A REASSESSMENT OF CHAO2 ESTIMATES FOR POPULATION MONITORING OF GRIZZLY BEARS IN THE GREATER YELLOWSTONE ECOSYSTEM

INTERAGENCY GRIZZLY BEAR STUDY TEAM

6 April 2021

Photo: Jake Davis

The research described in this report complied with current laws of the United States of America, was conducted in accordance with animal care and use guidelines, and was approved by Institutional Animal Care and Use Committees of the respective member agencies. 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.

Data contained in this report are preliminary or provisional and are subject to revision. They are being provided to meet the need for timely best science. The data have not received final approval by the U.S. Geological Survey and are provided on the condition that neither the U.S. Geological Survey nor the U.S. Government shall be held liable for any damages resulting from the authorized or unauthorized use of the data.

We thank R. Harris and J. Teisberg for their review of this report as part of the U.S. Geological Survey's Fundamental Science Practices. We thank 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.

Cover photo courtesy of Jake Davis (www.revealedinnature.com)

Suggested citation:

Interagency Grizzly Bear Study Team. 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

This report was authored by the members of the

Interagency Grizzly Bear Study Team:

U.S. Geological Survey
Frank T. van Manen, Michael R. Ebinger[1], Mark A. Haroldson

Wyoming Game and Fish Department
Daniel D. Bjornlie, Justin G. Clapp, Daniel J. Thompson

Montana Fish, Wildlife and Parks
Kevin L. Frey, Cecily M. Costello

Idaho Department of Fish and Game
Curtis Hendricks, Jeremy M. Nicholson

National Park Service
Kerry A. Gunther, Katharine R. Wilmot

U.S. Fish and Wildlife Service
Hilary S. Cooley, Jennifer K. Fortin-Noreus, Pat Hnilicka

U.S. Forest Service
Daniel B. Tyers

[1]Currrent affiliation: Montana Fish, Wildlife and Parks

Table of Contents **Page**

EXECUTIVE SUMMARY

The Yellowstone Ecosystem Subcommittee (YES) asked the Interagency Grizzly Bear Study Team (IGBST) to re-assess a technique used in annual population estimation and trend monitoring of grizzly bears in the Greater Yellowstone Ecosystem (GYE). This technique is referred to as the Chao2 approach and estimates the number of females with cubs-of-the-year (hereafter, females with cubs) and, in association with other demographic data, is used by the IGBST to produce annual population estimates. Females with cubs are an easily recognizable population segment, and trends for this reproductive segment of the population are assumed to be representative of trend for the entire population.

The overarching objective of the analyses presented in this report was to provide a more accurate representation of the GYE grizzly bear population using the current methodologies in place. Specifically, we addressed two limitations of the current Chao2 approach: 1) underestimation bias associated with a distance criterion used to differentiate annual sightings of females with cubs into unique individuals and 2) limitations of the model-averaging approach to effectively distinguish among potential future population trajectories (decline, stability, and growth).

The first issue addressed in this report is the underestimation bias associated with the rule set that Knight et al. (1995) developed to differentiate sightings of females with cubs into unique individuals (i.e., unique family groups). The rule set was originally designed to be conservative by reducing the risk of identifying more females with cubs than actually existed, primarily through use of a distance criterion of 30 km to separate sightings of unique females. This approach resulted in an underestimation bias, and previous research demonstrated that this bias increases with increasing number of females with cubs. Using location data from radio-marked females with cubs, we evaluated alternative distance criteria by simulating scenarios with varying numbers of true females with cubs and sightings. Findings from these analyses demonstrate that bias in estimates of females with cubs can be substantially reduced by changing the 30-km distance criterion in the rule set to 16 km, which produced relatively unbiased estimates. Findings also indicate, however, the importance of adaptability with regard to the distance criteria because of the complex relationships and biases among the various parameters involved in estimation of

unique females with cubs. The total number of annual sightings and the true number of females with cubs play particularly important roles. Whereas these analyses remind us that there is no perfect approach to estimating the number of females with cubs from sightings under various scenarios, they provide us with new tools to determine when and how to adapt the monitoring program.

The second issue we were tasked to investigate was the potential for improvement of the technique referred to as model-averaging, which serves to smooth relatively high variation in annual estimates. This technique was chosen by YES as the basis for monitoring the Yellowstone grizzly bear population, as described in the *2016 Conservation Strategy*. This choice was made in part because the technique has been well documented and population estimates derived from counts of females with cubs are conservative. Using simulations of population trends, we demonstrate why the model-averaging technique currently used cannot distinguish between plausible future trend scenarios. As a suitable alternative to model averaging, we propose the use of generalized additive models (GAMs). Using a suite of simulated trend dynamics relevant to management, we demonstrate GAM performance for tracking trends in females with cubs within the context of the annual monitoring program. We demonstrate the ability to not only document directional changes in population trend but also patterns of stabilization or resiliency after such changes. Furthermore, the proposed monitoring framework provides objective measures useful for early detection of directional changes in trend. The new framework is flexible, allowing retrospective analysis of Chao2-based estimates and future applications to time series of other population metrics, such as vital rates.

The aforementioned updates provide us with new tools to determine when and how to adapt the monitoring program. Within the context of current monitoring protocols and effort, and considering the full suite of simulations presented in this report and previous studies, the IGBST plans to incorporate the following changes to the population monitoring protocol: 1) modify the distance criterion, starting with 16 km under current sampling conditions and 2) revise the population monitoring framework using GAMs as the basis for smoothing of annual estimates and detecting trends and changes in trend.

Implementation of the 16-km distance criterion combined with use of GAM techniques would affect some of the population metrics (e.g., annual population size and

uncertainty, population trend, mortality rates) used to inform management responses. A primary consideration is that the 16-km distance criterion results in total population estimates derived from the Chao2 estimates that are greater than those we have reported in the past. This increase is due to a change in the implementation of the technique and more accurately represents the number of females with cubs in the GYE grizzly bear population. Additionally, interpretation of retrospective trend patterns may change due to the combination of a different distance criterion and enhanced trend monitoring based on the GAM approach we present here. Implementation will require relatively minor changes in the monitoring protocols described in Appendices B and C of the *2016 Conservation Strategy*. Finally, we note that the IGBST has ongoing investigations into the merits of an Integrated Population Model (IPM), for which annual Chao2-based estimates are important input data. The IGBST plans to continue those investigations using the 16-km distance criterion to derive Chao2 estimates.

SECTION I – PROBLEM STATEMENT

1. INTRODUCTION

The Yellowstone Ecosystem Subcommittee asked the Interagency Grizzly Bear Study Team (IGBST) to re-assess performance of techniques used in annual estimation and trend monitoring of the Greater Yellowstone Ecosystem (GYE) grizzly bear population. The IGBST uses the nonparametric Chao2 technique to annually estimate the number of females with cubs and derive total population size and monitor trend (Interagency Grizzly Bear Study Team 2012). Females with cubs are easily recognizable and estimates for this reproductive segment of the population are used by the IGBST for inference regarding size and trend of the entire population.

The estimation method involves several major steps. In the first step, sightings of females with cubs from systematic aerial surveys and opportunistic ground sightings are differentiated into a minimum count of unique family groups using a "rule set" that is primarily based on spatial, temporal, and litter size criteria (Knight et al. 1995). In the second step, an estimate of the total number of females with cubs (i.e., including females with cubs that are not sighted) is estimated based on sighting frequencies of unique family groups, using the Chao2 estimator (Chao 1989, Keating et al. 2002, Cherry et al. 2007).

Because annual variation in these estimates of females with cubs (N_{Chao2}) is relatively high due to both sampling and process variance, as a third step the IGBST developed and implemented a technique to address uncertainty in trend estimates and use all available data to provide annual estimates of the number of females with cubs. The chosen technique involved fitting linear and quadratic regressions to the time series of annual N_{Chao2} estimates 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). This approach provided a statistical mechanism to evaluate changes in trajectory for this population segment. Shifts in model weights for the linear and quadratic regressions of N_{Chao2} would be indicative of changes in trend. This approach also resulted in smoothing of annual population estimates, thus enhancing interpretation.

The model-averaged Chao2 technique was chosen by the Yellowstone Ecosystem Subcommittee as the basis for monitoring of the Yellowstone grizzly bear population, as described in the *2016 Conservation Strategy* (Yellowstone Ecosystem Subcommittee 2016). This choice was made in part because the model-averaging technique has been well documented, it has effectively tracked population trends, and population estimates derived from counts of females with cubs are conservative.

We assessed two areas of potential improvement in the current Chao2 estimation approach. First, underestimation bias associated with estimation of unique females with cubs has been documented (Schwartz et al. 2008), but approaches to correct this bias have not been fully investigated. Underestimation bias increases with higher densities of females with cubs and thus constrains our ability to detect changes in the size of this population segment, from which estimates of total population size are derived.

Second, it is desirable to have the capability to track changes in population size over time in various scenarios of population trend, particularly to inform policy-based decisions regarding management objectives and the use of mortality thresholds tied to different population levels. Although averaging of linear and quadratic models proved useful during the robust- to moderate-growth period of population recovery from smaller population sizes in the 1980s and 1990s, it has shown less utility for detecting changes in trend during the period for slower growth to stability occurring since the early 2000s. Model-averaging has shown little power to accurately distinguish among future population scenarios that may involve periods of decline, stability, or growth.

The overarching objective of the analyses presented in this report was to provide a more accurate representation of the GYE grizzly bear population using the current methodologies in place. We first explain the importance of addressing both issues. In Section II, we investigate alternatives to enhance the accuracy of estimates of females with cubs, and in Sections III and IV we explore options to improve the ability to detect changes in trend of those estimates. We apply these alternative techniques to empirical data in Section V and present a synthesis.

2. UNDERESTIMATION BIAS

Counts of distinct females with cubs from aerial and ground sightings have provided an important basis for monitoring the GYE grizzly bear population since 1975. Although they initially relied on subjective criteria, it was recognized that a minimum number of distinct females with cubs could be estimated each year if criteria were stringent enough (Knight et al. 1995). These criteria have evolved since 1975 and a rule set was designed to distinguish unique females with cubs (i.e., family groups) based on annual sightings as described in Knight et al. (1995). The protocol came to be known colloquially as the "Knight rule set" and was designed to reduce the probability of erroneously classifying multiple sightings of a single animal as being from multiple animals. In summary, the Knight et al. (1995) rule set distinguished sightings of unique females with cubs based on 3 primary criteria: 1) distance between sightings, 2) family group descriptions, and 3) time between sightings. Minimum distance for 2 groups to be considered distinct was based on annual ranges, travel barriers, and typical movement patterns. A movement index was calculated using standard diameter of annual ranges of all radiomarked females with cubs that were monitored during May 1–August 31 (Blanchard and Knight 1991). The mean standard diameter for all annual ranges of females with cubs was 15 km (SD = 6.7 km). Knight et al. (1995) estimated the average maximum travel distance as twice the standard diameter, or 30 km, and used this distance to distinguish sightings of unique females with cubs from repeat sightings of the same female. Given that the population was in an early phase of recovery and demographic data were limited, this was a purposely conservative approach. When applicable, family groups observed within 30 km of each other were distinguished by other factors, which are described in more detail in Knight et al. (1995) and Schwartz et al. (2008).

Schwartz et al. (2008) investigated bias due to the different components of the rule set by using simulations that generated "sightings" of females with cubs across the landscape from sampling actual telemetry data. The most relevant finding from that study was that the rule set returned increasingly negative-biased estimates (i.e., underestimates) as simulated number of unique females with cubs increased. With 10 true females with cubs, the rule set was negatively biased by 12%, but this bias increased to 48% for a true

simulated population of 100 females with cubs (see Schwartz et al. 2008 [page 550, Fig. 5]). As density increases, obtaining an unbiased estimate of the true number of females with cubs from sighting data is difficult because it becomes increasingly challenging to distinguish unique animals. In section II, we address the underestimation bias of the Chao2 estimator as identified in Schwartz et al. (2008) using simulations with alternative distance criteria under different scenarios of population size. We propose an alternative criterion that results in relatively unbiased estimates.

3. Model Averaging

Starting in 2007, the IGBST has used model averaging of annual N_{Chao2} estimates as a technique to 1) smooth annual variation and 2) monitor changes in trend. This approach involves annually fitting a linear and quadratic regression model to time series of N_{Chao2} estimates (natural log scale; starting with time period 1983–2006) (Chao 1989, Keating et al. 2002, Cherry et al. 2007) and using Akaike's Information Criterion (AIC_c) weights to evaluate the quantitative support for each model. The conceptual framework for this approach was provided by Harris et al. (2007), describing the use of an information-theoretic approach to detect a deviation in trajectory from the linear population increase documented at that time. This approach was based on the premise that the parameter estimate for the quadratic term would become negative if growth slowed, for example if the population reached carrying capacity (Harris et al. 2007:171). In 2011, the majority of the AIC_c weight shifted from the linear (AIC_c weight = 0.49) to the quadratic model (AIC_c weight = 0.51) for the first time, suggesting a slowing of growth had occurred for the female with cubs segment of the population.

The 2011 shift in AIC_c weight towards the quadratic model triggered a Biology and Monitoring Review by the IGBST (U.S. Fish and Wildlife Service 2007:8), which indicated that the change in trajectory most likely started in the early 2000s (Interagency Grizzly Bear Study Team 2012). Segmental regression of N_{Chao2} estimated for the period 1983–2020 corroborates this interpretation, with a significant slope parameter ($\beta = 0.052$, SE = 0.0096, $P < 0.001$) for the 1983–2001 segment but with less statistical evidence of a trend for the period 2002–2019 ($\beta = 0.014$, SE = 0.0075, $P = 0.075$; Fig. 1). Known-fate

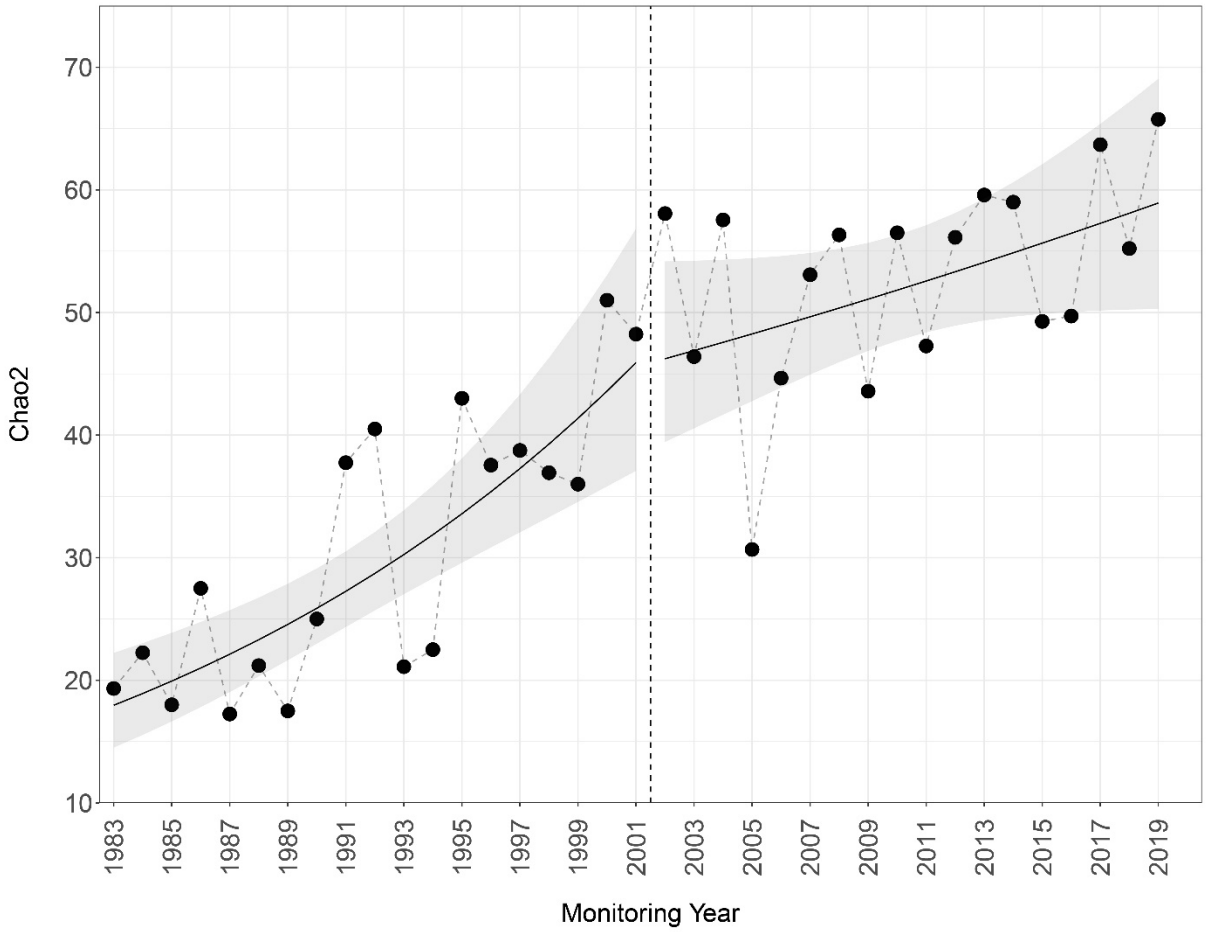

Fig. 1. Annual N_{Chao2} estimates (black circles and dashed line) of the number of female grizzly bears with cubs and fitted linear regression (solid black line; $\log(N_{\text{Chao2}} \sim \text{Year})$ with 95% confidence interval (grey), Greater Yellowstone Ecosystem, 1983–2019. Segmental regression was fitted to the time series (black solid line), and showed growth during 1983–2001 (equivalent to $\lambda = 1.052$), after which the growth rate slowed but remained positive for the period 2002–2019 (equivalent to $\lambda = 1.014$).

monitoring data also provided evidence of slowing of population growth and the potential role of density-dependent effects in the core of the ecosystem (i.e., the Recovery Zone) since the early 2000s (Interagency Grizzly Bear Study Team 2012, van Manen et al. 2016).

Although the AIC_c weighting technique proved useful for detecting a change from robust to modest growth during the period of population recovery, this approach has little power to accurately distinguish among future scenarios (i.e., stability, growth, or decline; Interagency Grizzly Bear Study Team 2005). During IGBST Demographic Workshops held in 2011–2012, it was recognized that additional candidate models needed to be considered in the future to allow for the possibility that the population indeed stabilized (Interagency Grizzly Bear Study Team 2012). Although this concern was noted previously, it has not been documented extensively. Therefore, we provide some background and results of simulation analyses to illustrate why the model-averaging technique will not be effective to detect future changes in population trend derived from counts of females with cubs.

We tested the current model-averaging protocols for future application by simulating N_{Chao2} estimates under various scenarios of increasing, stable, or decreasing estimates (see Appendix A for details). These simulations demonstrate the inability of relative AIC_c weights for the linear and quadratic regressions to distinguish between plausible future trend scenarios. The empirical data show that some AIC_c model weight shifted from the linear to quadratic model starting in 2007, and by 2011, model weight was greater for the quadratic model compared with the linear model (Figs. 2A and 2B). The model simulations starting in 2017 show that eventually all weight would remain on the quadratic model for the remainder of the time series. When we repeated the simulations with $\lambda = 0.978$ (corresponding to 90% annual survival of independent-age females; IGBST, unpublished data) for the period 2019–2041, AIC_c weights for the quadratic model showed an almost identical pattern as the $\lambda = 1.0$ scenario (Figs. 2C and 2D).

