Enhancements to population monitoring of Yellowstone grizzly bears

Frank T. van Manen[1,12], Michael R. Ebinger[1,11], Cecily M. Costello[2], Daniel D. Bjornlie[3], Justin G. Clapp[3], Daniel J. Thompson[3], Mark A. Haroldson[1], Kevin L. Frey[4], Curtis Hendricks[5], Jeremy M. Nicholson[5], Kerry A. Gunther[6], Katharine R. Wilmot[7], Hilary S. Cooley[8], Jennifer K. Fortin-Noreus[8], Pat Hnilicka[9], and Daniel B. Tyers[10]

[1]*U.S. Geological Survey, Northern Rocky Mountain Science Center, Interagency Grizzly Bear Study Team, 2327 University Way, Suite 2, Bozeman, MT 59715, USA*
[2]*Montana Fish, Wildlife and Parks, Kalispell, MT 59901, USA*
[3]*Wyoming Game and Fish Department, Large Carnivore Section, 260 Buena Vista, Lander, WY 82520, USA*
[4]*Montana Fish, Wildlife and Parks, Region 3 Headquarters Office, 1400 S 19th Avenue, Bozeman, MT 59718, USA*
[5]*Idaho Department of Fish and Game, Upper Snake River Region, 4279 Commerce Circle, Idaho Falls, ID 83401, USA*
[6]*National Park Service, Yellowstone Center for Resources, Bear Management Office, P.O. Box 168, Yellowstone National Park, WY 82190, USA*
[7]*National Park Service, Grand Teton National Park, P.O. Drawer 170, Moose, WY 83012, USA*
[8]*U.S. Fish and Wildlife Service, Grizzly Bear Recovery Program, University of Montana, 309 University Hall, Missoula, MT 59812, USA*
[9]*U.S. Fish and Wildlife Service, Mountain-Prairie Region, Lander Fish and Wildlife Conservation Office, 170 North 1st Street, Lander, WY 82520, USA*
[10]*U.S. Forest Service, Custer Gallatin National Forest, 10 East Babcock Street, Bozeman, MT 59715, USA*

Abstract: In the Greater Yellowstone Ecosystem, counts of female grizzly bears (*Ursus arctos*) with cubs-of-the-year (females with cubs) from systematic aerial surveys and opportunistic ground sightings are combined with demographic data to derive annual population estimates. We addressed 2 limitations to the monitoring approach. As part of a rule set, a conservative distance of >30 km currently is used as a threshold to assign sightings to unique females with cubs, resulting in underestimation bias. Using telemetry locations of females with cubs collected during 1997–2019, we created 1,000 data sets for each of 5 levels of simulated number of females with cubs, simulated sightings by selecting among these locations, and evaluated the classification performance of alternative distance criteria (12–30 km). Under all scenarios, 12–16-km criteria maximized classification performance and minimized estimation bias; the 16-km criterion was optimal for current conditions and sampling efforts. Our second objective was to test generalized additive models (GAMs) as a flexible trend analysis technique. We simulated 1,000 time series for each of 10 scenarios (10, 15, and 20% decline over periods of 5, 10, and 15 yrs, plus stability), applied GAMs, and assessed metrics associated with the posterior distribution of the instantaneous rate of change. We detected declines among >99.6% of replicates under the 15 and 20% decline scenarios and in 84.7–94.7% of replicates under the 10% decline scenario. From decline onset to first detection, periods ranged from 3.7 (20% decline over 5 yrs) to 11.1 (10% decline over 15 yrs), with 3.9–8.8 years mean duration of detection events. The GAM approach allows detection of directional changes in population trend, including early warning metrics, and stabilization after

[11]Present address: Montana Fish, Wildlife and Parks, Region 2, Ovando, Montana, USA
[12] email: fvanmanen@usgs.gov

such changes. Retrospective application of the 16-km distance criterion and GAMs resulted in higher population estimates and growth rates than are reported using current methods.

Key words: bias correction, females with cubs, Greater Yellowstone Ecosystem, grizzly bear, population monitoring, trend detection, *Ursus arctos*

DOI: 10.2192/ URSUS-D-22-00002.2 ***Ursus 33:article e17 (2022)***

Among bear populations, females with cubs of the year (hereafter, "females with cubs") are easily recognizable and counts for this reproductive segment of the population are used for population estimation or as ancillary data for population monitoring (e.g., Knight and Eberhardt 1984, Palomero et al. 1997). Use of such counts for monitoring purposes is based on the assumption that trends in this reproductive segment of the population are correlated with trends for the population as a whole (Eberhardt et al. 1999, Interagency Grizzly Bear Study Team 2006, Harris et al. 2007, Ordiz et al. 2007). Listed as a threatened population under the U.S. Endangered Species Act (1973, as amended) since 1975, 1 of 3 demographic recovery criteria established in the grizzly bear (*Ursus arctos*) recovery plan for the Greater Yellowstone Ecosystem specifies a minimum of 48 unique females with cubs annually within the area where monitoring takes place, the Demographic Monitoring Area (U.S. Fish and Wildlife Service 2017). Counts of females with cubs from systematic aerial surveys (twice/year; Jun–Aug) and opportunistic ground and aerial sightings collected from den emergence through 31 August have provided an important basis for monitoring the Yellowstone grizzly bear population since 1975. These count data are used by the Interagency Grizzly Bear Study Team (IGBST; established in 1973 and formalized via a Memorandum of Agreement in 1974, the IGBST is a science consortium of Federal, State, and Tribal agencies responsible for monitoring the grizzly bear population in the Greater Yellowstone Ecosystem) to estimate size and trend for this segment of the Yellowstone grizzly bear population and, in conjunction with additional demographic information, the size of the entire population. The annual estimation process involves 2 steps. First, sightings are differentiated into a minimum count of unique females with cubs using a rule set with criteria primarily based on litter size, time between sightings, and particularly distances among sightings (Knight et al. 1995). To develop the distance criterion, Knight et al. (1995) first estimated the mean standard diameter of annual ranges of females with cubs during 1 May–31 August (15 km; Blanchard and Knight 1991). They then doubled this diameter to 30 km (i.e., 2 × 15 km) and assumed this reflected a conservative estimate of the maximum distance between any 2 locations that a female with cubs would travel during the period from den emergence through 31 August. They established the rule that sightings of females with cubs lacking any other identifying characteristics (e.g., litter size) within this distance were categorized as repeat sightings of the same female and sightings beyond this distance were categorized as sightings of other females. Cub mortality was always possible, so no female with fewer cubs was considered distinct in a particular area unless she was seen on the same day as another female or unless both were radiomarked. Given that the population was in an early phase of recovery and demographic data were limited, these criteria represented a purposely conservative approach. However, using simulations, Schwartz et al. (2008) demonstrated that the Knight et al. (1995) rule set returned increasingly negative-biased estimates as the number of unique females with cubs increased. With higher grizzly bear densities currently existing in the Greater Yellowstone Ecosystem compared with several decades ago (see fig. 1S in Bjornlie et al. 2014), this underestimation bias is substantial and likely reduces the ability to accurately assess population trend.

The second step in the annual estimation process involves estimating the total number of females with cubs in the population, including those that have not been sighted. The nonparametric bias-corrected Chao estimate (Chao 1989) is used to obtain these estimates, which accounts for individual sighting heterogeneity based on sighting frequencies. These estimates are referred to as Chao2 estimates (N_{Chao2}) per Keating et al. (2002) and Cherry et al. (2007):

$$N_{\text{Chao2}} = m + \frac{\left(f_1^2 - f_1\right)}{2\left(f_2 + 1\right)},$$

where m represents the count of unique females with cubs and f_1 and f_2 represent females with cubs sighted once and twice, respectively (Chao 1989, Cherry et al. 2007).

Trend for the female with cub segment of the population is inferred from the time series of these annual Chao2 estimates. Annual variation in N_{Chao2} is relatively high be-

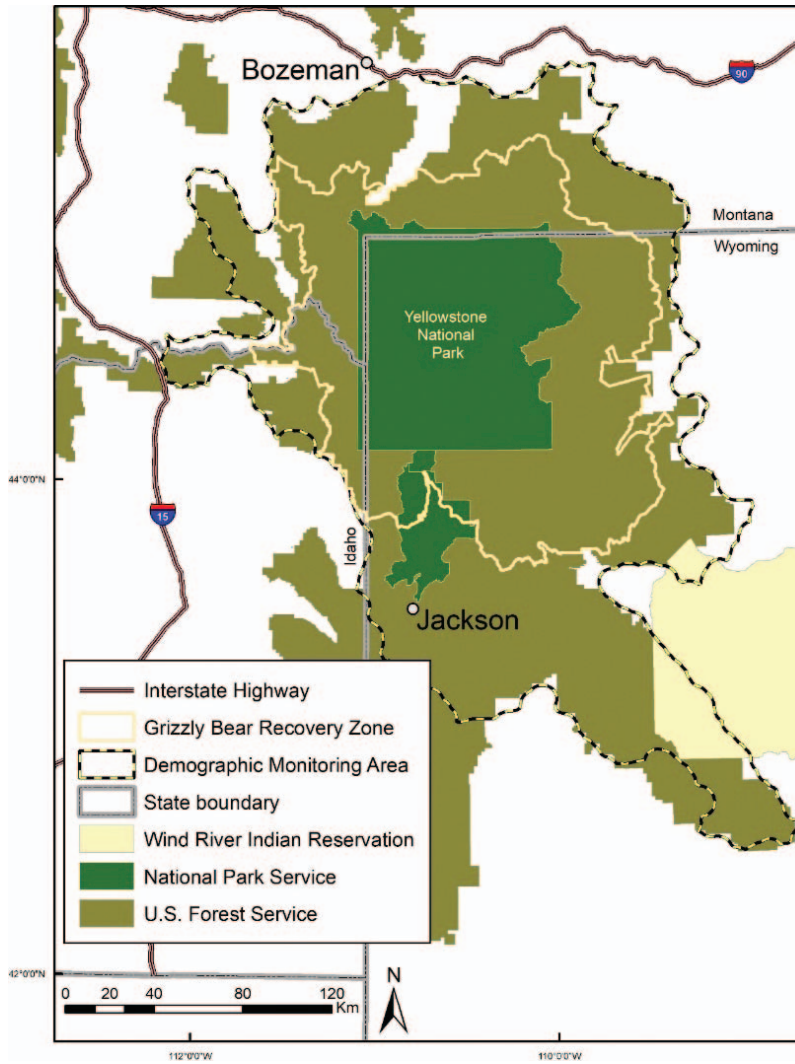

Fig. 1. Administrative boundaries relevant to grizzly bear (*Ursus arctos*) management in the Greater Yellowstone Ecosystem, Wyoming, Montana, and Idaho, USA.

cause of the sampling and process variance, so the IGBST developed and implemented a technique to address uncertainty in trend estimates and provide smoothed annual estimates of the number of females with cubs. The technique involved fitting linear and quadratic regressions to the time series starting in 1983 and using an information-theoretic approach to arrive at a model-averaged estimate for the endpoint of the time series (Harris et al. 2007). That approach provided a statistical mechanism to evaluate a change in trajectory for this population segment by monitoring the shift in model weight from the linear to the quadratic model. Although averaging of linear and quadratic models proved useful to detect a slowing of population growth in the early 2000s after almost 2 decades of robust growth, the approach has little power to accurately distinguish among future population scenarios that may involve periods of decline, stability, or growth (IGBST 2012, 2021). The 2016 Conservation Strategy for the Yellowstone grizzly bear population (a guiding document for management and monitoring of the population upon recovery and delisting) specified a management objective that reflects the mean population size during the period 2002–2014 (Yellowstone Ecosystem Subcommittee 2016), a period of relative stability after slowing of population growth in the early 2000s (van Manen et al. 2016). This management approach requires a trend

monitoring scheme that allows timely detection of any changes in population abundance.

