Gray whale southbound migration surveys 1967–2006: an integrated re-analysis

JEFFREY L. LAAKE[1], ANDRE E. PUNT[2], RODERICK HOBBS[1], MEGAN FERGUSON[1], DAVID RUGH[1] AND JEFFREY BREIWICK[1]

Contact e-mail: Jeff.Laake@noaa.gov

ABSTRACT

Between 1967 and 2007, 23 seasons of shore-based counts of the Eastern North Pacific (ENP) stock of gray whales (*Eschrichtius robustus*) were conducted throughout all or most of the southbound migration near Carmel, California. Population estimates have been derived from these surveys using a variety of techniques that were adapted as the data collection protocol evolved. The subsequent time series of estimates was used to evaluate trend and population status, resulting in the conclusion that the population was no longer endangered and had achieved its optimum sustainable population (OSP) level. We re-evaluated the data from all of the surveys using a common estimation procedure and an improved method for treatment of error in pod size and detection probability estimation. The newly derived abundance estimates between 1967 and 1987 were generally larger (–2.5% to 21%) than previous abundance estimates. However, the opposite was the case for survey years 1992 to 2006, with estimates declining from –4.9% to –29%. This pattern is largely explained by the differences in the correction for pod size bias, which occurred because the pod sizes in the calibration data over-represented pods of two or more whales and underrepresented single whales relative to the estimated true pod size distribution.

KEYWORDS: ABUNDANCE ESTIMATE; GRAY WHALES; WHALING – ABORIGINAL

INTRODUCTION

The National Marine Fisheries Service (NMFS) has conducted shore-based counts of the Eastern North Pacific (ENP) stock of gray whales (*Eschrichtius robustus*) in central California during December–February for 23 years with the first survey in 1967–1968 and the most recent in 2006–2007. Since 1974–1975 these surveys have been conducted from a cliff overlooking the ocean at Granite Canyon (36° 26' 41'' N), 13km south of Carmel. Prior surveys (1967–1974) were conducted at Yankee Point (36° 29' 30'' N), 6km north of Granite Canyon. The surveys have been conducted in this region because most gray whales migrate within 6km of land along this section of the coastline (Shelden and Laake, 2002), apparently due to the deep marine canyons north of Granite Canyon.

These survey data have been used to estimate abundance of the gray whale stock using various techniques (Buckland *et al.*, 1993; Hobbs *et al.*, 2004; Laake *et al.*, 1994; Reilly, 1981; Rugh *et al.*, 2008b; Rugh *et al.*, 2005). The resulting sequence of abundance estimates has been used to estimate the population's growth rate (Buckland and Breiwick, 2002; Buckland *et al.*, 1993), which resulted in removal of ENP gray whales from the US List of Endangered and Threatened Wildlife on 16 June 1994 (Federal Rule 59 FR 31095), and the more recent conclusion reported by Angliss and Outlaw (2008) and Angliss and Allen (2009) that the ENP gray whale stock was within its optimum sustainable population (OSP) range as defined by the US Marine Mammal Protection Act (MMPA).

Recently, Rugh *et al*. (2008c) evaluated the accuracy of various components of the shore-based survey method, with the focus on pod size estimation. They used a pair of observers working together to track one pod of whales at a time to evaluate error in pod size estimates made by the independent observers conducting the standard survey. They compared their correction factors to similar values constructed from aerial surveys in 1978–1979 (Reilly, 1981), 1992–1993 and 1993–1994 (Laake *et al.*, 1994), and from paired thermal sensors in 1995–1996 (DeAngelis *et al.*, 1997). The additive correction factors that had been used to compensate for bias in pod size estimates differed among the various data sets; in particular, the correction factors estimated by Laake *et al.* (1994) were substantially larger than those estimated by Reilly (1981). This was of concern because the 1987–88 abundance estimate (Buckland *et al.*, 1993) used the correction factors from Reilly (1981) and all subsequent estimates (1992–1993 to 2006–2007) used the correction factors from Laake *et al.* (1994). Also, the estimates for the surveys prior to 1987 in the trend analysis were scaled based on the abundance estimate from 1987–88. This meant that the first 16 abundance estimates used one set of correction factors, and the more recent seven abundance estimates used different (and larger) correction factors which would influence the estimated trend and population trajectory.

Additionally, there have been other subtle differences in analysis methods used for the sequence of abundance estimates. For example, the number of hours on watch has been reduced from 10 to 9 per day. Also, a pod was the sample unit used for fitting the migration curve for estimates prior to 1995, whereas whales were used (after correcting for bias in pod size estimates) subsequently. Thus, a re-evaluation of the analysis techniques and a re-analysis of the abundance estimates were warranted to apply a more uniform approach throughout the years. We have explored the additive correction factor for pod size bias developed by

[1] *National Marine Mammal Laboratory, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, 7600 Sand Point Way NE, Seattle, WA 98115.*
[2] *School of Aquatic and Fisheries Sciences, University of Washington, Seattle, WA 98195-5020.*

Reilly (1981) and show that it requires some strong assumptions that are unlikely to be met in practice. We devised a better approach with weaker assumptions and incorporated it into an analysis that was used to estimate abundance for all 23 surveys.

METHODS

Field survey methods

The survey data collection protocol has remained largely unchanged over the 40-year time span, but some refinements to the protocol have been made to reduce observer fatigue, collect more data, and provide more accurate data measurements (Table 1). During the survey, an observer scans the ocean (typically without binoculars) and locates passing whales that are visible when they blow, surface or dive showing their flukes. For all surveys, the sighting times, pod size estimates, and some measure of offshore distance were recorded. Also, start and end of watch effort and environmental conditions (e.g. Beaufort sea state (wind force) and visibility) were also recorded. In earlier years, observers may have searched a wide area, but since the late 1980s, there has been increasing emphasis on searching only the area directly west and north of the site. This has reduced confusion with sightings at great distances. In more recent years, when a whale was first seen, the time, horizontal angle, and reticle were recorded for the initial sighting and, if seen again, when the whale surfaced again near an imaginary line perpendicular to the coast (at a magnetic angle of 241°). This allowed calculation of travel speed and trajectory relative to the coast.

The primary shift in survey protocol occurred in 1987–1988 when several important changes were made (Table 1):

(1) Prior to 1987–1988, changes in environmental conditions (i.e. Beaufort sea state and visibility classification) were recorded only at the beginning of a watch and when a sighting occurred, or up to two more times during the watch if no sightings occurred during the watch. This approach precluded measuring the exact amount of time spent surveying at specific environmental conditions, which is important because these factors affect the observers' ability to detect whales. That was corrected starting in 1987–1988 when the survey protocol was changed to record the time and conditions whenever they changed, regardless of whether any sightings occurred.

(2) Offshore distance (perpendicular to the coast at the observer's location) prior to 1987–1988 was estimated visually without calibration, and the accuracy of these estimates is unknown. All subsequent measurements of distance were made with reticle readings etched in 7 × 50 binoculars. These marks provided quantification of the angle from the horizon to a sighting. Using an observer's eye height above the surface of the ocean (between 21 and 23m depending on which part of the research station bluff was used), the reticle measurements were converted to a radial distance from the observer to the whale (Lerczak and Hobbs, 1998). The distance offshore is computed from the radial distance and the horizontal angle measured with the

Table 1

Gray whale shore-based count locations, dates, and field methods. The index *y* for year refers to the year at the beginning of the survey (e.g. *y* = 1995 for the 1995–1996 survey). YP refers to Yankee Point and GC to Granite Canyon survey locations.

Year(y)	Location	Start date	End date	Watch periods per day[1]	Paired obs.	Distance data[2]	Visibility[3]	Pod size bias
1967	YP	18/12/1967	03/02/1968	2–5h each	–	Intervals	Sky/dist	–
1968	YP	10/12/1968	06/02/1969	2–5h each	–	Intervals	Sky/dist	–
1969	YP	08/12/1969	08/02/1970	2–5h each	–	Intervals	Sky/dist	–
1970	YP	09/12/1970	12/02/1971	2–5h each	–	Intervals	Sky/dist	–
1971	YP	18/12/1971	07/02/1972	2–5h each	–	Intervals	Sky/dist	–
1972	YP	16/12/1972	16/02/1973	2–5h each	–	Intervals	Sky/dist	–
1973	YP	14/12/1973	08/02/1974	2–5h each	–	Intervals	Sky/dist	–
1974	GC	10/12/1974	07/02/1975	2–5h each	–	Intervals	Sky/dist	–
1975	GC	10/12/1975	03/02/1976	2–5h each	–	Intervals	Sky/dist	–
1976	GC	10/12/1976	06/02/1977	2–5h each	–	Intervals	Sky/dist	–
1977	GC	10/12/1977	05/02/1978	2–5h each	–	Intervals	Sky/dist	–
1978	GC	10/12/1978	08/02/1979	2–5h each	–	Intervals	Vis codes	Aerial
1979	GC	10/12/1979	06/02/1980	2–5h each	–	Intervals	Vis codes	–
1984	GC	27/12/1984	31/01/1985	2–5h each	–	Intervals	Vis codes	–
1985	GC	10/12/1985	07/02/1986	3–3 or 3.5h each	–[4]	Intervals	Vis codes	–
1987	GC	10/12/1987	07/02/1988	3–3 or 3.5h each	✓	Reticles	Vis codes	–
1992	GC	10/12/1992	07/02/1993	3–3 or 3.5h each	✓	Reticles	Vis codes	Aerial
1993	GC	10/12/1993	18/02/1994	3–3h each	✓	Reticles	Vis codes	Aerial
1995	GC	13/12/1995	23/02/1996	3–3h each	✓	Reticles	Vis codes	Thermal[5]
1997	GC	13/12/1997	24/02/1998	3–3h each	✓	Reticles	Vis codes	Tracking
2000	GC	13/12/2000	05/03/2001	3–3h each	✓	Reticles	Vis codes	–
2001	GC	12/12/2001	05/03/2002	3–3h each	✓	Reticles	Vis codes	–
2006	GC	12/12/2006	22/02/2007	3–3h each	✓	Reticles	Vis codes	–

[1]1967–68 to 1984–85: two watch periods per day of 5 hours each, from 07:00–17:00; 1985–86 to 1992–93: three watch periods per day (07:00–10:30 hours, 10:30–13:30 hours, 13:30–17:00 hours); 1993–94 to 2006–07: three 3 hour watch periods (07:30–10:30 hours, 10:30–13:30 hours, 13:30–16:30 hours).
[2]Intervals were 0–¼ nautical miles (nmi), ¼–½ nmi, ¾–1 nmi, 1–1.5 nmi, 1.5–2 nmi, etc. Distances have been based on binocular reticles since 1987–88.
[3]No visibility codes were recorded prior to 1978–79. Instead observers recorded sky conditions and sometimes miles as an indication of visibility. Those values were translated to visibility codes 1–5 used through 1987–88. In 1992–93 observers began recording visibility in six subjective categories (1 = excellent; 6 = useless), a system used since.
[4]Small-scale trial double-observer study conducted for 6 days but not used in the analysis.
[5]Thermal data for pod size bias were not used in this analysis because pod and observer were not recorded.

binocular compass. During the 1987–1988 and 1992–1993 surveys, a reticle measurement was recorded only for the whale sighting closest to the 241° line. For all subsequent surveys, reticle readings were recorded for both the north and south sightings of a pod, if it was seen twice. This provided calculations of whale travel speed.

(3) Until 1987–1988, all surveys were conducted with a single observer on watch at a time, with the exception of a small test conducted in 1986 (Rugh *et al.*, 1990). To enable estimation of pods missed by an observer during the watch, a second concurrent independent observer was used throughout the 1987–1988 survey and for portions of the survey in all subsequent surveys. By matching the measurements of offshore distance, timing of the whale passage across the 241° line, and pod size, it was possible to assess which pods were seen in common and which were missed by one of the observers. This double-count approach follows standard capture-recapture methodology (Buckland *et al.*, 1993; Otis *et al.*, 1978).

Analysis methods

Past abundance estimates have been derived as the product of the count of pods and a series of multiplicative correction factors. Buckland *et al.* (1993) and Laake *et al.* (1994) used the following abundance estimator:

$$\hat{N} = m\bar{s}f_t f_n f_m f_s \ , \qquad (1)$$

where the observed number of pods (under acceptable visibility conditions), *m,* was multiplied by the mean pod size ($\bar{s}$) (i.e. $m\bar{s}$ is the total whale count) and correction factors for: (1) pods passing outside watch periods, f_t; (2) night travel rate, f_n ; (3) pods missed during watch periods, f_m ; and (4) bias in pod size estimation, f_s. Not included in these corrections are whales passing beyond the viewing range of the observers (only 1.28% of the population, according to Shelden and Laake (2002)) and whales passing the station well before or after the census, which is assumed to be a very small number. Estimates from 1995–1996 to 2006–2007 used the abundance estimator of Hobbs *et al*. (2004):

$$\hat{N} = \hat{W}f_t f_n \ , \qquad (2)$$

where $\hat{W}$ is an estimate of the number of whales that passed during the watch periods and includes corrections for both pod size bias (f_s) and pods missed by the observers during the watch (f_m).

The analysis method developed here is even more integrated than the method used by Hobbs *et al*. (2004), and the resulting abundance estimator can be expressed simply as:

$$\hat{N} = \hat{W}f_n \ , \qquad (3)$$

where $\hat{W}$ is an estimate of the number of whales that passed during the entire migration with corrections for pod size bias and missed pods but without differences in night vs. day passage rates. Although explicit multiplicative correction factors are not used, equivalent values for comparison to previous analysis were calculated.

Ideally, there would be data in each year to construct a year-specific value for each correction factor. However, there is no single year in which all of the data were collected to estimate each correction factor (Table 1). Despite this shortcoming, it is possible to estimate $f_{t,y}$ for each year, so a naïve estimate of abundance ($\tilde{W}_y$) can be constructed for each year (y):

$$\tilde{W}_y = m_y \bar{s}_y f_{t,y} \ , \qquad (4)$$

where $\tilde{W}_y$ is an estimate of whales passing during the migration with a correction only for whales that passed outside of the watch periods, $f_{t,y}$.