The primary conclusion from these simulations is that the AIC_c weighting approach as currently implemented cannot distinguish between population stability and decline. Extending the time period well beyond 2033 for these same simulations similarly showed the majority of the AIC_c model weight would remain with the quadratic model. We note this is not an unexpected result. Outcomes from IGBST demographic workshops in 2011–2012

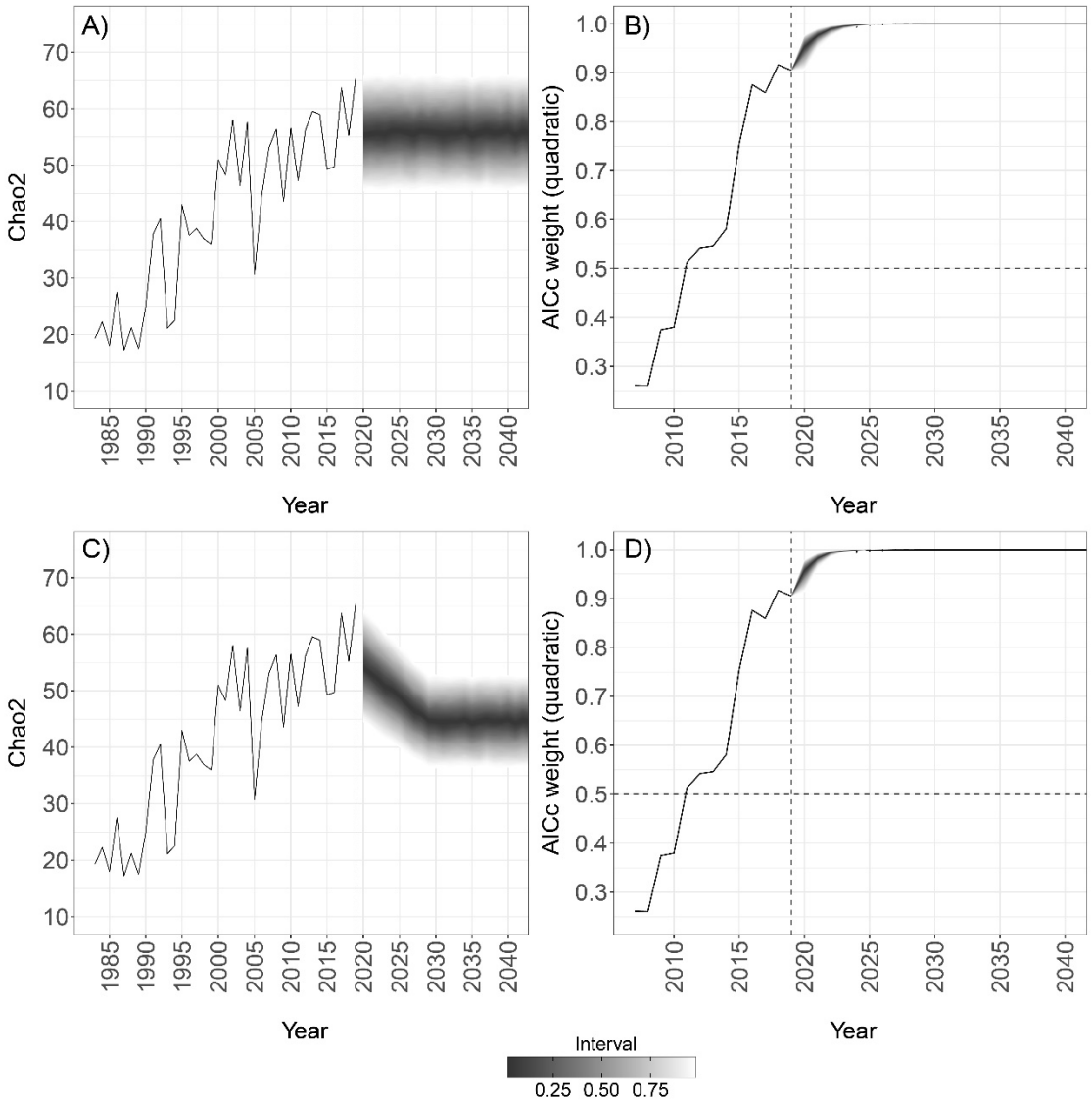

Fig. 2. Empirical (1983–2019) and simulated (2020–2041; right of vertical dashed line) N_{Chao2} estimates of the number of female grizzly bears with cubs in the Greater Yellowstone Ecosystem and associated AIC_c weights for the quadratic regression model. **A)** Simulated (n = 1,000 replicates) scenario with $\lambda = 1.000$. **B)** Annual AIC_c weight for the quadratic model associated with model-averaged values of N_{Chao2} estimates shown in A. **C)** Simulated (n = 1,000 replicates) scenario with $\lambda = 0.978$ over 10 years followed by $\lambda = 1.000$ through the year 2041. **D)** Annual AIC_c weight for the quadratic model associated with model-averaged values of N_{Chao2} estimates shown in C.

indicated that model averaging of linear and quadratic regressions would not accommodate the possibility of the population becoming stable, because the quadratic term imposes a declining trend during later years of the time series (Interagency Grizzly Bear Study Team 2012:28). Workshop participants agreed that alternative approaches would be required in the future. To address this concern, we present the use of generalized additive models, or GAMs, as an alternative. We provide an introduction to GAMs in Section III of this report and detail the alternative approach in Section IV.

SECTION II – CORRECTING UNDERESTIMATION BIAS: ALTERNATE DISTANCE CRITERIA

1. INTRODUCTION

Underestimation bias associated with N_{Chao2} estimates is primarily due to the use of a conservative rule set (Knight et al. 1995) to estimate the number of unique females with cubs and, to a much lesser extent, inherent characteristics of the Chao2 correction (Keating et al. 2002, Cherry et al. 2007). When separating sightings of unique females with cubs, the rule set was constructed to reduce the risk of identifying more individuals than existed. Based on simulations by Schwartz et al. (2008), negative bias increases with population size. The analyses of Higgs et al. (2013), who developed a mark-resight technique to address this underestimation bias, support this. The simulations of Schwartz et al. (2008) showed that the driving factor behind the bias was the distance criterion. Here, we extend the work of Schwartz et al. (2008) to allow updating the distance criterion in the rule set and enhance accuracy of the annual estimate of unique females with cubs.

2. METHODS

Construction of Simulated Datasets

Schwartz et al. (2008) developed a computer algorithm to automate application of the Knight et al. 1995 rule set consistent with the manual implementation by IGBST personnel. They then used location data of radio-marked 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, radio locations of bears from multiple years were overlaid on a map of the ecosystem as if they had all been produced in a single year, and bears were then randomly sampled from this "superpopulation" of observable bears. Because live trapping bears for radio-monitoring purposes is not feasible in some portions of the ecosystem, sets of known radio-monitoring locations were placed on the map to populate areas in which few radio-marked females had been located but were known to be occupied by adult female bears. The result was a rather uniform distribution of bear locations for the simulations to evaluate the Knight et al. 1995 rule set,

with the goal of producing realistic inter-sighting distances and associated dates and times, which are 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), but with slight modifications. First, to simulate observations of females with cubs, we compiled the most up to date location data: aerial telemetry locations (May 1–August 31) and ground sightings (prior to August 31) of radio-collared females with cubs collected annually during 1997–2019. This dataset included 17 years of data not included in Schwartz et al. (2008), allowing for evaluation of potential changes over time. Following the stochastic simulation procedures of Schwartz et al. (2008), we created 1,000 simulated datasets with "true" population sizes of females with cubs at each of 5 different levels of plausible sizes (N_{true} = 50, 60, 70, 80, and 90).

For each replicate, we allowed only one sample year to be chosen for any female with multiple years of data to prevent unrealistic spatial overlap (Schwartz et al. 2008). Similarly, because our location sample spanned more than 20 years, 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 because of unrealistic placement of home ranges directly on top of each other. Therefore, when selecting individuals, we required the activity center of a newly selected candidate female to be at least 1 km from any activity center of a female previously selected while still allowing two simulated females with cubs to have a high degree of spatial overlap.

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}]; range = 1.5–3.2; mean = 2.3). Thus, larger N_{true} resulted in a linearly increasing n. For example, based on 1,000 replicates, N_{true} = 50 and N_{true} = 90 females with cubs resulted in 105–159 and 189–287 simulated sightings, respectively. For simulated sightings, we retained the empirical day, month, time, and coordinate values from the telemetry records.

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. We then simulated changes in litter size caused by cub mortality by applying estimated daily cub survival rates (Interagency Grizzly Bear Study Team 2012) to the number of days between simulated sightings of the same female. Simulated sighting records were censored if complete litter loss occurred, because actual counts do not include females without cubs.

When observed females with cubs had previously been radio-collared, observers verified the telemetry frequency to determine the individuals' identification number. This information is included in the clustering algorithm and increases algorithm accuracy as these individuals cannot be mis-identified (Schwartz et al. 2008). To simulate collared females with cubs, we assigned a pseudo-collar identifier to a proportion of females with cubs in each replicate, based on a random sample from the distribution of empirical radio-monitored females with cubs on an annual basis (1997–2019; range = 3–13 females with cubs/year).

Evaluation of Distance Criteria

To assess the accuracy of different distance criteria, we used the same computer program as Schwartz et al. (2008) to cluster sightings into individuals, varying the distance threshold from 12 to 30 km in 2-km intervals (i.e., 10 distance criteria) and holding all other parameters equivalent, including setting the spatial extent to the area monitored by the IGBST (the Demographic Monitoring Area [49,931 km^2]). This resulted in 50,000 output datasets, with 1,000 simulated datasets of females with cubs for each of the 10 distance criteria and the 5 levels (N_{true} = 50, 60, 70, 80, and 90) of females with cubs (5 population levels × 10 distance criteria × 1,000 replicates each = 50,000).

To evaluate distance criteria at the individual bear level, we compared the unique identifier for true females to the predicted cluster identifier using a confusion matrix and associated classification metrics (Hossin and Sulaiman 2015). We calculated the number of lumping errors (2 or more different females with cubs classified as a single family group) versus splitting errors (a single family group classified as 2 or more different females with cubs) and the overall accuracy incorporating both types of errors. At the replicate level, we

compared the true number of females with cubs to the predicted number of clusters, or unique females with cubs. Because the Chao2 bias-correction uses frequencies of sightings, we also calculated the number of predicted females with cubs sighted once (f_1), twice (f_2), or more than twice ($f_{>2}$) to allow comparison of the true and predicted Chao2 bias-correction factor (Chao 1989, Cherry et al. 2007).

3. RESULTS

Sampling Frame Summary

The pool of data from which simulations were drawn, hereafter referred to as the sampling frame, contained 1,139 verified locations of females with cubs during 154 sampling years, representing 117 unique bears (Fig. 3). The number of locations per unique female varied from 1 to 36 ($\bar{x} = 7.4;\ \sigma = 4.1$). Median distance between locations 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 over time (1997–2019; $\hat{\beta}_{Year} = -0.05$; $P = 0.49$; adjusted $R^2 = -0.004$). The median diameter of the smallest circle encompassing all locations of an individual was 16.4 km, with 86% of individuals' minimum diameter calculations <30 km. Similar to inter-location distances, no trend was evident for this measurement over time (1997–2019; $\hat{\beta}_{Year} = -0.24$; $P = 0.16$; adjusted $R^2 = 0.007$).

Although it was necessary to pool individuals across years to meet sample size requirements for simulations, the empirical data contained 504 unique pairs of locations among different individual females with cubs within the same year. Whereas most of these pairings were located far from each other, approximately 9% (n = 44) had annual location centroids separated by less than the current distance criteria of 30 km (median centroid separation = 22 km). Within this subset of "neighboring" females, 48% had a minimum distance between their telemetry locations and those of their nearest neighbor (median = 4.0 km) that was smaller than the average distance between their own telemetry locations (median = 9.1 km). Although these results are not applicable to the overall population, they are demonstrative of a sizable proportion of females with cubs being located closer in space than the current Knight et al. (1995) 30-km distance criteria within the same season.

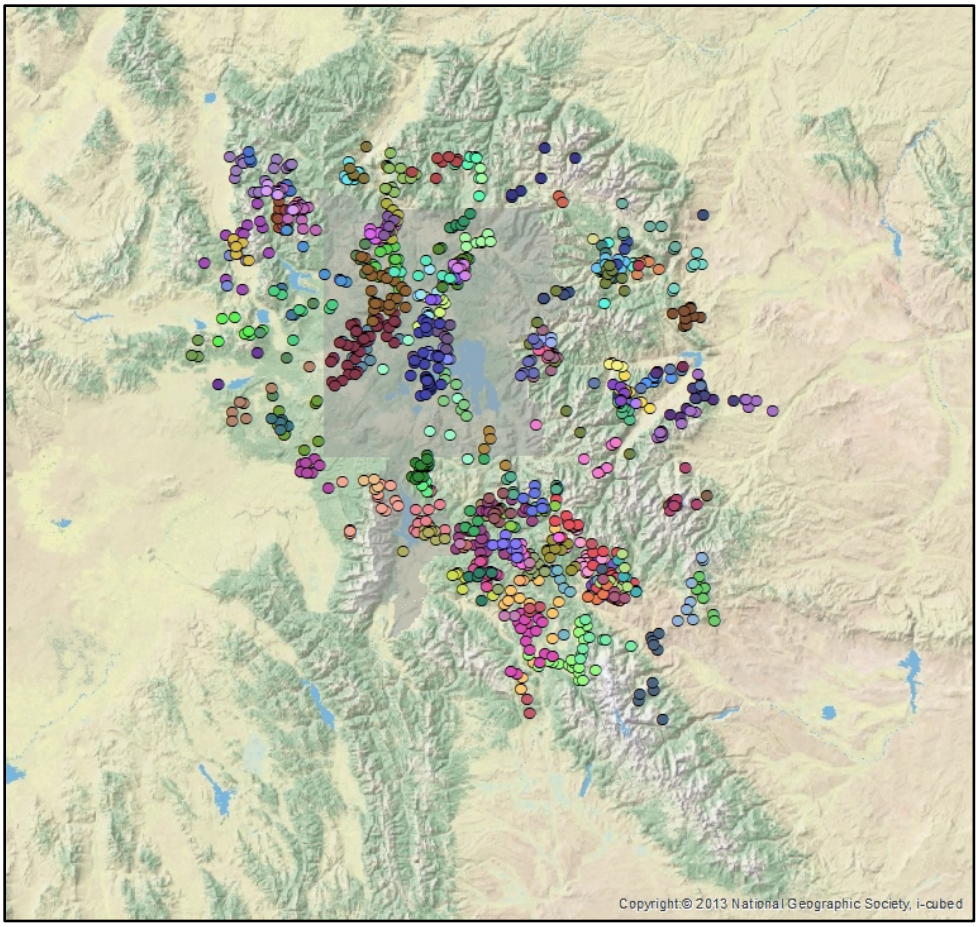

Fig. 3. Simulated sighting data frame containing 1,139 records of 154 sample years (represented by different colors) of female grizzly bears with cubs (sample year = telemetry data for 1 female with cubs for 1 year) representing 117 unique bears in the Greater Yellowstone Ecosystem. Records were composed of aerial telemetry locations (May 1–August 31) and ground sightings (prior to August 31) of females with cubs collected during 1997–2019. Yellowstone and Grand Teton National Parks are shown as grey polygons for spatial reference.

Simulated Datasets Summary

Simulated datasets varied by the number of unique females (N_{true}) and total observations (n; number of random draws). For any given range of total observations, fewer unique females resulted in a larger absolute number of locations (total observations – N_{true}) to be allocated to a smaller number of individuals. Thus, when controlling for total observations, the average and maximum locations per female, or sighting frequencies, decreased with increasing N_{true}. However, because total observations were determined by a multiplier of N_{true}, median sighting frequencies based on the full 1,000 replicates were stable (e.g., $\bar{x}_{50} = 2.27, \sigma_{50} = 1.29;\ \bar{x}_{90} = 2.28, \sigma_{90} = 1.30$; subscripts represent N_{true} level). The average minimum sighting frequency was always 1 (i.e., at least 1 individual was only sighted 1 time) and the average maximum number of times sighted ranged from 18 (N_{true} = 50) to 20 (N_{true} = 90). Sighting frequencies were dominated by f_1 and f_2 frequencies (i.e., observed once or twice, respectively), which on average accounted for ~64% of the total simulated sightings. Proportionally, average f_1 and f_2 frequencies were relatively constant across levels of N_{true} females ($\bar{f}_{1_{50}} = 0.337;\ \bar{f}_{2_{50}} = 0.307; \bar{f}_{1_{90}} = 0.339;\ \bar{f}_{2_{90}} = 0.307$). However, absolute means and standard deviations increased with increasing N_{true}, reflecting the larger sample sizes associated with larger N_{true} ($\bar{f}_{1_{50}} = 16.8, \sigma = 5.4;\ \bar{f}_{2_{50}} = 15.4, \sigma = 4.4;\ \bar{f}_{1_{90}} = 30.5, \sigma = 9.4;\ \bar{f}_{2_{90}} = 27.6, \sigma = 4.9$). Also, 95% of simulated f_1: f_2 ratios were ≤2.0, and within this subset, mean f_1: f_2 ratios were relatively constant across N_{true} levels ($\bar{x} = 1.12, \sigma = 0.01$).

At the intra-bear level, the average distance between an individual's relocations was invariant to the total unique number of females in the simulation ($\bar{x}_{50} = 9.6$ km, $\bar{x}_{90} = 9.5$ km). However, the larger sample sizes associated with higher unique female levels resulted in reduced variation across simulations ($\sigma_{50} = 1.23, \sigma_{90} = 0.84$). Similarly, the minimum diameter circle encompassing all of a female's simulated sightings was relatively constant across the levels of unique females ($\bar{x}_{50} = 16.2$ km, $\bar{x}_{90} = 16.1$ km) but showed decreasing variance with increasing N_{true} levels ($\sigma_{50} = 1.1$ km, $\sigma_{90} = 0.74$ km). This consistency was expected, as the spread of an individual's locations was entirely dependent on random sampling from its own set of points in the sampling frame. Conversely, metrics relating an individual's locations to other bears (i.e., inter-bear) were sensitive to the

number of unique females being simulated. The average distance from each individual's centroid location (i.e., home-range center) to that of their nearest neighbor's centroid decreased with increasing unique females ($\bar{x}_{50} = 13$ km, $\sigma_{50} = 8.6$; $\bar{x}_{90} = 9.2$ km, $\sigma_{90} = 6.4$). 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 ($\bar{x}_{50} = 3.8$ km, $\sigma_{50} = 2.3$; $\bar{x}_{90} = 6.9$ km, $\sigma_{90} = 3.6$). Because the area of the sampling frame (i.e., Demographic Monitoring Area) was fixed across all simulation replicates, these patterns reflect the increasing density of simulated females as N_{true} increases.

Alternative Distance Criteria

Identifying the number of unique females with cubs.—Impacts of alternative distance criteria in the Knight et al. (1995) rule set on classification performance are best understood at two distinct scales. At the broadest scale we considered the task of determining how many unique females with cubs (i.e., distinct family groups) are present in sets of annual observations. This corresponds to the "m" parameter in the Chao2 equation $\left(N_{\text{Chao2}} = m + \frac{(f_1^2 - f_1)}{2(f_2 + 1)}\right)$ and is the largest contributor to the estimate of N_{Chao2}. We refer to this as the "unique ID-level" (classification of individual sightings are not considered) and assessed the presence or absence of unique IDs of females with cubs in the simulation (true IDs) and modeled output (predicted IDs). This approach reduces classification to three distinct outcomes (Table 1): 1) true positives (female IDs correctly predicted); 2) false positives (female IDs erroneously predicted to be present when observations of a single female ID are split into multiple IDs); and 3) false negatives (female IDs that are present but missed because observations of multiple IDs are combined into one ID). Table 1 shows the true number of unique IDs is the sum of those correctly classified (true positives), plus those that were missed (false negatives). Therefore, only when the number of false positives equals false negatives are the correct number of unique IDs predicted. We illustrate these relationships graphically for $N_{true} = 70$ simulations in Fig. 4. These relationships highlight that a simple metric of performance such as accuracy (the fraction of predictions that are true) is not particularly useful for our assessment

Table 1. Outcomes for classification of simulated sightings at the unique ID-level of female grizzly bears with cubs. True positives represent unique IDs that are correctly predicted to be present. False negative events occur when observations of two or more true unique female IDs are erroneously combined into a single predicted female ID, resulting in missed female IDs. False positive events occur when observations of a single female ID are erroneously split into multiple IDs, resulting in IDs for females that do not exist. True negatives are not applicable for this classification assessment because unique IDs cannot be absent from both the simulated and predicted classes. Because the total number of true unique IDs is true positive + false negative cells, the correct predicted number of simulated IDs will occur if the number of false positive equal the number of false negative IDs.