In this study, we addressed the aforementioned limitations of the current population monitoring approach using simulation analyses and more flexible modeling techniques. Our objectives were to 1) identify alternatives to the 30-km distance criterion to produce unbiased counts of unique females with cubs, and 2) design more powerful and flexible trend analysis and smoothing techniques for abundance estimates of females with cubs derived from those counts.

Study area

The study area encompassed the Demographic Monitoring Area (49,931 km^2) of the Greater Yellowstone Ecosystem, within which demographic criteria for the Yellowstone grizzly bear population are currently monitored and evaluated (IGBST 2012). This area includes Yellowstone National Park, Grand Teton National Park, 5 adjacent national forests and other Federal lands, portions of the Wind River Indian Reservation, and State and private lands in Wyoming, Montana, and Idaho, USA (Fig. 1). The Greater Yellowstone Ecosystem is characterized by a high-elevation plateau with 14 mountain ranges >2,130 m, containing the headwaters of 3 continental-scale river systems (Missouri–Mississippi, Snake–Columbia, and Green–Colorado). Summers are short with average annual precipitation (51 cm) falling mostly as snow. Vegetation transitions from low-elevation grasslands through conifer forests at mid-elevations reaching alpine tundra around 2,900 m. Whereas the ability to obtain sightings varies depending on terrain features and vegetation cover, no trend in sightability was evident during 1986–2010 based on visual observations of females with cubs made during telemetry flights (van Manen et al. 2014).

Methods

Evaluating alternative distance criteria

Simulation framework. We developed a simulation framework to assess the performance of alternative distance criteria to correctly assign sightings to their respective true identities (IDs; Fig. 2). Schwartz et al. (2008) developed a computer algorithm to automate the application of the Knight et al. (1995) rule set consistent with its implementation. They then used location data of radiomarked females with cubs to simulate performance of the rule set under various hypothetical, but realistic, levels of "true" abundance of females with cubs. To accomplish the latter, they compiled a geospatial data layer with telemetry locations from multiple bear-years as if they had all been observed in a single year, and then randomly sampled from this superpopulation of observable bears. Live-trapping bears for radiomonitoring purposes is not feasible in some portions of the ecosystem; therefore, sets of known telemetry locations of females with cubs were placed on the data layer to populate areas in which few radiomarked females had been located but that were known to be occupied by adult female bears (Schwartz et al. 2008). The result was a representative distribution of bear locations for simulations to evaluate the Knight et al. (1995) rule set, with the goal of producing realistic inter-sighting distances and associated dates and times as crucial components of the rule set. They then took repeated samples ($n = 500$ simulations) of 10, 20, 40, 80, and 100 true females with cubs from this superpopulation to represent variability in samples obtained by chance through the sampling protocol.

For this study, we built on the general approach of Schwartz et al. (2008). We used aerial telemetry locations (1 May–31 Aug; individual bears located every 10–14 days) and ground sightings (prior to 31 Aug) of radiocollared females with cubs collected annually during 1997–2019 (*Supplemental Materials,* Fig. B.1). This data set was more recent compared with Schwartz et al. (2008) and enabled us to evaluate potential changes over time. Following the approach of Schwartz et al. (2008), we created 1,000 simulated data sets with true population sizes of females with cubs at each of 5 plausible levels ($N_{true} =$ 50, 60, 70, 80, and 90). We chose 50 as our minimum N_{true} to reflect the demographic recovery criterion of 48 unique females with cubs mentioned previously. We varied the total number of simulated sightings for each replicate as a ratio of N_{true}, based on empirical ratios of total sightings (n) and N_{Chao2} estimates for the period 1997–2019 (total sightings:unique females with cubs [n/N_{Chao2}]; mean = 2.3, range = 1.5–3.2). For simulated sightings, we retained the empirical day, month, time, and coordinate values from the telemetry data.

For each replicate, we allowed only 1 sample-year (1 sample-year = 1 year of location data for 1 female with cubs) to be chosen for any female with multiple years of data to prevent unrealistic spatial overlap (Schwartz et al. 2008). Similarly, our location sample spanned >20 years, so spatial overlap among different individuals could occur that is unrealistic. For example, if a female died and her home range was later occupied by a different female, randomly selecting both individuals may create an unrealistic dyad for evaluation of distance criteria. Therefore, when selecting individuals, we required the activity center of a newly selected candidate female to be ≥1 km

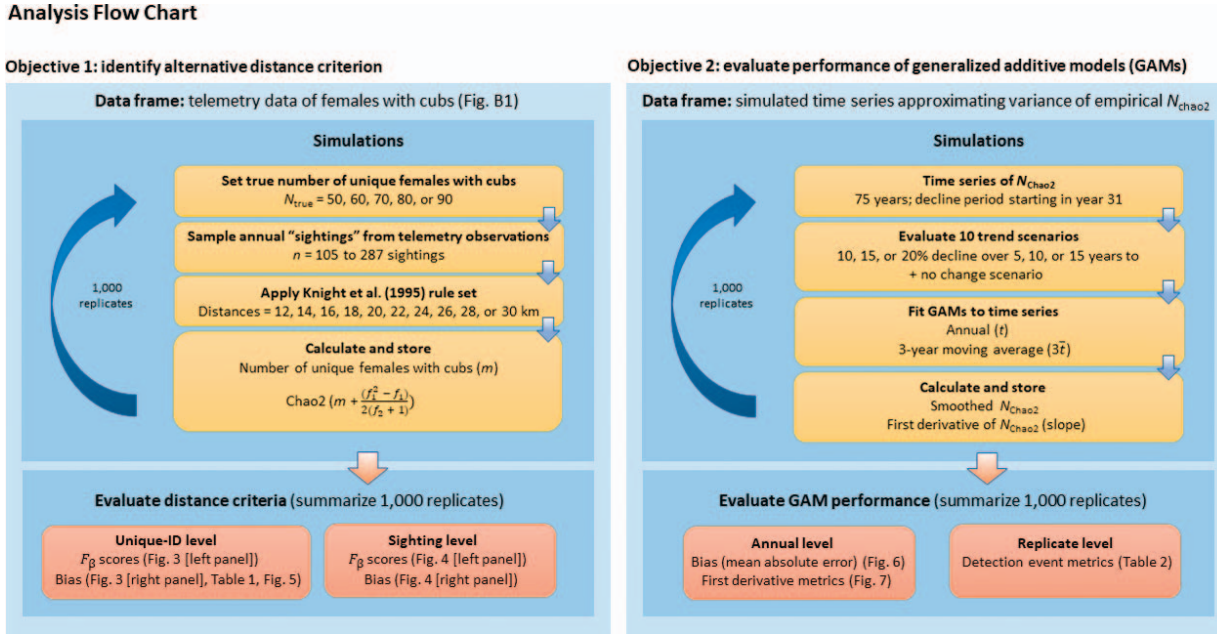

Fig. 2. Flow chart of analyses designed to evaluate and enhance techniques for monitoring the Yellowstone grizzly bear (*Ursus arctos*) population in the United States. See Results section for referenced figures and tables. See *Supplemental Materials* for Fig. B1.

from any activity center of a female previously selected while still allowing 2 simulated females to have a high degree of spatial overlap.

Litter size is also used in the clustering algorithm to distinguish sightings of unique females. Thus, we randomly assigned litter size to the earliest sighting of each female using discrete inverse transformation sampling (Devroye 1986) of empirical litter size data for the period 1997–2019 (Haroldson et al. 2020). We then simulated changes in litter size caused by cub mortality by applying estimated daily cub survival rates (IGBST 2012) to the number of days between simulated sightings of the same female. We censored simulated sightings if complete litter loss occurred because actual counts do not include females without cubs.

In field conditions, when observed females with cubs are radiomarked and are individually identified with telemetry, this information is included in the clustering algorithm and increases algorithm accuracy because these individuals cannot be misidentified (Schwartz et al. 2008). To simulate collared females with cubs, we assigned an identifier to a proportion of the females as being radiomarked in each replicate, based on a random sample from the distribution of empirical radiomonitored females with cubs on an annual basis (1997–2019; range = 3–13/yr).

Our simulation framework was designed to directly compare a true number of sighted bears with the estimate of *m,* but was not designed to test the efficacy of the Chao2 adjustment (i.e., $\frac{(f_1^2-f_1)}{2(f_2+1)}$). Thus, unsighted females were not simulated; therefore, inferences about N_{Chao2} are based on the premise of correctly assigning f_1 and f_2 sighting frequencies (i.e., the simulated f_1 and f_2 counts correctly capturing the Chao2 adjustment; Keating et al. 2002, Cherry et al. 2007, Schwartz et al. 2008).

Analysis. To assess the accuracy of different distance criteria, we used the computer program of Schwartz et al. (2008) to cluster sightings into individuals, by varying the distance threshold from 12 km to 30 km in 2-km intervals (i.e., 10 distance criteria) and holding all other parameters the same, including setting the spatial extent to the Demographic Monitoring Area. We chose 12 km as a lower bound for the range of distance criteria, based on the original 15-km annual home-range diameter documented by Knight et al. (1995) and more recent findings from Bjornlie et al. (2014), indicating that female home ranges in areas with higher bear densities have decreased in size. This approach resulted in 50,000 output data sets, with 1,000 simulated data sets of females with cubs for each of the 10 distance criteria and the 5 levels ($N_{\text{true}} = 50$, 60, 70, 80, and 90) of females with cubs (5 population levels $\times$ 10 distance criteria $\times$ 1,000 replicates each = 50,000).