Calibration data for pod size bias were collected during only five surveys (Table 1), so year-specific data were not available but the correction factor ($f_{s,y}$)was partially year-specific due to annual differences in the distribution of pod sizes. A year-specific value for missed pods ($f_{m,y}$) was computed for each of the last eight surveys (Table 1) because independent double-observer data were collected for all or portions of the survey such that each observer's detection probability could be estimated. Thus, for the last eight surveys a more 'complete' estimate of abundance with year-specific correction factors $f_{t,y}$, $f_{m,y}$, and $f_{s,y}$ but a constant night time correction factor was constructed. To construct comparable estimates for the first 15 surveys when these data were not available, a conventional ratio estimator (Cochran, 1977) was used with $\hat{W}_y$ and $\tilde{W}_y$ values for the last eight surveys and that estimated ratio was used to scale the naïve abundance estimates from each of the first 15 surveys.

Detail of each of the methods for handling pod size error, pods missed by the observer while on watch and estimation of abundance for each year are outlined below. All of the methods described here were implemented in the R (*R* Development Core Team, 2009) statistical computing environment. Both the data and the R code have been archived into an R package named ERAnalysis[3] that can be used with R to reconstruct the analysis and results presented here.

Pod size calibration

Estimates of the size of migrating gray whale pods are subject to error, with a tendency to undercount the number of whales in a pod because of the observer's oblique view from shore and the asynchrony of diving among whales in a pod. That is, multiple whales surfacing separately within a pod are often confused with a single whale surfacing multiple times. The magnitude and sign of the errors obviously depends on the true size of the pod. For example, it is possible that close, multiple dives of a single whale could be misconstrued as more than one whale in a pod, but by definition, underestimation cannot occur for a single whale. In contrast, a large pod of whales could be potentially counted as a single whale if the whales were close together and no more than one whale was observed at the surface simultaneously. The most reliable count of a pod occurs when all of the whales are observed at the start of a deep dive, when there is some synchrony to the group and each shows its fluke.

To address this source of error, two calibration methods were used (Table 2). In the first method, an aircraft was used to observe whale pods and count the number of whales in a pod while observers from shore recorded their independent

[3] *http://www.afsc.noaa.gov/nmml/software/eranalysis.php*

Table 2

Summary of gray whale pod size calibration data. Some observers provided estimates in more than one year and each pod was not observed by each observer. Only one or two estimates per pod were obtained via land tracking because they calibrated the single or double observers during the standard watch.

Year	Type	No. of pods	No. of observers	No. of observations
1978–79	Aerial	25	12	295
1992–93	Aerial	21	5	79
1993–94	Aerial	39	7	157
1997–98	Land tracking	111	10	192
Total		196	28	723

estimates of pod size. With the aerial view and relatively clear water, an accurate count of whales in a pod could be obtained, considered here to be the true pod size. Aerial surveys were conducted during the 1978–1979 southbound survey (Reilly, 1981) and during the 1992–1993 and 1993–1994 surveys (Laake *et al.*, 1994). To avoid the expense of an aircraft survey, another test of pod size estimation was conducted wherein pairs of observers tracked whales continuously through the viewing area with a theodolite or binoculars while observers on the standard watch maintained an independent effort (Rugh *et al.*, 2008c). The pod size measurements determined during the tracking were considered to be the true pod size and were later compared to the estimates of the observers conducting the standard watch. The aerial survey has the obvious advantage of providing a more reliable true pod size but was not as realistic because the shore-based observers were not conducting a standard watch and were focused solely on estimation of a single pod size. The tracking experiments more closely emulated pod size measurement for an observer conducting a standard watch, but the 'true' pod size measurement from the trackers may have not always been accurate because their view was similar to the shore observer. Pod size calibration data were also generated with paired thermal sensors in 1995–1996 (DeAngelis *et al.*, 1997). However, these data were not recorded such that each pod and observer could be identified (W. Perryman, Southwest Fisheries Science Center, National Marine Fisheries Service, pers. comm.), so these data were not considered in this analysis because it was not possible to evaluate those random effects.

It is important to examine the methodology of Reilly (1981) to understand the differences between the correction factors from these various data sources as reported by Rugh *et al*. (2008c). Initially we develop the notation and outline an alternative method with a much weaker assumption to be used in the re-analysis. Let S represent true pod size and s represent recorded pod size. With the survey data, we can measure $h(s)$, the distribution of observed (recorded) pod sizes, but we want to measure $f(S)$, the distribution of true pod sizes. If we knew the probability that an observer would record a pod of true size S as size s, $g(s|S)$, we could solve for $f(S)$ from the following convolution:

$$h(s) = \sum_S f(S) g(s|S) \; . \qquad (5)$$

For the calibration data, we know S. We measure the proportion of times observers record s for a pod of true size S, which provides a direct measurement of $g(s|S)$. Determination of $f(S)$ from equation (5) is a standard approach with discrete data for deriving the distribution of the true values (S) from the recorded values (s) and estimated calibration function, $g(s|S)$. (e.g. Heifitz et al. 1998). This approach does assume that $g(s|S)$ remains constant but $f(S)$ can vary annually, so the 'correction factor' expressed as the ratio of average true pod size to average recorded pod size ($\sum_S Sf(S) / \sum_s sh(s)$) will likely vary.

In contrast, Reilly (1981) constructed a set of adjustments, $c(s)$, from the pod size calibration data that were added to each recorded pod size s in the survey data. The c(s) were constructed by tabulating the values of S for each pod the observers recorded as size s and computing $c(s) = \bar{S} - s$. In the Appendix we provide the details to demonstrate that these additive adjustments are valid only if the distribution of true pod sizes selected for calibration $f^*(S)$ equals the distribution of true pod sizes during the survey $f(S)$. However, a simple thought experiment can demonstrate why the method could be substantially biased and hence is not appropriate in general. Consider, a survey in which $f(S) = 0.25$ for $S = 1, \ldots, 4$, but for the calibration experiment only pods of true size $S = 4$ were selected. That would lead to $c(s) = 4 - s$ because the average true size in the calibration data ($\bar{S}$) would always be 4 regardless of the value of s. Use of those data would lead to an estimate of 4 for the average pod size when the true value was 2 for the scenario we proposed. While such a pod selection strategy would never be chosen, it does demonstrate the potential bias that could occur if the distribution of selected pods for calibration did not match the true pod size distribution. While it may be possible to select pods randomly with regard to true size, the Reilly (1981) approach would require the pod size calibration data to be collected each year unless true pod size distribution never changed, which seems unlikely.

Differences in adjustment values, $c(s)$, for different calibration data sets as reported in Rugh *et al.* (2008c) can result from differences in either $f^*(S)$ or $g(s|S)$. If the differences reported by Rugh *et al.* (2008c) are due to differences in $g(s|S)$, that may reflect inherent variability in observer ability or variability due to inherent difference in the calibration pods (e.g. frequency and timing of surfacing, proximity of whales within a pod, and distance from observer). However, substantial bias could result if the differences are due to the selection of pods $f^*(S)$ during the different calibration experiments and $f(S)$ varies annually.

Four pod size calibration data sets (Table 2) were used to estimate g($s|S$), an Sxs calibration matrix with a row for each true value S and a column for each observed value s up to some reasonable maximum true pod size S^{max}. We used $S^{max} = 20$. If there were sufficient calibration data for all true pod sizes, a saturated multinomial model could be used with each cell estimated as the proportion of observations that were recorded to be size s that were in fact a true pod size S. However, the available calibration data were fairly sparse for true pod sizes >3 because most pods contain only 1–3 whales. Instead, a more parsimonious approach of fitting parametric distributions for $g(s|S)$ was chosen. We considered a truncated Poisson (for $s-1$)

$$g(s|S) = \frac{\alpha_S^{s-1} e^{-\alpha_S}}{(s-1)!\mu_S} \; , \qquad (6)$$

and a truncated discretised gamma distribution defined as:

$$g(s|S) = \int_{s-1}^{s} \frac{b_S^{a_S} x^{a_S-1} e^{-xb_S}}{\Gamma(a_S)\mu_S} dx \ . \quad (7)$$

Each of the distributions was truncated such that $s \le S^{max}$ (i.e. $\mu_S = \sum_{s=1}^{s^{\max}} g(s|S)$). The calibration function depends on S through the parameters. Models with separate parameters for S = 1,2,3 were considered because they represented the majority of the data, and we collapsed pods of true size >3 (4+). For $S > 3$, the log of the rate parameter (b_S in the gamma and α_S in the Poisson) of the distribution was expressed as a linear function of S. For the gamma shape parameter (a_S), four parameters, one for each S in the set S = 1,2,3,4+ were specified. The likelihood without any random effects is:

$$\mathcal{L}(\psi \mid s_{ij}) \propto \prod_i \prod_j g(s_{ij} \mid S_i) \ , \quad (8)$$

where ψ is the vector of parameters for the distributions, i indexes the pod, j indexes the observer and $g(s|S)$ is replaced with either of the parametric distributions. The dependence of $g(s|S)$ on ψ is implicit. As an example, the likelihood for a Poisson distribution is:

$$\mathcal{L}(\alpha_1,\alpha_2,\alpha_3,a,b \mid s_{ij}) \propto \prod_{S=1}^{3} \left(\frac{\alpha_S^{s-1} e^{-\alpha_S} / (s-1)!}{\mu_S} \right)^{n_{s,S}} \prod_{S>3} \left(\frac{e^{(s-1)(a+bS)} e^{-e^{(a+bS)}} / (s-1)!}{\mu_S} \right)^{n_{s,S}} \ , \quad (9)$$

where the parameter vector for this example is ψ = $(\alpha_1,\alpha_2,\alpha_3,a,b)$, $n_{s,S}$ is the number of observers that recorded size s when the true size was S and μ_S is the S-specific normalising sum over s = 1,…20 to ensure that the largest pod size s was less than or equal to S^{max} ($s \le S^{max}$).

The four calibration data sets (Table 2) were pooled and models fitted with a single set of S-dependent parameters. Models were also fitted with different S-dependent parameters for each of the four calibration data sets. In addition models with random effects for pod, observer and year (data set) were considered. The random effect was implemented by assuming a normal distribution $N(0,\sigma_\varepsilon^2)$ for the random effect (ε) on the log of the rate. Using the gamma distribution, a general likelihood for any single random effect was:

$$\mathcal{L}(\boldsymbol{a_S},\boldsymbol{b_S},\sigma|s_{ij},\, S_i) \propto \sum_k log \int_{-\infty}^{\infty} \left[\prod_{i\in I_k} \prod_{j\in J_k} \int_{s_{ij}-1}^{s_{ij}} \frac{e^{a_{S_i}\left(\log\left(b_{S_i}\right)+\varepsilon\right)} x^{a_{S_i}-1} e^{-xe^{\left(\log\left(b_{S_i}\right)+\varepsilon\right)}}}{\Gamma\left(a_{S_i}\right)\mu_{S_i}} dx \right] \frac{e^{-\varepsilon_k^2/2\sigma_\varepsilon^2}}{\sqrt{2\pi}\sigma_\varepsilon} d\varepsilon_k, \quad (10)$$

where the summation is over the k sets defined by the random effect (e.g. k = 1,....,n), i,j indexes the pods and observers within the respective sets I_k, J_k defined by the k^{th} random effect value, and $\boldsymbol{a_S}$ = (a_1,a_2,a_3,a_{4+}) and $\boldsymbol{b_S}$ = $(b_1,b_2,b_3,b_{4+} = e^{\beta_0+\beta_1 S})$. As an example, for a pod random effect k = 1,....,n = 196, $I_k = k$ and J_k is the set of observers that made estimates for the k^{th} pod. For the gamma random effect model $g(s \mid S)$ is:

$$g(s|S) = \int_{-\infty}^{\infty} \int_{s-1}^{s} \frac{e^{a_S\left(\log(b_S)+\varepsilon\right)} x^{a_S-1} e^{-xe^{\left(\log(b_S)+\varepsilon\right)}}}{\Gamma(a_S)\mu_S} dx \frac{e^{-\varepsilon^2/2\sigma_\varepsilon^2}}{\sqrt{2\pi}\sigma_\varepsilon} d\varepsilon \ . \quad (11)$$

Random effects models for the Poisson were constructed similarly. Each parametric distribution was fitted by solving for the maximum likelihood estimates using *optim* in R 2.9.1 (*R* Development Core Team, 2009); the most parsimonious model was selected using AIC.

Using the estimated $g(s|S)$ from the calibration data, allows derivation of an estimate of $f(S)$from the survey data for any year using a multinomial likelihood with either a saturated model (i.e. separate parameter for each value of S) or a parametric model for $f(S)$. The latter was chosen because it was more parsimonious and used a discretised gamma distribution:

$$f(S \mid \theta) = \int_{S-1}^{S} \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)} dx \ , \quad (12)$$

where $\boldsymbol{\theta}$ = (a,b). Other parametric models could be formulated for $f(S)$ but the gamma is sufficiently flexible to fit a variety of distribution shapes. To derive an estimate of $f(S)$ directly from the observed distribution of pod sizes $h(s)$, involves an assumption that the size of the pod did not influence the probability that the pod was seen. However, previous analyses (Buckland *et al.*, 1993; Hobbs *et al.*, 2004; Laake *et al.*, 1994; Rugh *et al.*, 2008b) show that larger pods are more likely to be seen. Consequently, an unbiased estimator for $f(S)$ from the observed data cannot be derived without accounting for detection probability.

Correcting for missed pods

From 1967 to 1985, a single observer searched and recorded migrating gray whale pods during the surveys. Beginning in 1987, two observers surveyed independently for all or some portion of the survey timeframe. These independent counts provided the mark-recapture framework (Buckland *et al.*, 1993) to estimate the proportion of pods that were missed by an observer by matching recorded pods based on offshore distance, timing, and pod size (Rugh *et al.*, 1993). The Appendix contains the details of the algorithm that was used to assess which pods were seen by both observers and which were missed by one of the observers. As part of that matching process pods seen in close proximity (time and offshore distance) by the same observer were linked (combined) for both observers prior to matching. Pods were linked to cope with situations in which one observer combined two close pods and the other observer recorded them as two separate pods. Estimated detection probability from the mark-recapture analysis and the abundance estimates were based on these linked pods. The notation n^* is used for the number of pods recorded by an observer and n ($\le n^*$) is used to denote the number of linked pods used in the analysis.