		Simulated ID (true)	
		Present	**Absent**
Predicted ID	**Present**	True positive (correct)	False positive (false female IDs)
	Absent	False negative (missed female IDs)	(not applicable)

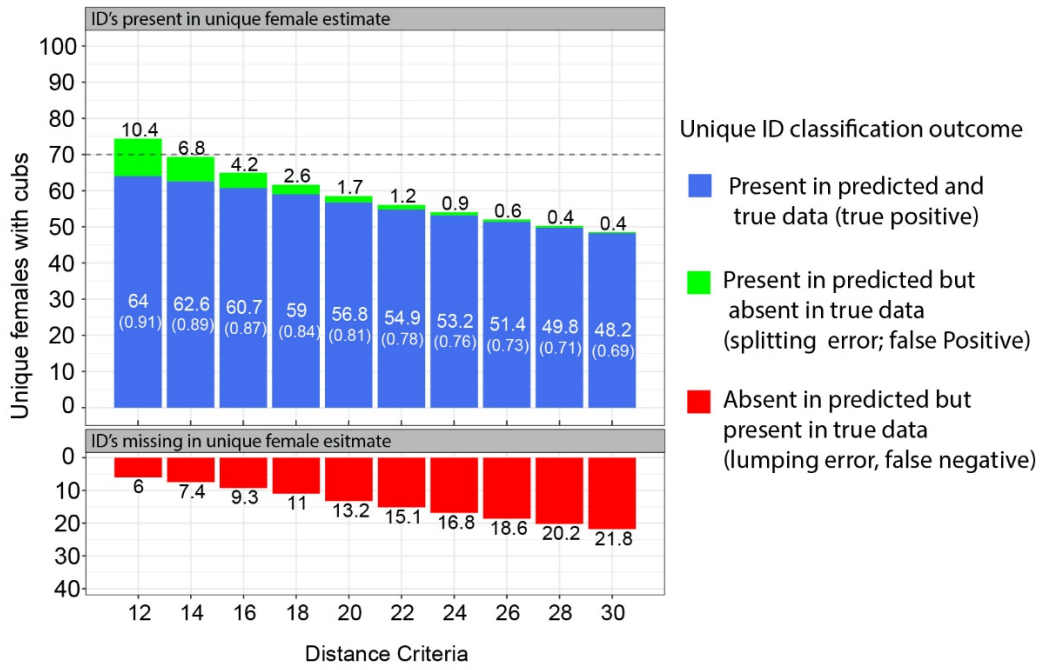

Fig. 4. Accuracy at the unique ID-level for N_{true} of 70 female grizzly bears with cubs based on simulations applying varying distance criteria (x-axis) to the Knight et al. (1995) rule set to differentiate unique females with cubs (y-axis) from simulated sightings. Blue bars represent the mean number of true bear IDs that were correctly identified in the predicted output (n = 1,000 replicates per distance criterion), with numbers in parentheses representing the mean proportion of total unique females correctly identified. Green bars represent mean number of erroneously identified female IDs when observations of a single ID are split into multiple IDs (false positive events). Red bars represent mean number of true unique IDs erroneously combined into a single predicted unique ID, resulting in missed bear IDs (false negative events). The mean total estimated unique females with cubs is represented by the sum of blue and green bars).

because it does not indicate whether false negatives or false positives are more common (Lever et al. 2016).

The concepts of false negative and false positive events (Table 1, Fig. 4) are combined with correct assignments of true positive in the following classification metrics (Ting 2011). At the unique-ID level, precision is the proportion of predicted female IDs that are correct:

$$\text{precision} = \frac{\text{true positive}}{\text{true positive + false positive}},$$

whereas recall is the proportion of true female IDs that are correctly identified:

$$\text{recall} = \frac{\text{true positive}}{\text{true positive + false negative}}.$$

Higher precision and recall are indicative of better classification performance, but generally come at a cost to one another (Table 2). For example, the 30-km distance criterion has high precision ($\bar{x} = 0.99$) because it rarely splits observations of a single female ID into multiple IDs (false positives). However, application of this criterion comes at the cost of relatively low recall ($\bar{x} = 0.69$) because of a higher average probability of erroneously combining true IDs into a single predicted ID, so that true female IDs are missed (false negatives). Conversely, the 12-km distance criterion had the highest overall mean recall ($\bar{x} = 0.92$), because on average it rarely misses a true female ID. However, using this shorter distance criterion comes at the cost of precision ($\bar{x} = 0.87$) because of higher prevalence of predicted female IDs that are not present in the set of true IDs.

The above examples and Table 2 highlight several important points related to evaluating distance criteria. First, as isolated classification metrics, precision and recall are incomplete, and must be interpreted relative to each other. Second, because their formulas only differ by the false negative and false positive terms in the denominator, the conclusion regarding balancing of error types holds true for precision and recall: when they are equal the correct number of simulated bears is predicted. Third, maximizing classification performance does not necessarily result in minimizing bias in *m*, the predicted numbers of unique females with cubs. Instead, the balancing of false negatives and false positives is more important than maximizing the number of true positives. Therefore, comprehensive

Table 2. Precision and recall global means (all levels of N_{true}) and means at the unique ID-level for N_{true} = 50 and N_{true} = 90 female grizzly bears with cubs. Simulations were based on empirical telemetry and ground sighting data of females with cubs from the period 1997–2019, for distance criteria of 12 to 30 km in 2-km steps, and n = 1,000 replicates for each combination of distance criterion and N_{true} level.

	Precision (true positives / all positives)			**Recall (true positives / all correct)**		
Distance criterion (km)	**Global mean**	**Mean N_{true} = 50**	**Mean N_{true} = 90**	**Global mean**	**Mean N_{true} = 50**	**Mean N_{true} = 90**
12	0.87	0.85	0.88	0.92	0.94	0.89
14	0.91	0.89	0.92	0.89	0.92	0.87
16	0.94	0.92	0.95	0.87	0.91	0.83
18	0.96	0.94	0.97	0.84	0.89	0.80
20	0.97	0.96	0.98	0.81	0.87	0.77
22	0.98	0.97	0.98	0.79	0.84	0.73
24	0.98	0.98	0.99	0.76	0.83	0.71
26	0.99	0.99	0.99	0.74	0.80	0.68
28	0.99	0.99	0.99	0.72	0.78	0.66
30	0.99	0.99	0.99	0.69	0.76	0.63

assessment of the distance criteria must integrate precision and recall and at the same time weigh these results against measures of bias. One such measure is the F_β score, which is an aggregative performance metric of precision and recall. It is bounded by 0 and 1, with higher values indicating better classification performance (Lever et al. 2016). The F_β score uses the parameter β to control the balance of precision and recall:

$$F_\beta = (1 + \beta^2)\left(\frac{\text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}\right).$$

As β decreases, precision is given greater weight. Because of the previously outlined positive aspects of balancing precision and recall, we set $\beta = 1$. Mean F_β scores were highest at distance criteria between 14 and 18 km (Fig. 5C) whereas mean absolute error (bias) was minimized at distance criteria between 12 and 16 km (Fig. 5D). The top-performing distance criteria for both measures decreased with increasing N_{true} level (Table 3). The differences between the top performing distance criteria for classification performance versus bias reflects the paradox that optimizing classification performance does not guarantee minimization of absolute bias. However, from a practical standpoint it is important to recognize these differences were always adjacent distance criteria (e.g., 14 versus 16 km) and differences in mean F_β scores were within 1% of each other while difference in mean bias were ≤5% of the N_{true} level (Table 3).

Whereas mean bias does reflect average performance, it is important to consider the variation around these means in terms of avoiding an overestimation bias in real-world applications of a single monitoring year. To reflect the likelihood of overestimation for each of the top-ranking distance criteria in Table 3, we also calculated the proportion of simulations with positive bias, positive bias greater than 5% of N_{true}, and positive bias greater than 10% of N_{true}. The top-ranking distance criteria based on minimizing *m* bias resulted in substantially higher proportions of overestimation than the top-ranking distance criteria based on classification performance (Table 4).

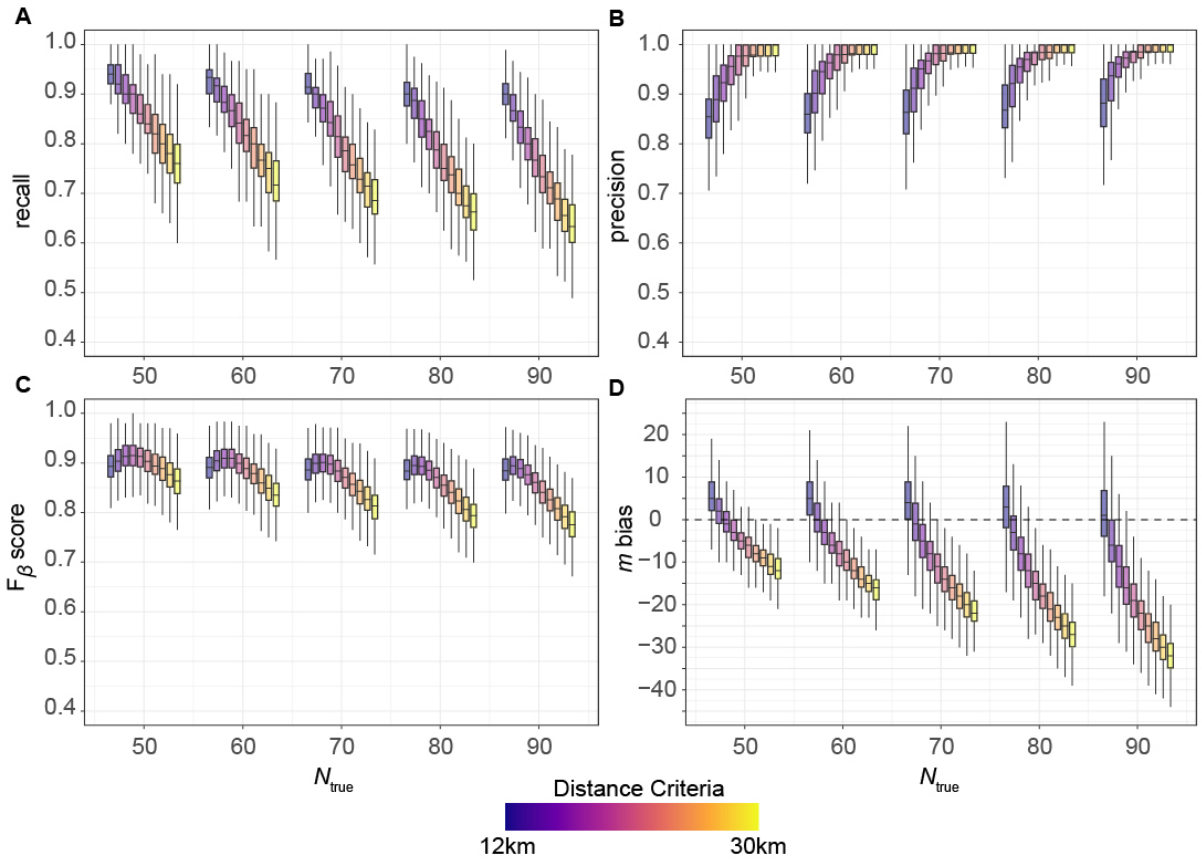

Fig. 5. Classification performance at the unique ID-level shown by **A)** recall, **B)** precision, **C)** F_β score, and **D)** predicted bias in the number of unique females (*m* bias) based on simulations applying varying distance criteria to the Knight et al. (1995) rule set to identify unique female grizzly bears 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 datasets.

Table 3. Top-ranking distance criteria based on mean classification performance (F_β score) and mean unique-ID estimation bias (predicted m – true m) for each N_{true} level based on simulations (n = 1,000 replicates for each N_{true} level) applying varying distance criteria to the Knight et al. (1995) rule set to identify unique female grizzly bears with cubs from sightings. For each top-ranking model the mean F_β score and mean m bias are shown for comparison.

	Maximizing classification performance			**Minimizing abs(estimation bias)**		
N_{true}	**Top-ranking distance criterion (km)**	**Mean F_β score**	**Mean bias in m**	**Top-ranking distance criterion (km)**	**Mean F_β score**	**Mean bias in m**
50	18	0.914	–2.845	16	0.912	–0.612
60	16	0.908	–2.911	14	0.904	0.608
70	16	0.900	–5.067	14	0.899	–0.667
80	14	0.895	–2.865	14	0.895	–2.865
90	14	0.894	–5.150	12	0.887	1.535

Table 4. Top-ranking distance criteria based on mean classification performance (F_β score) and mean unique-ID estimation bias (predicted $m - m_{\text{true}}$) for each N_{true} level based on simulations (n = 1,000 replicates for each N_{true} level) applying varying distance criteria to the Knight et al. (1995) rule set to identify unique female grizzly bears with cubs from sightings. For each top-ranking model, the mean m bias and proportion of simulations greater than 0, greater than 5%, and greater than 10% of N_{true} are shown.

	Maximizing classification performance					**Minimizing abs(estimation bias)**				
N_{true}	**Top-ranking distance criterion (km)**	**Mean bias in m**	**Prop. positive bias**	**Prop. bias >+5% N_{true}**	**Prop. bias >+10% N_{true}**	**Top-ranking distance criterion (km)**	**Mean bias in m**	**Prop. positive bias**	**Prop. bias >+5% N_{true}**	**Prop. bias >+10% N_{true}**
50	18	–2.845	0.18	0.08	0.02	16	–0.612	0.37	0.22	0.07
60	16	–2.911	0.21	0.08	0.02	14	0.608	0.50	0.28	0.11
70	16	–5.067	0.16	0.06	0.01	14	–0.667	0.41	0.25	0.10
80	16[a]	–7.935	0.08	0.02	0	14	–2.865	0.30	0.14	0.04
90	14	–5.150	0.23	0.11	0.02	12	1.535	0.53	0.36	0.17

[a]16 km was used in place of the top model of 14 km to demonstrate difference between distance criteria.

Correctly assigning sightings to IDs.—We also assessed classification performance at the location (sighting) level, reflecting the ability to correctly assign sightings to their respective true IDs. We used a multi-class confusion matrix where each row and column represent a unique ID (rows = predicted, columns = true). For multi-class classification problems where each observation can only be assigned to a single class label, false positives for one class will be false negatives for other classes, and vice-versa. As a result, average precision = recall = F_β score. Therefore, when summarizing classification performance at the sighting-level, mean F_β scores are sufficient.

As expected, mean F_β scores were lower at the sighting level than the unique-ID level; however, patterns related to distance criteria were similar to those at the unique-ID level (Fig. 6). Top-performing distance criteria ranged from 12 to 16 km, with smaller distance criteria performing better with increasing N_{true}. Similar to the unique-ID level, differences between means of the highest-ranked distance criteria and closest competitors were small (Fig. 6).

We focused our assessment at the sighting level on females sighted once (f_1) and twice (f_2), given that sighting frequencies greater than two are not used in the Chao2 equation. Ninety-five percent of simulated and 96% of predicted f_1:f_2 ratios were ≤2.0. Means and standard deviations of f_1:f_2 ratios decreased with increasing distance criteria and N_{true} level (Table 5). Simulated datasets showed a strong positive correlation between the f_1:f_2 ratio and the adjustment component to m $\left(\text{i.e.}, \frac{(f_1^2 - f_1)}{2(f_2 + 1)}\right)$ in the Chao2 equation at all N_{true} levels (Spearman′srank $\overline{r_s} = 0.93, \sigma = 0.01$). Therefore, we used the adjustment component of the Chao2 equation to quantify bias related to sighting frequencies because it is the direct application of f_1 and f_2 counts and interpretation in terms of Chao2 units is intuitive. We note 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. Distance criteria had a strong consistent pattern across all N_{true} levels, with effects over the range of distance criteria (12 versus 30 km) outweighing effects within distance criteria over the range of N_{true} (Fig. 7).

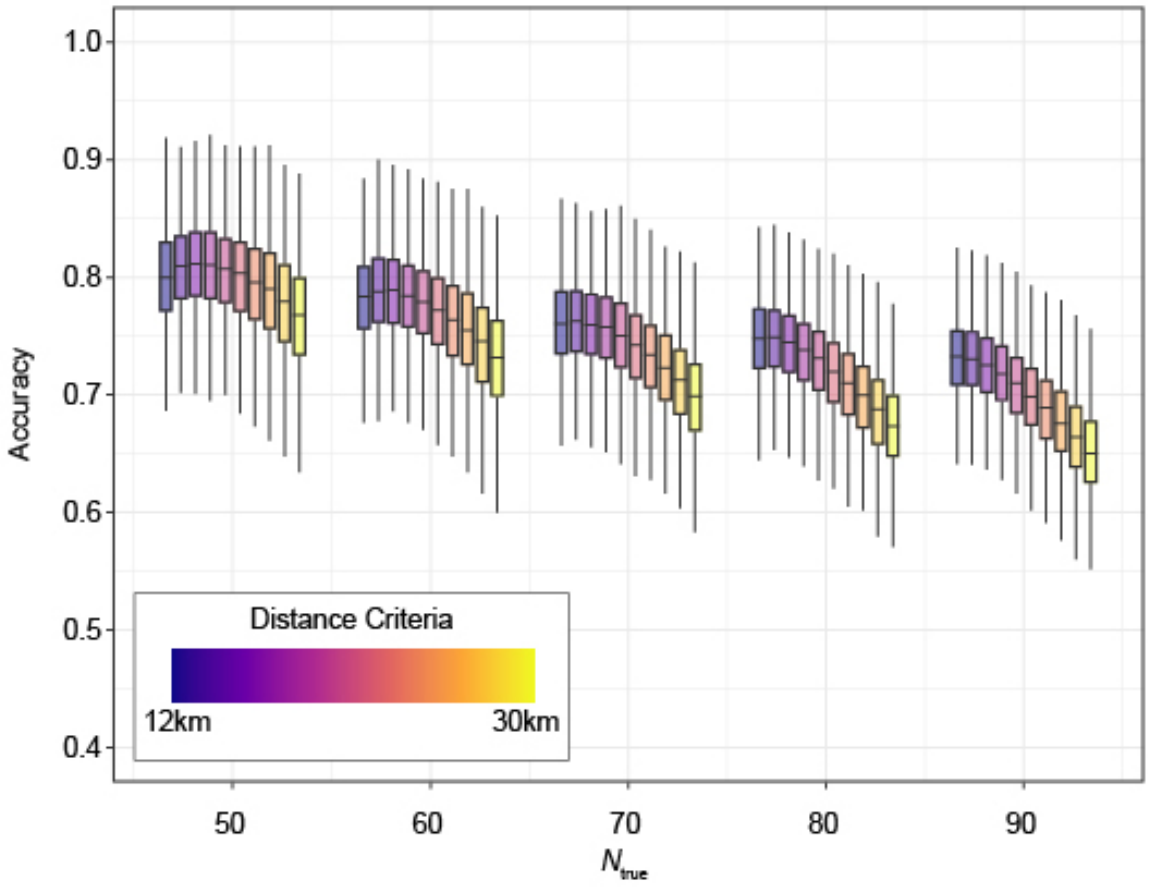

Fig. 6. Boxplots for F_β scores at the sighting based on simulations applying varying distance criteria to the Knight et al. (1995) rule set to identify unique female grizzly bears with cubs from sightings. For each N_{true} level, distance criteria range from 12 to 30 km in 2-km steps (indicated by color gradient and arranged from left to right). Each boxplot summarizes n = 1,000 simulated datasets based on micro-averaged F_β scores.

Table 5. Mean ($\bar{x}$) and standard deviation (σ) of the predicted f_1:f_2 ratios for N_{true} levels of 50, 70, and 90 based on simulations (n = 1,000 replicates for each combination of N_{true} level and distance criterion) applying varying distance criteria to the Knight et al. (1995) rule set to identify unique female grizzly bears with cubs from sightings. True f_1:f_2 mean ratios were 1.13 (SD = 0.39), 1.12 (SD = 0.36), and 1.12 (SD = 0.33) for N_{true} of 50, 70, and 90 respectively. Results are shown for f_1:f_2 ratios <2 (96% of data).

