Fecundity (eggs per female)	10	\$4 136 315	\$350 000
ADF&G acoustics	Hydroacoustic biomass (mt)	5	\$292 463	\$67 000
PWSSC acoustics	Hydroacoustic biomass (mt)	20	\$1 354 682	\$80 000
Aerial	Milt (mile-days)	33	\$627 980	\$16 600

Total survey cost was calculated as the sum of the annual cost of running the survey in each year of the time-series with adjustment for inflation.

Reference points

Three categories of reference points represented potential management quantities of interest, and were chosen for their ability to summarize the assumed dynamics of this population (Table 4). The biomass reference points were the initial (1980) year biomass and the forecast biomass in 2013 (final-year biomass). The two recruitment reference points were the mean recruitment for the pre-collapse period (1980–1991) and the mean recruitment for the post-collapse period (1992–2013)—recruitment in 2013 is a projection. The two mortality reference quantities were the mean non-fishery mortality for all ages in the pre-collapse period (1980–1991) and the mean non-fishery mortality for all ages in the post-collapse period (1992–2012). There is no projection for non-fishery mortality in 2013 because only pre-fishery run biomass is projected for 2013, which is the biomass of fish at the start of the modelled year before modelled fishery or non-fishery mortality.

Performance measures

Estimation performance was quantified by comparing the estimates of each reference point from the 100 simulations to the operating model-generated true values using 7 performance metrics (Table 5). These metrics quantify bias, precision, information-to-cost ratios, model reliability, and the probability of incorrect management decisions under each estimation model.

Table 4. The six reference points used in this study along with their mathematical equation.

Equation	Reference point	Equation
4.1	Initial year biomass in 1980	B_{1980}
4.2	Final-year biomass in 2013, the forecast	B_{2013}
4.3	Mean recruitment over the pre-collapse period (1980–1991)	$\overline{R_1} = \frac{\sum_{y=1980}^{1991} N_{y,\ 3}}{12}$
4.4	Mean recruitment over the post-collapse period (1992–2013)	$\overline{R_2} = \frac{\sum_{y=1992}^{2013} N_{y,\ 3}}{22}$
4.5	Mean mortality over the pre-collapse period (1980–1991)	$\overline{M_1} = \frac{\sum_y \left[\frac{\sum_{a=3}^{9^+} m_{y,a}}{7}\right]}{12}$
4.6	Mean mortality over the post-collapse period (1992–2012)	$\overline{M_2} = \frac{\sum_y \left[\frac{\sum_{a=3}^{8} (m_{y,a}+m_{y,a,d})+m_{9+}}{7}\right]}{21}$

In Equations (4.5)–(4.6), m_a represents age-specific (a) non-fishery related mortality. In Equation (4.6), $m_{d,a}$ represents age-specific (a) additional mortality because of increased disease, where d represents one of the two diseases, VHSV or *I. hoferi*.

Table 5. Performance measures using the true reference quantity θ and the estimate $\hat{\theta}$.

Equation	Performance measure	Equation
5.1	Bayesian posterior median	$\hat{\theta}_i = \text{median}_j\left(\hat{\theta}_{ij}\right)$
5.2	Relative error	$\text{RE}_i = \frac{\hat{\theta}_i - \theta_i}{\theta_i}$
5.3	Median relative error	$\text{MRE} = \text{median}_i\left(\frac{\hat{\theta}_i - \theta_i}{\theta_i}\right)$
5.4	Absolute relative error	$\text{ARE}_i = \left\lvert\frac{\hat{\theta}_i - \theta_i}{\theta_i}\right\rvert$
5.5	Median absolute relative error	$\text{MARE} = \text{median}_i\left(\left\lvert\frac{\hat{\theta}_i - \theta_i}{\theta_i}\right\rvert\right)$
5.6	Extreme tail probability	$\text{ETP} = \frac{\sum_{i=1}^{s}\left[P\left(\hat{\theta}_i > \theta_i\right) < 0.01 \text{ or } P\left(\hat{\theta}_i > \theta_i\right) > 0.99\right]}{i}$
5.7	Probability of an incorrect closure	$\text{PIC} = \frac{\sum_i (\hat{\theta}_i < L \text{ and } \theta_i > L)}{\sum_i (\theta_i > L)}$
5.8	Probability of an incorrect opening	$\text{PIO} = \frac{\sum_i (\hat{\theta}_i > L \text{ and } \theta_i < L)}{\sum_i (\theta_i < L)}$
5.9	Total probability of an incorrect management conclusion	$\text{PIM} = \frac{\sum_i \left[(\hat{\theta}_i < L \text{ and } \theta_i > L) + (\hat{\theta}_i > L \text{ and } \theta_i < L)\right]}{i}$
5.10	Information-to-cost ratio	$\text{ICR} = \frac{\Delta(\text{MARE})_h}{\psi_h}$

Subscript i is used to index the simulations and j to index the MCMC samples for that simulation. In Equations (5.6)–(5.9) the symbol L stands for the lower regulatory threshold or limit used to manage the population and s stands for 100, the total number of simulations conducted. In Equation (5.10), the symbol $\Delta(\text{MARE})_h$ stands for the difference between the MARE of the base case and the MARE of the case lacking the data from survey h and ψ_h stands for the total cost of collecting survey h (Table 3); each of the non-base cases have an associated information-to-cost ratio.