In each of the prior analyses of the gray whale survey data (Buckland *et al.*, 1993; Hobbs *et al.*, 2004; Laake *et al.*,

1994; Rugh *et al.*, 2008b), pod size was an important predictor for pod detection. A pod with more whales will involve more surfacings and will provide more obvious visual cues resulting in a greater number of opportunities for detection. In each of those prior analyses, the recorded pod size (s) was used as the covariate but this approach has a couple of disadvantages. When a pod was seen by both observers, disagreement between the recorded sizes was ignored in the analysis. In addition, recorded pod size s is not the best predictor for detection probability. For example, an observer might record a pod of three whales as a single whale if only one whale was at the surface at a time. Yet, one would expect far more surfacing events from asynchronous surfacing of a pod of three whales than a single whale, and would expect that it would be more likely to be detected than the single whale even though $s = 1$ in both cases. Detection probability was represented in terms of the true unknown size S and summed over the distribution of true pod sizes $f(S)$ which was simultaneously estimated from the data by including the pod size calibration matrix (eqn 11). Independent errors in pod size measurement were used when both observers detect a pod.

The additional notation ignores the year index to simplify the notation. Let,

x_{ij} = an indicator variable = 1 if the i^{th} of n pods is seen by the observer at the j^{th} station (j = 1,2) and 0 otherwise;

s_{ij} = recorded size of the i^{th} pod by the observer at the j^{th} station (j = 1, 2) if it was seen by the observer at the j^{th} station; and

$\gamma_j(\boldsymbol{C}_i,S)$ = probability that the observer at the j_{th} station (j = 1,2) sees the i^{th} pod which has a vector of associated covariates $\boldsymbol{C}_i$ and a true pod size S.

S is unknown, and the recorded pod size (s) is known only for observed pods. Either one or two estimates of pod size result if observers at one or both stations detect the pod. We sum over all possible values of S (1 to $S^{\max}$) weighting by the estimated probability distribution $f(S)$ and the estimated pod size calibration matrix $g(s|S)$. For each observed pod, we compose the vector of indicator variables (x_{i1}, x_{i2}) which has the possible observable values (1,0), (0,1) and (1,1). The vector (0,0) represents a pod that was missed by the observers at both stations.

Given that at least one observer detected the pod, the probability of observing the vector (x_{i1}, x_{i2}, s_{i1}, s_{i2}) for the i^{th} pod is:

$$p\left(x_{i1},x_{i2},s_{i1},s_{i2}\right)=\sum_S f(S)\prod_{j=1}^{2}g\left(s_{ij}|S\right)^{x_{ij}} \frac{\gamma_j\left(\boldsymbol{C}_i,S\right)^{x_{ij}}\left[1-\gamma_j\left(\boldsymbol{C}_i,S\right)\right]^{1-x_{ij}}}{1-\prod_{j=1}^{2}\left[1-\gamma_j\left(\boldsymbol{C}_i,S\right)\right]}, \quad (13)$$

Let $\boldsymbol{\theta}$ be the parameter vector for $f(S)$ and let $\boldsymbol{\varphi}$ be the parameter vector for the detection function γ. Then, the likelihood for the double-observer data, conditional on $g(s|S)$ is:

$$\mathcal{L}\left(\theta,\varphi|\psi,\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{s}_1,\boldsymbol{s}_2\right)= \prod_{i=1}^{n}p\left(x_{i1},x_{i2},s_{i1},s_{i2}\right), \quad (14)$$

where $n = n_1+n_2-n_3$ is the total number of pods seen by either observer, and $n_1$were seen by the primary observer, n_2 were seen by the secondary observer, and n_3 were seen by both observers. When there was only a single observer on watch, no information about γ can be derived, but the single observers' sightings for estimation of $f(S)$ can be used and γ will influence those measurements through the effect of S on detection. The conditional distribution for true pod size S for detected pods with covariates $\boldsymbol{C}$ is:

$$f\left(S|detected\right)=\frac{f(S)\gamma(\boldsymbol{C},S)}{\sum_S f(S)\gamma(\boldsymbol{C},S)} \quad (15)$$

The likelihood for the $n_1$observations by the single observer also conditional on $g(s|S)$ is:

$$\mathcal{L}\left(\theta,\varphi|\psi,s_1,\ldots,s_{n_1}\right)=\prod_{i=1}^{n_1}\frac{\sum_S f(S)\gamma\left(\boldsymbol{C}_i,S\right)g(s_i|S)}{\sum_S f(S)\gamma\left(\boldsymbol{C}_i,S\right)} \quad (16)$$

The two component likelihoods for the single- and double-observer data can be multiplied (or log-likelihoods summed) to derive the maximum likelihood estimates for the parameter vector ($\boldsymbol{\theta},\boldsymbol{\varphi}$). Pod size calibration data alone provide information about the $g(s \mid S)$ parameter vector $\boldsymbol{\psi}$ because there is no known true pod size contained in the double-observer data to assess bias.

A logistic distribution was used for the detection function $\gamma(\boldsymbol{C},S)$ and models considered with covariates $\boldsymbol{C}$ containing offshore (perpendicular) distance (km) with intervals (0–1, 1–2, 2–3, 3–4, 4+), and observer (each person). Additional models with Beaufort sea state or visibility as numeric covariates or visibility classified as Excellent–Good and Fair–Poor were then considered. The data from each of the eight years were analysed separately. The model that minimised Akaike Information Criterion (AIC) in each year was used but any models containing Beaufort sea state or visibility that showed an increase in detection probability with worsening environmental conditions were excluded.

Abundance estimation

With the correction for pod size bias and missed pods, we expanded the recorded number of whales during a watch to an estimate of the number of whales that actually passed during the watch. That prediction could be based on data from observers at both stations when two observers were on watch and a single observer when only one station was occupied. However, we chose to avoid this complication and used only the data from the observer at the designated primary station because in most years the additional data would not have improved precision very much. The predicted number of whales was based on a Horvitz-Thompson estimator ($1/p$), which provides an estimate of the number of pods (whales) that passed from those that were seen using the estimated detection probabilities. The reasoning for this estimator can be illustrated with a simple example. If one observes a pod and estimates its detection probability to be 0.5, then it is expected that one pod was missed for every pod that was seen, so the Horvitz-Thompson estimator results in a doubling of the observed number of pods (1/0.5 = 2).

The observed pod size was used with the correction for pod size bias and the estimate of $f_y(S)$ to make inference

about the probable true pod size S from the recorded size s using the conditional distribution:

$$f_y(S|s) = \frac{f_y(S)g(s|S)}{\sum_S f_y(S)g(s|S)} \quad , \tag{17}$$

where we now use index y for survey year to be explicit about which portions vary by year. Using this conditional distribution, the estimator for the number of pods passing during the j^{th} period of year y when the primary observer was searching (on watch) in year y from the n_{1jy} linked pods is:

$$\hat{P}_{jy} = \sum_{i=1}^{n_{1jy}} \sum_S f_y(S|s_{ijy}) \frac{1}{\gamma_y(\boldsymbol{C}_{ijy}, S)} \quad , \tag{18}$$

and the estimator for the number of whales is:

$$\hat{W}_{jy} = \sum_{i=1}^{n_{1jy}} \sum_S f_y(S|s_{ijy}) \frac{S}{\gamma_y(\boldsymbol{C}_{ijy}, S)} \quad . \tag{19}$$

Surveys were conducted for 9 to 10 hours a day, and it is known that whales migrate throughout the day and night (Perryman *et al.*, 1999). In addition, the environmental conditions can compromise sighting probability or become so poor that migrating whales are not visible to the observer and survey effort is suspended. Thus, it is also necessary to expand the estimate from the time observed to the total migration timeframe to account for whales that passed when no observers were surveying.

This second prediction component of the abundance estimate uses a migration curve fitted to the predicted number of whales passing when the observer was searching (on watch) to predict the total number passing including periods when the observer was not on watch (i.e. night time or poor visibility). The fitted migration curve is needed because the migration rate changes during the course of the survey (typically exhibiting a peak in mid-January) and because the amount of survey effort throughout the migration timeframe varies unpredictably due to varying visibility conditions. The timing and duration of those off-effort periods can severely impact the observed count of whales due to the variation in the migration rate (e.g. missing a day in mid-January has a greater impact than missing a day in early December).

For each survey year y, consider a sample of $j = 1, \ldots, m_y$ effort periods of length $l_{1y}, l_{2y}, \ldots, l_{m_y y}$ for time intervals that are not always consecutive such that $l_{jy} = t_{1jy} - t_{0jy}$, where the 0 and 1 indices represent the beginning and ending times of the interval. A curve can be fitted to the sequence of migration passage rates (whales/hour) $\hat{W}_{jy}/l_{jy}$, at the time mid-points ($t_{jy} = (t_{0jy} + t_{1jy}) / 2$). Following Buckland *et al.* (1993), we added an assumed value of 0 whales passing for day 0 and T to anchor the fitted curve when it was assumed whales did not pass. For each year a generalised additive model (GAM) was fitted with an assumed quasi-Poisson family for the $\hat{W}_{jy}$ $j = 1,\ldots,m_y$ with an offset of $\log(l_{jy})$ to account for varying length of observation period and to allow for over-dispersion. The function *mgcv* (version 1.5–5) (Wood, 2006) in R 2.9.1 (*R* Development Core Team, 2009) was used to fit the GAMs. The Poisson mean $\lambda_y(t) = e^{\xi_y(t)}$ used a log-link with a default smoother over time $\xi_y(t)$. This approach provides a much more flexible modelling technique than the normal-Hermite adjustment modelling of Buckland *et al.* (1993).

With a fitted migration curve, abundance was estimated by summing the expected value of the number of whales passing each day from time 0 to T_y:

$$\hat{W}_y = \sum_{t=0.5}^{T_y - 0.5} \hat{\lambda}_y(t) \quad . \tag{20}$$

For most years, $T_y = 90$ where the days are counted with the origin ($t = 0$) at 12:00 am 1 December. The only exceptions were 2000 and 2001 when the migration extended to $T_y = 100$ days. Buckland *et al.* (1993) constructed a multiplier as the integral of the migration model over the migration period (0, T_y) divided by the integral over the sampled periods:

$$f_{ty} = \frac{\int_0^{T_y} \lambda_y(u)du}{\sum_{j=1}^{m_y} \int_{t_{0jy}}^{t_{1jy}} \lambda_y(u)du} \quad , \tag{21}$$

and the multiplier was used to inflate the estimate of the whales passing during the sampled periods to the entire migration as follows:

$$\hat{W}_y = f_{ty} \sum_{j=1}^{m_y} \hat{W}_{jy} \quad . \tag{22}$$

The formulation for abundance (eqn 20) provided an easier way to formulate a variance and it provided nearly identical results as eqn 22.

For each of the eight survey years from 1987–1988 to 2006–2007, an estimate of abundance $\hat{W}_y$ (y indexes the year) was derived using the above methods. However, there were no double-count data prior to 1987, and there was almost no overlap in personnel during these two periods. Offshore distance was also not reliably measured prior to 1987. From prior analyses, it is known that detection of whales depends on the observer and offshore distance (Buckland *et al.*, 1993; Hobbs *et al.*, 2004; Laake *et al.*, 1994; Rugh *et al.*, 2008b; Rugh *et al.*, 2005). Thus, we could not use a common detection model from recent years and apply it to the earlier years because both distance and observer could not be used as covariates for years prior to 1987. As an alternative, we chose to construct a common total correction factor for a naïve estimate of abundance ($\tilde{W}_y$) was developed by fitting a GAM with a smooth over time $\tilde{\lambda}_y(t)$ for the observed count of whales $\tilde{W}_{jy} = \sum_{i=1}^{n^*_{1jy}} s_{ijy}$ in each of the m_y effort periods of length l_{jy} and predicting total abundance based on the sum of the predicted daily numbers of whales passing $\tilde{W}_y = \sum_{t=0.5}^{T_y - 0.5} \tilde{\lambda}_y(t)$. This was essentially the same process defined above but without any correction factors for missed pods, pod size bias, etc. A conventional ratio estimator (Cochran, 1977) was then constructed using the $\hat{W}_y$ and $\tilde{W}_y$ values for the eight surveys from 1987 to 2006:

$$\hat{R} = \frac{\sum_{y=1987}^{2006} \hat{W}_y}{\sum_{y=1987}^{2006} \tilde{W}_y} \quad , \tag{23}$$

The ratio was used as a multiplicative correction factor for the naïve estimates prior to 1987 ($y = 1967, \ldots, 1985$):

$$\hat{W}_y = \hat{R}\tilde{W}_y \tag{24}$$

Applying the ratio estimator to naïve abundance estimates for previous years, involves the assumption that the factors that affect detection of whales and bias in pod size measurement were similar on average across years. Survey data that were collected only when the conditions were such that the Beaufort sea state was 4 or less and visibility was fair or better (codes 1 to 4) were used to minimise variation due to environmental conditions. Data were filtered based on entire watch periods, because environmental conditions were not recorded continuously prior to 1987. If recorded environmental conditions exceeded the criterion for any sighting or effort period within the watch, all of the data for the watch were excluded. This filter was applied to all surveys, even though that was not necessary for the last eight surveys, because we thought that it was important to maintain a consistent treatment of the data to apply the ratio and to obtain a valid assessment of trend and population status.

Estimation of the variance-covariance matrix for the sequence of abundance estimates is complicated because there are three sources of estimation error: (1) Σ_1 includes variation from parameter estimation error for pod size ($\boldsymbol{\theta}$) and detection probability ($\boldsymbol{\varphi}$), (2) Σ_2 includes variation from parameter estimation error for the pod size calibration parameters ($\boldsymbol{\psi}$), and (3) Σ_3 includes variation from estimation error in fitting the GAM passage rate parameters and residual temporal variation in the number of migrating whales. The element-wise total of the three component matrices, each 23 × 23 (23 surveys), provides the variance-covariance matrix of the abundance estimates. We will use i = 1,...,23 and j = 1,...,23 to index the rows and columns of the elements of the covariance matrix. The estimates of abundance co-vary because the first 15 estimates depend on $\hat{R}$ which was computed from the last eight estimates, and the last eight estimates co-vary because they all used the same estimated set of pod size calibration parameters $\boldsymbol{\psi}$ for $g(s|S)$.