		Distance criterion (km)									
N_{true}	Parameter	12	14	16	18	20	22	24	26	28	30
50	$\bar{x}$	1.49	1.33	1.22	1.15	1.10	1.04	1.01	0.97	0.98	0.95
	σ	0.50	0.48	0.47	0.48	0.49	0.46	0.47	0.45	0.49	0.48
70	$\bar{x}$	1.40	1.25	1.12	1.06	1.00	0.94	0.91	0.89	0.89	0.88
	σ	0.45	0.42	0.39	0.40	0.41	0.40	0.40	0.43	0.47	0.48
90	$\bar{x}$	1.34	1.20	1.07	1.00	0.96	0.90	0.88	0.85	0.85	0.82
	σ	0.38	0.36	0.33	0.35	0.36	0.34	0.37	0.37	0.42	0.38

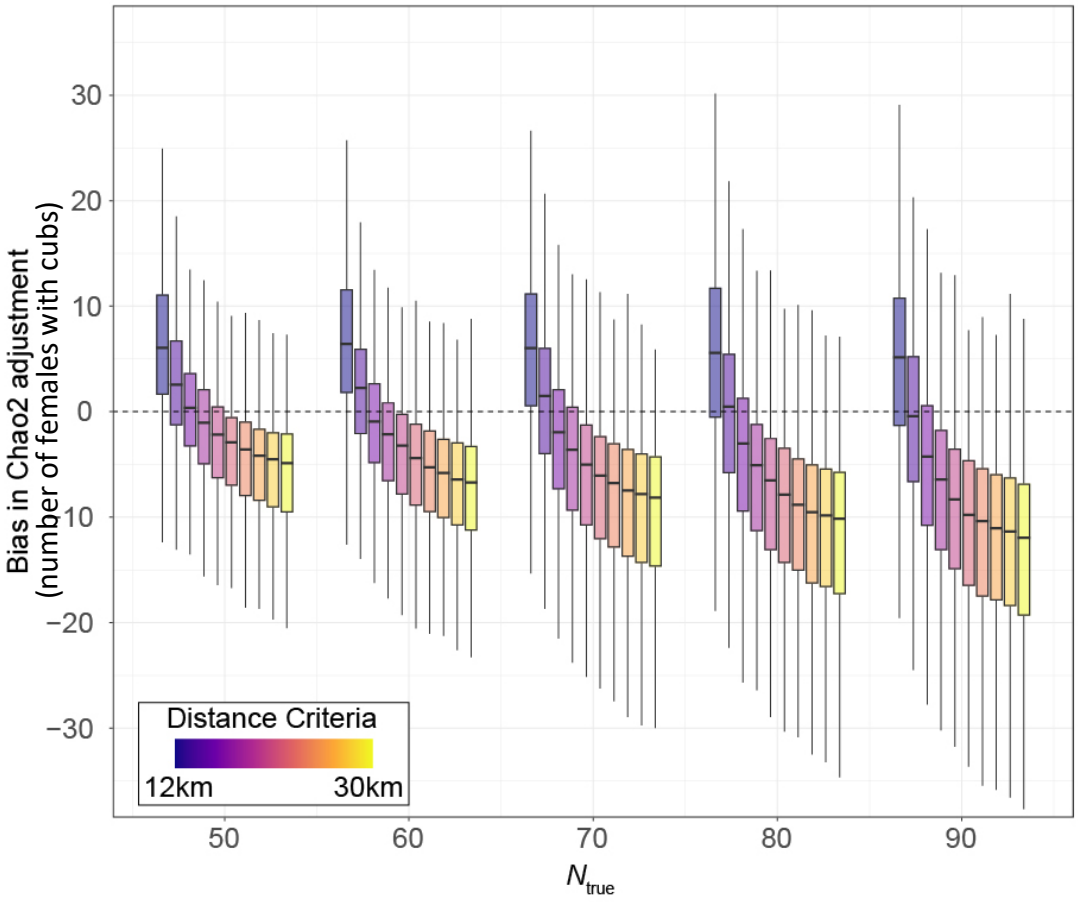

Fig. 7. Boxplots of bias (predicted—simulated; expressed as number of females with cubs) in Chao2 adjustment $\left(\frac{(f_1^2 - f_1)}{2(f_2 + 1)}\right)$ at the sighting level for all distance criteria and N_{true} levels based on simulations applying varying distance criteria to the Knight et al. (1995) rule set to identify unique female grizzly bears with cubs from sightings. For each N_{true} level, the 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 datasets.

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. 8). 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- to 16-km range best minimized bias of both m and the Chao2 adjustment (Fig. 8). The range of bias in the Chao2 adjustment for relatively 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 to 16 km, even when m is 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 is low, individual replicates varied widely, from −10 to +13 for f_1 and −17 to 14 for f_2 (Fig. 9). These biases are negatively correlated, and overestimation of f_1 tends to correspond to underestimates of f_2 and a positive bias in the adjustment component of N_{Chao2}, whereas underestimation of f_1 tends to correspond to overestimation of f_2 and a negative bias in the adjustment component of N_{Chao2} (Fig. 9).

Despite the challenges of simultaneously reducing bias of m, f_1, and f_2 estimates at smaller distance criteria, the improvements relative to the 30-km rule set are substantial and important to recognize. 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 -0.612, -0.082, and -0.348 respectively, but for the 30-km criterion were -11.62, -8.67, and -5.25, respectively.

These results confirm those of Schwartz et al. (2008) regarding the sensitivity of bias to changes in overall density of sightings (i.e., sightings/true females, or n/N ratios). Negative bias in both m and the Chao2 adjustment decreased with increasing n/N ratios (Fig. 10). The proportion of the overall Chao2 estimate accounted for by the Chao2 adjustment (versus m) increased with decreasing distance criteria at all levels of N_{true} (Table 6). This was expected, as reducing distance criteria, on average increases both f_1:f_2 and the proportion of total sightings that are either f_1 or f_2 (Table 6). Post-hoc regressions of Chao2 adjustments ~f_1:f_2 ratio for all distance criteria × N_{true} combinations confirm this,

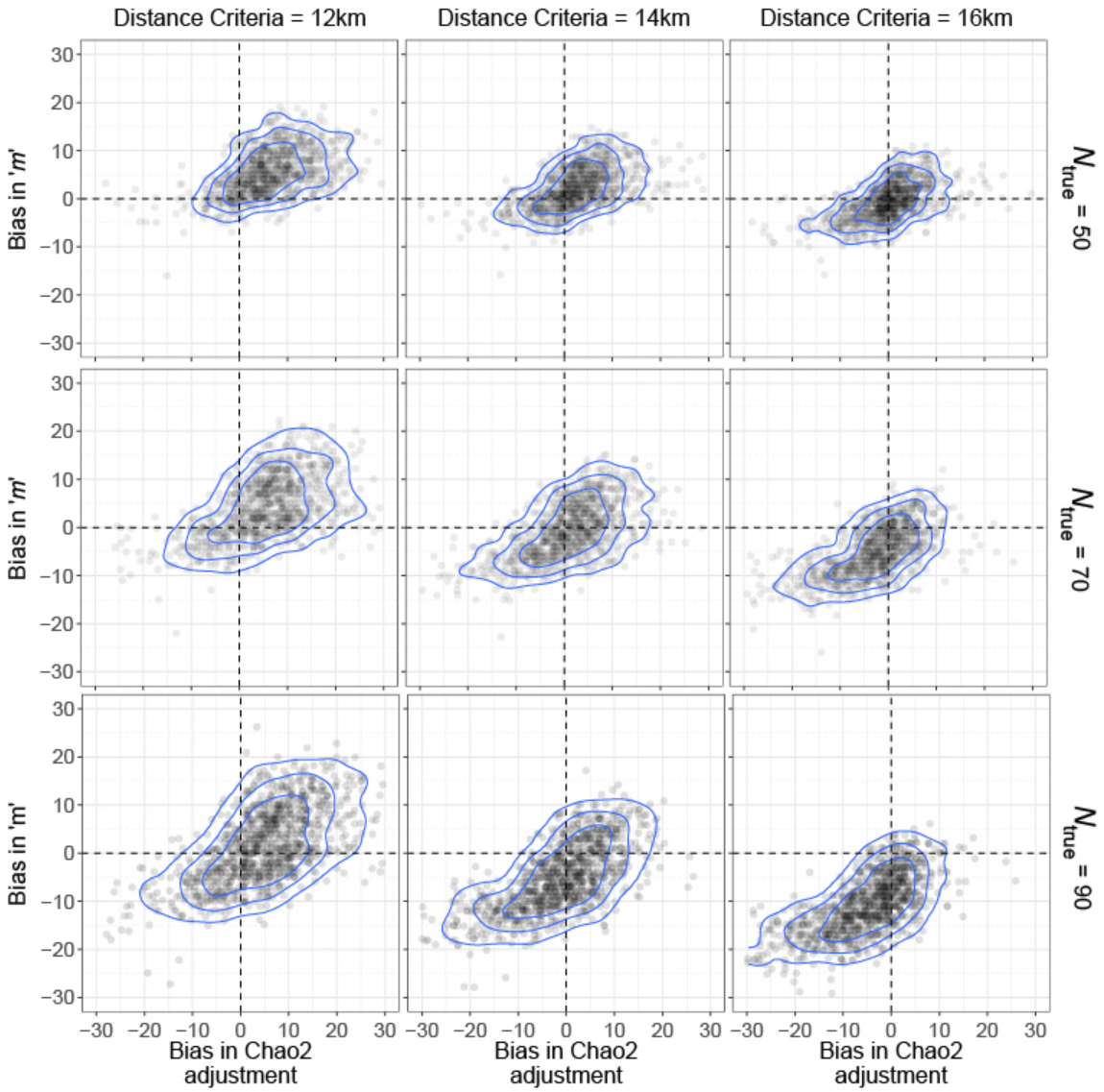

Fig. 8. Relationships between bias (expressed as number of unique females with cubs) in the parameter *m* and bias in the Chao2 adjustment (i.e., $N_{\text{Chao2}} - m$) of the estimator based on simulations (n = 1,000 replicates for each combination of distance criterion and N_{true} level), applying varying distance criteria to the Knight et al. (1995) rule set to identify unique female grizzly bears 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.

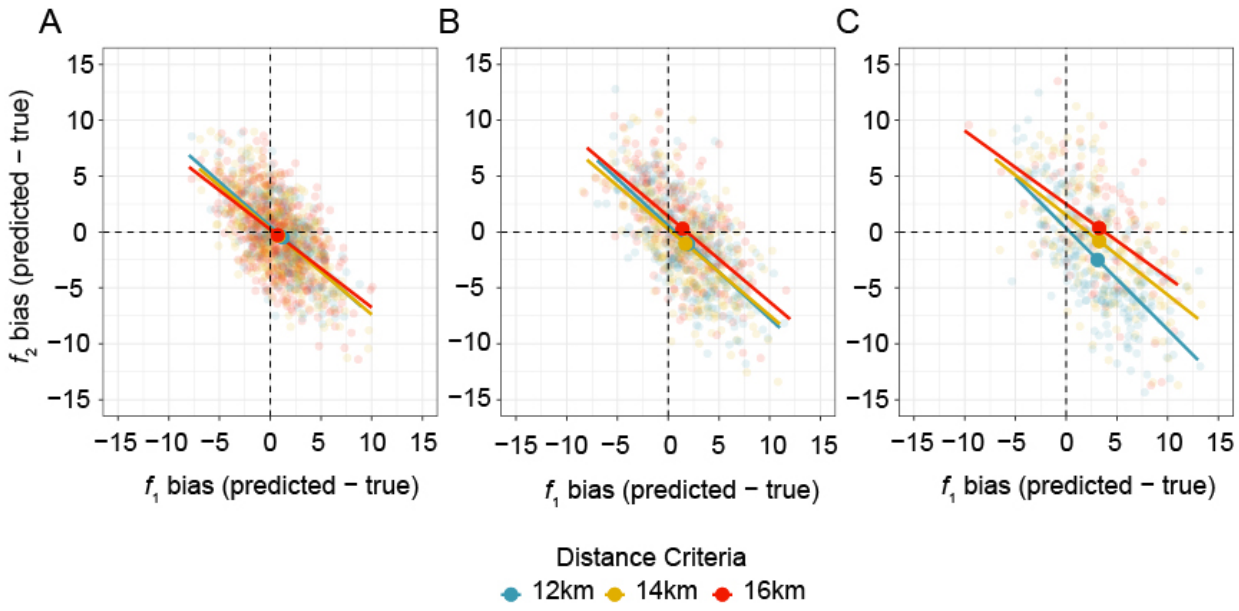

Fig. 9. Bias (predicted — true; expressed as number of females with cubs) in the number of single (f_1) and double (f_2) sightings based on simulations (n = 1,000 replicates for each distance criterion) applying the Knight et al. (1995) rule set to identify unique female grizzly bears with cubs, showing relationships between over- and underprediction of f_1 and f_2 when bias in m was +/- 2 of N_{true}, or true m. Colors represent 3 different distance criteria (12, 14, and 16 km). In each graph, the distance of the small, transparent circles to the intersection of the horizontal and vertical dashed lines (0, 0; no bias) represents the level of bias in terms of the number of f_1 (distance along horizontal axis) and f_2 (distance along vertical axis) sightings at the replicate level. Solid circles and trend lines show average relationships. **A)** Simulations using N_{true} = 50. **B)** Simulations using N_{true} = 70, and **C)** Simulations using N_{true} = 90.

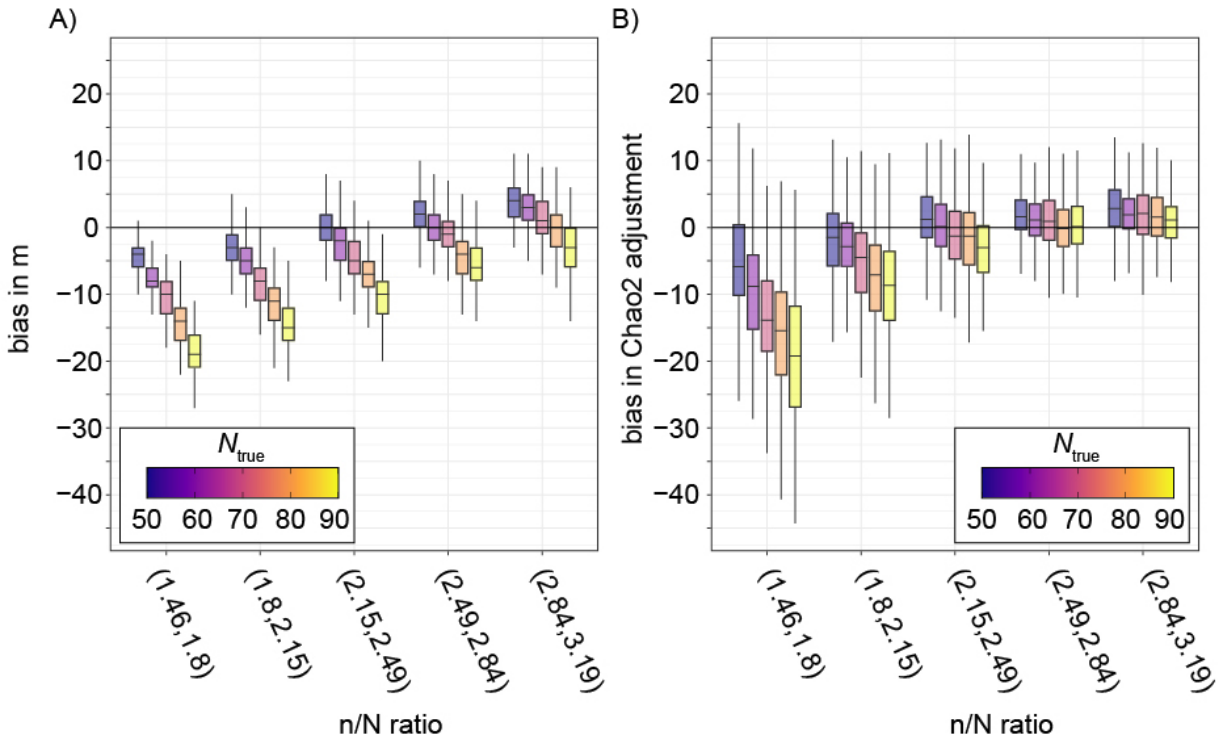

Fig. 10. Bias (predicted − true; expressed as number of unique females with cubs) in predicted *m* **(A)** and Chao2 adjustment to *m* **(B)** for distance criterion 16 km and N_{true} levels of 50, 60, 70, 80, and 90 based on simulations applying the Knight et al. (1995) rule set to identify unique female grizzly bears with cubs. The *x*-axis reflects binned values for *n*/*N* ratio with ranges of bins shown in parentheses. Within each bin, N_{true} levels increase from left to right (color gradient). Each boxplot summarizes *n* = 1,000 simulated datasets.

Table 6. Mean proportion of Chao2 estimate contained in the Chao2 adjustment (i.e., Chao2 adjustment/Chao2 estimate) and mean proportion of total sightings that are either f_1 or f_2 sightings for distance criteria of 12 to 30 km and N_{true} levels of 50, 70, and 90 based on simulations (n = 1,000 replicates for each combination of distance criterion and N_{true} level) applying the Knight et al. (1995) rule set to identify unique female grizzly bears with cubs from sightings.

	N_{true} = 50		N_{true} = 70		N_{true} = 90	
Distance criterion (km)	**Chao2 adjustment/ Chao2 estimate**	**Proportion f_1 or f_2**	**Chao2 adjustment/ Chao2 estimate**	**Proportion f_1 or f_2**	**Chao2 adjustment/ Chao2 estimate**	**Proportion f_1 or f_2**
12	0.22	0.72	0.20	0.70	0.19	0.67
14	0.18	0.68	0.17	0.65	0.16	0.62
16	0.16	0.65	0.14	0.61	0.13	0.58
18	0.14	0.62	0.12	0.57	0.11	0.54
20	0.12	0.59	0.11	0.54	0.10	0.50
22	0.11	0.56	0.09	0.51	0.08	0.48
24	0.10	0.54	0.08	0.49	0.08	0.45
26	0.09	0.52	0.08	0.46	0.07	0.43
28	0.09	0.50	0.07	0.44	0.06	0.41
30	0.08	0.48	0.07	0.42	0.06	0.39

with the predicted f_1:f_2 ratio as the main driver of variation in Chao2 adjustment (min = $r^2_{N_{\text{true = 50; 30 km}}} = 0.79$; max = $r^2_{N_{\text{true = 90; 12 km}}} = 0.92$).

4. DISCUSSION

In this section of the report, our primary task was to re-assess the Knight et al. (1995) rule set used to identify unique females with cubs, which provides the basis for the Chao2 estimation technique used to monitor the GYE grizzly bear population. Specifically, to increase the accuracy of the estimation approach, we examined alternative distance criteria used to differentiate annual sightings of females with cubs into unique individuals. Using location data from radio-marked females with cubs, we evaluated alternative distance criteria by simulating scenarios with varying numbers of true females with cubs and sightings. These 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 observations (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. Selecting a single optimal distance criterion from within the 12–16 km range requires a number of considerations, as outlined below.

To select the optimal distance criterion, we concentrated on reducing underestimation bias while limiting the risk of overestimation. We focused our insights on N_{true} levels of 60 and 70 unique females with cubs because empirical estimates of *m* (30-km criterion) and total observations for the period 2001–2019, when linked to simulation results, suggest this range is most relevant to contemporary conditions in the GYE (see Appendix B). Simulations show the 16-km distance criteria is relatively unbiased 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 more than 5% in only 3% (N_{true} = 60) and 0% (N_{true} = 70) of simulations. Although the 14-km distance criterion is less biased on average, it has higher proportions of simulations with >5% bias (Table 7A).

The Chao2 adjustment to *m* was also relatively unbiased at the 16-km distance criterion, under the assumption that the true classification (simulated sightings) produced

Table 7. Top-ranking distance criteria (12–16 km) and baseline 30-km distance criterion bias components for N_{true} = 60 and 70 female grizzly bears with cubs, and total observations within the empirical range of $n \leq 165$. Results are 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 females with cubs from sightings. **A)** m bias, 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
16	60	−3.9	0.03	0
	70	−8.3	0	0
30	60	−16.9	0	0
	70	−23.4	0	0

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	0
	70	−14.5	0	0

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
30	60	−26.5	0	0
	70	−37.9	0	0

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 7B).

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), 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 7C). We reiterate that the use of "known N_{Chao2}" relates to calculations using simulated m, f_1, and f_2 values, and not an actual measurement of the number of females with cubs in the population that are not seen, which we could not simulate.

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 a relatively unbiased estimate of females with cubs while reducing risk of overestimation. To better understand the implications of an alternate distance criterion, we applied the 30- and 16-km criteria to the empirical data of sightings of females with cubs for the period 1995–2019 in section V (Empirical Application).