We evaluated classification performance and bias associated with the 10 distance criteria at the unique ID level and the sighting level, allowing us to examine how assignments influenced the 2 components of the Chao2 equation. Accurate counts of m influence the primary and largest component of the equation. Accurate assignment of f_1 and f_2 sighting frequencies to each female ID influences the Chao2 adjustment ($\frac{(f_1^2-f_1)}{2(f_2+1)}$; Chao 1989, Cherry et al. 2007). Rather than evaluate strict accuracy of assignments, which would not indicate whether false negatives or false positives were more common, we evaluated classification performance by considering precision and recall (Lever et al. 2016). At the unique ID-level, we compared the presence or absence of unique IDs of females with cubs in the simulation (true IDs) with modeled output (predicted IDs). This involved 3 distinct outcomes: 1) true positive (true ID correctly predicted); 2) false positive (ID erroneously predicted to be present when sightings of a single true ID are split into multiple IDs); and 3) false negative (failure to predict true ID because multiple IDs are combined into a single ID). The true number of unique IDs is the sum of those correctly classified (true positives), plus those that were missed (false negatives). Only when the number of false positives equals false negatives are the correct numbers of unique IDs predicted. We used the F_β score, which is an aggregative performance metric of precision ($\frac{\text{true positive}}{\text{true positive} + \text{false positive}}$) and recall ($\frac{\text{true positive}}{\text{true positive} + \text{false negative}}$). The F_β score is bounded by 0 and 1, with higher values indicating better classification performance (Lever et al. 2016). The parameter β controls the balance of precision and recall, which we set to 1 for equal balance. At the sighting level, we assessed assignment of sightings to their respective true IDs by using a multiclass confusion matrix (Grandini et al. 2020). For multiclass classification where each observation can only be assigned to a single class label (i.e., female ID), average precision = recall = F_β score. Therefore, we used mean F_β scores to summarize classification performance at the sighting level.

To assess the likelihood of overestimation at the unique ID level, we calculated the proportion of simulations with positive bias >5 and >10% of N_{true}, the Chao2-adjustment value, and the resulting Chao2 estimates. Simulated data sets showed a strong positive correlation between the f_1:f_2 ratio and the Chao2 adjustment component in the Chao2 equation at all N_{true} levels (Spearman's rank $\bar{r}_s = 0.93$, $\sigma = 0.01$). Therefore, we estimated the true and predicted adjustment component of the Chao2 equation to quantify bias at the sighting level because it is the direct application of f_1 and f_2 counts, and interpretation in terms of Chao2 units is intuitive. We focused that analysis on the 3 top-performing distance criteria and, for comparison, the original 30-km criterion. We also focused on N_{true} levels of 60 and 70 unique females with cubs because empirical estimates of m based on the 30-km criterion and total sightings for 2001–2019, when linked to simulation results, suggest this range was most relevant to contemporary conditions (*Supplemental Materials,* Appendix A).

Evaluating performance of generalized additive models

Simulation framework. We used another simulation framework, separate from the previous analysis, to create realistic variation in trends of annual N_{Chao2} estimates (Fig. 2). Given that the 2016 Conservation Strategy specified a management objective that reflected the mean population size during the period 2002–2014 (Yellowstone Ecosystem Subcommittee 2016), we developed data sets to simulate a stable population experiencing a decline followed by a return to stability. We varied magnitude (10, 15, and 20%) and duration (5, 10, and 15 yrs) of the decline periods to reflect grizzly bear biology and the management approach. The combined duration × magnitude effect sizes correspond to constant population growth rates (λ) during decline years, ranging from 0.956 (20% decline over 5 yrs) to 0.993 (10% decline over 15 yrs; *Supplemental Materials,* Table B.1). We also included a null scenario of zero growth ($\lambda = 1.0$). Although we only simulated population declines to keep our analyses focused, we note that the ability to detect population increases is just as relevant to monitoring of the Yellowstone grizzly bear population. The flexibility of GAMs means that model performance would be the same and results would be equally applicable for equivalent scenarios of population increase.

To simulate annual N_{Chao2} values, we added "residual-noise" to the deterministic trends shown in Table B.1 (*Supplemental Materials*) with the goal of approximating process and sampling variance present in observed N_{Chao2} estimates. We used the empirical residuals from a regression of $N_{\text{Chao2}} \sim$ year from 2000 to 2018 data to parameterize residual noise and assumed a relatively stable true population during this period with noise equally distributed in positive and negative directions. We extracted the residuals from the regression and used their statistical properties (e.g., autocorrelation, standard deviation) to simulate an auto-regressive time series using the arima.sim() function in Program R (R Core Team 2019). We scaled simulated residuals by dividing by

the deterministic value (mean value of empirical N_{Chao2} estimates for 2000–2018; $N_{\text{Chao2}} = 55.8$). This allowed us to use the same simulated time series of noise for each scenario, regardless of the deterministic values. Scaled residuals were subsequently "unscaled" by multiplying by the deterministic time series value and adding it to the deterministic value to create stochastic time series.

We simulated 1,000 replicate time series of 75 years for each of the 10 scenarios (i.e., 1 null and 9 treatment scenarios). This period length allowed for stabilization, or "burn-in," time before the start of the decline period, and a postdecline stable period. We started declines in year 31 of the simulation. Stochastic noise of the simulated N_{Chao2} values resulted in variation across simulations leading up to the decline period (i.e., sometimes higher than deterministic values, sometimes lower), which added a realistic variance component to the simulations. To account for the observed increase of the empirical N_{Chao2} estimates for the Yellowstone grizzly bear population through the early 2000s (IGBST 2012), we set the first 10 years of each simulation to be an increasing linear trend, thus requiring the GAM models to make an initial "turn" from increasing to stable deterministic trends. Only contemporary trends are the target of our GAM application, so we chose a generic increase to challenge the model but were less concerned with exactly matching the empirical data for these first 10 years of the simulations. Thus, simulations started with a "burn-in" of 10 years of increase plus 20 years of stable population size, followed by a population decline lasting 5, 10, or 15 years and then a postdecline stable period of up to 40, 35, and 30 years, respectively.

Model parameterization. For each monitoring year of a simulation scenario, we fit a GAM to annual N_{Chao2} estimates (N_{Chao2_t}):

$$N_{\text{Chao2}_t} = \beta_0 + f(\text{year}_t) + \varepsilon_t,$$

where f is a smooth function of the covariate year from $t = 1$ to the current monitoring year and ε_t is a vector of error terms. We fit a similar model to 3-year moving averages of annual N_{Chao2} estimates ($N_{\text{Chao2}_{3\bar{t}}}$). We fit models with the mgcv package (Wood 2004) in Program R and evaluated model performance using raw and 3-year simple moving average ($\bar{x}_3$) of simulated N_{Chao2} values. We chose to include the moving average based on exploratory work and previous research showing that any reduction in sampling variance would increase power to detect trends (Harris et al. 2007). Although use of simple moving averages lags the data by half the size of the sample window (e.g., 1.5 yrs in our application) and delays the onset of changes in the input signal, the reduction in the signal-to-noise ratio generally outweighs this lag effect in model performance. Moving averages increase autocorrelation in the time series that could lead to overfitting if not accounted for; therefore, we modified the default GAM parameterization to protect against overfitting by upscaling the spline penalization. To provide a reasonable time-series for fitting models, we began model-fitting in simulation year 25 with 5 years of pre-impact before deterministic trends started. For each simulated monitoring year, we fit a GAM with N_{Chao2} or its 3-year moving average as the response variable and year as the predictor variable. Use of fitted models followed existing monitoring protocols of using only the monitoring year, or last year of a fitted model, for interpretation and not back-correcting estimates of previous years as time advances and more data become available. To account for the presence of autocorrelation in the data and protect against overfitting, we increased the effective degrees of freedom penalty by 30% ($\gamma = 1.3$). This produced a smoother fit (Kim and Gu 2004; Wood 2006, 2017) and balanced overfitting protection while still allowing the smoother to respond nonlinearly to changes in trend. Failure to account for such dependencies in the N_{Chao2} values could lead to overly complex model fitting and a greater probability of false-positive results (Simpson 2018). Following Wood (2011) and Simpson (2019), we used restricted maximum likelihood for parameter estimation. We set the smoother function to use univariate penalized cubic regression splines (Wood 2017).

Model outputs and inference. For each monitoring year ($n = 50$) and simulation replicate ($n = 1{,}000$), we fit GAMs and used the methods of Simpson (2018) to generate and store 1,000 posterior distributions of the smoothed N_{Chao2} estimates and the first derivatives $f'(\text{year}_t)$, or slopes, as a measure of the instantaneous rate of change (*Supplemental Materials,* Appendix C). We used the variation of estimated slopes from the control (no decline) scenario simulations to optimize the α-level based on rates of false-positive events, defined by first derivatives being significantly different from zero.

We calculated model bias as the mean absolute error of smoothed estimates (i.e., predicted GAM values) for each monitoring year at the replicate level. We selected mean absolute error over the more conventional root mean squared error because it retains the directionality of bias (positive vs. negative). We calculated this metric as follows:

$$\text{mean absolute error}_t = \text{fitted } N_{\text{Chao2}_t} - \text{deterministic } N_{\text{Chao2}_t},$$

where fitted N_{Chao2_t} and deterministic N_{Chao2_t} are the median of the smoothed N_{Chao2} posterior distribution and the simulated deterministic N_{Chao2} values, respectively, associated with year t. We output 3 metrics associated with the GAM first derivatives (f'(year)) that might be used to interpret trend. The first was the median of the posterior distribution as the point estimate for trend. The second was the proportion of the posterior distribution that was <0, representing the probability of decline (pd). The third was "trend state," for which we assigned a state of decline for years when $(1-\alpha)$ confidence intervals of first derivatives were different from zero and a state of no decline for years when the confidence interval contained 0.

Quantifying results across replications is challenging because measures of central tendency are relevant to the annual and replicate level. At the annual level, results reflect the average dynamics for a given year but do not account for time dependency present within a time series of a single replicate. Therefore, we also considered replicate-level results and explicitly accounted for the time dependency of each entire time series (i.e., any trend among annual data within each replicate, representing how monitoring data are used in practice) by using an indicator variable for "trend state." We defined detection of a decline event as ≥ 2 consecutive years in the decline state. For each scenario, we estimated the proportion of simulations where a decline event was detected, the lag (yrs) between the start of the decline and its detection, and the length of the detection event (yrs).

Results

Evaluating alternative distance criteria

Summary of simulation data. The simulated sighting data to evaluate alternative distance criteria contained 1,139 locations of females with cubs, representing 117 unique bears (*Supplemental Materials,* Fig. B.1). For some individual females, we had multiple years of location data while accompanied by cubs, resulting in 154 sampling years. The number of simulated sightings per unique female varied from 1 to 36 ($\bar{x} = 7.4$, $\sigma = 4.1$). Median distance between simulated sightings within the same individual averaged 9.1 km ($\sigma = 5.9$ km), ranged from 0.2 to 37 km, and lacked evidence of directional trend during 1997–2019 ($\beta = -0.05$, $P = 0.49$, adjusted $R^2 = -0.004$). The median diameter of the smallest circle encompassing all simulated sightings for unique individuals was 16.4 km, with 86% of individuals' minimum diameter <30 km; no trend was evident over the period 1997–2019 ($\beta = -0.24$, $P = 0.16$, adjusted $R^2 = 0.007$).

The average distance from each individual's centroid location to that of their nearest neighbor's centroid decreased with increasing number of unique females (e.g., $\bar{x}_{50} = 13$ km, $\sigma_{50} = 8.6$; $\bar{x}_{90} = 9.2$ km, $\sigma_{90} = 6.4$; subscript represents N_{true} level).