The delta method was used to estimate each of the variance-covariance matrices for abundance. The estimator can be represented in general as $\mathbf{D'\Sigma_\zeta D}$ where $\boldsymbol{\zeta}$ is a vector of k parameters, $\boldsymbol{\Sigma}_\zeta$ is the kxk variance-covariance matrix for $\boldsymbol{\zeta}$ and $\boldsymbol{D}$ is a kxm matrix of first derivatives of the quantities derived from $\boldsymbol{\zeta}$. For this specific case, m = 23 for the 23 estimates of abundance and k varied depending on the set of parameters in the variance component. For some of the parameters, the complex interaction of the parameters and the abundance estimators was such that it was only reasonable to estimate the derivative matrix $\boldsymbol{D}$ numerically, which meant computing each of the abundance estimates for each value of $\zeta_k \pm \delta\zeta_k$ (where δ = 0.001 and ζ_k is the maximum likelihood estimator of the k^{th} parameter) and estimating the rate of change (first derivative) for each abundance estimator.

For Σ_1, the variance-covariance matrix of the pod size ($\boldsymbol{\theta}$) and detection probability ($\boldsymbol{\varphi}$) parameters was obtained from the inverse of the Hessian matrix derived from the optimization of the log-likelihood, which was derived with the function *optim* in R 2.9.1 (*R* Development Core Team, 2009). The first derivative matrix was estimated by varying each parameter, which in turn would change the predicted number of whales passing in each watch, so each GAM model was refitted to predict the change in total abundance. The detection and pod size parameters for each of the 8 recent survey years were fitted separately so the covariances are all 0 ($\sigma_{ij} = 0$ for i = 16, …, 23 and j = 16, …, 23 and $i \neq j$). All other σ_{ij} were non-zero due to the use of R to scale the first 15 survey estimates.

For Σ_2, the variance-covariance matrix of the pod size calibration parameters ($\boldsymbol{\psi}$) were also obtained from the inverse of the Hessian matrix using the selected parametric distribution for S = 1, 2, 3, and 4+. The same general technique used for Σ_1 was used for this variance-covariance matrix except that the pod size calibration parameters affect both estimated detection probability ($\boldsymbol{\varphi}$) and pod size ($\boldsymbol{\theta}$) parameters and the fitted GAM model. For each of the pod size calibration parameters in $\boldsymbol{\psi}$, evaluating the first derivative numerically required optimising the likelihood for the detection and pod size model and then subsequently re-fitting the GAM and predicting each abundance.

For Σ_3, the variance components required the computation of the variance for the predicted total abundance from the fitted GAM. The smooth function derived using *mgcv* is represented as a matrix of linear predictors ($\mathbf{L}$) and parameters ($\boldsymbol{\beta}$). For year y, let $\boldsymbol{\Sigma}_{\mathbf{L}\boldsymbol{y}}$ be the variance-covariance matrix of the k parameters for the linear predictor and let $\mathbf{L}_\mathbf{y}$ be the T_yxk linear predictors for the GAM. Then the variance estimator for total abundance in year y (for $y \geq 1987$) is:

$$\hat{var}(\hat{W}_y) = \left(\lambda_y \boldsymbol{L}_y\right)' \Sigma_{\boldsymbol{L}_y} \left(\lambda_y \boldsymbol{L}_y\right) + c_y \hat{W}_y \ , \qquad (25)$$

where $\boldsymbol{\lambda}_y = e^{\boldsymbol{L}_y \boldsymbol{\beta}_y}$ is a vector of T_y predicted daily abundances of migrating whales, $\boldsymbol{\beta}_y$ is the vector of k parameters and c_y is the over-dispersion scale parameter of the fitted quasi-Poisson. A similar variance can be constructed for naïve abundance estimator $\tilde{W}_y$ for all surveys derived from fitting the GAM to the observed whale counts:

$$\hat{var}(\tilde{W}_y) = \left(\tilde{\lambda}_y \tilde{\boldsymbol{L}}_y\right)' \tilde{\Sigma}_{\boldsymbol{L}_y} \left(\tilde{\lambda}_y \tilde{\boldsymbol{L}}_y\right) + \tilde{c}_y \tilde{W}_y \ , \qquad (26)$$

For σ_{ii}, i = 1,…,15, the diagonal elements $\hat{var}(\tilde{W}_y)$ for $y <$ 1987 are estimated using the delta method:

$$\hat{var}(\hat{W}_y) = \tilde{W}_y^2 \sigma_R^2 \ (k+1) + \hat{R}^2 \hat{var}\left(\tilde{W}_y\right) \ , \qquad (27)$$

where σ_R^2 is the variance of the ratio estimator $\hat{R}$ (Cochran, 1977) for the k = 8 surveys. The first term is the prediction variance for $\hat{R}$ and the second term includes variance for the naïve abundance estimator. For the off-diagonal elements i = 1,...,15 and j = 1, …,15 and $i \neq j$, $\sigma_{ij} = \tilde{W}_{yi} \hat{W}_{yj} \sigma_R^2$. For i = 1, ...,15 and j = 16,...,23,

$$\sigma_{ij} = \sigma_{ji} = \tilde{W}_{y_i} \left(\frac{\hat{var}(\hat{W}_{y_j})}{\tilde{W}_{y_j}} - \frac{\hat{W}_{y_j}^2}{\tilde{W}_{y_j}^3} \hat{var}(\tilde{W}_{y_j}) \right) . \qquad (28)$$

Night time differential

For surveys conducted during 1994–1996, Perryman *et al.* (1999) demonstrated that the night time passage rate was 28% higher during the latter half of the migration (> 15 Jan.). Using this as the median migration date (f = 0.5; 50% migrated before and 50% after), based on a 9-hour day and 15-hour night, Rugh *et al*. (2005) estimated a multiplicative correction factor of 1.0875 with a standard error of $f \times 15$ / 24 × 0.116 after correcting the typographical errors in Perryman *et al*. (1999). Here, a 14-hour night is assumed to avoid the minor but complicating adjustment that would be

needed to account for the 10-hour survey from 1967 to 1987 and 9-hour survey since 1992. A constant night time correction factor of f_n = 1.0817 (SE = 0.0338) was applied to each of the 23 estimates to create the final abundance estimates

$$\hat{N}_y = f_n \hat{W}_y \tag{29}$$

The adjusted variances and covariances in the matrix $\boldsymbol{V}$ are:

$$\hat{var}\left(\hat{N}_y\right) = \hat{var}\left(f_n \hat{W}_y\right) = (f_n \hat{W}_y)^2 \left(\left(\frac{0.0338}{1.0817} \right)^2 + \frac{\hat{var}(\hat{W}_y)}{(\hat{W}_y)^2} \right) \tag{30}$$

and

$$\hat{cov}(\hat{N}_{y_i}, \hat{N}_{y_j}) = f_n^2 \hat{cov}(\hat{W}_{y_i}, \hat{W}_{y_j}) \tag{31}$$

Where $\hat{var}(\tilde{W}_y)$ are the diagonal elements of $\Sigma_1+\Sigma_2+\Sigma_3$ and are the off-diagonal elements.

RESULTS

Naïve abundance estimates

GAMs were fitted to the observed passage rates (whales/hour) over time for each survey year (Fig. 1), using the recorded data from the primary observer during survey periods in which Beaufort sea state never exceeded 4 and visibility was fair or better (1 to 4). With the fitted GAMs, naïve estimates of abundance were computed (Table 3), that ranged from 7,000 to nearly 16,000. Without corrections for error in pod size, missed pods, or a night time differential, the naïve estimates would expectedly be lower than the true abundance.

Pod size calibration

Pod size calibration data were collected on 196 pods in four years (Table 2). The distribution of pods included 69, 56, 28, and 26 of true sizes S = 1 to 4, and an additional 8,6,2,1 pods of true sizes of 5, 6, 8, and 10, respectively. For each pod, as few as 1 and as many as 12 observers estimated a size for the pod (Table 2).

The more flexible gamma model provided a better fit than the Poisson (Table 4). A gamma mixed-effects model with a random effect for pod (eqn 10) was the most parsimonious (Table 5). A random pod effect captured the apparent variation amongst whale pods in the whale's behaviour, spatial separation of whales and synchronicity in surfacing of whales in a pod. As expected, pod size was typically underestimated with some small (usually <0.1) probability of overestimation (Fig. 2).

Correcting for missed pods

There were two independent observers throughout the 1987–1988 survey, so the number of matched observations was considerably greater than for the other survey years that had only partial double counts (Table 6). The average detection rate for the primary observer, ignoring any covariates, ranged from 0.70 to 0.81 across years (Table 6); thus, it can crudely be estimated that 20 to 30% of the pods that passed through the viewing area during watch periods with adequate visibility were missed by the observer at the primary station.

The fitted detection probability models (Table 7) demonstrated that the observers were most likely to miss pods of single whales and whales at offshore distances greater than 4km. There was also considerable variation among observers. For example, observers #6 and #10 in 2001 had respective detection probabilities of 0.91 and 0.71 for pods with two whales at the intermediate distances of 1 to 2km. With the exception of the 1995–1996 survey, observers were most likely to detect pods between 1 to 2km which was the corridor where most whales passed (Shelden and Laake, 2002). Pods within 1km were less likely to be detected because of the observer's focus farther offshore and because whales were in view for less time when travelling closer to shore. Visibility was an important predictor only in 1987 and 1993 and Beaufort sea state only in 2006 (Table 7).

Expected pod size $E(S)$ from the fitted survey-specific gamma pod size distributions, ranged from 1.72 to 2.63 whales per pod and was on average 11% (range: 3.9 – 18.8%) greater than the year-specific observed mean size of linked pods ($\bar{s}$) (Table 7). The computed $E(S)$ adjusts for two sources of bias $\bar{s}$ in with opposite directions. Inclusion of pod size calibration data $g(s|S)$ increased $E($S) relative to and accounting for size-biased detection of pods (i.e. larger pods are easier to see) decreased $E(S)$.

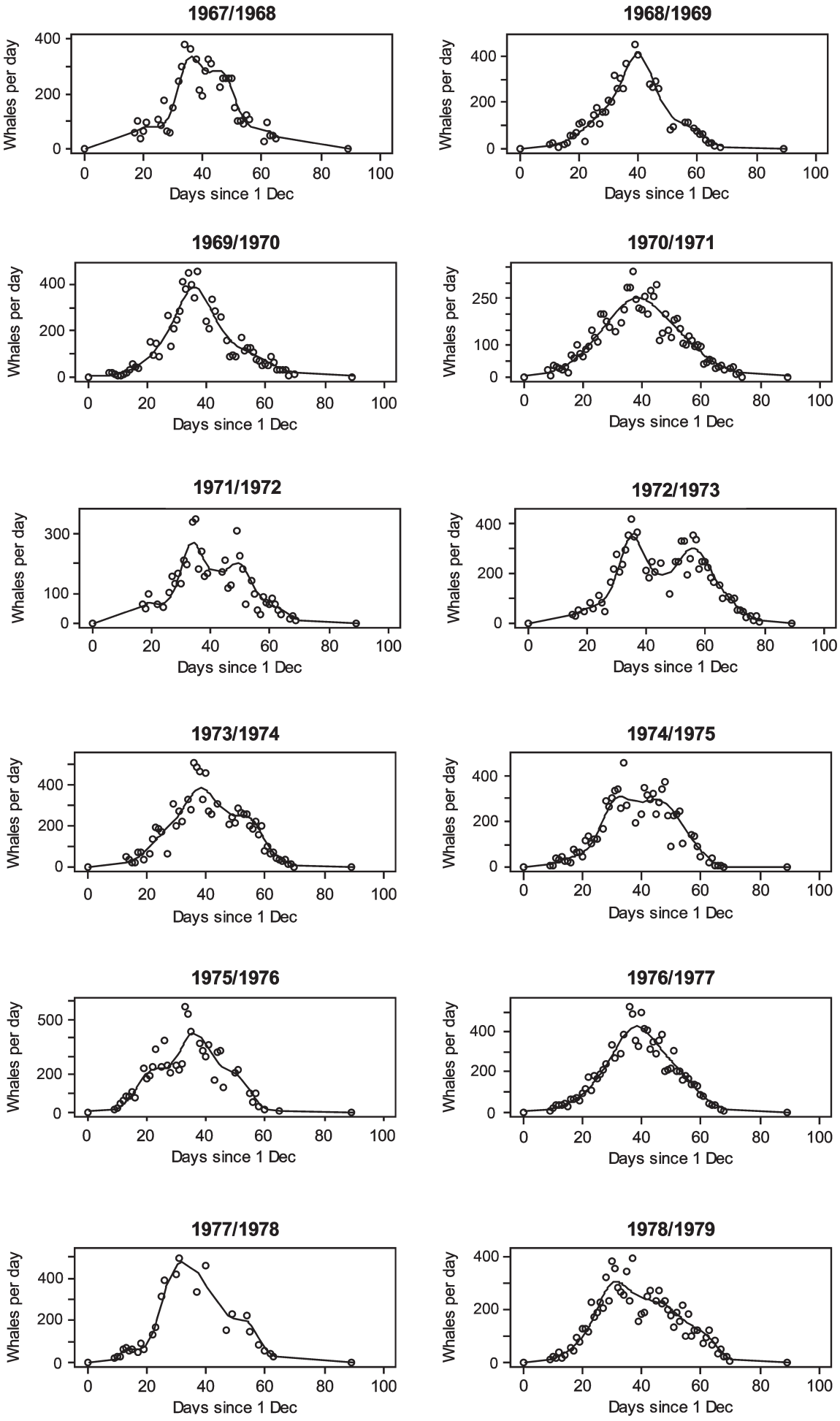
Abundance estimation

Whale passage rates (whales/hour) were estimated within each watch interval using the year-specific fitted models for pod size and missed pods (eqn 19), based on the observations from the primary observer after linking pods to correspond with the linking process for matched pods (Table 8). A year-specific GAM (Fig. 3) was fitted to the estimated whale passage rates to estimate total abundance ($\hat{W}_y$) (eqn 20) based on the daytime passage rate (Table 8). The ratio estimate $\hat{R}$ (eqn 23) was used to correct the naïve abundance estimates (eqn 24) for the 15 surveys from 1967 to 1985. Then all of the year-specific estimates were multiplied by the nighttime correction factor to obtain the final abundance estimate $\hat{N}_y$ (eqn 29) for each year (Table 9).

The newly derived abundance estimates (Fig. 4) between 1967 and 1987 were generally larger (–2.5% to 21%) than those reported by Rugh *et al.* (2008a). However, the opposite was the case for survey years 1992 to 2006 with estimates declining from –4.9% to –29%. This pattern is largely explained by the differences in the correction for pod size bias (Table 9) which occurred because the distribution of pod sizes from the calibration data over-represented pods of two or more whales and underrepresented single whales relative to the estimated true pod size distribution (Fig. 5).