SECTION III – GENERALIZED ADDITIVE MODELS AS AN ALTERNATIVE TO MODEL AVERAGING

1. GENERALIZED ADDITIVE MODELS

Generalized Additive Models (hereafter GAMs; Hastie and Tibshirani 1986, 1990; Wood 2017) are semi-parametric extensions of generalized linear models (GLMs; McCullagh and Nelder 1989) and are one of the most popular and powerful modeling tools currently in use by ecologists (Pedersen et al. 2019). GAMs are often described as "data-driven" rather than "model-driven" because data determine the relationship between response and predictor variables rather than an assumed functional relationship (Guisan et al. 2002). Whereas linear models relate the mean of the response to a linear combination of predictor variables, GAMs describe these relationships via 'smoothed' functions, whose shape need not be known a priori. As extensions of GLMs, GAMs have the desirable properties of linear models, but are more flexible, easier to interpret, and excel when relationships are non-linear and non-monotonic (Guisan et al. 2002, Shalizi 2019).

Here, we provide brief background information on GAMs relevant to the applied setting of a wildlife monitoring program using annual count data (i.e., female grizzly bears with cubs in the GYE). Trend analyses for the GYE grizzly bear population do not include analyses of multiple covariates, and therefore we limit our discussion of GAMs to the univariate setting (i.e., count ~ year), emphasizing ease of use and interpretation compared with current trend monitoring. We use example data to explain GAMs, the rationale of the proposed method, and to illustrate how GAMs can enhance inference. We designed these datasets to demonstrate potential scenarios of population change relevant to the GYE grizzly bear monitoring program, so they are hypothetical and should not be used for biological interpretations.

2. GENERALIZED ADDITIVE MODELS AS EXTENSIONS OF LINEAR MODELS

Application of GAMs is most easily understood in terms of differences and similarities compared with linear models and the current smoothing approach used by the IGBST in the monitoring program, i.e., model averaging. A linear regression model to

estimate a trend of annual counts (count_t) for each year (year_t) with t = 1, 2, … , current monitoring year is:

$$\text{count}_t = \beta_0 + \beta_1 \text{year}_t + \varepsilon_t,$$

where β_0 is a constant, or intercept, and β_1 is the rate of change, or slope, of the fitted line given the data. The GAM version of the above linear model is

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

where $f(\text{year}_t)$ is a smoothing function replacing the linear fixed effect ($\beta_1 \text{year}_t$) from the previous equation (Simpson 2018). The main challenge when fitting GAMs is achieving the optimal degree of smoothness of $f()$ to capture the underlying trend without being overly sensitive to underlying noise (Jones and Almond 1992). This optimal degree can be achieved by incorporating a "wiggliness penalty" into the objective function that is minimized during model fitting; we refer to Wood (2017) and Simpson (2018) for further details. When applied properly, this penalty serves to avoid model overfitting while still allowing the smoother to respond nonlinearly to changes in trend. In fact, the penalization process can reduce complexity of the GAM to a linear trend if the data support it. Therefore, a GAM can estimate a linear response without making the explicit assumption of linearity a priori. When penalization reduces the GAM to a linear model, it can match the results of model averaging (i.e., when model support is dominated by the linear model; Fig. 11A). Similarly, when data are increasingly supported by the quadratic model, and thus show a more curvilinear response, the fit of the $f(\text{year}_t)$ term will be practically identical to the fit based on model averaging. Thus, for time series that can be adequately described by a combination of linear and quadratic regressions, such as the N_{Chao2} data, the smoothed responses of GAMs pose little departure from the fitted results of model averaging (Figs. 11A and 11B). However, as nonlinear complexity of a time series increases, the ability of the smoother term $f(\text{year}_t)$ to accommodate this complexity gives GAMs a substantial advantage over model averaging (Figs. 11C and 11D).

Whereas model averaging could incorporate higher-order polynomials into the competitive model suite, such an approach would introduce additional complications, like which order of polynomials to include (Bolker 2008, Simpson 2018). It would also increase difficulty of interpreting AIC_c model weights as the number of candidate models increases

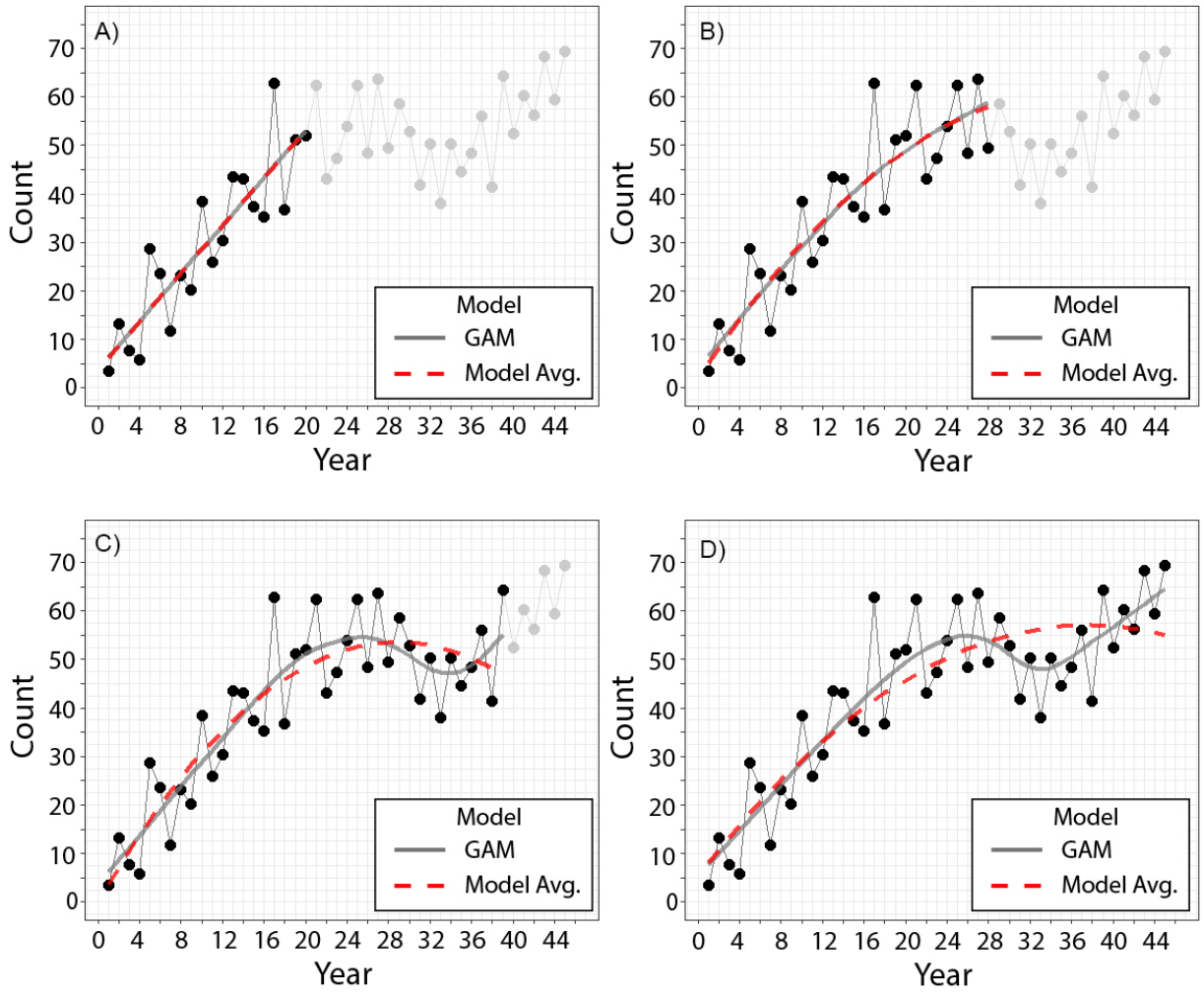

Fig. 11. Theoretical time series of count data and fitted smooths of female grizzly bears with cubs at **A)** monitoring year 20, **B)** monitoring year 28, **C)** monitoring year 39, and **D)** monitoring year 45. Black circles show data from year 1 to the monitoring year and light grey circles show future values (not yet observed) for each panel. The grey solid line is a fitted generalized additive model (GAM) and the dashed red line is the fitted model-averaged (linear and quadratic models) smooth (AIC_c weight for quadratic = 0.18 [A], 0.76 [B], 1.0 [C], 1.0 [D]). Note the near identical fits of the GAM and model-averaged approach in panels A and B, but increasing divergence between the two approaches in panels C and D as nonlinear complexity increases.

beyond two. GAMs do not suffer from these issues and can handle response patterns of varying complexity, ranging from simple linear trends to highly nonlinear patterns within a single model structure. This feature is important in wildlife management applications because trajectories of animal populations can take many different forms and are often triggers for changes in management strategies. Therefore, the flexible model structure of GAMs provides the basis for a more robust population monitoring system and science-based decision-making.

3. Model Interpretation

The main difference in model interpretation between simple linear models and their GAM equivalent is shifting from a single parameter value to a continuous function that cannot be expressed as a single number. Because $f(\text{year}_t)$ is a function, it cannot be expressed in the convenient way of a constant slope estimate (e.g., β_1). Therefore, GAMs rely to a large extent on visual interpretation of the fitted smooth and its associated uncertainty, which are easily expressed in terms of the partial response function $f()$ or predicted values for the fitted response. From a practical application perspective, a fitted GAM is not different from that of model averaging: both involve interpreting a best-fit prediction and a $1 - \alpha$ confidence interval as a measure of uncertainty. Furthermore, other aspects of model inference, such as relative changes in trend, are simplified under the GAM framework and, unlike model-averaging, GAMs provide model-level summaries including estimated degree of freedom (an index of smoother complexity), significance test for model terms, adjusted R^2, deviance explained, and standardized model diagnostics.

First Derivatives

Considering the example time series shown in Fig. 11A (year_t = 20), where the GLM ($\text{count}_t = \beta_0 + \beta_1\text{year}_t + \varepsilon_t$) and GAM ($\text{count}_t = \beta_0 + f(\text{year}_t) + \varepsilon_t$) have nearly identical fits, one advantage of the univariate linear model is the clarity of the slope estimate for the fitted line: β_1 = 2.456. As a slope, $\hat{\beta}_1$ is the estimated change in y (count) over the change in x (year) and is the rate of change of the fitted linear trend with respect to x: $\hat{\beta}_1 = \frac{\Delta y}{\Delta x}$. Large values of Δx and Δy reflect an average rate of change, but as Δx and Δy

become increasingly small, i.e., approaching 0, we begin to measure the instanteous rate of change, hereafter refered to as the first derivative, or $f'(x)$. Recognizing this equivalency in terminology between β and f' values, and conceptualizing parameter estimates as first derivatives, is important because it facilitates interpreting β values not as scalar terms, but as continuous functions describing the change in count with changing year. For univariate linear models, this equivalency is trivial because the slope is constant and $f'(x)$ is the same across years. However, for nonlinear responses, $f'(x)$ can change along the gradient of a covariate, and therefore provides a continuous measure of (instantaneous) slope. Calculating derivatives for GAM responses is not easily available analytically, but can be approximated using the finite difference method (Simpson 2018).

Analytical Tools

Beyond the ability to fit more complex trends in Chao2 estimates than model averaging, GAMs provide additional analytical tools to inform decision making. First, it provides a continuous estimate (with uncertainty) describing how the smoothed Chao2 estimate changes over time. Thus, for any given monitoring year, instead of inferring changes in trend from changes in AIC_c weights, i.e., model averaging, GAMs provide a more direct and interpretable estimate of changes in trend. Furthermore, interpretation of the first derivate can be restricted to the monitoring year the same way smoothed estimates are, whereas AIC_c weights are based on the entire time series of Chao2 estimates. Second, GAMs offer access to the powerful tool of posterior simulation (Simpson 2018), providing additional insights into parameter uncertainty. This approach allows the uncertainty in parameter values to be represented as a probability distribution whereby parameter values that are more consistent with the data have higher probabilities than those less consistent, providing a more complete picture of estimated uncertainty given the data (Albers et al. 2018, Kruschke 2018; see Appendix C for details on posterior simulation). Although consistent with confidence intervals, probability distributions go beyond the concept of being "inside" or "outside" an interval and provide a more transparent picture of differences in certainty within the confidence interval. Moreover, their continuous nature allows probabilistic statements, which can provide a more intuitive interpretation of estimates and their associated uncertainty (Hespanhol et al. 2019; Fig. 12).

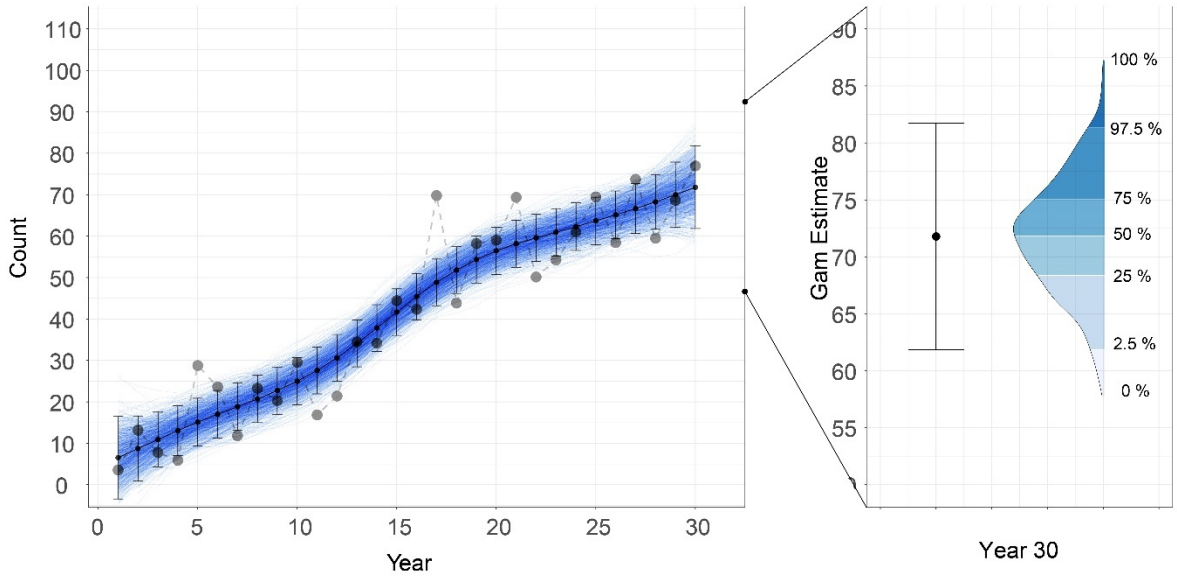

Fig. 12. A) Hypothetical data (grey circles connected by dashed line) and fitted generalized additive model (GAM; solid black line) with 95% confidence intervals (error bars) and posterior simulation fits (n = 1,000; thin blue lines) of the number of female grizzly bears with cubs over 30 years. Simulations show the density of posterior fits by intensity of blue lines. **B)** vertical bar and point show the fitted estimate and 95% CI for year 30. The distribution to the right shows plausible estimates of smoothed values consistent with the fitted model in A (posterior simulation, n = 1,000) for model year 30 with color indicating quantile. We note the following key points regarding inference for model year 30: (1) the median, 2.5th and 97.5th quantiles (i.e., 95% of data) of the distribution are equivalent to point and tails of the 95% confidence interval; (2) inference based on the confidence interval (significance test) is restricted to being inside or outside the tails (e.g., 65), whereas the distributional representation provides the same information but allows for probabilistic statements in reference to threshold values. For example, although the confidence interval provides no statistical evidence that the point estimate is significantly greater than a reference value of 65, approximately 90% of the posterior simulations were greater than this value, providing substantial statistical support for the interpretation that the number of females with cubs in year 30 is greater than 65.

The approach can be applied to the smoothed GAM estimates (predicted values) and first derivatives (rate of change parameter) and thus serves as a unifying framework for additional model inference.

Monitoring Changes in Trend over Time

We approached trend monitoring using the first derivative posterior distributions of fitted GAMs with the primary purpose to better inform decision making. Interpretation focuses not only on model outputs for a given monitoring year, but also their relation to outputs of recent years. To quantify the ability to detect change, we used the significance of the first derivative estimate, defined as the confidence intervals not containing zero. Harris et al. (2007) selected the use of AIC_c weights rather than hypothesis tests on parameter estimates because of low statistical power associated with the high annual variation in the Chao2 time series. This conclusion has not changed and determining significance based on $\alpha = 0.05$ may be too restrictive in applied management situations. Therefore, we optimized the use of $(1 - \alpha)$ confidence intervals greater than $\alpha = 0.05$ to increase power of detecting a change event, while still producing an acceptable false positive rate. First derivative confidence intervals were calculated using varying α levels from 0.05 to 0.20, and we quantified the false-detection rate under the control simulations (n = 1,000) where the deterministic Chao2 value did not change over time.

Using the optimized α-level of 0.15 (see Section IV), we calculated the first year after the simulated decline where confidence intervals for the first derivative did not overlap zero. We recorded the duration of the detection event as the number of consecutive years in this state once detected and the year and magnitude of maximum decline. To provide managers with additional tools for interpreting slope dynamics, we used the proportion of the posterior distribution less than zero (probability of decline; pd) to capture shifts not accounted for by the confidence interval significance tests. Relative changes in pd from year to year provide a straightforward interpretation of increasing or decreasing rates of change and relevant measures of early warning and recovery. For example, first derivates may be significantly less than zero, but trending back towards non-negative values (Fig. 13).

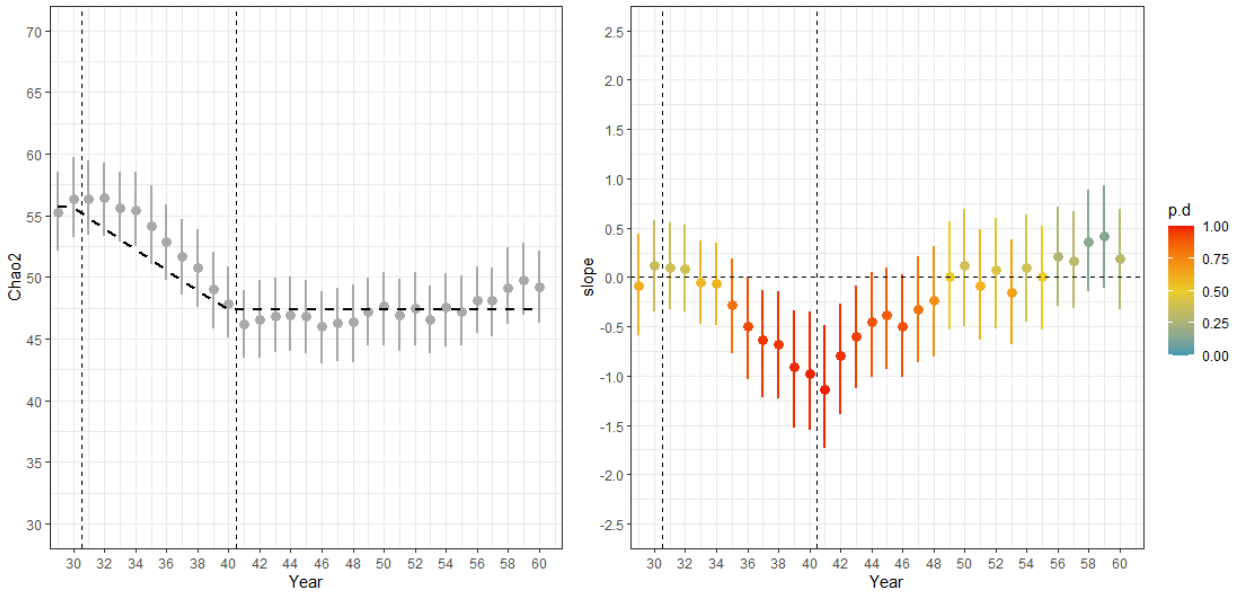

Fig. 13. A) Annual fitted smooth GAM estimates and ~95% confidence intervals for a simulated population of female grizzly bears with cubs exhibiting a trajectory of stable to decline and returning to stable. Dashed line indicates the deterministic trend (simulated Chao2 values not shown). Vertical dashed lines show the start and end of the decline period (years 31–40). **B)** First derivative estimates of fitted GAM (same input data as panel A) and 85% confidence intervals (based on adjusted α-level of 0.15 producing adequate false positive rate in control simulations; see Section IV for details). Color gradient indicates the proportion of the posterior distribution <0 (i.e., declining). First significant decline detection occurs in year 36 and lasts through year 43. Sustained relative changes in *p*d (probability of decline) suggest an early warning of change in year 35 and a change in trend after year 41, with sustained reduction in slopes. Interpretation must account for the fact that each year's estimates do not contain future data (i.e., analyses are not retrospective).