The number of nearest neighbors with centroids within 30 km also increased with N_{true} levels, showing an 81% increase from 50 to 90 unique females ($\bar{x}_{50} = 3.8$, $\sigma_{50} = 2.3$; $\bar{x}_{90} = 6.9$, $\sigma_{90} = 3.6$). The spatial extent of the sampling frame (i.e., Demographic Monitoring Area) was fixed across all simulation replicates, so these patterns reflect increasing density of simulated females as N_{true} increases. Additional summary statistics are in *Supplemental Materials,* Appendix D.

Evaluation of alternative distance criteria. At the unique-ID level, mean F_β scores were highest at distance criteria between 14 and 18 km (Fig. 3 left panel), whereas mean bias was minimized between 12 and 16 km (Fig. 3 right panel). The top-performing distance criteria for both measures decreased with increasing N_{true} level, but those maximizing classification performance (Fig. 3 left panel) did not always match those minimizing bias (Fig. 3 right panel). However, the top-performing criteria were always within 1 distance increment (e.g., 14 vs. 16 km); differences in mean F_β scores were within 1% of each other and differences in mean bias were ≤5% of the N_{true}.

At the sighting level, mean F_β scores were lower than the unique-ID level, although patterns related to distance criteria were similar (Fig. 4 left panel). Top-performing distance criteria ranged from 12 to 16 km for classification and bias, with smaller distance criteria performing better with increasing N_{true}. Differences between mean F_β of the top-performing distance criteria and its closest competitors were small (Fig. 4). Distance criteria showed distinct and consistent pattern across all N_{true} levels, with effects over the range of distance criteria (12 vs. 30 km) outweighing effects within distance criteria over the range of N_{true} (Fig. 4 right panel). Again, this assessment of bias in the Chao2 adjustment is reflective of the bias relative to the simulated frequencies (i.e., perfect clustering of all locations assigned to the correct unique ID), not bias associated with females not observed.

Correlations between the bias in m and the bias in Chao2 adjustment were moderate ($r_s = 0.44$–0.70) but strengthened with increasing numbers of simulated females with cubs (Fig. 5). Scatterplot patterns indicated the relationship between bias in m and the Chao2 adjustment were similar across different distance criteria within levels of N_{true}. However, distance criteria in the 12–16-km range best minimized bias of both m and the

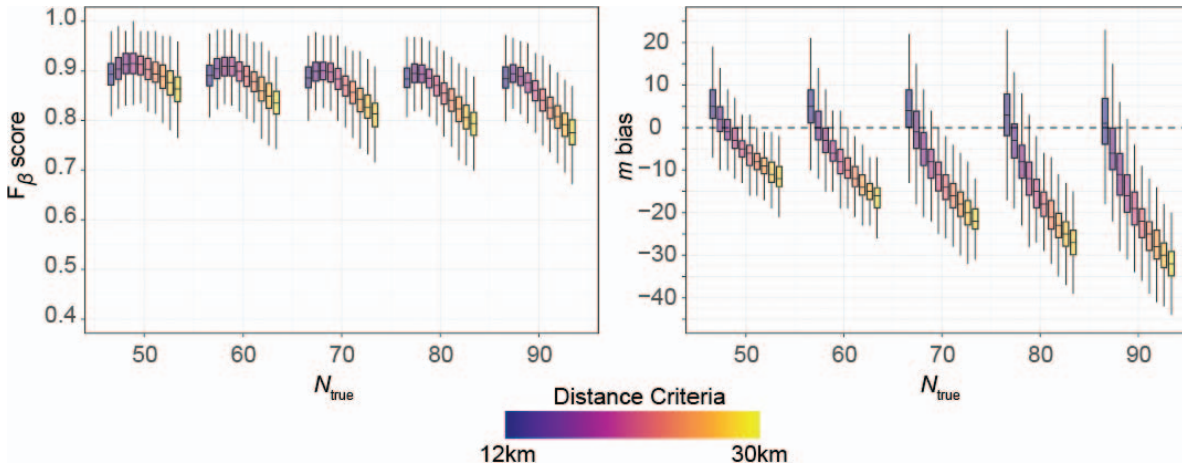

Fig. 3. Classification performance at the unique ID-level shown by (left panel) F_β score and (right panel) predicted bias in the number of unique females (*m* bias) based on simulations applying varying distance criteria to the Knight et al. (1995) rule set to identify unique female grizzly bears (*Ursus arctos*) with cubs from sightings. For each N_{true} level, distance criteria range from 12 to 30 km in 2-km steps (indicated by color gradient and arranged from left to right). Each boxplot summarizes $n = 1{,}000$ simulated data sets.

Chao2 adjustment (Fig. 5). The range of bias in the Chao2 adjustment for unbiased ranges of *m* highlights the additional challenge of simultaneously estimating the f_1 and f_2 sighting frequencies compared with only *m*. For example, for distances of 12 km to 16 km, even when *m* was predicted with reasonable accuracy (e.g., within ±2 females with cubs of the true value), although mean bias of f_1 and f_2 sighting frequencies was low, individual replicates varied from −10 to 13 for f_1 and −17 to 14 for f_2 (Fig. B.2). These biases were negatively correlated, and overestimation of f_1 generally corresponded to underestimates of f_2 and a positive bias in the adjustment component of N_{Chao2}, whereas underestimation of f_1 corresponded to overestimation of f_2 and a negative bias in the Chao2 adjustment (Fig. B.2). Despite the challenges of simultaneously reducing bias of m, f_1, and f_2 estimates

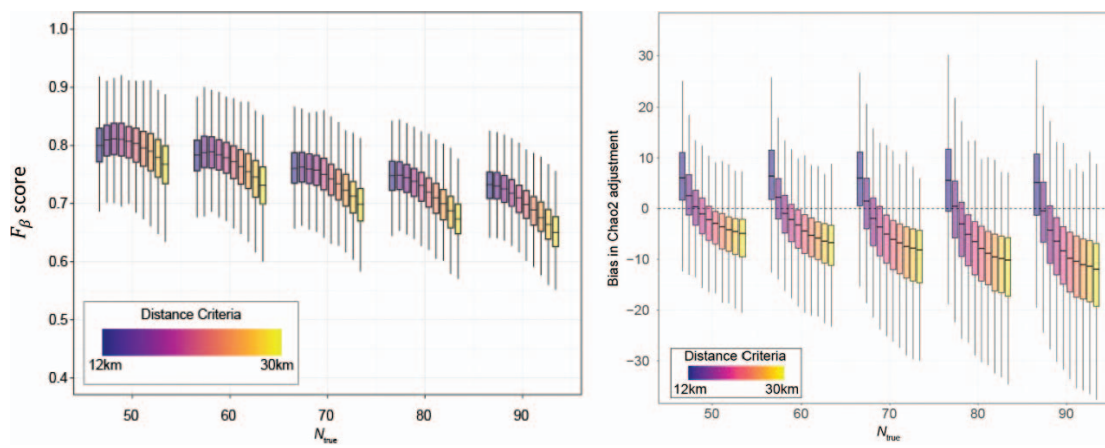

Fig. 4. Classification performance at the sighting level based on simulations applying varying distance criteria to the Knight et al. (1995) rule set to identify unique female grizzly (*Ursus arctos*) bears with cubs from sighting. (left panel) F_β score and (right panel) predicted bias (predicted−simulated; expressed as no. of females with cubs) in the Chao2 adjustment ($\frac{(f_1^2 - f_1)}{2(f_2+1)}$) of the Chao2 equation. For each N_{true} level, distance criteria range from 12 to 30 km in 2-km steps (indicated by color gradient and arranged from left to right). Each boxplot summarizes $n = 1{,}000$ simulated data sets based on micro-averaged F_β scores.

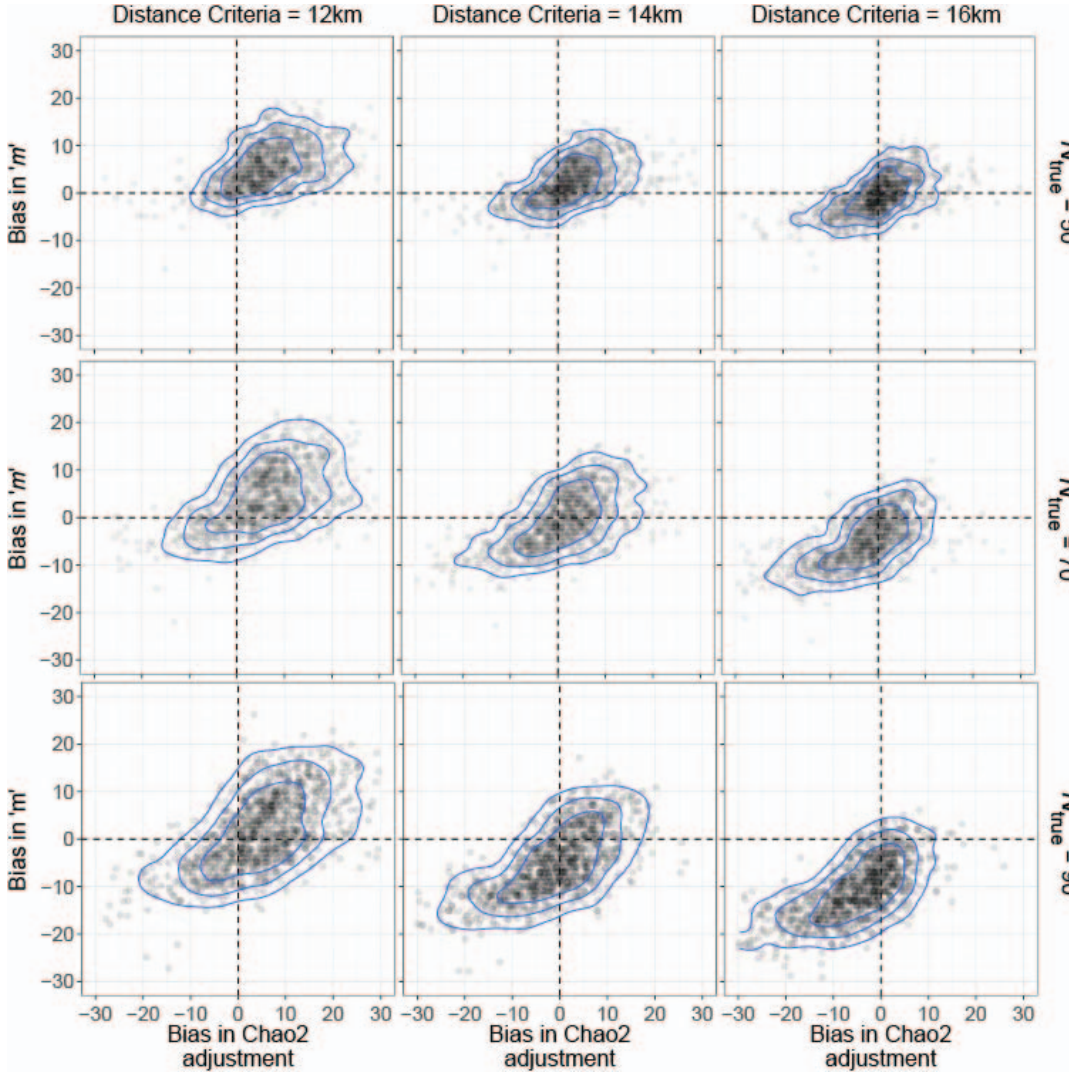

Fig. 5. Relationships between bias (expressed as no. of unique females with cubs) in the parameter *m* and bias in the Chao2 adjustment (i.e., $N_{Chao2} - m$) of the estimator based on simulations ($n = 1{,}000$ replicates for each combination of distance criterion and N_{true} level), applying varying distance criteria to the Knight et al. (1995) rule set to identify unique female grizzly bears (*Ursus arctos*) with cubs. Results are shown for distance criteria of 12, 14, and 16 km (columns) within each of 3 simulated levels of true females with cubs ($N_{true} = 50$, 70, and 90; rows). Blue contour lines represent 50th, 75th, and 90th isopleths, respectively.

at smaller distance criteria, improvements relative to the 30-km rule set were substantial. For example, even at the lowest N_{true} level of 50, where the 30-km distance criterion performed best, differences between 30 and 16 km were large: mean bias of m, f_1, and f_2 using the 16-km distance criterion were −0.6, −0.1, and −0.4, respectively, but for the 30-km criterion were −11.6, −8.7, and −5.3, respectively.