DISCUSSION

When the southbound gray whale surveys were initiated in 1967 and a single observer searched and counted passing whales, those researchers had not anticipated that such a complicated process was needed to estimate abundance of the gray whale population. However, the data collection and estimation processes had to be adapted to account for the apparent deficiencies and biases resulting from variable environmental conditions, the limits of human visibility and cognition, and vagaries in whale behaviour as the survey process was evaluated (Perryman *et al.*, 1999; Reilly, 1981;


1967/1968
1968/1969
1969/1970
1970/1971
1971/1972
1972/1973
1973/1974
1974/1975
1975/1976
1976/1977
1977/1978
1978/1979
Whales per day
Days since 1 Dec

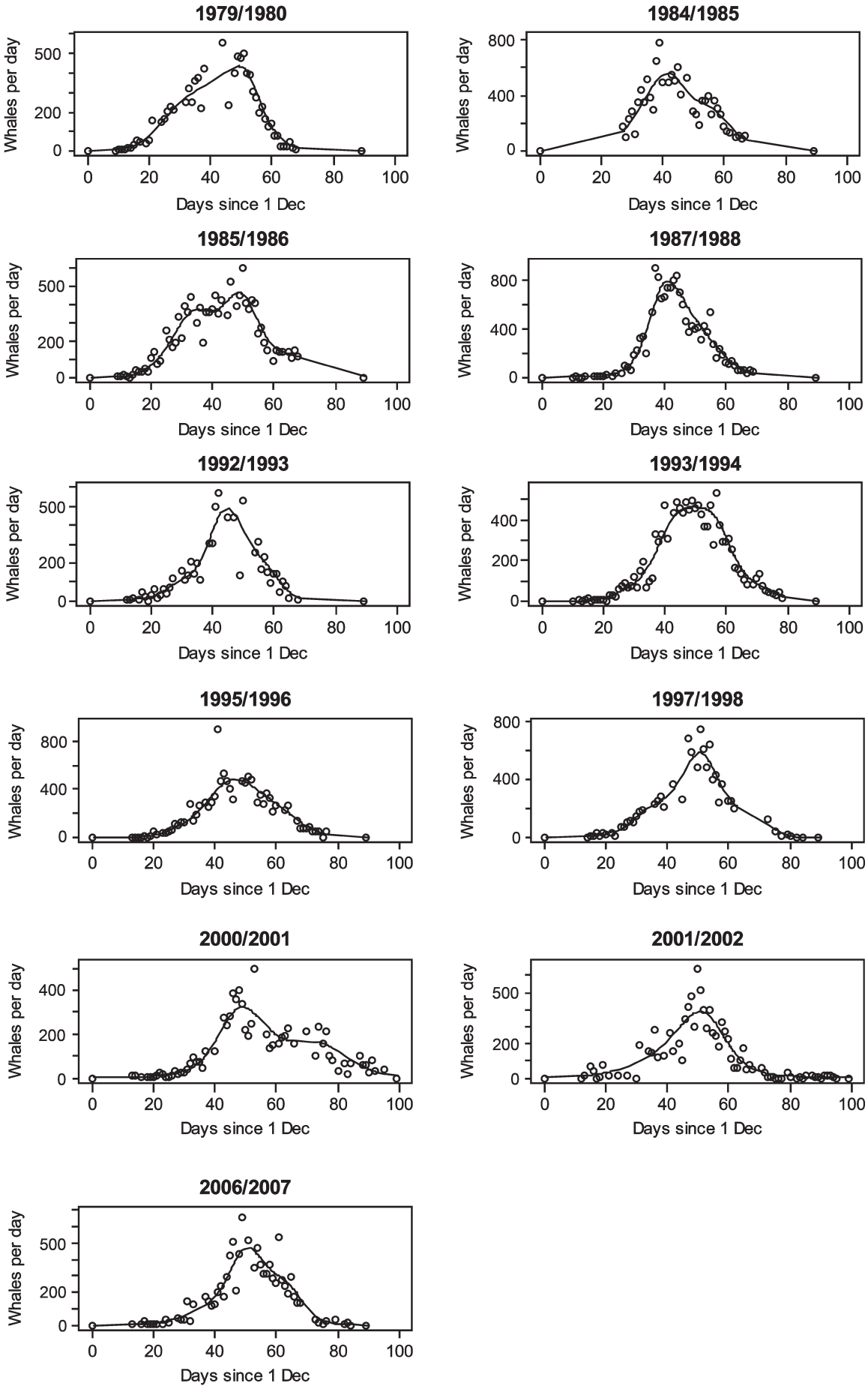


Fig. 1. Observed whale passage rates expressed as whales per day (circles) and fitted GAM model for the 23 southbound gray whale surveys during 1967–1968 to 2006–2007. The shift to later migration timing since 1992 is evident in this series of plots.

Table 3

Recorded number of pods and whales passing during acceptable effort periods of the southbound gray whale surveys from 1967 to 2006. Naïve abundance ($\widetilde{W}_y$) was estimated by smoothing observed whale passage rates (whales/hr) over time within each survey using a GAM (Fig. 1) and predicting total number of whales passing during the migration without applying correction factors.

Year	Number of pods	Number of whales	Average pod size	Effort (hours)	Naïve abundance
y	n^*_{1y}	$\sum_{i=1}^{n^*_{1y}} s_{iy}$	$\bar{s} = \sum_{i=1}^{n^*_{1y}} s_{iy} / n^*_{1y}$	$\sum_{j=1}^{m_y} l_{jy}$	$\widetilde{W}_y$
1967	903	2,202	2.44	303.0	8,558
1968	1,072	2,290	2.14	380.0	9,273
1969	1,236	2,626	2.12	465.0	9,276
1970	1,463	2,951	2.02	594.7	8,140
1971	859	1,885	2.19	345.0	7,062
1972	1,539	3,365	2.19	465.0	11,068
1973	1,497	3,139	2.10	425.0	11,074
1974	1,508	3,068	2.03	475.0	9,746
1975	1,188	2,462	2.07	293.5	11,195
1976	1,992	4,087	2.05	519.0	11,713
1977	657	1,211	1.84	195.0	12,453
1978	1,726	3,474	2.01	516.4	9,805
1979	1,457	2,998	2.06	376.3	12,596
1984	1,736	4,006	2.31	268.0	14,978
1985	1,840	4,119	2.24	456.5	14,609
1987	2,370	4,991	2.11	441.0	15,934
1992	1,002	1,772	1.77	297.5	10,438
1993	1,925	3,522	1.83	462.4	13,195
1995	1,439	2,669	1.85	304.0	13,741
1997	1,564	2,531	1.62	284.1	14,507
2000	1,089	1,869	1.72	399.0	10,571
2001	1,194	2,030	1.70	390.2	9,808
2006	1,254	2,568	2.05	310.0	11,484

Rugh *et al.*, 1993; Rugh *et al.*, 2008c; Shelden and Laake, 2002; Swartz *et al.*, 1987). Ideally, we would have all of the data needed to construct independent year-specific estimates that accounted for all of the potential biases affecting the counts. However, there is no way to obtain those data for the early surveys. Even when the data needs were apparent, budgets were not always sufficient to collect the data in each year. Thus, compromises have been necessary to construct a complete time series of abundance estimates.

One of those compromises was incorporation of a 'correction' for error and bias in observers' estimation of the size of pods. Corrections are based on calibration data from aircraft and intense effort by dedicated shore-based teams. However, these data were not collected for each survey. In hindsight, both the method proposed by Reilly (1981) and

Table 4

Model selection results for pod size calibration data. The rate model ~size + True:plus represents the structure with separate rates for S = 1, 2, 3 and a linear model (intercept + slope × S) for S>3 (k = 5 parameters). Each of the Gamma models also contained four shape parameters for sizes S = 1, 2, 3, >3. The most parsimonious model (smallest AIC_c – small sample version of AIC) is shown in bold.

	Poisson		Gamma	
Rate model	AIC_c	k	AIC_c	k
Fixed: ~size + True:plus	1,548.12	5	1,532.64	9
Fixed: ~year*(size + True:plus)	1,514.95	20	1,466.23	36
Fixed: ~size + True:plus, Random:pod	1,506.32	6	**1,454.21**	**10**
Fixed: ~size + True:plus, Random:observer	1,542.96	6	1,517.07	10
Fixed: ~size + True:plus, Random:year	1,536.89	6	1,517.94	10

Table 5

Parameter estimates for the gray whale pod size calibration data. The estimates are based on a discrete gamma distribution that includes a pod random effect on the rate parameter (b_S) and fixed effects for the rate (b_S) and shape (a_S) parameters based on true size of the pod.

	Estimate	Standard error
log (σ_ε)	–0.9361	0.0089
S = 1; log(b_1)	1.0040	0.2875
S = 2; log(b_2)	1.6177	0.0090
S = 3; log(b_3)	1.2783	0.2070
S > 3; log(β_0)	1.6714	0.1873
S > 3; log(β_1)	–0.1998	0.0085
S = 1; log(a_1)	0.4934	0.3361
S = 2; log(a_2)	1.7361	0.0089
S = 3; log(a_3)	1.8518	0.1920
S > 3; log(a_{4+})	1.1586	0.1644

Table 6

Number of pods seen by observers at primary and secondary station and by both observers upon completion of linking and matching for watch periods with double observers during acceptable environmental conditions (as determined by assessment of observer at primary station). Linking of pods in close proximity reduced number of pods by 1.1% to 4.6%. Linking and matching used the scoring algorithm with the defined weights as described in the Appendix.

Year	Seen by primary (n_1)	Seen by secondary (n_2)	Seen by both (n_3)	Primary detection rate (n_3/ n_2)
1987	2,258	2,296	1,710	0.745
1992	323	301	228	0.757
1993	719	697	532	0.763
1995	401	378	305	0.807
1997	748	788	588	0.746
2000	657	677	513	0.758
2001	603	691	483	0.699
2006	395	405	303	0.748

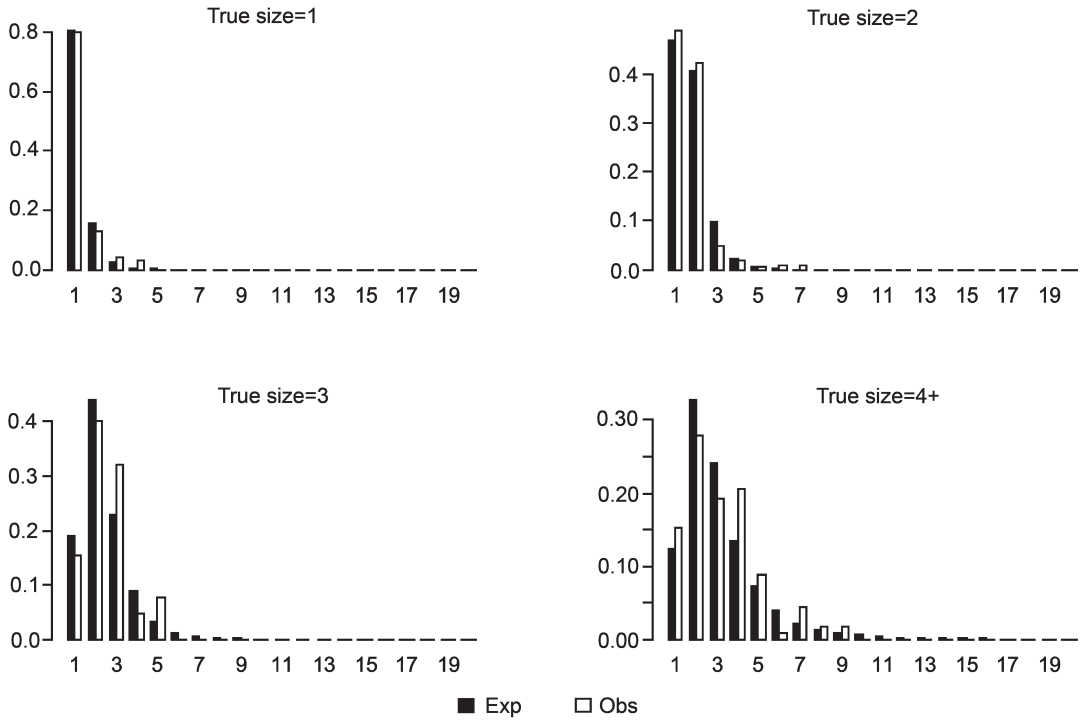


Fig. 2. For true pod sizes S = 1,2,3,4+, probability distributions for recorded (observed) pod sizes (s) and expected values from the gamma model with random pod effects for calibration data (Table 3).

the change in data selection for pod size bias (Buckland *et al.*, 1993; Hobbs *et al.*, 2004; Laake *et al.*, 1994; Rugh *et al.*, 2008b; Rugh *et al.*, 2005) were not optimal choices. At the very least all of the pod size calibration data should have been pooled to estimate a common correction factor for the entire time series. Here we have devised a more robust estimation approach for handling pod size bias, and we used all of the calibration data, with the exception of the thermal imaging data of DeAngelis *et al.* (1997).

Re-evaluation of the correction for pod size bias and the other changes made to the estimation procedure yielded a substantially different trajectory for population growth.

Table 7

Parameter estimates (standard errors in parentheses) for the gamma distribution of pod size (S), the expected pod size ($E(S)$) and detection probability parameters from the selected model for each year for the eight most recent southbound gray whale surveys. Parameters for the gamma distribution are on the log-scale (e.g., for 1987 shape = exp(0.422) and rate = 1/scale = exp(–0.326). Parameters for detection probability are on logit scale. Intercept represents observer #1 for pod of size 0 at distance < 1km and for either Vis <4 or Beaufort = 0 depending on model. For example, detection probability for observer #3 with pod size = 2 at distances between 2–3km in 1987 with visibility <4 was: 1/(1+exp(0.310+0.087+0.172–0.553×2). Observers are arbitrarily numbered and different for each year. Average pod size $\overline{s}$ here is for linked primary pods (Table 8).