(i) *Bias: relative error (RE) and median relative error (MRE).* The median (across MCMC samples) of the respective posterior distribution was used as the estimated value for each reference point ($\hat{\theta}_i$), and the RE of each reference point was calculated as the relative difference between the true value from the operating model (θ_i) and the estimated value ($\hat{\theta}_i$) for each simulation i. The median (MRE) and 90% interval of the distribution of RE values were used to quantify bias and imprecision across simulations for each reference point.

(ii) *Bias and precision: absolute relative error (ARE).* Absolute relative error quantifies the absolute relative difference between the estimated value and the true value from the operating model for each reference point, and is an indicator of both bias and precision. The median (MARE) and 90% interval of the ARE values were used to describe the distribution of ARE values across simulations.

(iii) *Model reliability: extreme tail probability (ETP).* Extreme tail probability detects failures of a Bayesian estimation procedure as how often the true reference value is within the extreme tails of the posterior distribution (Gelman *et al.*, 2004). For this study, the lower 0.01 and the upper 0.99 percentiles were chosen as the extremes of the posterior distribution. For a well-performing estimation model it is expected that the true value of each reference point will fall outside of the 98% credible interval in no more than 2% of the simulations. This constraint was relaxed and an ETP rate of >5% was considered an indicator of poor estimation performance because results are based on only 100 simulations.

(iv) *Probability of incorrect management conclusions (PIM)*. The probability of an "incorrect closing" (i.e. not allowing a fishery when one should be allowed given the management regulations) measures a certain way each estimation model can fail under a particular management prescription by comparing the posterior median of the forecast from each estimation model to the management limit and tracking how often the forecast from the estimation model is below the limit when the true forecast biomass (from the operating model) is above the limit [Equation (5.7)]. The probability of an "incorrect opening" measures the probability that the fishery would be opened as a result of sufficiently high estimated biomass when the true biomass level is below the management threshold [Equation (5.8)]. Incorrect fishery openings are ecologically important because they may result in higher exploitation rates than the management strategy prescribes. The total probability of incorrect management conclusions (PIM) is the number of combined occurrences of incorrect fishery closures and openings for an estimation model summed over all data simulations [Equation (5.9)].

(v) *Information-to-cost ratio (ICR)*. Each of the estimation models differs from the base case in that it lacks data from a single survey (Table 2). A metric was developed to quantify the return on money invested in each survey, which weighted the improvement in model performance through each survey's data by that survey's total collection costs [Equation (5.10)], and is termed the information-to-cost ratio. Improvement in model performance was quantified by the decrease in the MARE when that survey's data were added back to the assessment model. The ICR provides a relative ranking of those surveys that lead to a higher return on investment towards estimating accurate and precise management reference points.

Results

Accuracy and precision of biomass

For initial year biomass (B_{1980}), the largest biases were positive, and resulted when the diver survey data were omitted (MRE of 0.18). The next most positive bias occurred when the data from the disease survey were omitted (MRE of 0.16; Figure 2a). The remaining four cases (including the base case) resulted in low, approximately equal levels of negative bias in B_{1980} (MREs of -0.02 or less). The results for MARE of initial year biomass closely track those of MRE for all cases (Figure 3a).

In general, all cases resulted in more biased (in the same direction) and less precise estimates of the forecast biomass, B_{2013}, than the initial biomass (Figures 2a and b and 3a and b). The largest bias and imprecision in the forecast biomass resulted from omitting the disease survey (MARE of 0.50), followed by omitting the diver survey (MARE of 0.29), and the upper bound of the 90% ARE intervals for these cases are 1.29 and 1.10, respectively. The remaining cases (including the base case) estimated B_{2013} with levels of MARE ≤ 0.10 and the upper bound of the 90% ARE intervals $\leq$ 0.60 (Figure 3b).

Accuracy and precision of recruitment

The case omitting the diver survey resulted in the largest bias and imprecision for the two recruitment reference points: MARE and 90% intervals of 0.38 (0, 0.86) for $\overline{R_1}$ and 0.3 (0, 0.93) for $\overline{R_2}$, and the MRE for this case reveals positive bias (Figure 2c and d). The remaining cases resulted in similarly low levels of bias and imprecision for $\overline{R_1}$ and for $\overline{R_2}$ (MARE and 90% intervals ≤ 0.17 (0, 0.40)), with the latter reference point estimated with less accuracy and precision than the former; the results for the RE and MRE of the recruitment reference points closely track those of the MARE for these cases (Figure 2c and d).