The proportion of simulations with positive bias >5 or >10% of N_{true} indicated trade-offs among distance criteria for minimizing mean bias while also reducing risk of overestimation (Table 1). This pattern was similar for m, the Chao2 bias adjustment, and overall Chao2 estimates, but most distinct for the latter.

Evaluating performance of generalized additive models

Model performance. Monitoring year estimates for the $N_{Chao2_{3i}}$ model under the null model scenario (i.e., no simulated decline) were unbiased, with >85% of

Table 1. Measures of bias for top-ranking (12–16-km) and reference (30-km) distance criteria for simulations where $N_{true} = 60$ and 70 female grizzly bears (*Ursus arctos*) with cubs and total sightings within the empirical range of $n \leq 165$. Results are based on 1,000 simulation replicates for each combination of distance criterion and N_{true} level using the Knight et al. (1995) rule set to identify unique females with cubs from sightings. (A) *m* bias (no. of unique females with cubs), and proportion of simulations >+5 and >+10% of N_{true}. (B) Chao2-adjustment bias, and proportion of simulations >+5 and >+10% of mean known Chao2 adjustment (simulation estimate). (C) Chao2 bias, and proportion of simulations >+5 and >+10% of mean known Chao2 (simulation estimate). The 5% adjustment of the known Chao2 was approximately 3.6 ($N_{true} = 60$) and 4.4 ($N_{true} = 70$).

(A) *m* bias

Distance criterion	N_{true}	Mean bias	Proportion, bias >+5% N_{true}	Proportion, bias >+10% N_{true}
12	60	3.7	0.53	0.30
	70	0.4	0.21	0.04
14	60	–0.4	0.19	0.04
	70	–4.3	0.03	0.00
16	60	–3.9	0.03	0.00
	70	–8.3	0.00	0.00
30	60	–16.9	0.00	0.00
	70	–23.4	0.00	0.00

(B) Chao2-adjustment bias

Distance criterion	N_{true}	Mean bias	Proportion, bias >+5% N_{true}	Proportion, bias >+10% N_{true}
12	60	4.6	0.60	0.39
	70	1.9	0.41	0.23
14	60	0.3	0.35	0.14
	70	–3.1	0.18	0.06
16	60	–3.0	0.14	0.04
	70	–7.0	0.08	0.01
30	60	–9.6	0.00	0.00
	70	–14.5	0.00	0.00

(C) Chao2 bias

Distance criterion	N_{true}	Mean bias	Proportion, bias >+5% N_{true}	Proportion, bias >+10% N_{true}
12	60	8.3	0.69	0.59
	70	2.3	0.48	0.31
14	60	–0.2	0.39	0.24
	70	–7.4	0.12	0.05
16	60	–6.8	0.12	0.04
	70	–15.2	0.03	0.00
30	60	–26.5	0.00	0.00
	70	–37.9	0.00	0.00

monitoring years ($n = 50{,}000$) within 2 N_{Chao2} units from the deterministic, or true, N_{Chao2} value (mean absolute error = 0.029; $\sigma = 1.33$). The N_{Chao2_t} model showed a slight positive bias under the null model scenario (mean absolute error = 0.484, $\sigma = 1.33$). These differences occurred mostly during the pre-impact phase and were associated with larger lag effects because of higher variance in the annual N_{Chao2_t} values subsequent to the initial increase prior to stabilization. For the 9 treatment scenarios, the fitted bias varied as a function of the interaction of effect size and duration. General patterns in bias reflected the lag time required for models to distinguish declines from annual variation in N_{Chao2}. Smoothed estimates during the decline were positively biased and bias increased with the size and steepness of simulated declines. Similarly, during the postdecline stabilization period, there was a transition to a period of negative bias when responding to the change from decline to stabilization. Compared with the N_{Chao2_t} models, variance reduction associated with the 3-year moving averages of the $N_{Chao2_{3\bar{t}}}$ models resulted in less bias during impact phases and faster returns during the subsequent stable period with bias levels equivalent to the null model scenario (Fig. 6).

Trend detection. We present trend detection results only for the $N_{Chao2_{3\bar{t}}}$ model because of its smaller bias compared with the N_{Chao2_t} model. For all scenarios, the annual probability of decline, *p*d, increased within the first 2–3 years of the decline period, indicating that

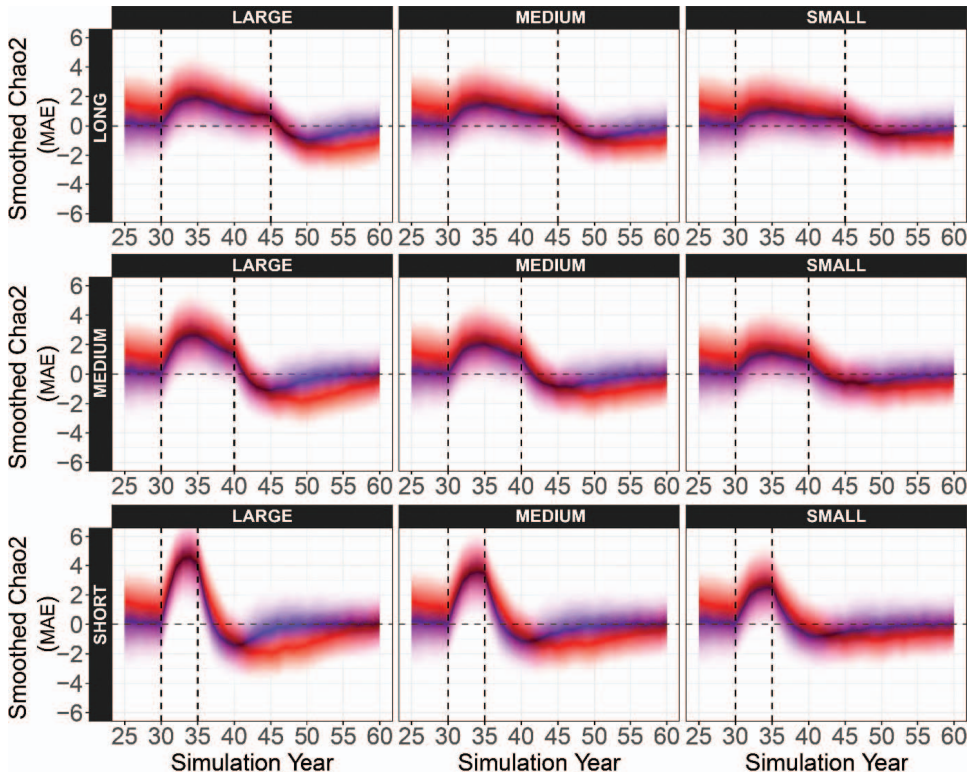

Fig. 6. Mean absolute error (MAE) for $N_{\text{Chao2}t}$ (annual; blue) and $N_{\text{Chao2}_{3\bar{t}}}$ (3-yr moving average; red) fitted estimates of female grizzly bears (*Ursus arctos*) with cubs as a function of simulation year. Columns show magnitude of the impact (large = 20%, medium = 15%, small = 10% decline) and rows show duration of impact period (long = 15 yrs, medium = 10 yrs, short = 5 yrs). Dashed vertical lines indicate the start and end of the impact period and dashed horizontal line indicates reference bias of 0.

existence of a declining trend was rapidly detected. Temporal dynamics based on median *p*d values closely tracked the different scenarios, with shorter durations and larger magnitudes resulting in faster shifts and larger probabilities. Peak values for annual medians ranged from 0.845 (decline = 10% over 15 yrs; $\lambda = 0.993$) to 0.999 (decline = 20% over 5 yrs; $\lambda = 0.956$). On average, median *p*d values reached maximum levels within 3 years of the end of the decline period for all durations of decline, except for the short duration of 5 years, which reached maximum values 1 or 2 years after completion of the decline period (Fig. 7). This pattern reflects that changes postdecline were more pronounced with shorter impact periods because it is difficult for models to fully capture these rapid dynamics.

At the annual level, all impact scenarios except for the most gradual decline (10% over 15 yrs; $\lambda = 0.993$) showed support for significant negative slopes when averaged across replicates (Fig. 7). Although the scenario of 10% decline over 15 years lacked power to detect slope significance, it is unlikely that trends would go undetected. For example, during the simulated decline phase, median posterior slope estimates were less than the previous year's estimates during 56% of simulation years, and the number of consecutive years under this pattern (current year's slope < previous year's slope) averaged 4.5 years. When coupled with inference on the probability of decline, *p*d, which averaged 0.72 during the impact period, the temporal trends provide substantial additional inference of changing conditions despite the lack of statistical significance.