	1987	1992	1993	1995	1997	2000	2001	2006
Gamma shape	0.422 (0.060)	–0.073 (0.161)	–0.070 (0.100)	–0.063 (0.111)	–0.598 (0.131)	0.089 (0.127)	–0.095 (0.131)	–0.106 (0.106)
Gamma rate	–0.326 (0.062)	–0.347 (0.147)	–0.474 (0.094)	–0.545 (0.106)	–0.674 (0.118)	–0.280 (0.122)	–0.366 (0.125)	–0.685 (0.102)
$E(S)$	2.626 (0.044)	1.886 (0.067)	2.060 (0.051)	2.176 (0.066)	1.724 (0.047)	1.995 (0.058)	1.885 (0.056)	2.340 (0.075)
$E(S)/\overline{s}$	1.188	1.054	1.079	1.127	1.039	1.115	1.065	1.104
(Intercept)	–0.310 (0.183)	–0.044 (0.730)	0.579 (0.427)	1.840 (0.583)	0.267 (0.336)	–0.458 (0.429)	1.050 (0.534)	0.867 (0.495)
Podsize	0.553 (0.063)	0.747 (0.260)	0.938 (0.189)	0.438 (0.141)	0.553 (0.151)	0.908 (0.192)	0.485 (0.141)	0.343 (0.104)
Distance 1–2km	0.289 (0.138)	0.528 (0.440)	0.012 (0.273)	–0.660 (0.483)	0.476 (0.281)	0.656 (0.352)	0.277 (0.401)	0.274 (0.350)
Distance 2–3km	–0.172 (0.147)	–0.183 (0.438)	–0.391 (0.278)	–1.310 (0.498)	–0.035 (0.278)	0.328 (0.357)	–0.261 (0.404)	–0.327 (0.355)
Distance 3–4km	–0.702 (0.203)	–0.683 (0.488)	–0.713 (0.367)	–1.740 (0.570)	–0.223 (0.315)	–0.361 (0.438)	–0.944 (0.448)	–0.788 (0.479)
Distance >4km	–1.840 (0.288)	–1.790 (0.704)	–1.410 (0.506)	–2.580 (0.754)	–0.825 (0.385)	–0.793 (0.676)	–1.340 (0.548)	–1.380 (0.621)
Observer 2	0.483 (0.137)	–0.219 (0.651)	–0.827 (0.302)	–0.552 (0.395)	0.978 (0.397)	–0.845 (0.424)	–0.580 (0.407)	0.121 (0.300)
Observer 3	–0.087 (0.128)	0.317 (0.615)	–0.478 (0.334)	–0.307 (0.373)	0.340 (0.295)	0.048 (0.295)	–0.776 (0.443)	0.278 (0.318)
Observer 4	0.136 (0.115)	–0.192 (0.607)	–1.340 (0.331)	–0.360 (0.344)	0.246 (0.284)	–0.865 (0.237)	–0.635 (0.390)	0.142 (0.314)
Observer 5	0.156 (0.116)	0.060 (0.613)	–0.840 (0.302)	–0.747 (0.376)	0.528 (0.301)	0.090 (0.286)	–1.100 (0.376)	–0.546 (0.419)
Observer 6	0.416 (0.136)	0.182 (0.634)	–1.550 (0.339)	–1.000 (0.560)	–0.262 (0.172)	–0.052 (0.295)	0.051 (0.414)	0.220 (0.299)
Observer 7	0.120 (0.172)	–0.574 (0.603)	–0.451 (0.354)	–0.748 (0.364)	–0.236 (0.276)	–0.553 (0.207)	–0.542 (0.424)	–1.110 (0.299)
Observer 8	0.282 (0.166)		0.076 (0.605)	0.640 (0.465)	0.129 (0.229)	–0.706 (0.235)	–1.200 (0.406)	0.473 (0.424)
Observer 9	0.237 (0.171)				–0.481 (0.227)	–0.017 (0.385)	0.030 (0.437)	1.170 (0.641)
Observer 10					0.247 (0.339)	–0.079 (0.255)	–1.410 (0.420)	
Observer 11							–0.690 (0.466)	
Observer 12							–0.591 (0.433)	
Observer 13							–0.659 (0.418)	
Observer 14							–0.956 (0.426)	
Vis >3	–0.345 (0.106)		–0.316 (0.165)					
Beaufort								–0.128 (0.125)

Table 8

For recent eight gray whale surveys from 1987 to 2006, number of pods and linked pods seen by the primary observer, average linked pod size, naïve abundance, estimated abundance (without night-time correction) and ratio estimate for correction factor for estimates from surveys prior to 1987.

Year	Number of pods	Number of linked pods	Average linked pod size	Naïve abundance	Abundance	Ratio
Y	n^*_{1y}	n_{1y}	$\bar{s} = \sum_{i=1}^{n_{1y}} s_{iy} / n_{1y}$	$\widetilde{W}_y$	$\widehat{W}_y$	$\widehat{W}_y / \widetilde{W}_y$
1987	2,370	2,262	2.21	15,934	24,883	1.562
1992	1,002	991	1.79	10,438	14,571	1.396
1993	1,925	1,848	1.91	13,195	18,585	1.408
1995	1,439	1,388	1.93	13,741	19,362	1.409
1997	1,564	1,522	1.66	14,507	19,539	1.347
2000	1,089	1,043	1.79	10,571	15,133	1.432
2001	1,194	1,150	1.77	9,808	14,822	1.511
2006	1,254	1,213	2.12	11,484	17,682	1.540
Ratio						1.450
SE						0.030

Previously, the peak abundance estimate was in 1998 followed by a large drop in numbers (Rugh *et al.*, 2008c). Now the peak estimate is a decade earlier (Table 9; Fig. 4), and the predicted population trajectory has remained flat and relatively constant since 1980 (Fig. 4).

The correction for night time differential migration rate should be revisited and more data should be collected to evaluate within-year and annual variation in day and night migration rates described by Perryman *et al*. (1999). The assessment of population growth will be improved by collection of data in each survey that provides survey-specific correction factors. Incorporation of thermal imaging and land tracking in each survey would provide survey-specific estimates for pod size calibration and night time differential. In addition, independent double-observer data should continue to be collected as part of the survey protocol to provide survey-specific measures of detection probability for pods.

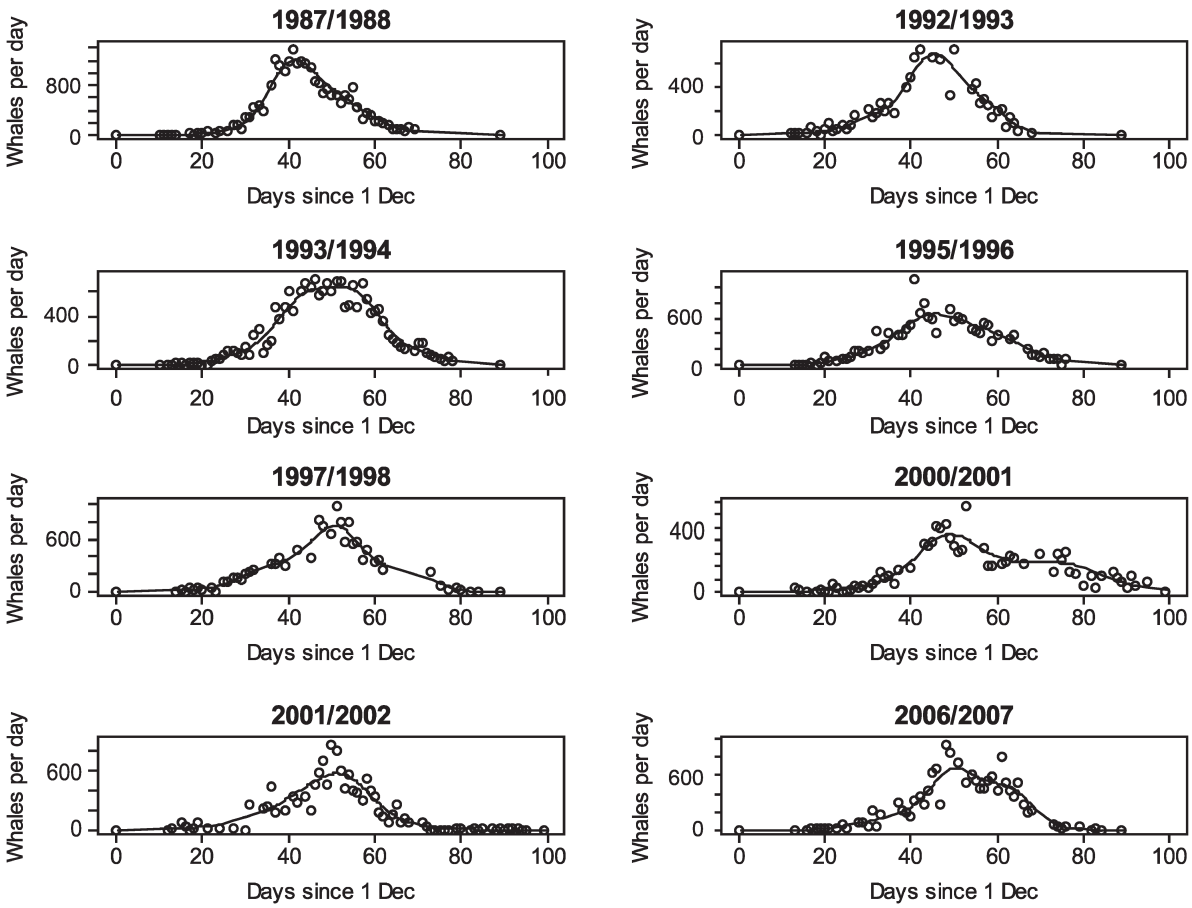


Fig. 3. Estimated number of whales passing per day during watch periods (circles) from year specific models for detection probability and pod size, and fitted GAM model (line) for the eight southbound gray whale surveys during 1987 to 2006.

Table 9

Current and previous gray whale abundance estimates and coefficient of variation (CV = standard error/estimate) constructed from southbound migration surveys conducted from 1967–68 to 2006–07. Ratio of current to previous estimates shows proportional change which is largely explained by f_s ratio which is $E(S)/\overline{s}$ from Table 7 divided by f_s, the pod size correction from previous surveys.

	Current		Previous				
Year	$\widehat{N}_y$	$cv(\widehat{N}_y)$	$\widehat{N}_y$	$cv(\widehat{N}_y)$	Ratio	f_s	f_s ratio
1967–68	13,426	0.094	13,776	0.078	0.975	–	–
1968–69	14,548	0.080	12,869	0.055	1.130	–	–
1969–70	14,553	0.083	13,431	0.056	1.084	–	–
1970–71	12,771	0.081	11,416	0.052	1.119	–	–
1971–72	11,079	0.093	10,406	0.059	1.065	–	–
1972–73	17,365	0.080	16,098	0.052	1.079	–	–
1973–74	17,375	0.082	15,960	0.055	1.089	–	–
1974–75	15,290	0.084	13,812	0.057	1.107	–	–
1975–76	17,564	0.086	15,481	0.060	1.135	–	–
1976–77	18,377	0.080	16,317	0.050	1.126	–	–
1977–78	19,538	0.088	17,996	0.069	1.086	–	–
1978–79	15,384	0.080	13,971	0.054	1.101	–	–
1979–80	19,763	0.083	17,447	0.056	1.133	–	–
1984–85	23,499	0.089	22,862	0.060	1.028	–	–
1985–86	22,921	0.082	21,444	0.052	1.069	–	–
1987–88	26,916	0.058	22,250	0.050	1.210	1.131[1]	1.050
1992–93	15,762	0.068	18,844	0.063	0.836	1.430[2]	0.737
1993–94	20,103	0.055	24,638	0.060	0.816	1.420[2]	0.760
1995–96	20,944	0.061	24,065	0.058	0.870	1.399[3]	0.806
1997–98	21,135	0.068	29,758	0.105	0.710	1.516[4]	0.685
2000–01	16,369	0.061	19,448	0.097	0.842	1.486[4]	0.750
2001–02	16,033	0.069	18,178	0.098	0.882	1.485[4]	0.717
2006–07	19,126	0.071	20,110	0.088	0.951	1.361[5]	0.811

[1]Buckland *et al*., 1993, [2]Laake *et al.*, 1994, [3]Hobbs *et al.*, 2004, [4]Rugh *et al.*, 2005, [5]Rugh *et al.*, 2008a.

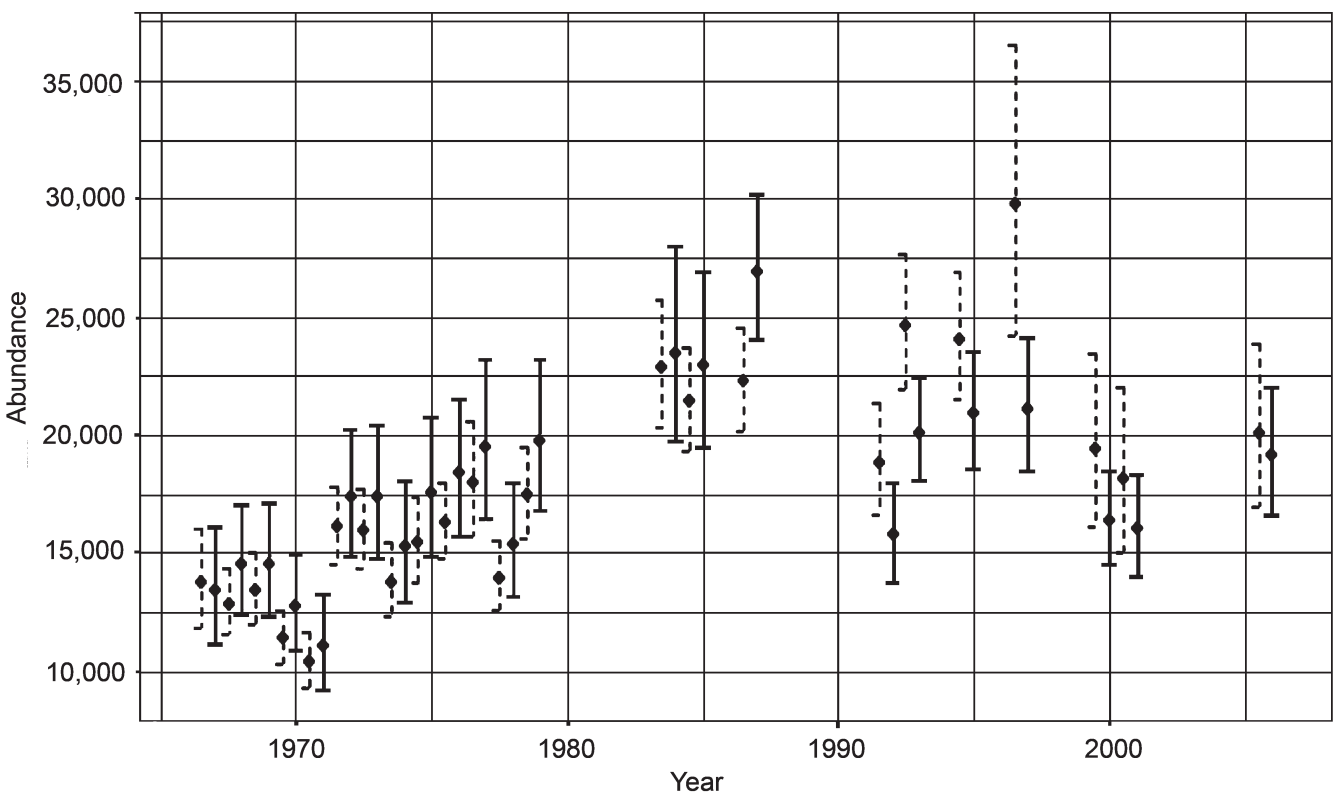


Fig. 4. Abundance estimates with 95% log-normal confidence intervals for previous estimates (dashed line) taken from Rugh *et al.* (2008a) and current estimates (solid line).