Accuracy and precision of non-fishery mortality

All cases had similarly low values of bias and imprecision for mortality in the pre-collapse period ($\overline{M_1}$; all MARE <0.03 and the upper bound of the 90% ARE intervals ≤ 0.10), and the results for the REs closely track those of the AREs (Figures 2e and 3e). Results for non-fishery mortality in the pre-collapse period are not shown for the case omitting the disease survey (Table 2) because this mortality is fixed; results for this case merely reflect the variation across simulations and should not be compared with the other case results.

The cases omitting the diver surveys resulted in the poorest performance, and omitting the disease survey resulted in the next poorest performance, for estimating mean mortality in the second period ($\overline{M_2}$), with MARE values of 0.08 and 0.06, respectively. However, the former cases resulted in only a 0.01–0.02 increase in average bias and precision compared with the remaining four cases.

Which cases lead to high probabilities of model failure (ETP)?

The case omitting the disease survey resulted in the highest ETP for all reference points except $\overline{M_1}$ (not estimated; see description of this case) and $\overline{R_1}$ (Table 6). This case resulted in 56 and 43% probabilities of model failure when estimating $\overline{M_2}$ and B_{2013}, respectively. Omission of the diver survey resulted in the highest probability of model failure when estimating $\overline{M_2}$ and $\overline{R_1}$, with probabilities 36 and 11%, respectively. For all cases (including the base), $\overline{M_2}$ was estimated least reliably of all reference points, with ETP values ranging from 8 to 56%.

Which cases lead to high probabilities of incorrect management conclusions (PIM)?

Omitting the disease and the diver surveys resulted in the highest total probability of incorrect management conclusions (49 and 41%, respectively); the remaining cases led to total probabilities between 18 and 25% (Table 7, bottom row). Omitting the disease survey resulted in a 77% probability of an incorrect fishery opening and omitting the diver survey lead to a 58% such probability; both cases resulted in considerably higher probabilities of incorrect openings, rather than incorrect closures. The remaining cases resulted in probabilities of incorrect closures of $\geq$24%, and probabilities of incorrect openings of $\leq$23%, i.e. each of these cases had higher probabilities of leading to an incorrect fishery closure, rather than an incorrect fishery opening.

Which surveys provide the highest return on estimating the forecast quantity?

The disease survey resulted in the highest return on investment (0.22 ICR in thousands) when estimating B_{2013}, and the aerial survey resulted in the next best return on investment, but its ICR was 100 times lower than for the disease survey (Figure 4a). The PWSSC hydroacoustic survey had relatively the lowest return on