Summarizing results at the replicate level using the state variable of decline versus no decline for each year of a time series, detection of simulated declines was high, with declines detected in >99.6% of replicates under the medium (15%) and large (20%) decline scenarios. For small magnitude scenarios (10%), declines were detected in 84.7% (15-yr duration) to 94.7% (5-yr duration) of

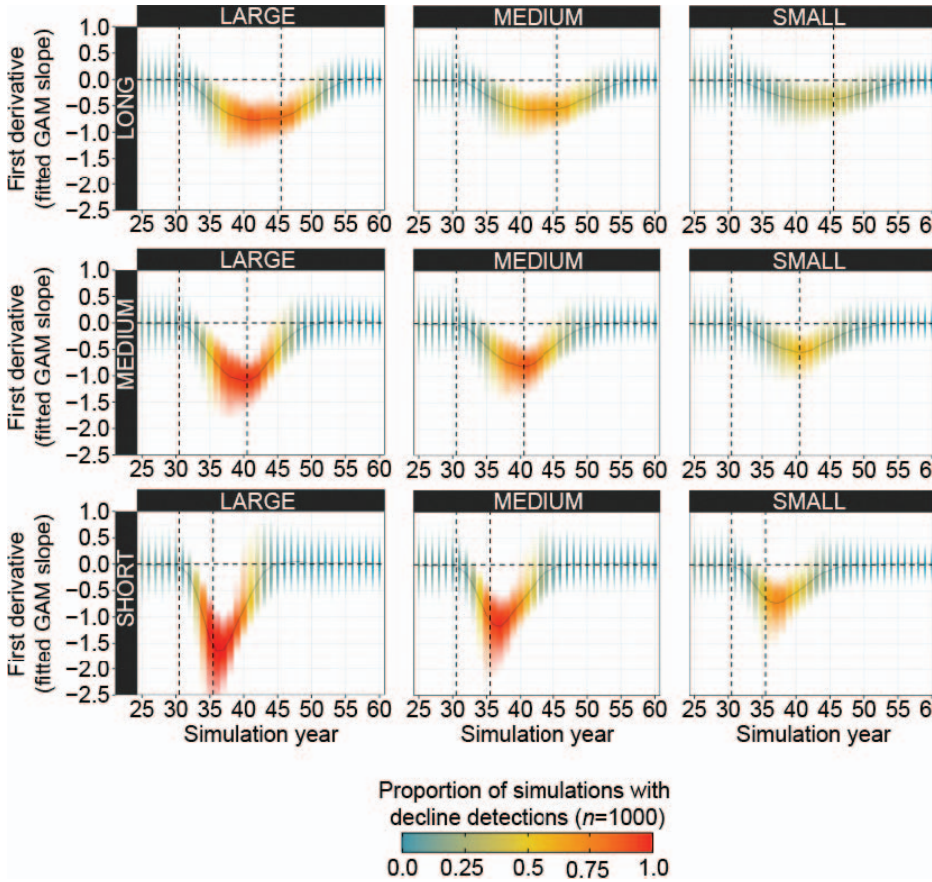

Fig. 7. Trend dynamics for $N_{\text{Chao2}_{3i}}$ first derivative (slope) posterior distributions of number of female grizzly bears (*Ursus arctos*) with cubs, for 9 treatment scenarios. Columns show magnitude of the impact (large = 20%, medium = 15%, small = 10% decline) and rows show duration of impact period (long = 15 yrs, medium = 10 yrs, short = 5 yrs). Black dashed lines indicate the start and end of the simulated decline periods. Density strips associated with each year reflect the distribution of posterior medians across replicates (n = 1,000). The width of each density strip reflects the average *p*d value, scaled such that *p*d = 1.0 is reflected by adjacent years having no space between their density strips. Color gradient indicates the proportion of simulations (n = 1,000/scenario) in a decline "state" where confidence intervals for ≥2 consecutive years do not contain zero.

replicates. The mean number of years from decline onset to year of first detection ranged from 3.7 (20% decline over 5 yrs) to 11.1 (10% decline over 15 yrs), and mean detected duration of events (i.e., consecutive years in detect state; range = 3.9–8.8 yrs) was correlated with interaction of decline duration × magnitude (Table 2). Patterns for detecting the return to stabilization postdecline were similar to decline detection and symmetrical around the peak support for decline for the 15- and 10-year decline scenarios (Fig. 7). Five-year scenarios showed slight asymmetry around the peak, with a longer and more linear return toward negligible levels. For all scenarios, rebounding trends were evident well before state transition from decline to no decline occurred, based on the relative change in *p*d and sustained increases in the median posterior distribution (Fig. 7).

Discussion

Our goals for this study were to enhance techniques for estimation and trend analysis of the number of female grizzly bears with cubs, which provides an important component for monitoring and management of the Yellowstone grizzly bear population. In the first part of our study, we addressed a previously documented underestimation bias due to application of a conservative,

Table 2. Change detection metrics for 9 scenarios of decline for simulated time series of N_{Chao2} estimates of female grizzly bears (*Ursus arctos*) with cubs, based on significance of first derivative of generalized additive models and 3-year simple moving averages ($N_{Chao2_{3y}}$). We simulated 1,000 replicate time series for each scenario, each with a length of 75 years and with population decline starting in year 31 of the simulation. Detection of the decline event was defined as ≥2 consecutive years with first derivatives statistically different from 0. Mean and standard deviation (SD) for lag to detect reflect the number of years post–simulation decline before a detection and its variation. Mean length of detection events represent the mean number of consecutive years in each event and the mean slope estimate gives an indication of effect size.

Decline duration[a]	Decline magnitude[b]	Proportion with detected declines	Mean lag to detect (yrs)	SD lag to detect	Mean duration of detection event (consecutive yrs in detect state)	Mean slope estimate (first derivative) at first detection
Long	Large	1.00	6.91	2.54	8.29	−0.75
Long	Medium	0.99	8.74	3.80	5.76	−0.64
Long	Small	0.85	11.05	5.45	3.88	−0.53
Medium	Large	1.00	5.36	1.71	8.83	−0.92
Medium	Medium	1.00	6.53	2.49	6.76	−0.74
Medium	Small	0.90	8.66	4.51	4.33	−0.60
Short	Large	1.00	3.65	1.09	7.80	−1.21
Short	Medium	1.00	4.36	1.48	6.92	−0.93
Short	Small	0.95	6.02	3.70	4.70	−0.70

[a]Long = 15 yrs, medium = 10 yrs, short = 5 yrs.
[b]Large = 20%, medium = 15%, small = 10%.

30-km distance threshold to identify sightings as belonging to unique females with cubs. Our simulations indicated that distance criteria <30 km increased classification accuracy and reduced bias associated with m and the Chao2 adjustment to m. Top-performing distance criteria varied with the number of unique females with cubs being simulated (N_{true}), the number of sightings (n), and their ratio (n/N_{true}). Distance criteria in the range of 12–16 km minimized bias and maximized classification performance at the unique ID and sighting levels under all simulation scenarios. Considering our objective to reduce underestimation bias while limiting the risk of overestimation, selecting a single optimal distance criterion from within the 12–16-km range requires additional considerations.

The 16-km distance criteria exhibited little bias with low risk of overestimation. On average, use of the 16-km criterion underestimated m by −3.9 ($N_{true} = 60$) to −8.3 ($N_{true} = 70$) females, and overestimated m by >5% in only 3% ($N_{true} = 60$) and 0% ($N_{true} = 70$) of simulations. Although the 14-km distance criterion was less biased on average, it had higher proportions of simulations with >5% bias (Table 1A). The Chao2 adjustment to m was also unbiased at the 16-km distance criterion, under the assumption that the true classification (simulated sightings) produced the correct N_{Chao2} to account for females not seen (Keating et al. 2002, Cherry et al. 2007, Schwartz et al. 2008). Using the benchmark of 5% of the mean simulated Chao2 estimates, the proportion of simulated Chao2 adjustments based on the 16-km distance criterion exceeded this benchmark in fewer than 14% ($N_{true} = 60$) and 8% ($N_{true} = 70$) of simulations (Table 1B). Finally, the combined estimation bias of m and the known Chao2 adjustment averaged −6.8 ($N_{true} = 60$) and −15.2 ($N_{true} = 70$) using the 16-km distance criterion. These represent 26% and 40% reductions in the Chao2 bias compared with the 30-km rule set of −26.5 ($N_{true} = 60$) and −37.9 ($N_{true} = 70$), respectively. When total annual sightings were restricted to the empirical range ($n < 165$; *Supplemental Materials,* Appendix A), the 16-km-based Chao2 estimates remained conservative, with biases exceeding the 5% benchmark (+3.6 and +4.4) in fewer than 12% ($N_{true} = 60$) and 3% ($N_{true} = 70$) of simulations (Table 1C). Higher numbers of females with cubs may occur in the future, and we expect this would result in more annual sightings. Such a change would be gradual and become apparent in the monitoring data. Under such conditions, it may be necessary to reevaluate whether a shift in the optimal distance criterion is warranted. However, under current sampling regimes our simulations indicate the 16-km distance criterion provides an unbiased estimate of females with cubs, while reducing risk of overestimation. Hence, we recommend implementation of the 16-km distance criterion in the rule set and Chao2 estimation.

Previous efforts to address the same underestimation bias that motivated our study involved the use of a latent multinomial model with mark–resight data of females with cubs (Higgs et al. 2013). The mark–resight estimates

produced unbiased estimates of females with cubs, but precision was low (e.g., 95% interquartile for 2019 mark–resight was 37–114; Haroldson et al. 2020). Based on Peck (2016), who used simulations to investigate the ability of the mark–resight technique to detect changes in trend, the technique was deemed insufficient for effective monitoring of population trend (Haroldson et al. 2017). However, collection of mark–resight data continued because it provided unbiased estimates. Estimates of females with cubs using the 16-km distance criterion (Fig. 8A) align closely with estimates based on mark–resight data (Haroldson et al. 2020). The mark–resight data are independent from the female with cub sightings, thus supporting our conclusion that the 16-km distance criterion improves the accuracy of Chao2 estimates.

Using telemetry data of females with cubs from different European brown bear populations, Ordiz et al. (2007) developed distance–time criteria to identify unique females with cubs and establish minimum estimates. They found that prior to 1 July, 2 observations of females with cubs 30 days apart were unlikely to be of the same family group if >13 km apart in Sweden and >15 km and >7 km apart for released and native bears, respectively, in southern and central Europe. Our finding that 16 km represents a reliable threshold as a distance criterion is in line with those results, particularly considering that our analyses apply to a longer time period (den emergence through 31 Aug). Distance criteria to identify unique females with cubs have been applied other brown bear populations, most notably in Spain (Palomero et al. 1997, 2007). However, females with cubs are not commonly used as the primary sample unit to study and monitor demographics of bear populations; low sightability associated with terrain and vegetation characteristics is often a limiting factor to effective application. Where sightability is not a limitation, our results indicate that accurate minimum counts or estimates of females with cubs can be obtained if optimal distance criteria to assign sightings to unique individuals can be identified. When telemetry data are available, simulations similar to those we present here can be helpful to fine-tune such distance criteria.

Simulations of Chao2 time series indicated that GAMs effectively addressed the limitations of model-averaging for estimation and trend detection of the Yellowstone grizzly bear population. Furthermore, applying GAMs within a more robust statistical framework to assess population trend substantially enhanced inference. This framework not only improved trend detection but also provided early indication of impending change or return to previous state. Together, the *p*d, point estimate, and state variables indicating decline or no decline provide a comprehensive set of tools for interpreting and communicating N_{Chao2} trends. For example, when confidence intervals indicated a slope different from 0, relative differences in *p*d and slope estimates indicated more detailed temporal dynamics, such as whether the current-year estimate is past the peak of a decline and returning to nonsignificant slopes.

As expected, models showed bias associated with periods of change. However, this was not because of an inherent bias of the GAMs, but limitations due to high annual variation in N_{Chao2} estimates and only using the monitoring year of a fitted model for inference (Harris et al. 2007). Statistical advancements alone cannot overcome these limitations and these biases are inherent in monitoring female grizzly bears with cubs from sightings (Brodie and Gibeau 2007). Biases showed predictable patterns relative to fitted slope estimates and duration of declines, which can aid in interpretation of model results.