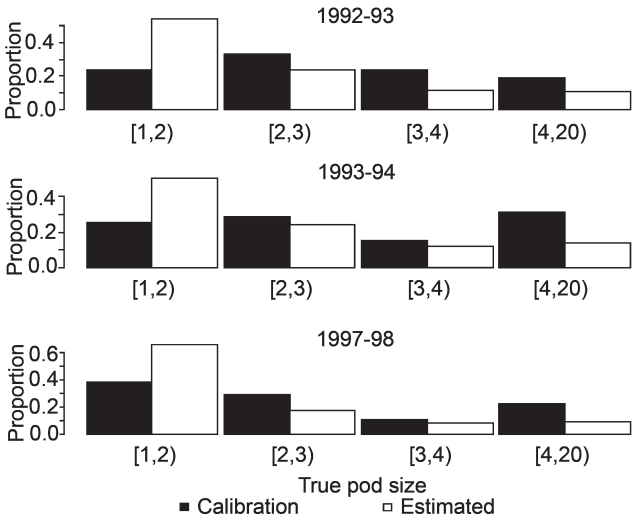


Fig. 5. Pod size distributions for calibration data (light) and estimated true pod size distribution using estimated parameters for gamma distribution (see Table 7). Calibration data from 1978–1979 are not shown because it was not possible to derive estimates of the true pod size distribution with the survey data in that year.

ACKNOWLEDGMENTS

We thank the numerous observers who spent countless hours searching and recording data on southbound gray whales for the last 40 years and Marcia Muto who helped pour through data records locating and correcting errors in the data. We thank John Durban, Steve Buckland and Steve Reilly for reviews of drafts of this paper. A previous version of this manuscript was NOAA Technical Memorandum NMFS-AFSC-203 (available at: *http://www.afsc.noaa.gov/Publications/AFSC-TM-AFSC-203.pdf*).

REFERENCES

Angliss, R.P. and Allen, B.M. 2009. Alaska marine mammal stock assessments 2008. *NOAA Tech. Mem. NMFS-AFSC*- 193: 258pp. NTIS No. PB2009-109548.

Angliss, R.P. and Outlaw, R.B. 2008. Alaska marine mammal stock assessments 2007. *NOAA Tech. Mem. NMFS-AFSC*- 180: 252pp. NTIS No. PB2008-112874.

Buckland, S.T. and Breiwick, J.M. 2002. Estimated trends in abundance of eastern Pacific gray whales from shore counts (1967/68 to 1995/96). *J. Cetacean Res. Manage.* 4(1): 41–48.

Buckland, S.T., Breiwick, J.M., Cattanach, K.L. and Laake, J.L. 1993. Estimated population size of the California gray whale. *Mar. Mammal Sci.* 9(3): 235–49.

Cochran, W.G. 1977. *Sampling Techniques*. 3rd ed. John Wiley and Sons, New York. i–xvi+428pp.

DeAngelis, M.L., Martin, T. and Parryman, W.L. 1997. Pod size estimates studies through thermal sensors. Contract rep. to Natl. Mar. Mammal Lab., NMFS, NOAA, 7600 Sand Pt. Way, NE, Seattle, WA 98115. 7pp.

Hobbs, R.C., Rugh, D.J., Waite, J.M., Breiwick, J.M. and DeMaster, D.P. 2004. Abundance of gray whales in the 1995/96 southbound migration in the eastern North Pacific. *J. Cetacean Res. Manage.* 6(2): 115–20.

Laake, J.L., Rugh, D.J., Lerczak, J.A. and Buckland, S.T. 1994. Preliminary estimates of population size of gray whales from the 1992/93 and 1993/94 shore-based surveys. Paper SC/46/AS7 presented to the IWC Scientific Committee, May 1994, Puerto Vallarta, Mexico (unpublished). 13pp. [Paper available from the Office of this Journal].

Lerczak, J.A. and Hobbs, R.C. 1998. Calculating sighting distances from angular readings during shipboard, aerial, and shore-based marine mammal surveys. *Mar. Mammal Sci.* 14(3): 590–99. [See Errata. 1998. *Mar. Mammal Sci.* 14(4):903].

Otis, D.L., Burnham, K.P., White, G.C. and Anderson, D.R. 1978. Statistical inference from capture data on closed animal populations. *Wildl. Monogr.* 62: 1–135.

Perryman, W.L., Donahue, M.A., Laake, J.L. and Martin, T.E. 1999. Diel variation in migration rates of eastern Pacific gray whales measured with thermal imaging sensors. *Mar. Mammal Sci.* 15(2): 426–45.

R Development Core Team. 2009. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0 [Available at: http://www.R-project.org].

Reilly, S.B. 1981. Population assessment and population dynamics of the California gray whale (*Eschrichtius robustus*). PhD. Thesis, University of Washington. 265pp.

Rugh, D., Breiwick, J., Hobbs, R., Shelden, K. and Muto, M. 2008a. Eastern North Pacific gray whale abundance in the winter of 2006–2007. Paper SC/60/BRG6 presented to the IWC Scientific Committee, June 2008, Santiago, Chile. 12pp.

Rugh, D., Breiwick, J., Muto, M.M., Hobbs, R., Shelden, K., D'Vincent, C., Laursen, I.M., Reif, S., Maher, S. and Nilson, S. 2008b. Report of the 2006–2007 census of the eastern North Pacific stock of gray whales. *AFSC Processed Report* 2008–03: 157pp. Alaska Fisheries Science Center, NOAA, Natl. Mar. Fish. Serv., 7600 Sand Point Way NE, Seattle WA 98115.

Rugh, D.J., Breiwick, J.M., Dahlheim, M.E. and Boucher, G.C. 1993. A comparison of independent, concurrent sighting records from a shore-based count of gray whales. *Wildl. Soc. Bull.* 21(4): 427–37.

Rugh, D.J., Ferrero, R.C. and Dahlheim, M.E. 1990. Inter-observer count discrepancies in a shore-based census of gray whales ({IEschrichtius robustus}). *Mar. Mammal Sci.* 6(2): 109–20.

Rugh, D.J., Hobbs, R.C., Lerczak, J.A. and Breiwick, J.M. 2005. Estimates of abundance of the Eastern North Pacific stock of gray whales 1997 to 2002. *J. Cetacean Res. Manage.* 7(1): 1–12.

Rugh, D.J., Muto, M.M., Hobbs, R.C. and Lerczak, J.A. 2008c. An assessment of shore-based counts of gray whales. *Mar. Mammal Sci.* 24(4): 864–80.

Shelden, K.E.W. and Laake, J.L. 2002. Comparison of the offshore distribution of southbound migrating gray whales from aerial survey data collected off Granite Canyon, California, 1979–96. *J. Cetacean Res. Manage.* 4(1): 53–56.

Swartz, S.L., Jones, M.L., Goodyear, J., Withrow, D.E. and Miller, R.V. 1987. Radio-telemetric studies of gray whale migration along the California coast: a preliminary comparison of day and night migration rates. *Rep. int. Whal. Commn* 37: 295–99.

Wood, S.N. 2006. *Generalized Additive Models: an introduction with R*. Chapman and Hall, Boca Raton, Florida. 391pp.

Date received: September 2011
Date accepted: September 2011

APPENDIX

Additive pod size correction factor

We will use the following notation to describe the methodology of Reilly (1981):

S = true pod size

s = recorded pod size

$f(S)$ = probability distribution of true pod sizes

$h(s)$ = probability distribution of recorded pod sizes

$g(s|S)$ = probability that an observer will record a group of true size S as size s.

$f^*(S)$ = probability distribution of true sizes in the calibration data

From the calibration data, the probability that a group is of true size of S given that it was recorded as size s is:

$$f^*(S|s) = \frac{f^*(S)g(s|S)}{\sum_S f^*(S)g(s|S)} .$$

With the method of Reilly (1981), the calibration data are used to construct a set of adjustments, $c(s)$, which are added to the recorded pod size s

$$c(s) = \sum_S (S-s) f^*(S|s) = \left[\sum_S S f^*(S|s)\right] - s ,$$

to get the estimate of the average group size

$$\hat{\bar{S}} = \sum_s \left[s + c(s)\right] h(s) ,$$

which can also be written as:

$$\hat{\bar{S}} = \sum_s \left[s + \sum_S (S-s) f^*(S|s) \right] h(s) =$$
$$\sum_s h(s) \sum_S S f^*(S|s) = \sum_s h(s) E_{f^*}[S|s] .$$

Differences in adjustment values, $c(s)$, for different calibration data sets as reported in Rugh *et al.* (2008c) can result from differences in either $f^*(S)$ or $g(s|S)$. If the differences reported by Rugh *et al.* (2008c) are due to differences in $g(s|S)$ that may reflect inherent variability in observer ability or variability due to inherent differences in the calibration pods (e.g. frequency and timing of surfacing, proximity of whales in pod, distance from observer). However, if the differences are due to the selection of pods $f^*(S)$ during the different calibration experiments and $f(S)$ varies annually, substantial bias could result with the correction method of Reilly (1981).

The method of Reilly (1981) will be unbiased as long as $f^*(S) = f(S)$ (i.e. calibration distribution was selected to match the true distribution). That assumption could hold if passing pods could be selected randomly for calibration. However, use of the calibration data beyond the year in which they were collected would not be warranted unless $f(S)$ was the same in each year. While that may be possible, it is a strong assumption that is not necessary with the analysis method we describe here.

Instead of trying to ensure equality ($f^*(S) = f(S)$), the calibration data should be viewed like a regression problem in that pods should be selected to provide a best estimate of $g(s|S)$. In general, one would want the selection of pods to balance both $f(S)$ and the variance of $g(s|S)$ to minimise the uncertainty. For example, if $g(1|1)$ was nearly 1.0, then one would not need many calibration pods of size 1 and instead may select more pods of size 2 or more even if most pods were of size 1 (e.g. mode of $f(S)$ was at $S = 1$).

Matching and linking criterion

Two observers searched for gray whales at the same time and recorded their data independently to provide a measure of how many pods were missed during the watch. From the separate independent data records, we needed to decide which pods were seen by both observers and which were missed by one or the other. We have used the term 'matching' for this process of comparing observer records. The observers had a working definition for a gray whale pod as a group of whales that were within a body length of each other. However, errors were quite possible with whales in a pod surfacing at different times, and what one observer treated as a single pod could have been recorded as more than one pod by the other observer. Thus, the matching process also had to consider this possibility, so prior to matching we used a 'linking' process whereby the proximity of all sightings from a given observer were compared to each other, and any pods that were sufficiently close were merged. The records of these 'linked' (merged) pods were then 'matched' by comparing their proximity and pod size. For instance, if one observer recorded a pod of two whales and a second observer saw the same whales but recorded them as two pods of single whales each, then the linking process would merge the two whales, providing a good match between the two observers' records. An underlying assumption in this system is that there are no false positives, that is, no one records a sighting unless there truly is a whale there, and the sighting data (time and location) are accurate enough to make a match.

We used a linking/matching criterion that was a modified version of the criterion described by Rugh *et al.* (1993). The criterion constructs a score based on a comparison of crossing times (t241), distance offshore (d241), and pod sizes (s) (Fig. A1). The time and distance computations assume that whales travelled parallel to the coast at a constant speed of 6km/hour. The t241 is the time the pod would cross an imaginary line perpendicular to the location of the observer on shore (241° magnetic). It is computed from the last (most southerly) time and location of the pod by projecting, either forward or backward, the time needed to travel the distance from the last location to the 241° line. The d241 is the perpendicular distance from shore to the projected point on the 241° line where the whale pod crossed; this is estimated via a simple trigonometric calculation from the distance and angle to the most southerly location. The score function can be represented as:

$$score_{ij} = f\left[W_t \left| t241_i - t241_j \right|, \frac{W_d \left| d241_i - d241_j \right|}{\max(d241_i, d241_j)} \right]$$
$$+ W_s \left| s_i - s_j \right| ,$$

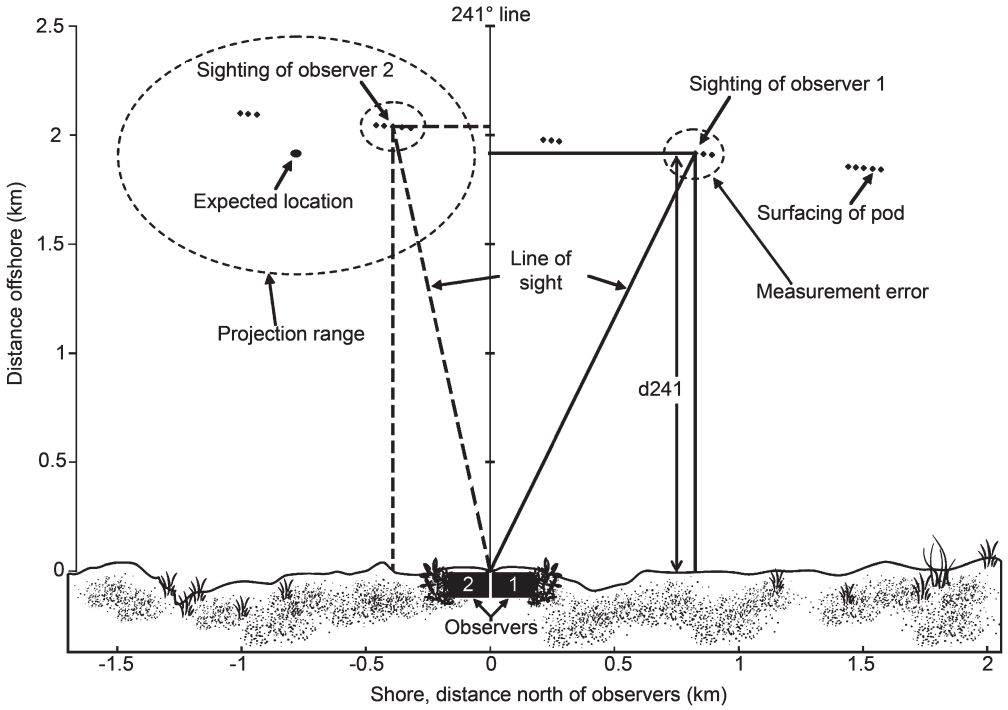


Fig. A1. Observers search from adjacent sheds (#2 and #1). As a pod passes offshore, each observer independently records time, magnetic angle, and vertical reticle. From these data, the sighting distance is calculated. The distance from shore and travel distance are calculated using trigonometry. The expected location at the time of the second sighting is estimated from the time difference and the assumption of parallel travel at 6Km/hr and the difference in t241 times is the parallel distance between these points divided by 6 km/hour. The projection range ellipse is a 95% probability area calculated from the fitted distributions for speed and deviation from parallel travel using the time difference.

where

(1) i and j are the indexes of the i^{th} and j^{th} pods of a single observer record for linking or the i^{th} and j^{th} pods recorded by independent observers for matching,

(2) the function f was a sum in Rugh *et al*. (1993) but here we have used a square root of the sum of the squared arguments, and

(3) W_t, W_d and W_s are defined weights for the time difference, distance difference, and pod size (s) difference.