4. SUMMARY

Fitting GAMs to grizzly bear count data and basing trend inference on the continuous rate of change based on the first derivative estimate provides a flexible means of overcoming the limitations of model averaging outlined in Section I, without deviating from the original intentions of model averaging. From a statistical standpoint, GAMs provide more flexibility to accommodate, identify, and interpret any future changes in population size. This addresses a key limitation of the model-averaging approach, which was only designed to detect a slowing of population growth. Our proposed alternative to assess trend through GAM first derivatives relies on a direct measure of the fitted model's rate of change, which is invariant to trend patterns. Furthermore, whereas the model-averaging approach relies on data from the entire time series, first derivatives are continuous, and therefore inference can be focused on any portion of the time period, including recent time periods that are likely more relevant for decision making. Although it requires a conceptual shift, application of GAMs, first derivatives, and inferential tools such as the probability of decline (*p*d) improve transparency and communication of model outputs.

Finally, under this proposed framework there is a clear separation between the scientific product of trend assessment (e.g., $1 - \alpha$ confidence interval, *p*d threshold) and policy-based management objectives (Wagner 2013). We do not present specific threshold criteria or management actions, which are outside IGBST's monitoring and science support role (van Manen et al. 2014). However, in proposing application of the methodology, we recognize the need to aid in the development of threshold criteria such as first derivative confidence interval size (α-level) and the proportion of posterior distributions less or greater than zero that can be used as guidance for assessing potential revision of the criteria established in the *2016 Conservation Strategy*. In Section IV, we use the methods proposed here with simulated population time series to provide this guidance.

SECTION IV – EVALUATING GAM PERFORMANCE WITH SIMULATED DATA

1. METHODS

Simulation Framework

We used a simulation framework to create realistic dynamics in trends of annual N_{Chao2} estimates. We developed the simulations within the context of a stable population experiencing short-term perturbations, followed by a return to stability. We modeled post-decline responses as a return to stability, but these simulations should not be interpreted as a limitation of the models or management responses. We simulated declines by varying magnitude and duration, with 3 levels of decline (10, 15, and 20%) and 3 durations (5, 10, and 15 years). The combined duration × magnitude effect sizes correspond to constant population growth rates (λ) ranging from 0.953 (20% decline over 5 years) to 0.993 (10% decline over 15 years). We also included a null scenario of zero growth (λ = 1.0; Table 8). 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 GYE grizzly bear population and important for management decisions. Indeed, because of the flexibility of GAMs, 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 (Table 8) intended to mimic process and sampling variance present in observed N_{Chao2} estimates. We used the empirical residuals from a regression of N_{Chao2} ~ year from 2000–2018 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 the 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_{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 then "unscaled" by multiplying by the deterministic time series value and adding it to

Table 8. Parameters for simulation scenarios of declines in the number of female grizzly bears with cubs with varying magnitude (10, 15, and 20%) and duration (5, 10, and 15 years). Approximate mortality and constant population growth (λ) rates corresponding to the combined duration × magnitude levels are shown for reference.

Decline scenario	Decline duration (years)	Decline magnitude (%)	Approximate mortality rate	Approximate lambda (λ)
Duration = null Magnitude = null	0	0	7.6	1.000
Duration = short Magnitude = small	5	20	12.3	0.979
Duration = short Magnitude = medium	5	15	11.1	0.968
Duration = short Magnitude = large	5	10	9.9	0.956
Duration = medium Magnitude = small	10	20	10.2	0.990
Duration = medium Magnitude = medium	10	15	9.4	0.984
Duration = medium Magnitude = large	10	10	8.8	0.978
Duration = long Magnitude = small	15	20	9.2	0.993
Duration = long Magnitude = medium	15	15	8.8	0.989
Duration = long Magnitude = large	15	10	8.4	0.985

the deterministic value to create stochastic time series. We simulated 1,000 replicate time series for each of 10 scenarios (i.e., 1 null and 9 treatment scenarios), each with a length of 75 years. This length allowed for stabilization, or "burn-in," time before the start of the decline period, and a post-decline stable period. We started declines in year 31 of the simulation. Stochastic noise of the simulated N_{Chao2} values resulted in variation across simulations of N_{Chao2} values 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 GYE grizzly bear population through the early 2000s (Interagency Grizzly Bear Study Team 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.

Model Fitting and Data Structure

To assess the relationship between year and N_{Chao2} in the GAM framework, we fit a single covariate model using the general structure we introduced in Section III:

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

where f is a smooth function of the covariate year. We fit models with the mgcv (Wood 2004) package in program R.

We 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). Use of a 3-year moving average is easily understood and provides a modest amount of variance reduction. However, there are several important considerations in the use of simple moving averages. First, its use lags the data by half the size of the sample window (e.g., 1.5 years in our application), delaying the onset of changes in the input signal. However, the reduction in the signal-to-noise ratio generally outweighs this lag effect in model performance (see Smoothed Estimates; p. 50). Second, moving averages increase autocorrelation in the time series that could lead to

overfitting if not accounted for. Accordingly, we modified default GAM parameterization to protect against overfitting by upscaling the spline penalization (see Model Parameterization).

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 fitted 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 the IGBST 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.

Model Parameterization

For each monitoring year of a simulation scenario, a GAM was fitted on annual N_{Chao2} estimates (N_{Chao2_t}) or 3-year moving averages ($N_{\text{Chao2}_{3\bar{t}}}$):

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

or

$$N_{\text{Chao2}_{3\bar{t}}} = \beta_0 + f(\text{year}_t) + \varepsilon_t,$$

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. 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 increased the "penalty" per increment in the degrees of freedom, producing a smoother fit (Kim and Gu 2004; Wood 2006, 2017) and resulted in a reasonable balance of 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 the suggestions of Wood (2011) and Simpson (2019), we used restricted maximum likelihood (REML) for parameter estimation. We set the smoother function to use univariate penalized cubic regression splines (Wood et al. 2017).

Model Outputs and Inference

For each monitoring year (n = 50) and simulation replicate (n = 1,000) we fitted GAMs and used the methods of Simpson (2018) to generate 1,000 posterior simulations of smoothed N_{Chao2} estimates and first derivates $f'(\text{year}_t)$ as outlined in Appendix C. Posterior distributions were extracted and stored for fitted model estimates and first derivatives, or slopes, for each monitoring year. 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 (see Section III).

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 versus 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 evaluated the ability of the models to assess trend dynamics in two ways. First, the proportion of the posterior distribution of the GAM first derivative ($f'(\text{year})$) that is less than 0 (pd; section III) serves as a metric of trend existence, directionality, and magnitude. Conceptually, the role of the pd index is equivalent to the current use of AIC_c weights by alerting biologists to a possible change in system state (Harris et al. 2007). However, unlike AIC_c weights, pd is a probabilistic statement incorporating estimate uncertainty, which can be easily interpreted. Second, we assessed if the $(1 - \alpha)$ confidence intervals of the fitted GAM first derivative contained the deterministic slope value of zero (i.e., no change).

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 include replicate-level results, explicitly accounting for the time dependency of each entire time series by using an

indicator variable for "trend state." We assigned a state of decline for years with first derivatives different from zero, representing support for a statistically significant decline, and a state of no decline for years when the confidence interval contained 0. Contiguous years of the same state were considered belonging to the same event, allowing reporting of year of event change and duration (e.g., duration of decline). 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 indicate a slope that is significantly different from 0, the relative differences in *p*d and slope estimates can indicate more detailed temporal dynamics in trend. It is easy to infer if the current years estimate is past the peak of a decline, and if relative changes in *p*d and slope estimates reflect a trend returning to non-significant slopes (Fig. 13B; years 41–47).

2. RESULTS

Smoothed Estimates

Monitoring year estimates for the $N_{\text{Chao2}_{3\bar{t}}}$ model under the null model scenario (i.e., no simulated decline) were relatively unbiased, with over 85% of 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 $= 0.484, \sigma = 1.33$). As expected, these differences occurred mostly during the pre-impact phase and were associated with larger lag effects due to higher variance in the annual N_{Chao2_t} values subsequent to the initial increase prior to stabilization. For the 9 treatment scenarios, the degree and dynamics of 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 increased with the size and speed of simulated declines (Table 9). Similarly, during the post-decline stabilization period, models showed a transition to a period of negative bias as models responded to the abrupt change from decline to stabilization. Compared with the N_{Chao2_t} models, variance reduction associated with the 3-year moving averages of the

Table 9. Bias associated with smoothed estimates using generalized additive models of simulated time series of estimates of female grizzly bears with cubs (N_{Chao2}), based on 9 scenarios of population decline with varying levels of duration and magnitude for 3-year simple moving averages ($N_{\text{Chao2}_{3\bar{t}}}$). Results reflect the period of impact (decline). Null model results are presented for reference. We simulated 1,000 replicate time series for each scenario, each with a length of 75 years and with population declines starting in year 31 of the decline simulations.

Model	Decline duration (years)	Decline magnitude (%)	Approximate λ	Mean bias	Standard deviation bias	0.025 quantile	0.975 quantile
Null	0	0	1.0	0.03	1.33	−2.51	2.55
$N_{\text{Chao2}_{3\bar{t}}}$	15	20	0.985	1.18	1.44	−1.64	3.98
$N_{\text{Chao2}_{3\bar{t}}}$	15	15	0.989	0.92	1.42	−1.87	3.65
$N_{\text{Chao2}_{3\bar{t}}}$	15	10	0.993	0.64	1.40	−2.11	3.33
$N_{\text{Chao2}_{3\bar{t}}}$	10	20	0.978	1.91	1.56	−1.18	4.84
$N_{\text{Chao2}_{3\bar{t}}}$	10	15	0.984	1.50	1.50	−1.49	4.31
$N_{\text{Chao2}_{3\bar{t}}}$	10	10	0.990	1.05	1.46	−1.84	3.81
$N_{\text{Chao2}_{3\bar{t}}}$	5	20	0.956	3.73	1.83	0.00	7.04
$N_{\text{Chao2}_{3\bar{t}}}$	5	15	0.968	2.89	1.71	−0.58	5.97
$N_{\text{Chao2}_{3\bar{t}}}$	5	10	0.979	2.00	1.59	−1.20	4.91

$N_{\text{Chao2}_{3\bar{t}}}$ 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. 14).

Trend Detection

We provide detailed results of each step of the trend detection procedure to help interpretation. Because we assessed 10 scenarios, we restricted results to the $N_{\text{Chao2}_{3\bar{t}}}$ model because of its smaller bias compared with the N_{Chao2_t} model.

Optimizing the α-Level.—At the annual level, the null treatment (no change in deterministic trend) rate of false-positive significant slopes (increasing or declining trend) ranged from 0.01 at α = 0.05 to 0.06 at α = 0.20. These errors reflect the relatively low probability of false detections for a given year over the entire simulated time series. False-positive rates at the replicate level captured if there was ever a false detection during the entire time span (years 30–75) and consider events as blocks of consecutive years in the same state (i.e., detection or no detection). Replicate-level false positive rates were 0.05, 15.5, 32.4, and 48.8 at α-levels of 0.05, 0.10, 0.15, and 0.20 respectively. Replicate-level rates must be carefully interpreted in relation to the simulated time period. For example, even at the highest α-level of 0.20, although 48.8% of null model simulations having a false positive event seems high, when accounting for the 45-year period, and all false-positive events, the expected frequency of a false positive event is low, only once every 65 years. The year of first detection of false-positive change events varied widely across the simulation time series, regardless of α-level ($\sigma_{\alpha = 0.05} = 12.4$; $\sigma_{\alpha = 0.20} = 12.5$) and with relatively short mean durations (2.7 to 3.0 years) reflecting a random and transient dynamic.

Ultimately, the α-level is a tunable parameter, reflecting a manager's comfort balancing "costs" associated with falsely signifying change versus failing to detect change. Given the short-term nature of the false positive events, we believe these costs are relatively low compared with those of failing to detect a change that is occurring. Therefore, we evaluated GAM change detection using α = 0.15, which results in an annual false-positive rate of 0.039 and at the replicate-level an expected false event detection frequency of once every 109 years.

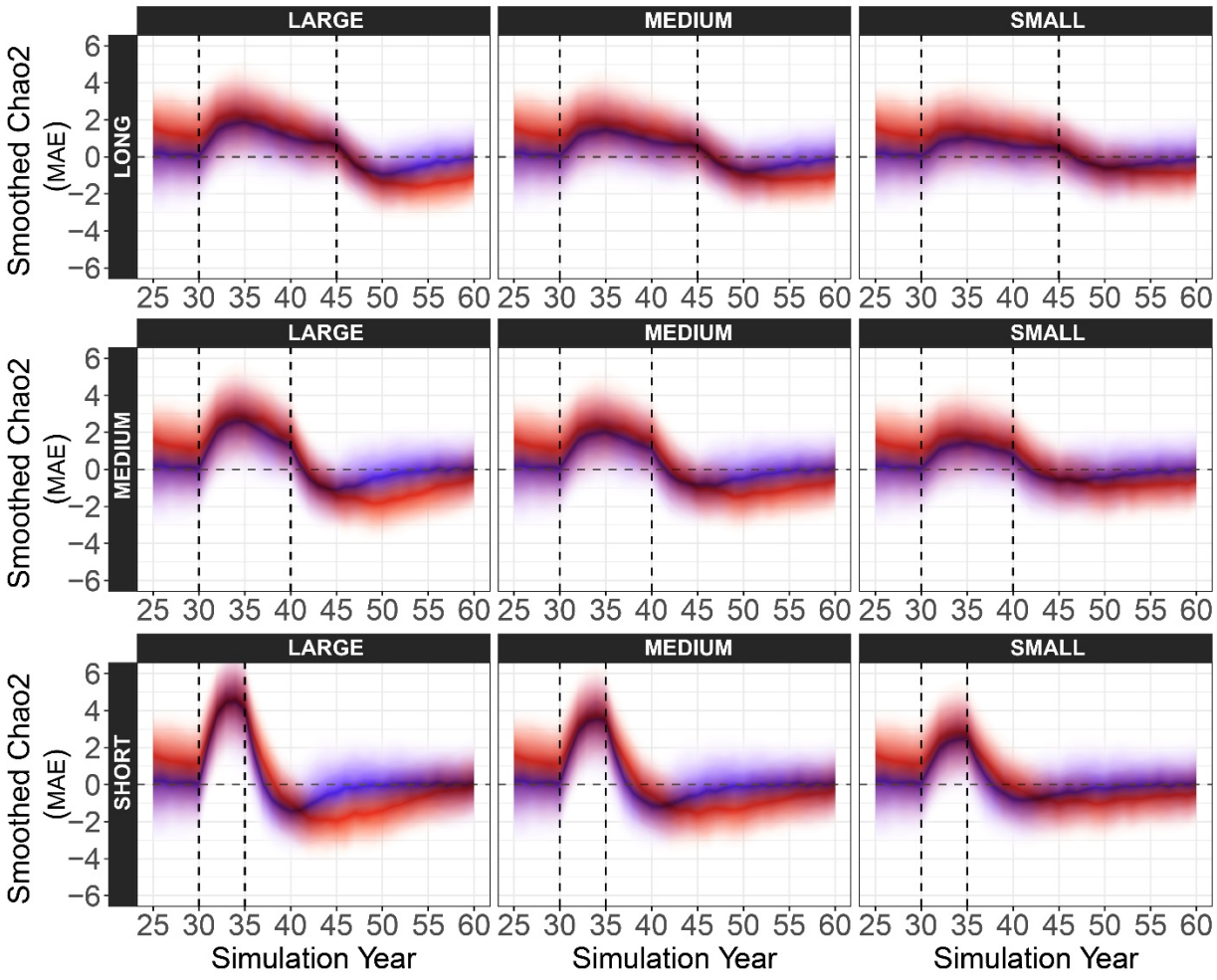

Fig. 14. Mean absolute error (MAE) for N_{Chao2_t} (annual; blue) and $N_{\text{Chao2}_{3\bar{t}}}$ (3-year moving average; red) fitted estimates of female grizzly bears with cubs as a function of simulation year. Columns show magnitude of the impact (large = 20%, medium = 15%, small = 10% decline) and rows show duration of impact period (long = 15 years, medium = 10 years, short = 5 years). Dashed vertical lines indicate the start and end of the impact period and dashed horizontal line indicates reference bias of 0.

Treatment Scenarios: Probability of Decline (*p*d).—First derivative posterior distributions varied with modeled growth rate and duration of the simulated impact period. For all scenarios, the annual probability of decline (*p*d) increased within the first 2 to 3 years of the start of a decline, indicating that existence of a declining trend was rapidly assessed. 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 of decline. 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 ≤3 years of the end of the decline period for all but the short (5 years) durations of decline, which reached maximum values within the first two years after the decline. This pattern reflects that the short-term effects post decline were more pronounced with shorter impact durations, as it is difficult for models to fully capture these rapid dynamics. These patterns are relevant as temporal patterns in *p*d shifts provide important information for interpreting short-term dynamics, particularly initial shifts from stable periods and attenuation after peaks (Fig. 15).

Treatment Scenarios: Support for Slope Significance.—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. 15). Although that scenario lacked power to detect slope significance, it is unlikely that the 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 4.47 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 inference of changing conditions despite the lack of statistical significance, an important feature for applied management.

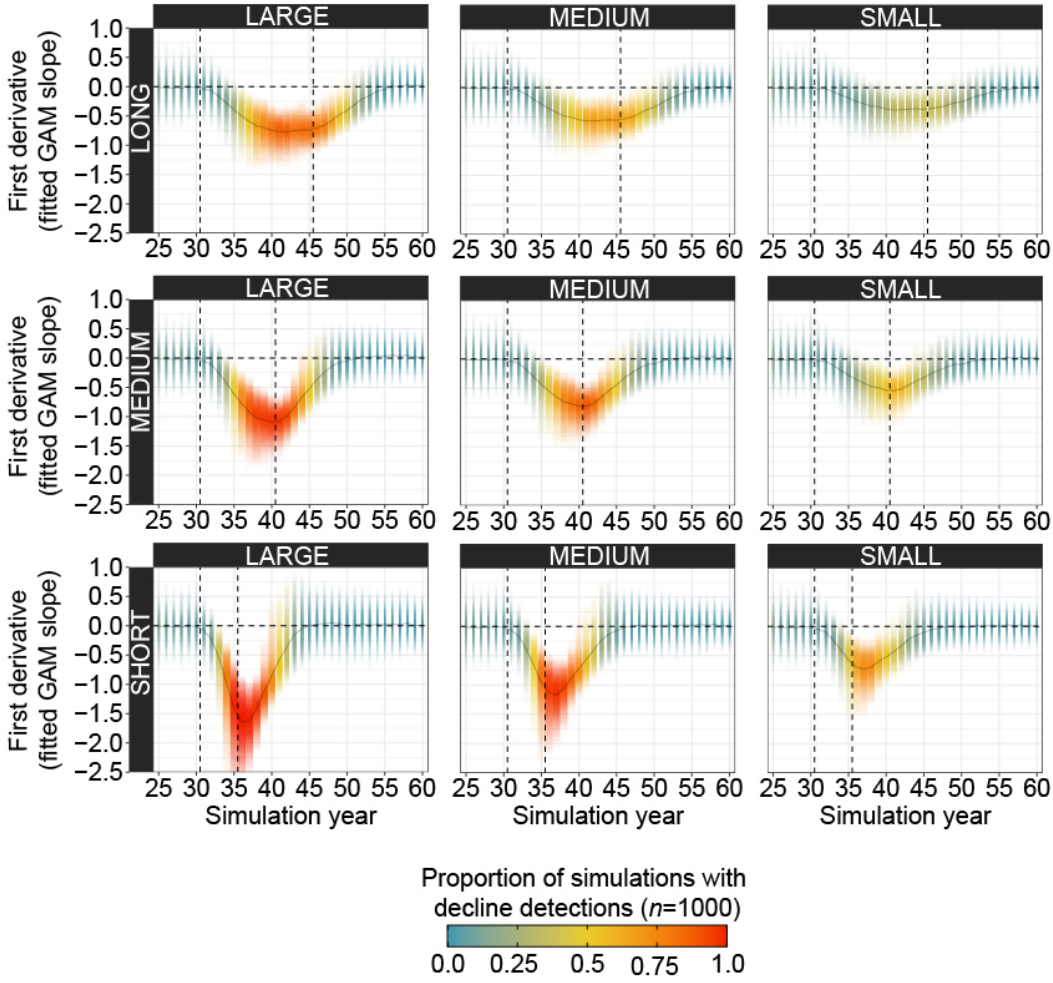

Fig. 15. Trend dynamics for $N_{\text{Chao2}_{3\bar{t}}}$ first derivative (slope) posterior distributions of number of female grizzly bears with cubs, for 9 treatment scenarios. Columns show magnitude of the impact (large = 20%, medium = 15%, small = 10% decline) and rows show duration of impact period (long = 15 years, medium = 10 years, short = 5 years). 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). Width of each density strip reflects the average pd value, scaled such that pd = 1.0 is reflected by adjacent years having no space between their density strips. Color gradient indicates the proportion of simulations (n =1,000 per scenario) in a decline "state" where confidence intervals for ≥2 consecutive years do not contain zero.