Posterior inferences of the first derivative slope estimates were responsive to all decline scenarios, and detected change during almost all of 15% and 20% decline scenarios. For the smallest declines (10%), high detection rates (≥90%) were achieved within the first few years after the onset of decline for all but the most gradual decline (15 yrs), which still had an overall detection rate of 0.85. This lower detection rate reflects the challenges of differentiating between high annual variation in N_{Chao2} and gradual declines over a longer time period. However, even when significance is not achieved, the likelihood of mistaking a gradual trend remains low because temporal dynamics between null model simulations of no growth and the most gradual declines were fundamentally different. For example, sustained directional trends in *p*d and first derivative estimates provide clear inference that gradual changes are taking place regardless of statistical significance. These patterns would serve as early indications of significant future change, or at least increasing evidence of a sustained gradual effect. In either case, comparisons of the smoothed N_{Chao2} estimates over relevant time scales would allow evaluation of whether a meaningful effect had taken place. Additionally, temporal patterns in *p*d shifts provide important information for interpreting short-term dynamics, particularly initial shifts from stable periods and attenuation after peaks. These findings highlight the value of trend assessment involving the synthesis of a suite of trend detection metrics, rather than the result of a single metric as currently applied with the use of AIC_c weights. This new framework can be applied to the entire time series of N_{Chao2} estimates for retrospective analysis and future population monitoring. Furthermore, the GAM-based framework and metrics can easily be applied to monitoring programs of other populations

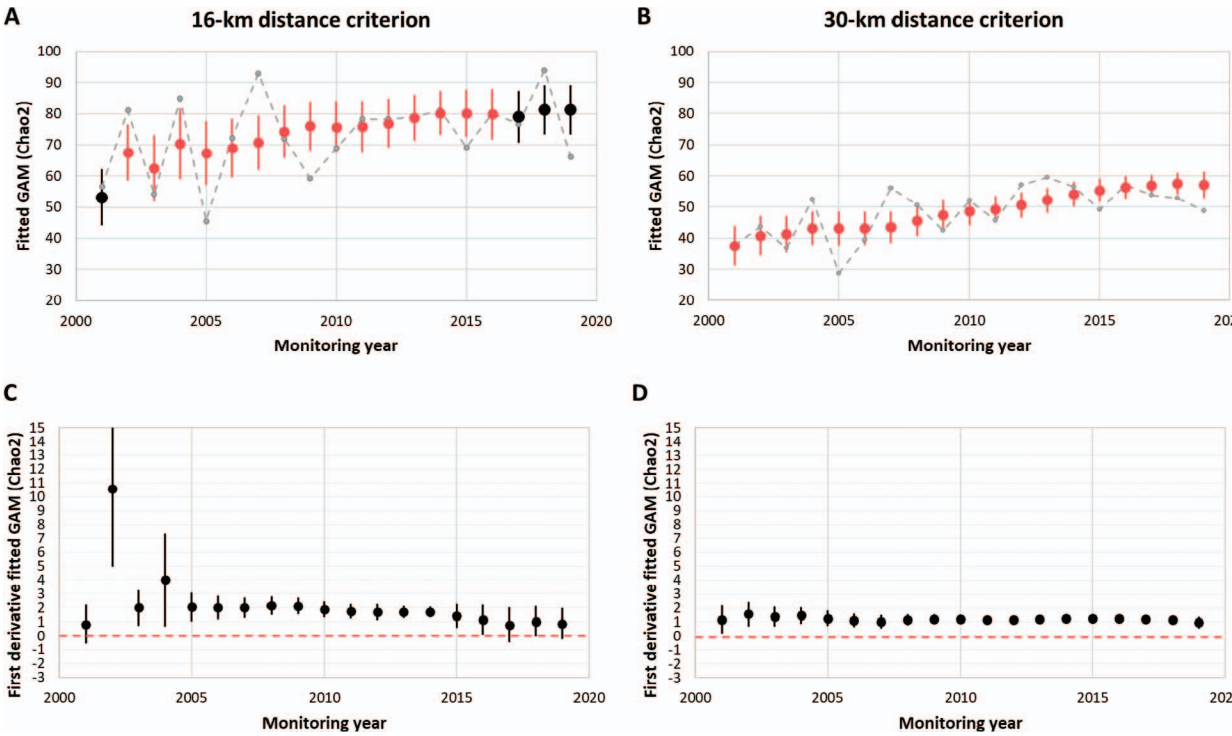

Fig. 8. Estimates of N_{Chao2} derived from the number of unique female grizzly bears (*Ursus arctos*) with cubs in the Greater Yellowstone Ecosystem (Demographic Monitoring Area) during 2001–2019, based on application of the Knight et al. (1995) rule set using 16-km (left panels) and 30-km (right panels) distance criteria. (A) and (B) number of estimated females with cubs using fitted generalized additive model (GAM) estimates of 3-year moving averages of N_{Chao2} estimates ($N_{Chao2_{3y}}$). Each annual estimate is the endpoint of a time series with data starting in 1992, allowing for 10 years of initial data to fit the GAMs (i.e., first estimate is for 2001), and reflecting practical implementation of these techniques. Large circles show the median and black vertical lines show the upper (0.975) and lower (0.025) quantiles for the region containing 95% of posterior simulation values; red circles indicate significant increase based on the first derivative (rate of change) of N_{Chao2}, whereas black circles indicate no significant change based on first derivative values (see panels C and D). Raw annual N_{Chao2} estimates (small gray circles connected by gray dashed line) are shown for reference. (C) and (D) first derivative (rate of change) of N_{Chao2}; black circles indicate median of posterior distribution and vertical black lines show the upper (0.975) and lower (0.025) quantiles for the region containing 95% of posterior simulation values; where vertical black lines intersect the red dashed (zero) line, the rate of change was not significant. Data associated with these graphs are available in *Supplemental Materials* (Appendix E).

or species. Indeed, the value of GAM-based techniques for long-term wildlife monitoring programs were previously recognized (e.g., greater sage-grouse [*Centrocercus urophasianus*]; Fedy and Aldridge 2011) but have not received widespread use among bear biologists. In combination with tools such as the first derivative and probability of decline, these techniques offer powerful tools for trend detection.

As with any analyses involving simulations, there are a number of caveats to these findings. First, we emphasize that our conclusions were based on obtaining unbiased average estimates from simulations with different levels of known females with cubs. The empirical data were equivalent to a single simulation run and, by chance, could represent a time series that diverges from central tendencies. We accounted for this statistical reality in our recommendations of the 16-km distance criterion, but this approach still does not guarantee an absence of overestimation during a single year. This potential for overestimation is one reason that smoothing of these time series is important, such as using GAMs and 3-year moving averages. Second, similar to the limitations that Schwartz et al. (2008) identified regarding their analyses, the sampling frame we generated used data that were not specifically collected to evaluate the distance criteria. For example, we had to combine multiple years of data to create a sampling

frame that allowed adequate "sampling" of annual sightings for the simulations and assume it was reflective of how sightings are collected in any given year. Although this is a reasonable assumption, it may not be entirely accurate. Third, we focused on the distance criterion in the rule set of Knight et al. (1995) because of its overarching implications on the outcome (Schwartz et al. 2008). However, there are other criteria in the rule set that we did not explicitly investigate because of their limited role (e.g., time between sightings, upper Grand Canyon of the Yellowstone River and paved roads as movement barriers). Finally, we emphasized scenarios we deemed relevant to managers and represented realistic changes in population trends. Although actual population scenarios will differ, our purpose was to understand the effectiveness of the proposed population monitoring tools to capture this range of dynamics.

We applied the 16-km distance criterion and the proposed GAM approach for smoothing and trend detection to empirical estimates of females with cubs for 1996–2019 to demonstrate how the findings of this study would be implemented in the monitoring program and enhance interpretation (Fig. 8). During 2019, for example, the median posterior smoothed Chao2 estimate using the 16-km distance criterion was 81.2 (95% CI = 73.3–89.2) females with cubs (Fig. 8A). The median posterior first derivative of the fitted GAM was 0.88 (95% CI = −0.25–2.01); although the 95% confidence interval overlapped 0, there was 94% support for the slope being >0 (i.e., increasing population trend), providing useful inference for managers beyond that of statistical significance. For comparison, we also applied the 30-km distance criterion to these same data and, as expected, estimates of N_{Chao2} were lower (Fig. 8B). The higher values under the 16-km criterion were due to larger estimates of m but also to the Chao2 adjustment representing a greater proportion of the overall N_{Chao2} estimate: the mean Chao2 adjustment for this time period represented 14% of the estimate for the 30-km distance criterion, but 27% of the estimate for the 16-km criterion. Similar to our simulation findings, the greater proportion associated with the Chao2 adjustment is partly a function of an increase in f_1 frequencies relative to f_2 frequencies when shifting the distance criterion from 30 km to 16 km. That a more accurate distance criterion results in a greater proportion of individuals that are only sighted once is not unexpected given the low sighting probabilities of this secretive carnivore. The rate of change of N_{Chao2} estimates based on the first derivative was positive for all years and both distance criteria, with larger estimates and higher growth rates indicated for the 16-km criterion, although estimates were not statistically significant in several years (Fig. 8C and 8D). First derivatives indicated greater sensitivity to change for the 16-km criterion compared with the 30-km criterion.

Management implications

The 16-km distance criterion results in estimates that are greater than previously reported, and our simulations and comparison with independent mark–resight data suggest they more accurately represent the number of females with cubs in the Yellowstone grizzly bear population. For example, the 2019 N_{Chao2} estimate of 81 females with cubs using the 16-km distance criterion and GAMs is 40% greater than the 2019 estimate of 58 females with cubs based on the 30-km distance criterion and model-averaging (Haroldson et al. 2020). Implementation of the 16-km distance criterion combined with use of GAM techniques would affect several population metrics that are derived from the N_{Chao2} estimates and are used to inform management responses (e.g., total population size and uncertainty, population trend, mortality rates). To illustrate, the 2019 estimate for total population size of 737 using the 30-km distance criterion would be equivalent to 1,029 bears under the 16-km criterion. In turn, the higher population estimates result in more accurate depictions of total mortality, and both have implications for evaluation of population metrics specified in the 2016 Conservation Strategy for the Yellowstone grizzly bear population (Yellowstone Ecosystem Subcommittee 2016:36). Implementation of the 16-km criterion and GAMs would require relatively minor changes in the monitoring protocols described in appendices of the 2016 Conservation Strategy (Yellowstone Ecosystem Subcommittee 2016).

The IGBST has ongoing investigations into the merits of an integrated population model, for which annual Chao2-based estimates are important input data, and plans to continue those investigations using the 16-km distance criterion. Finally, we note that the findings from this work emphasize that high inter-annual variation of N_{Chao2} estimates constrains population monitoring. Of course, variation over time is inherent and expected for any wildlife population. However, variation of N_{Chao2} estimates is in part driven by substantial sampling variance. Future monitoring efforts should strive to develop strategies to reduce this source of variation.