All pods were scored against all other pods within an effort period. If the score was less than a maximum allowable score value, then the sightings met the criterion for linking/matching.

For linking, the pod size weight was set to zero. Pods were linked iteratively to allow for the potential that a pod was split into more than two separate pods. The pair of pods with the lowest score was merged into a single pod with the average t241 and d241 and the pod sizes summed to create a single pod replacing each subset. This was then repeated until no pair of pods met the criterion. For matching, the candidate matches were ranked by score with the lowest being the best match. The best match was recorded and the two pods in the match were removed from further matching. This process continued until there were no more candidate matches that met the criterion. The weights were scaled so that the matching maximum score was set to 1.0. The linking criterion was set to a lower value to limit the risk that a legitimate match could be lost due to the averaging of distance and time in merging pods.

The weights account for two types of errors involved in estimation of t241 and d241, measurement errors and projection errors. Measurement errors result from errors in measuring the horizontal angle, the angle below the horizon (via reticles), and the event time. These errors were estimated from comparisons between tracking teams and standard watch observers (Rugh *et al.*, 2008c). The frequencies reported in table 2 of Rugh *et al.* (2008c) were fitted by integrating the normal distribution between +0.5 and –0.5 of the horizontal degree difference and minimising the squared difference between the reported and the predicted frequency. The standard deviation for the error was estimated at 2.23°, which is consistent with the statement in Rugh *et al.* (2008c) that 95% of measurements differed by 3° or less. Reported frequencies of discrepancies in reticle measurements (Table 3 of Rugh *et al.*, 2008c) were fitted by integrating the normal distribution between +0.05 and –0.05 of the reticle difference and minimising the squared difference between the reported frequency and the predicted frequency. The standard deviation for the error was estimated at 0.14 reticles, which is consistent with the statement in Rugh *et al.* (2008c) that 95% of measurements differed by 0.4 reticles or less. Rugh *et al.* (2008c) found time precision to be limited to 45 seconds for the same surfacing of a pod which may include sequential surfacings of the pod members. Rugh *et al*. (2008c) reported time differences of less than 10 seconds for matches between tracked whales and standard watch data where the locations matched exactly (same angle and reticle), suggesting that it was the same whale surfacing. Transforming these measurement errors, the standard deviation for the error in t241 was 0.55 minutes at 1km offshore and 1.35 minutes at 3km of shore, and the standard deviations for the error in d241 were 0.032km and 0.319km respectively. When the d241 was compared between pods, this resulted in a 3.2% difference at 1km and 10.6% difference at 3km.

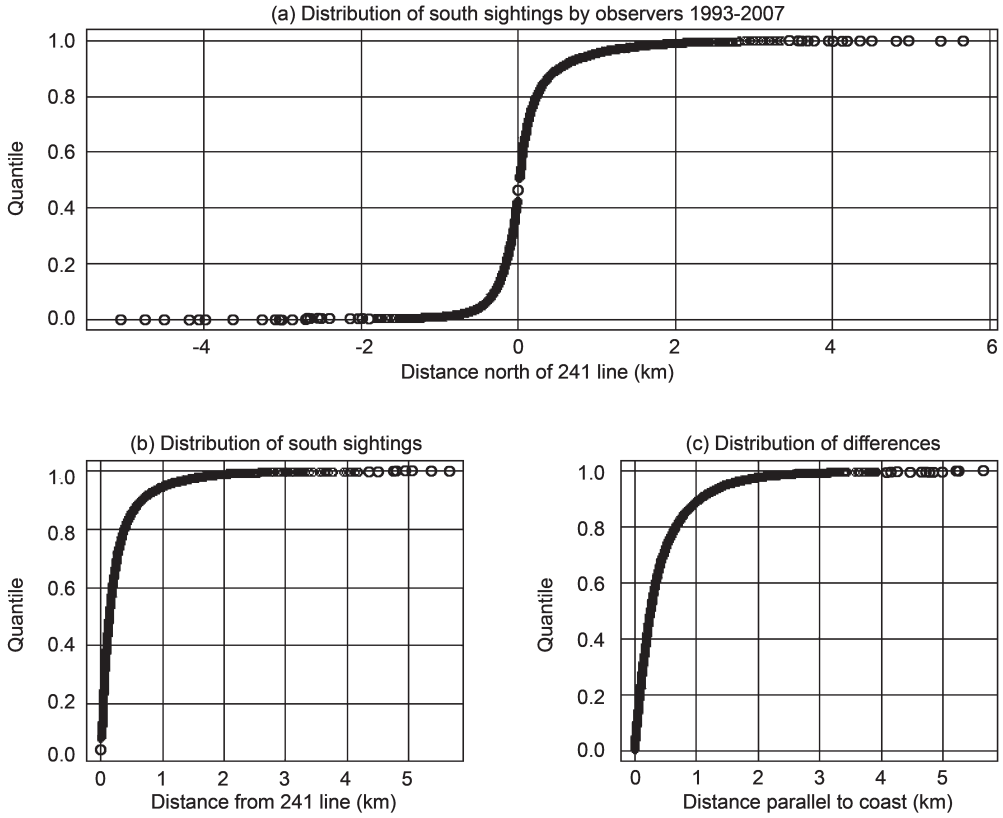


Fig. A2. (a) Distance north from the 241° line to the location of south sightings for all observers 1993–2007. (b) Absolute distance from 241° line. Note that 95% of south sightings fall between within 1 km and 99% within 2 km. (c) Distribution of differences between random pairs of sightings when sightings were drawn at random from the distribution of south sightings. Note that 90% of expected comparison distances between sightings were 1 km or less.

Projection errors resulted from differences between the actual speed and direction of a pod and the assumptions of 6km/hour and parallel travel (Fig. A1). The most southerly sightings were clustered around the 241° line with the median = 0.00km, mean = 0.079km (north) and standard deviation = 0.488 (Fig. A2a). Projection distance regardless of direction was zero (on the 241° line) for 8% of south sightings and 95% within 1km and 99% within 2km (Fig. A2b).

Travel speed was estimated directly from the sighting data using the travel time between north and south sightings. The sighting data incorporates the measurement error into the projection error. A subset of sightings was selected that have both north and south data, with a south sighting between –1.0km and +0.5km and a travel distance from north to south of 1.0 to 2.5km with a minimum time difference of 6 minutes and no other pods with t241 within 5 minutes. The south distance was chosen to insure that the travel occurred near the 241° line, the travel distance and minimum time were chosen to limit the effect of measurement errors. Only pods with no other recorded pods near were chosen to limit the effect of improperly linked sightings. Significant relations between speed and survey date and speed and pod size were found, but neither contributed significantly to reducing the variance. The average speed was 6.19km/hour (sd = 1.55, var = 2.41). The distribution of bearings relative to the 241° line was estimated from a similar data set except that all sightings with a minimum time difference of 3 minutes and travel distance between 0.02 and 2.5km were used. These were binned into 0.2km travel distance bins centered on the even tenths of a km and the mean deviation and variance about the track perpendicular to the 241° line were calculated. A linear fit of the mean deviation with the distance travelled yielded a significant but small trend shoreward of less than 30 meters/km travelled (Table A1). Two models for the change in variance were considered: (1) a 'random walk' in which the whales continually made small changes in heading as they proceeded south so that variance would increase linearly with distance, and (2) a fixed heading in which the square root of the variance would increase linearly with distance travelled. Of the two, the fixed heading model provided a better fit (Table A1).

The probability that a sighting by one observer was correctly matched to a sighting of the same pod by a second observer was estimated from the distribution of bearing and speed and applying the matching to the distribution of possible distances between sightings of the same group. Assuming that the distance between the sighting locations was the result of chance and observer behaviour rather than whale behaviour (e.g. sightings of faster pods are more likely to be farther apart), then the cumulative distribution of possible distances between sightings was determined by random draws of pairs from the distribution of south sightings (Fig. A2c). The projection errors were much greater than the measurement errors; consequently, it was not necessary to include the measurement errors explicitly in the choice of the weights.

While there are three measurements involved with each sighting, the determination of a match is reduced to a two dimensional comparison by relating the difference in time and distance parallel to the coast (and perpendicular to the 241° line) assuming a fixed speed of 6km/h and accepting a range of difference in the t241 times to allow for variation

Table A1

Parameter estimates for deviation from travel parallel to the coastline (perpendicular to the 241° line) in kilometres difference in d241 per kilometre of travel parallel to the coast.

Model	Mean(deviation km) = a + b(travel dist km)		Variance(deviation km) = a + b(travel dist km)		SD(deviation km) = a + b(travel dist km)	
Parameter	a	b	a	B	a	b
Estimate	0.037	-0.029	0.006	0.050	0.139	0.092
SE	0.011	0.007	0.014	0.009	0.020	0.014
t	3.41	-3.89	0.47	5.33	6.83	6.68
Pr(>|t|)	0.00665	0.00299	0.65201	0.00034	0.00005	0.00005
R-squared	0.56		0.71		0.80	
F-statistic:	15.2	P = 0.0030	28.4	P = 0.00034	44.6	P = 0.00006

Table A2

Comparison table for weights used in matching criterion. Weights were scaled so that the probability of matching in each dimension was equal.

				Standard model		Alternate model	
Probability of matched by t241	Probability of matched by d241	Probability of matched	W_t	W_d	Probability of one other pod	W_d	Probability of one other pod
99%	99%	98%	0.11	3.02	79	1.9	60%
98%	98%	96%	0.16	3.66	66	2.25	44%
97%	97%	95%	0.18	3.95	61	2.38	40%
95%	95%	90%	0.27	5.06	45	2.86	27%
89%	89%	80%	0.46	6.66	27	3.56	15%

in speed. The range of time differences and consequently speeds that meet the criteria can be related to the distribution of distances between sightings (ignoring pod size and assuming travel parallel to the coast) by rewriting the difference in the t241 times in terms of the difference in time and difference in distance to the 241° line. Likewise the extremes of the deviations from parallel travel can be estimated assuming that speed was 6 km/hour.

$$S_{slow} = \frac{\Delta x}{\frac{\Delta x}{s} + \frac{K}{W_t}}$$

$$S_{fast} = \begin{cases} \dfrac{\Delta x}{\frac{\Delta x}{s} - \frac{K}{W_t}} & if\ \Delta x > \dfrac{Ks}{W_t} \\ \infty & otherwise \end{cases}$$

$$\text{Standard: } \Delta y_{near} = \frac{K}{W_d} y_1; y_1 \geq y_2,$$

$$\Delta y_{off} = \frac{\frac{K}{W_d} y_1}{1 - \frac{K}{W_d}}; y_1 < y_2, \text{ Alternative: } \Delta y = \pm \frac{K}{W_d},$$

where, S_{slow} and S_{fast} are the extremes of the distribution speed perpendicular to the 241° line; Δx is the difference in the distance perpendicular to the 241° line between the two sightings, note that S_{fast} is undefined until Δx is sufficient to make the denominator positive; K is the maximum allowable score for a match or link; and S is the speed used for the projection, in this case 6km/hour. Δy is the maximum allowable difference in the deviation distance parallel to the 241° line between the two sightings, with y_1 being the distance offshore of the northern of the two sightings and y_2 the southern. The standard version was described in Rugh *et al.* (1993) and was intended to account for the greater measurement error with distance offshore resulting from reticle measurements by allowing a larger deviation in the offshore direction and wider range with distance offshore. The alternative ignores the measurement error and uses a constant width.

The probability that two sightings of the same pod, at a given distance apart, are matched is estimated as the product of the probabilities that the speed and deviation fall into each of these ranges. Integrating over the distribution of distances gives the approximate probability that a match will be made. Note that this analysis ignores the discrete nature of the measurement errors and as a consequence will favour the alternative to some extent. However, it is satisfactory to optimise the parameters for the standard method and to estimate the potential for improvement of matching efficiency by using the alternative.

The probability of overmatching or mismatching is approximated by the likelihood that at least one other sighting falls within that range. The linking algorithm is modified to count the number of groups that could be matched. To fully estimate the probability of mismatching we would need to include a model of the probability of a second sighting of the pod being matched having a higher score as well, and the probability of overmatching would include the probability that the pod was missed by the second observer.

While there clearly is a trade off between the certainty of correctly matching the same pod and the risk of overmatching, the risk of under matching has the potential to result in an overestimate of abundance and a conservative analysis would limit this risk. We used the weights at the 95% probability of a match (0.18 and 3.95) as the best compromise while acknowledging that the rate of missed pods may be underestimated by 50%. This analysis suggests that the alternate model would reduce the risk of overmatching by about one-third; however simulations with a discrete measurement error structure are required to determine the actual matching rate.