Treatment Scenarios: Event (Change) Detection.—Whereas annual summaries of the *p*d and support for rate of change significance (i.e., confidence intervals of first derivative not containing zero) provide useful measures of central tendency for each model year, they do not maintain the time-dependent structure inherent in the model simulations. For example, the most gradual decline scenario (10% over 15 yrs; $\lambda = 0.993$) did not achieve median levels of significance during any year. However, when examining time series at the level of individual replicates, 74.6% of simulations showed support for slopes significantly different from zero for two consecutive years at some point during the 15-year decline. This was a result of support for decline significance being modest (median posterior of first detection event = -0.51), short-lived event duration (median duration= 3 years), and staggered year of first detection ($\sigma = 7.03$). Thus, when averaged annually (i.e., across replicates for each year) support for significance is lacking. Although this effect is less pronounced for other scenarios, we address this overall issue by 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 >99.6% of replicates detecting decline events under the medium (10%) and large (15%) decline scenarios. For small magnitude decline scenarios (5%), detection probability ranged from 84.7% (15-year duration) to 94.7% (5-year duration) of replicates. The mean number of years from decline onset to year of first detection ranged from 3.7 (20% decline over 5 years) to 11.1 (10% decline over 15 years), and mean duration (range = 3.9–8.8 yrs) was correlated with interaction of decline duration × magnitude (Table 10).

Patterns for detecting the return to stabilization post decline were similar to decline detection and symmetrical around the peak support for decline for the 15- and 10-year decline scenarios. Five-year scenarios showed slight asymmetry around the peak with a longer and more linear return towards practically 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.

Table 10. Change detection metrics for 9 scenarios of decline for simulated time series of N_{Chao2} estimates of female grizzly bears with cubs, based on significance of first derivative of generalized additive models and 3-year simple moving averages ($N_{\text{Chao2}_{3\bar{t}}}$). 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; p(detect) is the proportion of simulation with at least 2 consecutive years of statistically significant first derivatives. Mean and standard deviation 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]	p(detect)	Mean lag to detect (years)	Standard deviation lag to detect	Mean length of detection event (consecutive years 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 years, medium = 10 years, short = 5 years.
[b]Large = 20%, medium = 15%, small = 10%.

3. DISCUSSION

Our findings demonstrate that GAMs are a suitable alternative for model-averaging procedures currently used by the IGBST to smooth annual variation in population estimates and assess changes in population trend based on N_{Chao2} estimates. As a smoother, we demonstrated the ability of GAMs to track trends and respond to not only declines, but subsequent stabilization of the population. As expected, models showed bias associated with periods of change. However, this was not because of an inherent bias of the GAMs, but a limitation of high annual variation in N_{Chao2} estimates and only using the monitoring year of a fitted model for inference (Harris et al. 2007). Thus, inference is limited by variation inherent in the data and statistical advancements alone cannot overcome this limitation. The aforementioned biases are inherent in monitoring female grizzly bears with cubs from sightings in the GYE, indicating the importance of understanding these biases when interpreting model outputs. 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 (Table 10, Fig. 15). For the smallest declines (10%), high detection rates (≥90%; Table 10) 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 point estimates provided 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 whether a meaningful effect had taken place (Fig. 15). 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.

As with any analysis involving simulations, there are important caveats. First, simulations inherently do not encompass all reality, and alternative scenarios may occur that we have not modeled. We focused on scenarios that we deemed relevant to managers and represent realistic dynamics of 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. Second, the simulations may create a false impression that the proposed monitoring tools are complex and difficult to apply. Whereas communicating results across 10 different scenarios (9 treatment and 1 null) and thousands of replications is challenging, implementation in a monitoring program is actually simple. With only a single time series, GAMs are applied as proposed here and evaluated on an annual basis as new data are added. We demonstrate such an application with empirical data in Section V.

SECTION V – EMPIRICAL APPLICATION

1. INTRODUCTION

In this section we demonstrate the combined application of using an alternate distance criterion (Section II) and replacing model averaging with GAMs for smoothing and trend detection (Sections III and IV) using empirical counts of females with cubs. We note that the 30-km time series presented here are different from those previously reported by the IGBST because 1) in contrast to model averaging starting in 2007, we retrospectively fitted GAMs for the entire time period of 1997–2019 and 2) estimates presented here were derived solely based on applying the Knight et al. (1995) rule set using the computer program from Schwartz et al. (2008), rather than the manual application of the rule set deployed historically and combined manual and computer applications since the development of the Schwartz et al. (2008) program code. The latter was necessary for running the thousands of simulations. Manual application of the rule set typically results in higher counts of *m* because of sighting subtleties that could not be programmed into the computer code. Although annual differences between the 2 approaches were usually small (average difference of 1.75 unique females with cubs during 1997–2019), higher manual counts were more evident in early years of that time period.

2. Estimates of *m* and N_{Chao2}

The time series based on estimates of *m* for the 16-km criterion showed stronger positive growth and for a longer time period compared with the 30-km criterion, with the growth rate peaking in 2008, and then slowing but with positive growth for the remainder of the period (Fig. 16). For N_{Chao2} estimates, the rate of change for the 1997–2019 period was positive for both distance criteria, with larger estimates and slightly higher growth rates indicated by the first derivative for the 16-km criterion compared with the 30-km criterion; growth rates were more similar in recent years (Fig. 17). This mirrors our simulation findings that smaller distance criteria resulted in the Chao2-adjustment representing a greater proportion of the overall N_{Chao2} estimate: the adjustment represents 13.0% of the estimate for the 30-km distance criterion, but 26.0% of the estimate for the

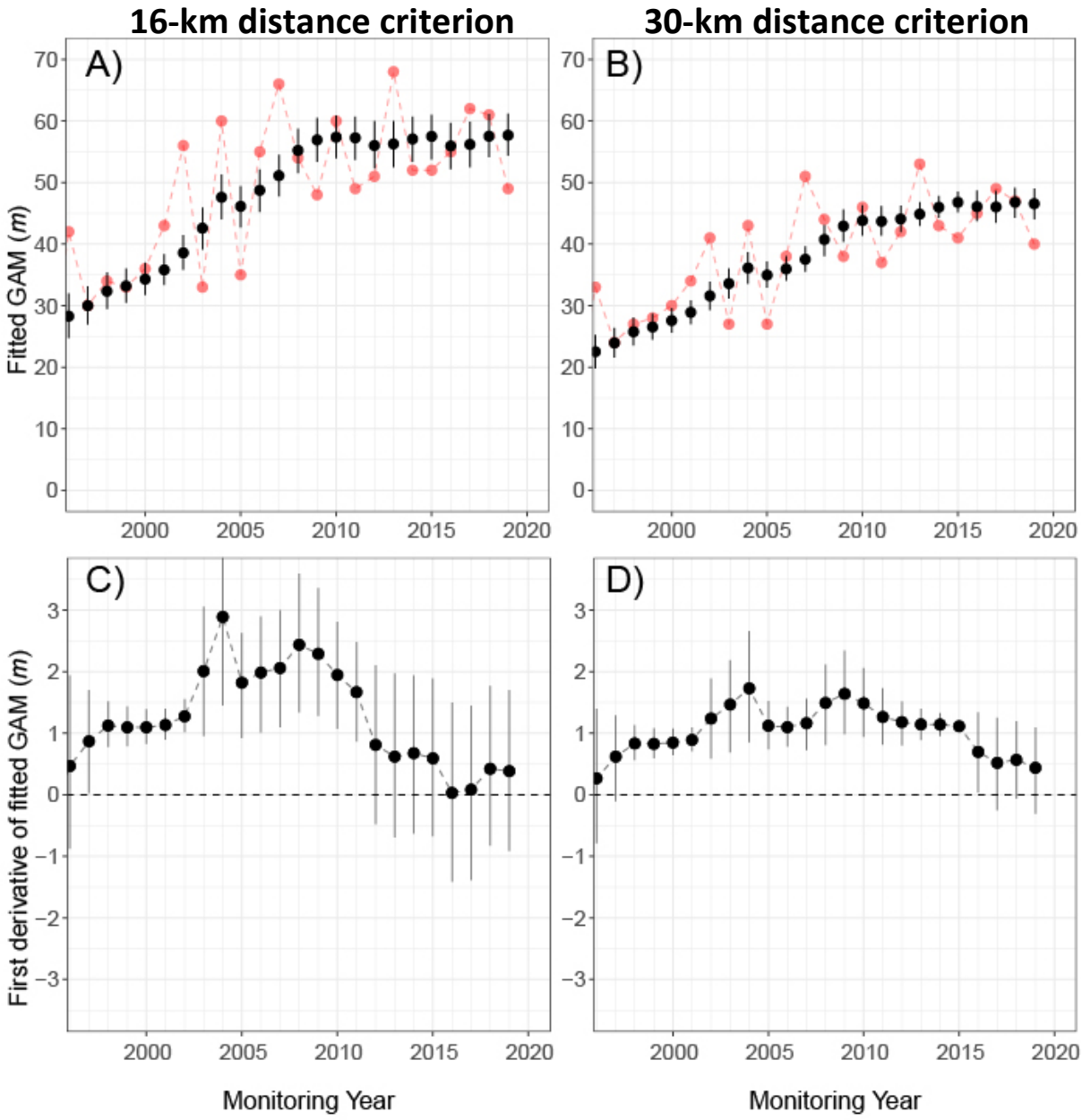

Fig. 16. Estimates of unique female grizzly bears with cubs (*m*; i.e., not including the Chao2 adjustment) based on annual sightings in the Greater Yellowstone Ecosystem (Demographic Monitoring Area) during 1995–2019, based on application of the Knight et al. (1995) rule set using 16-km (left panels) and 30-km (right panels) distance criteria. **A)** and **B)** Number of observed females with cubs (*m*) using fitted GAM estimates based on 3-year moving averages. Black 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. Raw annual *m* estimates (red circles connected by dashed line) are shown for reference. **C)** and **D)** First derivative (rate of change) of *m;* 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.

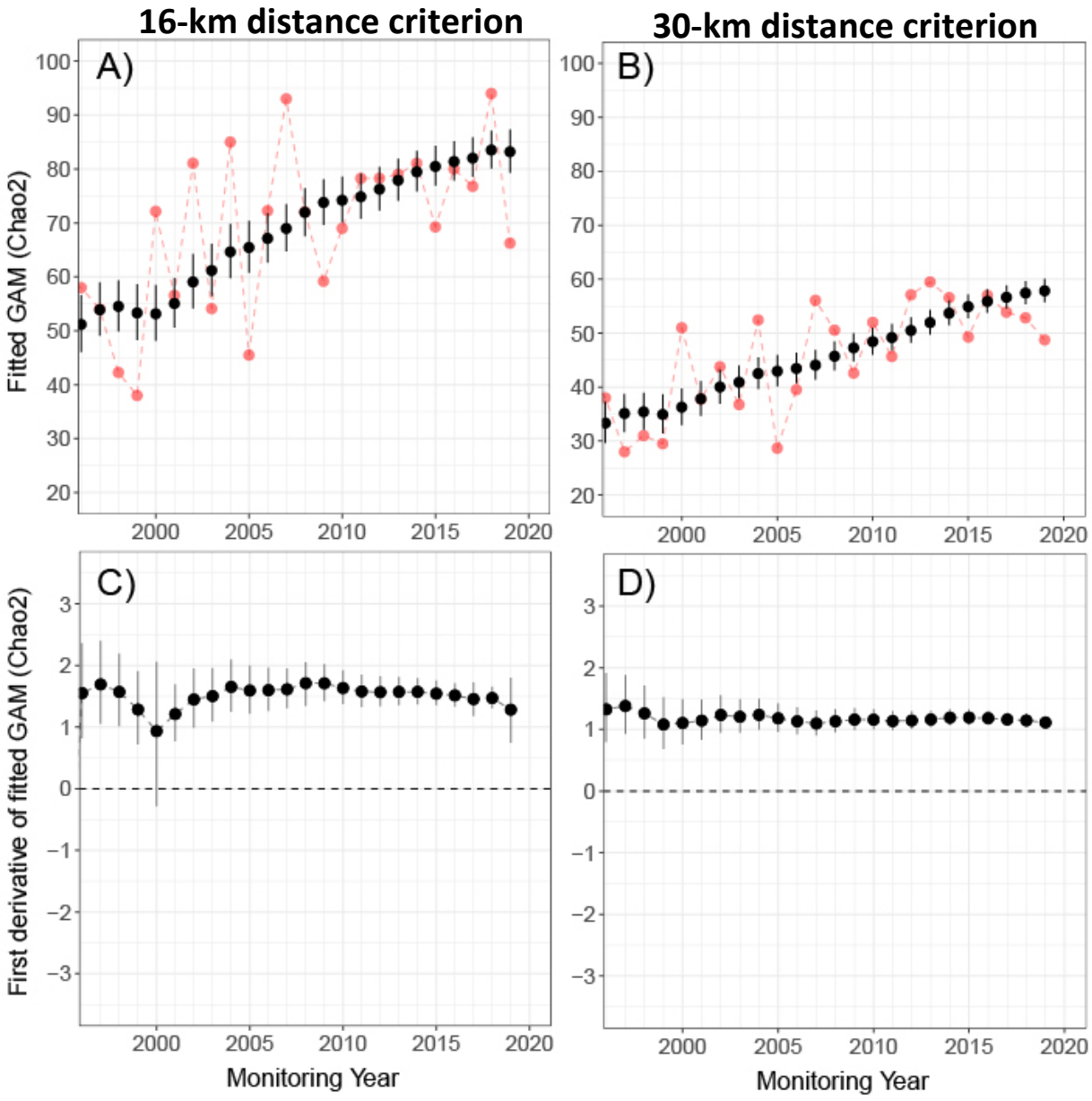

Fig. 17. Estimates of N_{Chao2} derived from the number of unique female grizzly bears with cubs (m; see Fig. 16) in the Greater Yellowstone Ecosystem (Demographic Monitoring Area) during 1995–2019, based on application of the Knight et al. (1995) rule set using 16-km (left panels) and 30-km (right panels) distance criteria. **A)** and **B)** Number of estimated females with cubs using fitted GAM estimates of 3-year moving averages of N_{Chao2} estimates ($N_{\text{Chao2}_{3\bar{t}}}$). Black 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. Raw annual N_{Chao2} estimates (red circles connected by 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.

16-km criterion. This is primarily a function of an increase in f_1 frequencies relative to f_2 frequencies when shifting the criterion from 30 to 16 km (Fig. 18). These levels were approximately 1.5 times higher than those observed in the simulated datasets (Table 6). This finding likely reflects that simulated data based on VHF locations cannot fully capture the observation process for unmarked bears, but there is no evidence of bias as a function of distance criteria.

3. DISCUSSION

Our primary motivation for exploring alternative distance criteria was to obtain unbiased estimates of numbers of female grizzly bears with cubs in the GYE. Findings from the simulation analyses demonstrate that relatively unbiased estimates of m and N_{Chao2} can be obtained by modifying the 30-km distance criterion in the rule set. Within the context of current monitoring protocols and effort, and considering the full suite of simulations presented in Section II and previous studies, we plan to update our monitoring protocols and change the distance criterion in the rule set from its current level of 30 to 16 km. By virtue of producing relatively unbiased estimates, time series of m and N_{Chao2} using the 16-km distance criterion will be more sensitive to true changes compared with estimates based on the 30-km criterion that are increasingly constrained as population size increases (Figs. 16 and 17; Schwartz et al. 2008).

In terms of trend detection, our evaluations based on simulated N_{Chao2} time series in Section IV and the application to empirical data presented in Figs. 16 and 17 show that GAMs effectively address the limitations of model averaging for estimation and trend detection of the GYE grizzly bear population. Furthermore, applying GAMs within a more robust statistical framework to assess population trend substantially improves our ability to monitor the population. This framework not only enhances trend detection but also adds an early indication of impending change or return to previous state. Additional advantages of this new framework are that it can easily be applied to the entire time series of N_{Chao2} estimates for the GYE, thus allowing retrospective analysis, and can be applied to future monitoring techniques based on any time series of population estimates.

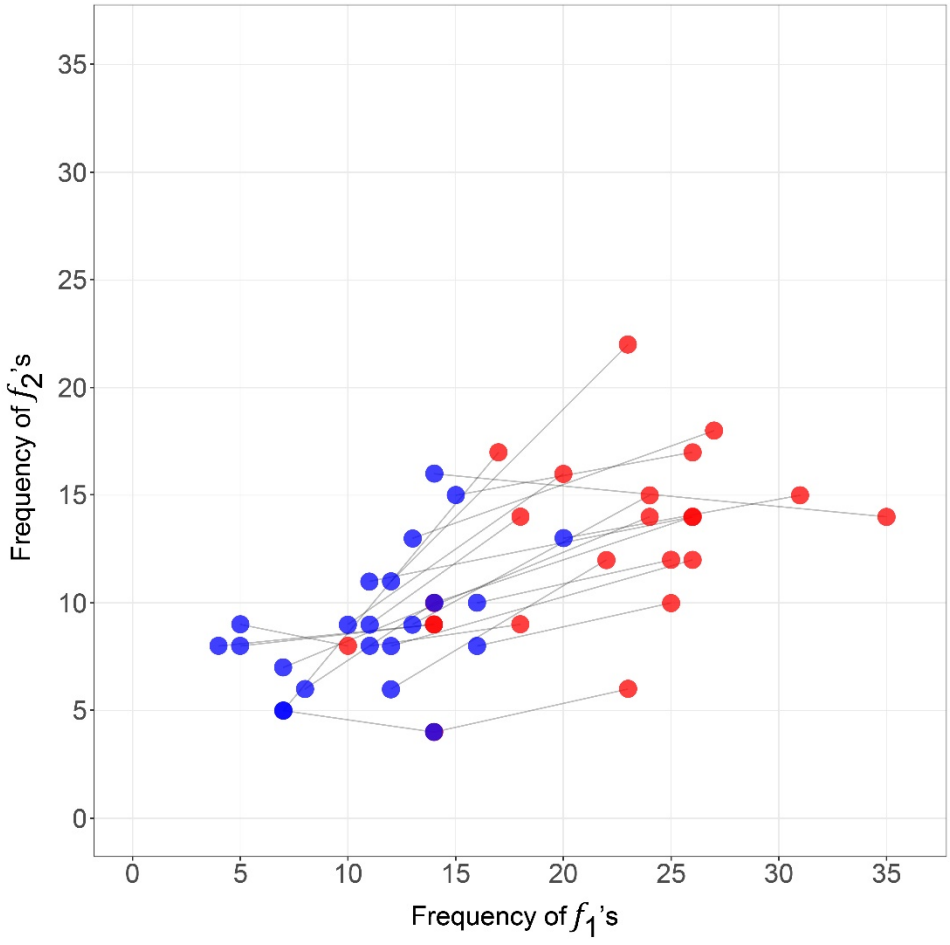

Fig. 18. Empirical data (1997–2019) of the annual frequency of unique female grizzly bears with cubs with 1 (f_1) and 2 (f_2) sightings based on application of 16-km (red circle) and 30-km (blue circle) distance criteria of the Knight et al. (1995) rule set. Lines connect the two independent estimates for the same year (year values not shown). The relatively flat slopes of lines indicate a greater increase in f_1 frequencies versus f_2 frequencies when changing the distance criterion from 30 to 16 km.