Acknowledgments

We appreciate the constructive comments from Associate Editor M. Obbard and 2 anonymous reviewers, which substantially improved the manuscript. We thank

R. Harris, J. Teisberg, and J.D. Clark for their review of a report and drafts of this manuscript as part of the U.S. Geological Survey's Fundamental Science Practices. We are grateful to the partner agencies of the Interagency Grizzly Bear Study Team for their continued support of our research and monitoring efforts: U.S. Geological Survey, National Park Service; U.S. Fish and Wildlife Service; U.S. Forest Service; Wyoming Game and Fish Department; Montana Fish, Wildlife and Parks; Idaho Department of Fish and Game; and the Eastern Shoshone and Northern Arapaho Tribal Fish and Game Department. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S., State, or Tribal Governments.

Literature cited

BJORNLIE, D.D., F.T. VAN MANEN, M.R. EBINGER, M.A. HAROLDSON, D.J. THOMPSON, AND C.M. COSTELLO. 2014. Whitebark pine, population density, and home-range size of grizzly bears in the Greater Yellowstone Ecosystem. PLoS ONE 9(2):e88160.

BLANCHARD, B.M., AND R.R. KNIGHT. 1991. Movements of Yellowstone grizzly bears. Biological Conservation 58: 41–67.

BRODIE, J. F., AND M. L. GIBEAU. 2007. Brown bear population trends from demographic and monitoring-based estimators. Ursus 18:137–144.

CHAO, A. 1989. Estimating population size for sparse data in capture–recapture experiments. Biometrics 45:427–438.

CHERRY, S., G.C. WHITE, K.A. KEATING, M.A. HAROLDSON, AND C.C. SCHWARTZ. 2007. Evaluating estimators of the number of females with cubs-of-the-year in the Yellowstone grizzly bear population. Journal of Agricultural, Biological, and Environmental Statistics 12:195–215.

DEVROYE, L. 1986. Non-uniform random variate generation. Chapter 2: General principles in random variate generation. Springer-Verlag, New York, New York, USA. http://luc.devroye.org/rnbookindex.html. Accessed 15 Sep 2022.

EBERHARDT, L.L., R.A. GARROTT, AND B.L. BECKER. 1999. Using trend indices for endangered species. Marine Mammal Science 15:766–785.

FEDY, B.C., AND C.L. ALDRIDGE. 2011. The importance of within-year repeated counts and the influence of scale on long-term monitoring of sage-grouse. Journal of Wildlife Management 75:1022–1033.

GRANDINI, M., E. BAGLI, AND G. VISANI. 2020. Metrics for multi-class classification: An overview. arXiv:2008.05756v1. https://arxiv.org/pdf/2008.05756.pdf. Accessed 15 Sep 2022.

HAROLDSON, M.A., B.E. KARABENSH, F.T. VAN MANEN, AND D.D. BJORNLIE. 2020. Estimating number of females with cubs. Pages 12–22 in F.T. van Manen, M.A. Haroldson, and B.E. Karabensh, editors. Yellowstone grizzly bear investigations: Annual report of the Interagency Grizzly Bear Study Team, 2019. U.S. Geological Survey, Bozeman, Montana, USA. https://www.sciencebase.gov/catalog/file/get/6266a697d34e76103cce5808?f=__disk__87%2F3c%2Fcb%2F873ccbe0529fe4471c289a2f442420dcfac7059a. Accessed 15 Sep 2022.

———, F.T. VAN MANEN, AND D.D. BJORNLIE. 2017. Estimating number of females with cubs. Pages 15–24 in F.T. van Manen, M.A. Haroldson, and B.E. Karabensh, editors. Yellowstone grizzly bear investigations: Annual report of the Interagency Grizzly Bear Study Team, 2016. U.S. Geological Survey, Bozeman, Montana, USA. https://www.sciencebase.gov/catalog/file/get/6266a697d34e76103cce5808?f=__disk__a1%2F87%2Fe7%2Fa187e75c128cd250f77e0e549d1081f5c071191e. Accessed 15 Sep 2022.

HARRIS, R.B., G.C. WHITE, C.C. SCHWARTZ, AND M.A. HAROLDSON. 2007. Population growth of Yellowstone grizzly bears: Uncertainty and future monitoring. Ursus 18: 168–178.

HIGGS, M.D., W.A. LINK, G.C. WHITE, M.A. HAROLDSON, AND D.D. BJORMLIE. 2013. Insights into the latent multinomial model through mark–resight data on female grizzly bears with cubs-of-the-year. Journal of Agricultural, Biological, and Environmental Statistics 18:556–577.

[IGBST] INTERAGENCY GRIZZLY BEAR STUDY TEAM. 2006. Reassessing methods to estimate population size and sustainable mortality limits for the Yellowstone grizzly bear: Workshop document supplement. U.S. Geological Survey, Northern Rocky Mountain Science Center, Montana State University, Bozeman, Montana, USA.

———. 2012. Updating and evaluating approaches to estimate population size and sustainable mortality limits for grizzly bears in the Greater Yellowstone Ecosystem. Interagency Grizzly Bear Study Team, U.S. Geological Survey, Northern Rocky Mountain Science Center, Bozeman, Montana, USA. https://www.sciencebase.gov/catalog/file/get/6266a697d34e76103cce5808?f=__disk__d5/b2/f9/d5b2f9d6bd27fce053e6b7087826ae8052ca40d7. Accessed 15 Sep 2022.

———. 2021. A reassessment of Chao2 estimates for population monitoring of grizzly bears in the Greater Yellowstone Ecosystem. Interagency Grizzly Bear Study Team, U.S. Geological Survey, Northern Rocky Mountain Science Center, Bozeman, Montana, USA. https://www.sciencebase.gov/catalog/file/get/6266a697d34e76103cce5808?f=__disk__f0/42/5e/f0425e8a8e4ad709c8d1f03846b6549e755299ef. Accessed 15 Sep 2022.

KEATING, K.A., C.C. SCHWARTZ, M.A. HAROLDSON, AND D. MOODY. 2002. Estimating numbers of females with cubs-of-the-year in the Yellowstone grizzly bear population. Ursus 13:161–174.

KIM, Y.J., AND C. GU. 2004. Smoothing spline Gaussian regression: More scalable computation via efficient approximation. Journal of the Royal Statistical Society B 66:337–356.

KNIGHT, R.R., B.M. BLANCHARD, AND L.L. EBERHARDT. 1995. Appraising status of the Yellowstone grizzly bear population by counting females with cubs-of-the-year. Wildlife Society Bulletin 23:245–248.

———, AND L.L. EBERHARDT. 1984. Projected future abundance of the Yellowstone grizzly bear. Journal of Wildlife Management 48:1434–1438.

LEVER, J., M. KRZYWINSKI, AND N. ALTMAN. 2016. Points of significance: Logistic regression. Nature Methods 13: 541–542.

ORDIZ, A., C. RODRÍGUEZ, J. NAVES, A. FERNÁNDEZ, D. HUBER, P. KACZENSKY, A. MERTENS, Y. MERTZANIS, A. MUSTONI, S. PALAZÓN, P.Y. QUENETTE, G. RAUER, AND J.E. SWENSON. 2007. Distance-based criteria to identify minimum number of brown bear females with cubs in Europe. Ursus 18: 158–167

PALOMERO, G., F. BALLESTEROS, C. NORES, J.C. BLANCO, J. HERRERO, AND A. GARCÍA-SERRANO. 2007. Trends in number and distribution of brown bear females with cubs-of-the-year in the Cantabrian Mountains, Spain. Ursus 18:145–157.

———, A. FERNÁNDEZ-GIL, AND J. NAVES. 1997. Reproductive rates of brown bears in the Cantabrian Mountains, Spain. International Conference on Bear Research and Management 9:129–132.

PECK, C.P. 2016. Defining and assessing trend using mark–resight estimates for the number of female grizzly bears with cubs-of-the-year in the Greater Yellowstone Ecosystem. Final report to the Interagency Grizzly Bear Study Team, Department of Mathematical Sciences, Montana State University, Bozeman, Montana, USA.

R CORE TEAM. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

SCHWARTZ, C.C., M.A. HAROLDSON, S. CHERRY, AND K.A. KEATING. 2008. Evaluation of rules to distinguish unique female grizzly bears with cubs in Yellowstone. Journal of Wildlife Management 72:543–554.

SIMPSON, G.L. 2018. Modelling palaeoecological time series using generalised additive models. Frontiers in Ecology and Evolution 6:149.

———. 2019. Gratia: Graceful 'ggplot'-based graphics and other functions for GAMs fitted using 'mgcv'. R package version 0.2-8. https://CRAN.R-project.org/package=gratia. Accessed 15 Sep 2022.

[ESA] U.S. ENDANGERED SPECIES ACT OF 1973, as amended, Pub. L. No. 93-205, 87 Stat. 884 (1973). https://www.govinfo.gov/content/pkg/STATUTE-87/pdf/STATUTE-87-Pg884.pdf. Accessed 9 Dec 2022.

U.S. FISH AND WILDLIFE SERVICE. 2017. Final rule; availability of final Grizzly Bear Recovery Plan Supplement: Revised demographic criteria. 50 CFR 17. Federal Register 82(125):30502–30633. https://www.govinfo.gov/content/pkg/FR-2017-06-30/pdf/2017-13160.pdf. Accessed 15 Sep 2022.

VAN MANEN, F.T., M.R. EBINGER, M.A. HAROLDSON, R.B. HARRIS, M.D HIGGS, S. CHERRY, G.C. WHITE, AND C.C. SCHWARTZ. 2014. Re-evaluation of Yellowstone grizzly bear population dynamics not supported by empirical data: Response to Doak & Cutler. Conservation Letters 7:323–331.

———, M.A. HAROLDSON, D.D. BJORNLIE, M.R. EBINGER, D.J. THOMPSON, C.M. COSTELLO, AND G.C. WHITE. 2016. Density dependence, whitebark pine decline, and changing vital rates of Yellowstone grizzly bears. Journal of Wildlife Management 80:300–313.

WOOD, S.N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99:673–686.

———. 2006. Generalized additive models: An introduction with R. CRC Press, Boca Raton, Florida, USA.

———. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society B 73:3–36.

———. 2017. Generalized additive models: An introduction with R. Second edition. CRC Press, Boca Raton, Florida, USA.

YELLOWSTONE ECOSYSTEM SUBCOMMITTEE. 2016. 2016 Conservation Strategy for the grizzly bears in the Greater Yellowstone Ecosystem. Yellowstone Ecosystem Subcommittee, Interagency Grizzly Bear Committee, Missoula, Montana, USA. https://igbconline.org/document/161216_final-conservation-strategy_signed-pdf. Accessed 15 Sep 2022.

Received: January 21, 2022
Accepted: August 5, 2022
Associate Editor: M. Obbard

Supplemental materials

Appendix A: Linking Empirical and Simulation Data

Appendix B: Supplemental Tables and Figures

Appendix C: Posterior Simulation for Evaluation of GAMs

Appendix D: Summary Statistics of Simulated Data Sets

Appendix E: Numerical Data for Fig. 8