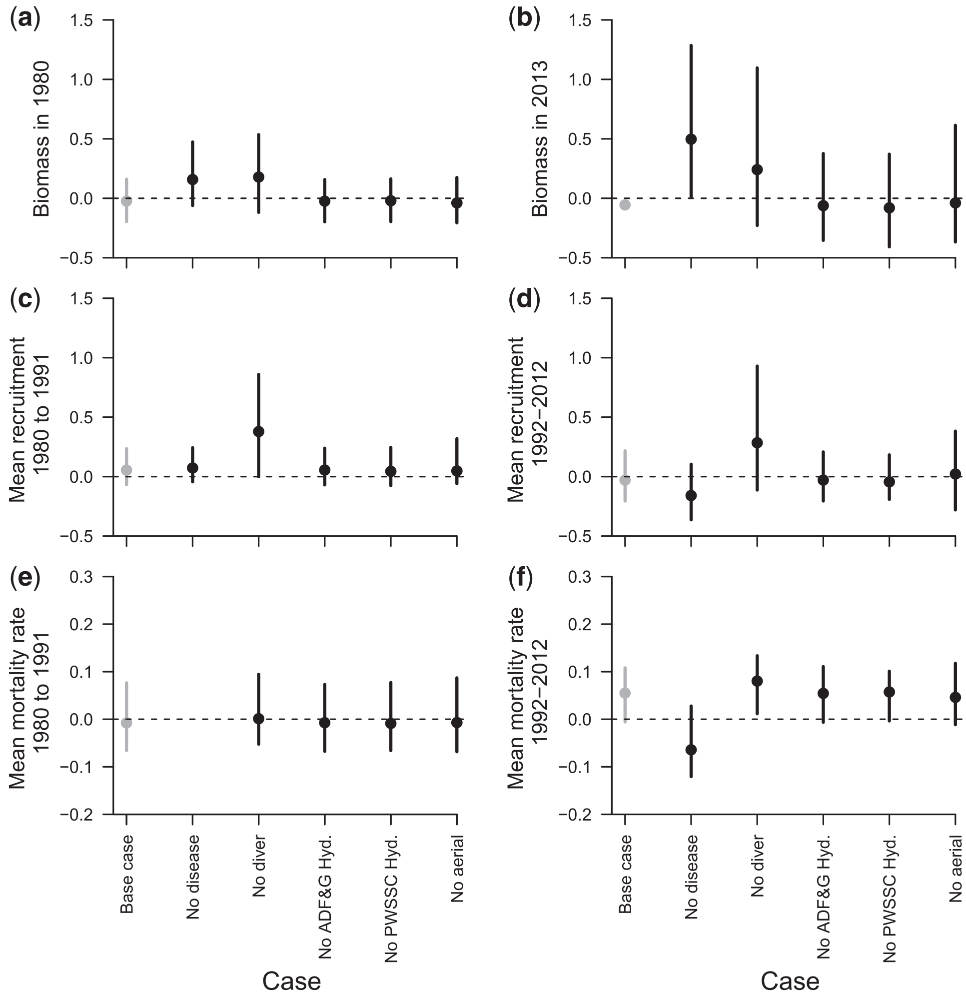

Figure 2. Relative error (RE) by case: each panel presents the results across cases by reference point, so panels (a)–(f) correspond to reference points (4.1)–(4.6). Filled-in points represent the median and lines illustrate the 90% interval of the RE across simulations. In each panel, the base case is grey. See main text for why no results show in (e) for "No disease".

investment with an ICR six times lower than that for the aerial survey.

Discussion

Removing the oldest dataset (the diver survey) resulted in the largest positive bias across biomass, mean recruitment, and mean mortality. The diver survey took place during 1984–1997 and, in the base case, is used as an absolute index of biomass alongside three relative biomass indices. When this absolute index was removed, the model overestimated biomass compared with the scale provided by the remaining indices because the suite of data used by the herring model lacks adequate information from the catch age-composition to estimate mortality. It should be noted, notwithstanding the diver survey, it would be generally expected that older data would be identified as having low value in the present time, although older data would have been valuable in the past.

Using the information from the data spanning 1980–1992, the herring biomass appears to have increased while total catches were increasing, which could be explained by a large biomass. Then, once the biomass was low (after 1992), the fishery was closed and the catch time-series ended, except for a limited number of years of small catches. Therefore, the model lacks adequate contrast in the catch data to estimate the scale of the biomass so that when high biomass levels are proposed during the MCMC iterations, the model can scale recruitment up and select low q values for the remaining time-series to fit the trends in the indices. These results reveal that providing the assessment model with an absolute index of abundance is necessary to keep the scale of the biomass close to the scale of the data.

Removing the information contained in the disease survey led to underestimation in mean recruitment and mean mortality during the post-collapse period ($\overline{M_2}$) and to overestimation of the forecast biomass (Figure 2). Underestimation of $\overline{M_2}$ was because of the assumptions made to develop the estimation model for this case (Table 2): 1992–1993 were the only years that involved instantaneous mortality rates greater than the assumed background rate for ages 3–8. As discussed in the Methods section, this case was testing an assessment model that lacked disease data and therefore did not integrate the assumption of increased mortality in the years following the collapse. Thus, the model lacking the disease survey information estimated additional mortality in 1992–1993 at a high enough rate to reduce the biomass to fit the sudden decrease in the biomass indices. However, without additional mortality in the remaining years, the model was forced to estimate recruitment levels low enough to fit