There are a number of caveats to these findings. First, we emphasize that our conclusions are based on obtaining relatively unbiased average estimates from simulations with different levels of known females with cubs (i.e., N_{true}). The empirical data are the equivalent of a single simulation run and thus could, by chance, represent a time series away from central tendencies. We account 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 why smoothing of these time series data is important, and why we plan to use GAMs and 3-year moving averages. Second, the simulation framework was designed to directly compare a true number of "sighted" bears to the estimate of *m*. 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 captures the Chao2 adjustment (Schwartz et al. 2008, Keating et al. 2002, Cherry et al. 2007). Third, 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. Although the sampling frame based on telemetry and ground sightings was data-driven and based on reasonable assumptions, the simulations are only approximations. 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. Finally, we focused on the distance criterion in the rule set because of its overarching implications on the outcome; however, there are other criteria in the rule set that also play a role, which we did not explicitly investigate.

4. IMPLICATIONS

There are several important implications associated with implementation of a new distance criterion and use of GAM techniques. A primary consideration is that the 16-km distance criterion results in total population estimates derived from the Chao2 estimates that are greater than those we have reported in the past. While this increase is due to a

change in the implementation of the technique, it also more accurately represents the number of females with cubs in the GYE grizzly bear population. For example, the estimate of 82 females with cubs mentioned in the previous paragraph using the 16-km distance criterion is 41% greater than the 2019 model-averaged N_{Chao2} estimate of 58 females with cubs based on the 30-km distance criterion in the current rule set (Haroldson et al. 2020). Total population estimates derived from the N_{Chao2} estimates would increase accordingly: the 2019 estimate of 737 would be the equivalent to a total population size over 1,000. In combination with switching from manual- to computer-based application of the Knight et al. (1995) rule set, underestimation may also have contributed to an artificial flattening of population trend since the early 2000s. Although the IGBST documented slowing of population growth based on independent data for vital rates as well (Interagency Grizzly Bear Study Team 2012), trend data based on N_{Chao2} estimates and the 30-km distance criterion may have overestimated the flattening of the population trajectory (Haroldson et al. 2020).

Implementation of the 16-km distance criterion combined with use of GAM techniques would affect some of the population metrics (e.g., annual population size and uncertainty, population trend, mortality rates) used to inform management responses. Implementation would require relatively minor changes in the monitoring protocols described in Appendices B and C of the *2016 Conservation Strategy*. Additionally, we note that the IGBST has ongoing investigations into the merits of an integrated population model (IPM), for which annual Chao2-based estimates are important input data. The IGBST plans to continue those investigations using the 16-km distance criterion to derive Chao2 estimates.

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 adapt monitoring strategies to reduce this source of variation, and the IGBST continues to investigate approaches for such improvements.

LITERATURE CITED

Albers, C., H. Kiers, and D. van Ravenzwaaij. 2018. Credible confidence: a pragmatic view on the frequentist vs Bayesian debate. Collabra: Psychology 4:31.

Blanchard, B. M., and R. R. Knight. 1991. Movements of Yellowstone grizzly bears. Biological Conservation 58:41–67.

Bolker, B. M. 2008. Ecological Models and Data in R. Princeton University Press, Princeton, New Jersey, USA.

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

Guisan, A., T. C. Edwards, Jr., and T. Hastie. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modeling 157:89–100.

Hall, P., and A. R. Padmanabhan. 1992. On the bootstrap and the trimmed mean. Journal of Multivariate Analysis 41:132–153.

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://prd-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/atoms/files/2019 IGBST Annual Report %28FINAL FINAL%29_Sec.pdf

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.

Hastie T. J., and R. J. Tibshirani. 1986. Generalized additive models. Statistical Science 1:297–310.

Hastie T. J., and R. J. Tibshirani. 1990. Generalized additive models. Taylor & Francis, New York, New York, USA.

Hespanhol, L., C. S. Vallio, B. T. Saragiotto, and L. C. M. Costa. 2019. Understanding and interpreting confidence and credible intervals around effect estimates. Brazilian Journal of Physical Therapy 23:290–301.

Higgs, M. D., W. A. Link, G. C. White, M. A. Haroldson, and D. D. Bjornlie. 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.

Hossin, M., and M. N Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process 5:1–11.

Interagency Grizzly Bear Study Team. 2005. Reassessing methods to estimate population size and sustainable mortality limits for the Yellowstone grizzly bear. Interagency Grizzly Bear Study Team, U.S. Geological Survey, Northern Rocky Mountain Science Center, Montana State University, Bozeman, Montana, USA.

Interagency Grizzly Bear Study Team. 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://prd-wret.s3-us-west-2.amazonaws.com/assets/palladium/production/s3fs-public/atoms/files/GYEGBMonMortWksRpt2012%282%29_2.pdf

Jones, K., and S. Almond. 1992. Moving out the linear rut: the possibilities of generalized additive models. Transactions of the Institute of British Geographers 17:434–447.

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.

Killick, R., and I. A. Eckley. 2014. Changepoint: an R package for changepoint analysis. Journal of Statistical Software 5:1–19.

Killick, R., K. Haynes, and I. A. Eckley. 2016. Changepoint: an R package for changepoint analysis. R package version 2.2.2. https://CRAN.R-project.org/package=changepoint

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.

Kruschke, J. K. 2018. Rejecting or accepting parameter values in Bayesian estimation. Advances in Methods and Practices in Psychological Science 1:270–280.

Lever, J., M. Krzywinski, and N. Altman. 2016. Points of significance: logistic regression. Nature Methods 13:541–542.

McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. Second edition, Chapman and Hall, London, UK.

Pedersen, E. J., D. L. Miller, G. L. Simpson, and N. Ross. 2019. Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ 7:e6876.

R Core Team. 2019. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rosenbaum, B., M. Raatz, G. Weithoff, G. F. Fussmann, and U. Gaedke. 2019. Estimating parameters from multiple time series of population dynamics using Bayesian inference. Frontiers in Ecology and Evolution 6:234.

Rosner, B. 1983. Percentage points for a generalized ESD many-outlier procedure. Technometrics 25:165–172.

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.

Shalizi, C. R. 2019. Advanced data analysis from an elementary point of view." https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/

Simpson, G. L. 2018. Modelling palaeoecological time series using generalised additive models. Frontiers in Ecology and Evolution 6:149.
Simpson, G. L. 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
Ting, K. M. 2011. Precision and recall. *In* C. Sammut and G. I. Webb, eds. Encyclopedia of Machine Learning. Springer, Boston, Massachusetts, USA.
U.S. Fish and Wildlife Service. 2007. Grizzly bear recovery plan supplement: revised demographic criteria for the Yellowstone ecosystem. 72 FR 11377.
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.
van Manen, F. T., M. A. Haroldson, D. D. Bjornlie, M. R. Ebinger, D. J. Thompson, C. M. Costello, and G. C. White. 2016. Density dependence, whitebark pine, and vital rates of grizzly bears. Journal of Wildlife Management 80:300–313.
Wagner, T., B. J. Irwin, J. R. Bence, and D. B. Hayes. 2013. Detecting temporal trends in freshwater fisheries surveys: statistical power and the important linkages between management questions and monitoring objectives. Fisheries 38:309–319.
Wood, S. N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99: 673–686.
Wood, S. N. 2006. Generalized additive models: an introduction with R. CRC Press, Boca Raton, Florida, USA.
Wood, S. N. 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.
Wood, S. N. 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. http://igbconline.org/wp-content/uploads/2016/03/161216_Final-Conservation-Strategy_signed.pdf

APPENDIX A

Simulation Data for Testing Model-Averaging Protocol

Using empirical data from the period of relative stability (2000–2016), we extracted residuals $\left(N_{\text{Chao2}_t} - \left(\frac{\sum_{2000}^{2016} N_{\text{Chao2}_t}}{n}\right)\right)$ as a time series representing an unknown combination of process and sampling variance of the N_{Chao2} estimates. To simulate future scenarios, we first standardized the residuals by size estimates and then used a time series bootstrap of an autoregressive integrated moving average (ARIMA(1,0,0)) model to create new simulated time series of variation. We then added simulated residual time series to deterministic scenarios of future growth to create time series of N_{Chao2} estimates with variation similar to the empirical data (see section IV.1 for additional simulation details). For each simulation replication, we applied the current monitoring protocols, beginning with model averaging in 2007 and continuing annually for each monitoring year in the time series. To best approximate the true monitoring environment, we used the empirical N_{Chao2} data for 1983–2019 and simulated N_{Chao2} data for 2020–2041.

APPENDIX B

Linking Empirical and Simulation Data

We simulated females with cubs ranging from $N_{\text{true}} = 50$ to $N_{\text{true}} = 90$ in steps of 10. Our intention was to provide a range of simulated known females with cubs that covers the current true number and plausible future values under the assumption of continued population growth. Although the true number of females in the population is unknown, this broad range of N_{true} values is well supported by empirical data and previous simulations (Schwartz et al. 2008, Higgs et al. 2013, Haroldson et al. 2020). Because the total number of observation simulated was determined by a randomly selected multiplier applied to N_{true}, higher values of N_{true} on average resulted in larger total observations (n). This is a reasonable assumption as the observation flights used to monitor females with cubs are based on standardized flight time in bear observation units, and the presence of more females with cubs would results in a greater number of observations. Therefore, to provide additional context for inference relative to contemporary data, we explored total sightings of female grizzly bears with cubs matching that of the empirical data for the period 2001–2019.

Based on a changepoint analysis (Killick and Eckley 2014, Killick et al. 2016) of estimated females with cubs, we identified 2001 as the optimal breakpoint in the time series and thus focused on the empirical data for the 2001–2019 period. For this time period, approximately 90% of the distribution of annual total sightings was between 65 and 160 so we focused on the density distribution within this range of simulated total sightings (we excluded 2007 and 2010 as outlier years based on a Rosner's Test [1983]; sightings in both years were abnormally high due to a single roadside bear being observed 63 and 55 times, respectively).

Simulation results using the 30-km distance criterion of the Knight et al. (1995) rule set indicated underestimation bias in estimates of females with cubs (m) at all 5 levels of N_{true} (Fig. 5D, Table B.1). These data indicate a substantial range of bias, varying almost 4-fold across the simulated range of true females with cubs (e.g., $x_{50(high)} = -9.56; x_{90(high)} = -37.00$). Thus, the degree of underestimation bias using the 30-km distance criterion depends on the empirical true number of females with cubs, which is

unknown. However, we can gain useful insights by synthesizing information from the union of empirical and simulated datasets.

Empirical estimates of *m* (i.e., number of unique females with cubs based on sightings only; this estimate does not include the Chao2 adjustment) within the Demographic Monitoring Area have been stable to slightly increasing since 2001 (Fig. B.1). To account for non-normality, and the low outlier year of 2005, we calculated a 10% trimmed bootstrap mean and confidence interval (Hall and Padmanabhan 1992) to quantify the central region of empirical estimates of *m* for 2001–2019. The mean value of *m* was 44.2, for which we used a 99% CI (38.5–49.4) to be most conservative. Given the range of empirical total sightings and predicted number of females with cubs when using the biased 30-km distance criterion, the frequency of simulation predictions within this central region of empirical estimates provides insights into the likelihood of the true value of *m*: for the scenario of 60 true females with cubs, 94% of simulation replicates were within this range, followed by 70 females with cubs (83%), and 50 females with cubs (61%; Table B.2). Scenarios with 80 and 90 females with cubs had much lower proportions of replicates within the range of empirical estimates (Table B.2). These findings indicate a realistic true estimate of the number of observed females with cubs (*m*) for the 2001–2019 period, when using the 30-km criterion, is most likely in the range of 60–70. We provide this information as an additional guide to help distinguish model evaluation for current conditions, as well as future conditions under the assumption of continued population growth.

Table B.1. Mean and standard deviation of bias in estimated number of female grizzly bears with cubs (predicted $m - N_{true}$) for 5 levels of true number of females with cubs (N_{true}) and applying the 30-km distance criterion for estimation. We stratified results by equal-interval categorical ranges (low, n = 65–96; medium, n = 97–128; and high, n = 129–160) and a pooled (all) category (n = 65–160). Sample sizes indicate the number of replicates meeting the total sighting criteria. Total simulated sightings (n) were restricted to the range of annual empirical sightings (n = 65–160).

True number of females with cubs (N_{true})	Range of number of sightings (n)	Sample size (replicates)[a]	Mean m bias	Standard deviation m bias
50	Low	296	−13.46	2.64
	Medium	417	−11.73	2.62
	High	287	−9.56	2.60
	All	1,000	−11.62	3.01
60	Low	33	−20.33	3.15
	Medium	415	−17.92	2.86
	High	354	−15.65	2.85
	All	802	−17.02	3.15
70	Low	0		
	Medium	163	−25.09	2.87
	High	379	−22.88	3.07
	All	542	−23.55	3.17
80	Low	0		
	Medium	21	−32.14	2.50
	High	327	−29.68	3.44
	All	348	−29.83	3.44
90	Low	0		
	Medium	0		
	High	111	−37.00	3.57
	All	111	−37.00	3.57

[a]Available replicates out of 1,000 with total sightings within 90% of the empirical range of sightings (n = 65–160) for the period 2001–2019.

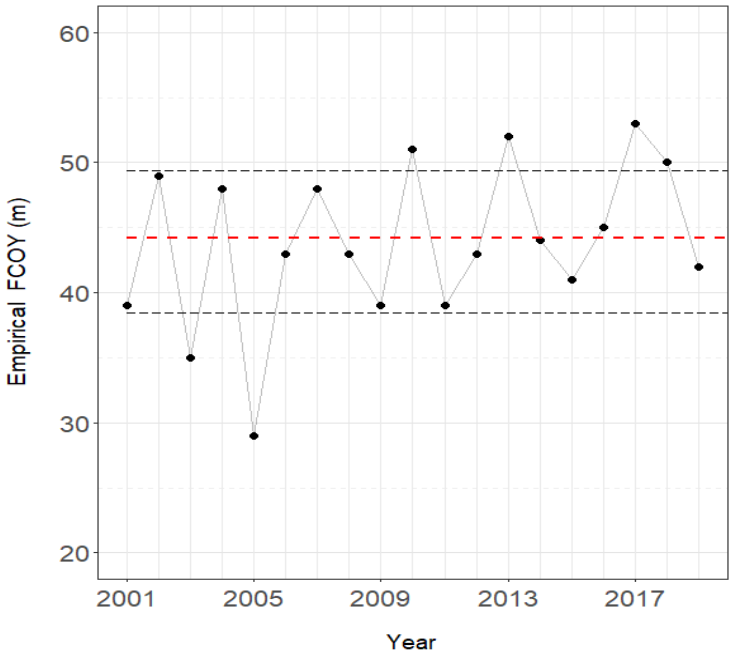

Fig. B.1. Empirical estimates of unique female grizzly bears with cubs from sightings in the Greater Yellowstone Ecosystem for the period 2001–2019, based on the 30-km distance criterion of the Knight et al. (1995) rule set. Red and black dashed lines show the 10% trimmed bootstrapped mean and 99% confidence interval, respectively.

Table B.2. Number of simulation replicates with estimates of number of female grizzly bears with cubs within the 99% CI of empirical estimates for the period 2001–2019.

True number of females with cubs (N_{true})	Number of replicates[a]	Number of replicates with estimates within 99% CI of empirical data	Proportion of replicates with estimates within 99% CI of empirical data	Minimum predicted number of females with cubs	Maximum predicted number of females with cubs
50	1,000	607	0.61	27	48
60	802	757	0.94	32	52
70	542	452	0.83	35	55
80	348	146	0.42	40	61
90	111	16	0.14	44	61

[a]Available replicates out of 1,000 with total annual sightings within 90% of the empirical range of sightings (n = 65–160) for the period 2001–2019.

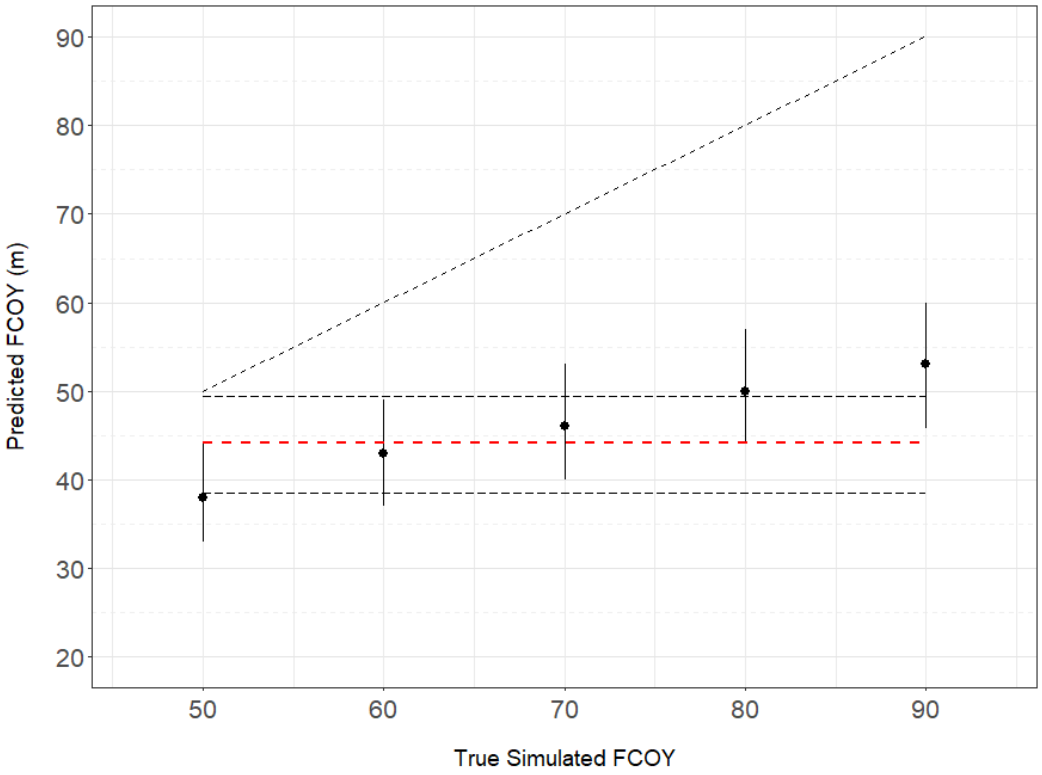

Fig. B.2. Simulation results to identify unique female grizzly bears with cubs from sightings using the Knight et al. (1995) rule set and the 30-km distance criterion for 5 simulated levels of true number of females with cubs (N_{true} = 50, 60, 70, 80, and 90). Solid black circles and lines show means and 95% CI for each true level. The diagonal dashed black line shows the unbiased relationship between truth and prediction, with regions above indicating overestimation bias and regions below indicating underestimation bias. The horizontal red dashed line and 2 black dashed lines represent the mean and corresponding 99% CI of empirical estimates of females with cubs for the period 2001–2019. Total annual sightings were restricted to 90% of the empirical range (n = 65–160) for the period 2001–2019. See Fig. 5 of Schwartz et al. (2008) for comparison.

APPENDIX C

Posterior Simulation for Evaluation of GAMs

We take a Bayesian perspective to enhance inference regarding uncertainty in model parameters by positing a distribution of possible parameter values that are consistent with a fitted model, but incorporate the uncertainty of the model fit. Parameters are modeled as being randomly chosen from this distribution, and given the statistical model allows simulations from the distribution $\hat{\beta}|Y$ (Wood 2017). Referred to as posterior inference in the Bayesian literature, this approach allows the uncertainty in parameter values to be represented as a probability distribution, or posterior distribution (Simpson 2018). Parameter values that are more consistent with the data have higher probabilities than those less consistent, providing a more complete picture of estimated uncertainty given the data (Albers et al. 2018, Kruschke 2018).

The process of fitting GAMs involves finding estimates for the coefficients of the underlying "basis functions"; we do not cover this concept here, but it is described in detail in Wood (2017). Together, these coefficients are multivariate normal distributions, with mean vector and covariance matrix specified by the fitted model (Simpson 2018). A single multivariate random draw from this distribution generates a new set of β_j estimates, which represent a slightly altered overall smooth that is consistent with the fitted model, but also represent only one realization of the model uncertainty (Rosenbaum et al. 2019). With a sufficiently large number of random draws, a distribution of plausible fits consistent with the fitted model can be generated to supplement conventional point estimate and confidence interval summary of parameter estimates, a process referred to as posterior simulation (Wood 2017, Simpson 2018). Compared to the dichotomy of being inside or outside a conventional $(1 - \alpha)$ confidence interval, posterior simulation provides the exact proportion of the posterior distribution that is less (or greater) than a critical value (e.g., slope = 0), providing enhanced interpretation and communication of model uncertainty. The approach can be applied to the smoothed GAM estimates (predicted values) and first derivatives (rate of change parameter), and thus serves as a unifying framework for model inferences and outputs.