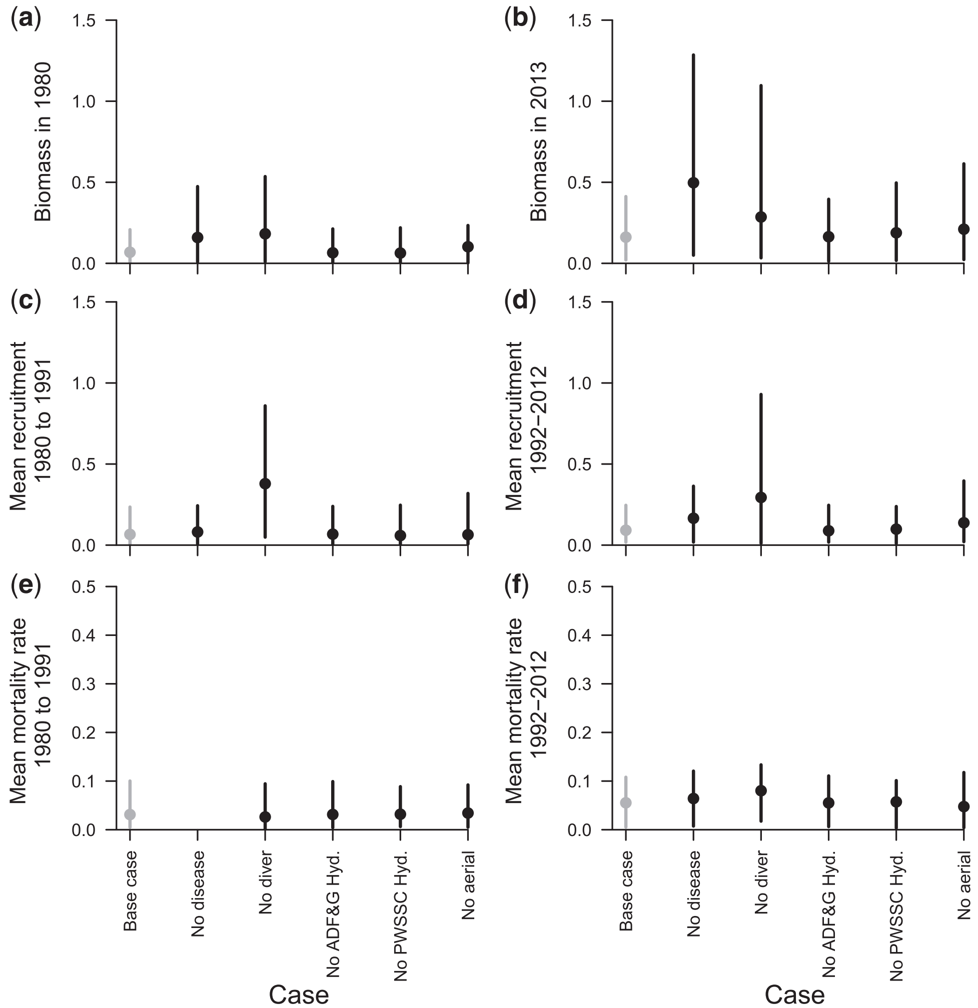

Figure 3. Absolute relative error (ARE) by case: each panel presents the results across cases by reference point, so panels (a)–(f) correspond to reference points (4.1)–(4.6). Filled-in points represent the median and lines illustrate the 90% interval of the ARE across simulations. In each panel, the base case is grey. See main text for why no results show in (e) for "No disease".

Table 6. Values of extreme tail probability (ETP) by reference point and estimation model.

Reference point	Base	No. disease	No. diver	No. ADF&G hyd.	No. PWSSC hyd.	No. aerial
B_{2013}	0.05	0.43	0.04	0.03	0.05	0.02
$\overline{M_1}$	0.02	NA	0.03	0.01	0.02	0.03
$\overline{M_2}$	0.18	0.56	0.36	0.13	0.15	0.08
$\overline{R_1}$	0.05	0.07	0.11	0.04	0.04	0.06
$\overline{R_2}$	0.01	0.12	0.10	0.01	0.00	0.02
Mean ETP	0.05	0.27	0.11	0.04	0.05	0.04

High values of ETP translate to a high probability of model failure and values of ETP $\leq$0.05 imply model reliability. The last row displays the mean ETP across reference points for each case.

Table 7. Probability of incorrect fishery closures (PIC), probability of incorrect fishery openings (PIO), and total probability of incorrect management conclusions (PIM) resulting from implementing the estimation model associated with each case.

Metric	Base	No. disease	No. diver	No. ADF&G hyd.	No. PWSSC hyd.	No. aerial
PIC	0.24	0.03	0.13	0.24	0.37	0.29
PIO	0.15	0.77	0.58	0.15	0.15	0.23
PIM	0.18	0.49	0.41	0.18	0.23	0.25

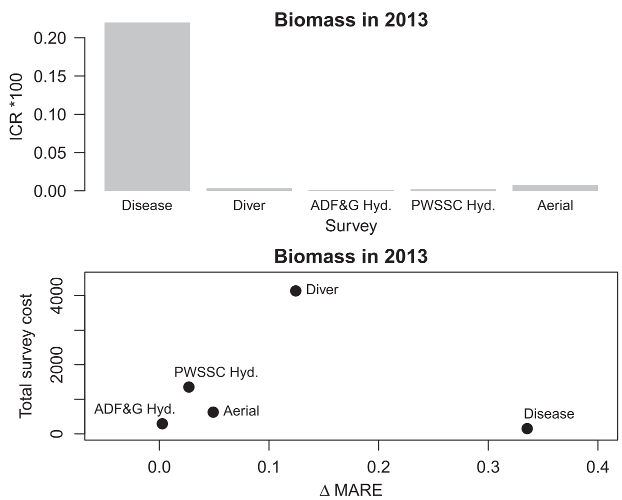

Figure 4. The top panel shows the information-to-cost ratio (ICR) for each survey for the estimate of the forecast biomass, B_{2013}. The bottom panel shows the cost of each survey programme plotted by the improvement in the MARE for the forecast biomass because of the addition of that survey's data.

the sustained low biomass levels, leading to negative bias in $\overline{R_2}$. The positive bias in the forecast biomass for this case is explained by the passage of 10 years since the large mortality events in 1992–1993, combined with low mortality; the population accrued three generations of herring with very low mortality rates and this was sufficient time for the biomass to begin to exceed that of the operating model.

Removal of either of the two hydroacoustic surveys or the aerial survey data resulted in approximately the same relatively moderate increase in bias and imprecision across all reference points, including the forecast biomass, although the duration of the three surveys varied greatly (Table 3). All three surveys display trends that are in agreement so that when one is removed, the other two provide enough information on scale and relative biomass to adequately estimate the reference points, on average. Obviously, if all three of these surveys supplying trend data were removed simultaneously, this would result in a highly inaccurate stock assessment. Thus, it must be emphasized important to note that because we only modelled a leave-one-out set of scenarios, it cannot be concluded that costs could have been saved by removing two or more surveys simultaneously.

Which survey data are most important for estimating accurate and precise forecasts?

All cases resulted in more biased and less precise estimates of B_{2013} compared with B_{1980} because of incomplete cohorts in the age-composition data making it difficult to estimate recruitment and mortality in the later years. That being said, including the diver and the disease survey is best for minimizing ARE across all reference points, but especially in the forecast biomass (Figure 3b). Using results from ETP, omitting the disease survey presented the only significant challenge for the estimating reliable posteriors of the forecast biomass (Table 6).

Omitting which survey data could lead to incorrect fishery openings?

This question is vital for managers to ensure that their models do not consistently support setting a higher harvest rate than is expected from the management rules. Removing the disease survey or the diver survey from the assessment model led to the highest probabilities of incorrect fishery openings when using the posterior median of the forecast biomass to set the harvest rate; retaining these surveys in the assessment model would be advised under conservative management guidelines.

Omitting which survey data could lead to incorrect fishery closures?

This question is important for managers to ensure their assessment models do not consistently limit the amount of commercial fishing, resulting in lost revenue and impairing stakeholder relations. Removing the PWSSC hydroacoustic survey or the aerial survey resulted in the highest probabilities of an incorrect fishery closure, therefore retaining the PWSSC hydroacoustic and the aerial surveys would be economically important.

What are the trade-offs between cost of data and model accuracy?

Comparing the separate rankings of information, cost, and ICR shows trade-offs between cost of data and accuracy in estimation of reference points. Simply using how expensive it was to run each survey in the most recent year, the disease survey and the aerial survey were least expensive and the PWSSC hydroacoustic survey and the diver survey were the most expensive. Therefore, an analysis that only considered cost would support continuing the disease and the aerial surveys and would support discontinuing either the PWSSC hydroacoustic survey or the diver survey. However, the diver survey was one of the two most informative surveys when considering accuracy and precision of the forecast biomass, and the aerial survey was among the least informative. On the other hand, the aerial survey was important for ensuring against forecasts that lead to incorrect fishery closures, meaning the aerial survey provides return to commercial economic interests outside of the agency.

When considering the ICR, the disease survey unequivocally provided the highest return on investment, with the aerial survey providing the next highest return on investment (Figure 4, top panel). Although the diver survey was highly informative for estimating accurate and precise reference points—even more informative than the disease survey for estimating mortality and recruitment—the associated costs were high enough that the resulting return on investment was among the lowest returns.

Further exploring the trade-offs between removing the disease survey or the diver survey

Removing either the diver survey or the disease survey results in similar total probabilities of drawing an incorrect management conclusion (Table 7 bottom row), but the cost savings associated with removing the diver survey are much greater compared with the disease survey. Therefore, removing the diver survey results in similar probabilities of incorrect management conclusions, but results in much greater cost savings than removing the disease survey. When looking at MARE and ARE, removing the disease survey would result in the least accurate forecast biomass, but removal of the diver survey would result in the least accurate recruitment and mortality estimates.

When considering precision and accuracy per dollar invested (ICR), the disease survey is the most valuable and the diver,

Table 8. The best (**) and next best (*) surveys using five criteria that are important for management.

Metric	Criterion	Disease survey	Diver survey	ADF&G hyd. survey	PWSSC hyd. survey	Aerial survey
Cost	Cheap	**	–	–	–	*
MARE	Informative	**	*	–	–	–
ICR	High return	**	–	–	–	*
ETP	Reliability	**	*	–	–	–
PIM	Correct management	**	*	–	–	–

Surveys were ranked by how cheap they are, using collection cost in the most recent survey year (Table 3, final column), how informative they are, using Δ(MARE) (Figure 4, *x*-axis of bottom panel), and by their information-to-cost ratio (ICR) relating to the estimate of the forecast biomass, B_{2013} (Figure 4, top panel). Surveys were also ranked by their importance for model reliability (mean ETP: final row of Table 6) and their importance to making correct management decisions (PIM: final row of Table 7).

PWSSC hydroacoustic, and ADF&G hydroacoustic surveys are the least valuable. However, an estimation model without the diver survey would lead to incorrect fishery openings with >50% probability. Therefore, using only the results from the ICR could be considered to be supporting a management approach that places short-term economic gain above cautionary and conservative management. In addition, eliminating the diver survey would have, on average, resulted in 25% positive bias in the forecast biomass and could lead to posterior medians of the forecast biomass that are twice as large as the true biomass levels.

General methodology and future directions

Few marine programmes explicitly consider the value of information gained from a particular research programme compared with other uses for the same money (Hansen and Jones, 2008). Yet such studies often add great value. For example, Hansen and Jones (2008) demonstrate that management of invasive lampreys is best conducted by reducing the money spent determining which streams should be treated for lampreys, which allows more streams to be treated. In another study, fisheries affected by limits on the accidental catch of harbour porpoises would make more money by investing in abundance surveys of the porpoises to reduce uncertainty, and the resulting improvement in precision would increase the limits on porpoise by-catch, more than paying for the survey (Bisack and Magnusson, 2014). Similarly, studies on the value of two types of pre-fishery salmon surveys (Link and Peterman, 1998), and on the value of research into the stock structure of orange roughy (McDonald *et al.*, 1997), both provided strong evidence that such studies more than paid for themselves through increasing fishing revenue.

It is crucially important to assess the value of individual data collection programmes, as we did here. By assessing the value of each survey using several criteria, we also allow managers to perform multi-objective comparisons similar to the summary of results in Table 8. Managers of other stocks and species can adopt their own set of pertinent performance metrics into the methods outlined here to perform a tailored evaluation of their sampling programmes.

Although there are analytical advantages to omitting a single survey at a time, an obvious extension of this method would be to omit more than one survey at a time. Under this extension, caution should be used when interpreting results because it may be difficult to identify the cause of the resulting change in estimation performance. Also, care needs to be taken when doing this to ensure that subsets of surveys are grouped together using practical or reasonable criteria.

The results of this simulation study are based on interpreting the past usefulness of an entire survey as evidence of its future utility. Future work could expand the leave-one-out approach to simulate population dynamics into the future to project the continued value of conducting each survey, including testing different management rules, and assessing the effects of alternative operating models i.e. a full management strategy evaluation. However, the results from such an approach would rely on the ability of the study designer to predict and model the management loop and the resulting impact on the population (as done by McDonald *et al.*, 1997, for example). Simulating the future value of survey data in this way would have been difficult for the population in this study because future projections would ordinarily assume that herring stocks fluctuate substantially, but this particular population has remained at low levels for several decades, thus adding great uncertainty to predictions of future stock status. Such an approach may be more appropriate for a more stable fish population.

A third extension of this method would be to test the value of the frequency with which a survey is conducted by creating scenarios that omit years or periods of data from an individual survey and comparing estimation performance (e.g. Ono *et al.*, 2015). This extension of the current study design would give managers additional options for scheduling survey effort in ways that maintain model performance.

Conclusion

Exploring the trade-offs between the cost of running each survey and the deterioration in model performance through exclusion of that survey's data, provided a ranking of which historical sampling programmes have provided the most valuable data for forecasting biomass. The disease survey (which is relatively cheap and collects an index of additional mortality through disease) and the diver survey (which is relatively expensive and collects an absolute index of abundance) were the most valuable sampling programmes. This work is directly useful to Pacific herring researchers by providing critical information about how to prioritize monitoring efforts going forward. Furthermore, this work presents a general framework for evaluating the return on investment in surveys. Future researchers can use their own set of pertinent performance metrics and adopt the methods outlined in the second chapter to perform a tailored, multi-objective evaluation of their own sampling programmes.

Acknowledgements

This work was supported by the Exxon Valdez Oil Spill Trustee Council (13120111-Q to MLM and TAB); and partially supported by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement (NA15OAR4320063, Contribution No. 2019-1015 to MLM); and the Richard C. and Lois

M. Worthington Endowed Professorship in Fisheries Management (to TAB).

Disclaimer

The funders encourage publications but did not have any role in design, data, writing, or the decision to submit this particular article for publication.

Author contributions

TAB and MLM designed the study; MLM and TAB conducted the analyses and wrote the paper; and MLM, TAB., and AEP discussed the results and contributed to the manuscript.

References

Bisack, K. D., and Magnusson, G. 2014. Measuring the economic value of increased precision in scientific estimates of marine mammal abundance and bycatch: harbor porpoise *Phocoena phocoena* in the Northeast U.S. gill-net fishery. North American Journal of Fisheries Management, 34: 311–321.

Brown, E. D., and Baker, T. T. 1998. Injury to Prince William Sound Herring Following the Exxon Valdez Oil Spill. Exxon Valdez Oil Spill State/Federal Natural Resource Damage Assessment Final Report (Fish/Shellfish Study Number 11). Alaska Department of Fish and Game, Division of Commercial Fisheries Management and Development, Cordova, Alaska.

Bureau of Labor Statistics. 2015. National Consumer Price Index, CPI Inflation Calculator. Bureau of Labor Statistics, U. S. Department of Labor. http://www.bls.gov/data/inflation_calculator.htm (last accessed 1 April 2015).

Csepp, D. J., Vollenweider, J. J., and Sigler, M. F. 2011. Seasonal abundance and distribution of pelagic and demersal fishes in southeastern Alaska. Fisheries Research, 108: 307–320.

Deriso, R. B., Quinn, T. J. II, and Neal, P. R. 1985. Catch-age analysis with auxiliary information. Canadian Journal of Fisheries and Aquatic Sciences, 42: 815–824.

Fournier, D., and Archibald, C. P. 1982. A general theory for analyzing catch at age data. Canadian Journal of Fisheries and Aquatic Sciences, 39: 1195–1207.

Fournier, D. A., Skaug, H. G., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., *et al.* 2012. AD model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optimization Methods and Software, 27: 233–249.

Funk, F. C., and Sandonne, G. J. 1990. Catch-age analysis of Prince William Sound, Alaska, herring, 1973–1988. Fishery Research Bulletin No. 90-01. Alaska Department of Fish and Game, Division of Commercial Fisheries, Juneau, Alaska.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. 2004. Bayesian Data Analysis. Chapman & Hall/CRC, Boca Raton, FL.

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *In* Bayesian Statistics, 4, pp. 169–193. Ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. Oxford University Press, Oxford.

Hansen, G. J. A., and Jones, M. L. 2008. The value of information in fishery management. Fisheries, 33: 340–348.

Hose, J. E., McGurk, M. D., Marty, G. D., Hinton, D. E., Brown, E. D., and Baker, T. T. 1996. Sublethal effects of the *Exxon Valdez* oil spill on herring embryos and larvae: morphological, cytogenetic, and histopathological assessments, 1989–1991. Canadian Journal of Fisheries and Aquatic Sciences, 53: 2355–2365.

Hulson, P-J. F., Hanselman, D. H., and Quinn, I. I., T. J. 2012. Determining effective sample size in integrated age-structured assessment models. ICES Journal of Marine Science, 59: 281–292.

Hulson, P-J. F., Miller, S. E., Quinn, T. J. II, Marty, G. D., Moffitt, S. D., and Funk, F. 2008. Data conflicts in fishery models: incorporating hydroacoustic data into the Prince William Sound Pacific herring assessment model. ICES Journal of Marine Science, 65: 25–43.

Link, M. R., and Peterman, R. M. 1998. Estimating the value of in-season estimates of abundance of sockeye salmon (*Oncorhynchus nerka*). Canadian Journal of Fisheries and Aquatic Sciences, 55: 1408–1418.

Magnusson, A., and Hilborn, R. 2007. What makes fisheries data informative? Fish and Fisheries, 8: 337–358.

Marty, G. D., Okihiro, M. S., Brown, E. D., Hanes, D., and Hinton, D. E. 1999. Histopathology of adult Pacific herring in Prince William Sound, Alaska, after the Exxon Valdez oil spill. Canadian Journal of Fisheries and Aquatic Sciences, 56: 419–426.

Maunder, M. N., and Punt, A. E. 2013. A review of integrated analysis in fisheries stock assessment. Fisheries Research, 142: 61–74.

McDonald, A. D., Smith, A. D. M., Punt, A. E., Tuck, G. N., and Davidson, A. J. 1997. Empirical evaluation of expected returns from research on stock structure for determination of total allowable catch. Natural Resource Modeling, 10: 3–29.

Muradian, M. L., Branch, T. A., Moffitt, S. D., and Hulson, P-J. F. 2017. Bayesian stock assessment of Pacific herring in Prince William Sound, Alaska. PLoS One, 12: e0172153.

Norcross, B. L., Hose, J. E., Frandsen, M., and Brown, E. D. 1996. Distribution, abundance, morphological condition, and cytogenetic abnormalities of larval herring in Prince William Sound, Alaska, following the *Exxon Valdez* oil spill. Canadian Journal of Fisheries and Aquatic Sciences, 53: 2376–2387.

Ono, K., Licandeo, R., Muradian, M. L., Cunningham, C. J., Anderson, S. C., Hurtado-Ferro, F., Johnson, K. F., *et al.* 2015. The importance of length and age composition data in statistical age-structured models for marine species. ICES Journal of Marine Science, 72: 31–43.

Pearson, W. H., Deriso, R. B., Elston, R. A., Hook, S. E., Parker, K. R., and Anderson, J. W. 2012. Hypotheses concerning the decline and poor recovery of Pacific herring in Prince William Sound, Alaska. Reviews in Fish Biology and Fisheries, 22: 95–135.

Quinn, T. J. II, Marty, G. D., Wilcock, J., and Willette, M. 2001. Disease and population assessment of Pacific herring in Prince William Sound, Alaska. *In* Proceedings of the 18th Lowell Wakefield Fisheries Symposium, Herring 2000: Expectations for a New Millenium, Anchorage, AK, February 23–26, 2000, AK-SG-01-04, pp. 363–379. Ed. by F. Funk, J. Blackburn, D. Hay, A. J. Paul, R. Stephensen, R. Toreson, and D. S. Witherell. University of Alaska Sea Grant College Program, Fairbanks, AK.

Thomas, G. L., and Thorne, R. E. 2003. Acoustical-optical assessment of Pacific herring and their predator assemblage in Prince William Sound, Alaska. Aquatic Living Resources, 16: 247–253.

Thorne, R. E., and Thomas, G. L. 2008. Herring and the "Exxon Valdez" oil spill: an investigation into historical data conflicts. ICES Journal of Marine Science, 65: 44–50.

Wetzel, C. R., and Punt, A. E. 2011. Model performance for the determination of appropriate harvest levels in the case of data-poor stocks. Fisheries Research, 110: 342–355.

Yin, Y., and Sampson, D. B. 2004. Bias and precision of estimates from an age-structured stock assessment program in relation to stock and data characteristics. North American Journal of Fisheries Management, 24: 865–879.

Handling editor: Stan Kotwicki