Received: 17 March 2019 | Accepted: 21 August 2019

DOI: 10.1111/2041-210X.13302

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

When can we trust population trends? A method for quantifying the effects of sampling interval and duration

Hannah S. Wauchope[1] | Tatsuya Amano[1,2,3] | William J. Sutherland[1] | Alison Johnston[1,4]

[1]Conservation Science Group, Department of Zoology, University of Cambridge, Cambridge, UK

[2]Centre for the Study of Existential Risk, University of Cambridge, Cambridge, UK

[3]School of Biological Sciences, The University of Queensland, Brisbane, Australia

[4]Cornell Lab of Ornithology, Cornell University, Ithaca, NY, USA

Correspondence
Hannah Wauchope
Email: hsw34@cam.ac.uk

Funding information
T .A. was supported by the Grantham Foundation for the Protection of the Environment, the Kenneth Miller Trust and the Australian Research Council Future Fellowship, Grant/Award Number: FT 180100354; Cambridge; Cambridge Department of Arcadia

Handling Editor: Nigel Yoccoz

Abstract

1. Species' population trends are fundamental to conservation. They are used to determine the state of nature, and to prioritize species for conservation action, for example through the IUCN red list. It is crucial to be able to quantify the degree to which population trend data can be trusted, yet there is not currently a straightforward way to do so.
2. We present a method that compares trends derived from various samples of 'complete' population time series, to see how often these samples correctly estimate the sign (i.e. direction) and magnitude of the complete trend. We apply our method to a dataset of 29,226 waterbird population time series from across North America.
3. Our analysis shows that, for waterbirds, if a statistically significant ($p < .05$) trend is detected, even from only a few years, it is likely to reliably describe the sign (positive or negative) of the complete trend, but is unlikely to accurately match the percentage change in population per year. If no significant trend is detected, a many-years long sample is required to be confident that the population is truly stable. Furthermore, an insignificant trend is more likely to be missing a decline rather than an increase in the population. Sampling infrequently, but regularly, was surprising reliable in determining trend sign, but poor at determining percentage change per year.
4. By providing percentage estimates of reliability for combinations of sampling regimes and lengths, we have a means to determine the reliability of species population trends. This will increase the rigour of large-scale population analyses by allowing users to remove time series that do not meet a reliability cut-off, or weighting time series by reliability, and could also facilitate planning of future monitoring schemes. While the specific values estimated by our analysis might not be applicable to other taxa or systems, the methods are easily transferable, and we provide the tools to do so.

KEYWORDS
generation length, large-scale analyses, monitoring, population trends, survey, time series, Type M error, Type S error

1 | INTRODUCTION

Many crucial conservation decisions rely on knowing the overall trend of a species or population. This information underpins IUCN red-list classifications (Rodrigues, Pilgrim, Lamoreux, Hoffmann, & Brooks, 2006), many national threatened-species ranking systems (e.g. NESP Threatened Species Recovery Hub, 2018; U.S. Fish & Wildlife Service, 2018) and can convey to policy makers the state of nature globally, regionally and locally (Collen et al., 2009; Gärdenfors, 2001). It is important that decision makers appreciate the extent to which they can trust the apparent trend of a population, both to ensure that at-risk species are not ignored and to avoid misallocating conservation resources towards species that are not actually at risk. The reliability of population trends are poorly understood (McCain, Szewczyk, & Knight, 2016; Wilson, Kendall, & Possingham, 2011), especially when data on variability and measurement error are not available. In addition, many large-scale analyses and policy recommendations (e.g. WWF, 2016) rely on aggregating trends across numerous populations with little guidance on how to weight trends by their likely veracity.

Estimating the trend of a population requires a series of counts over time (typically years, as considered here). Linear, or nonlinear, models are then fit to estimate yearly change (% change per year, e.g. Farmer, Hussell, & Mizrahi, 2007), or modelled counts are compared between years at the start and end of a time period (% change relative to a baseline, e.g. Collen et al., 2009). The number of years of data, sampling frequency, degree of measurement error and population variability all affect the reliability of the derived trend (Magurran et al., 2010). When data are available on measurement error and population variability, power analyses are recommended to estimate degree of reliability in trend estimates (Johnson, Barry, Ferguson, & Müller, 2014; Magurran et al., 2010). Although power analyses are useful where sufficient data are available, there is often insufficient information, especially when assessing many populations, or using existing count data.

Previous studies have attempted to quantify reliability of trends using both simulated and real data. Simulated studies conclude that longer time-scales are needed for better trend estimates, and that there are high margins of error when detecting small population declines (Connors, Cooper, Peterman, & Dulvy, 2014; Fox et al., 2018; Prozt, Peterman, Dulvy, Cooper, & Irvine, 2012; Tománková, Boland, Reid, & Fox, 2013; Wilson et al., 2011). Studies working with real data on diverse taxa have found that populations exhibiting a particular trend across one time-interval often show an opposing trend in later years (Dunn, 2002; Keith et al., 2015). Others have assessed the number of years needed to reliably estimate a trend with a certain percentage of accuracy. For example White (2019) found that for a trend to be accurate to within 2% change per year, samples needed to be anywhere from 5 to 30 years in length, depending on the taxa. Others have estimated the number of years required for an accurate estimate to be between 10 (Rueda-Cediel, Anderson, Regan, Franklin, & Regan, 2015 for a snail species) and 21 years (Reynolds, Thompson, & Russell, 2011 for brown bears). These investigations are useful for gaining an approximate idea of reliability, but do not provide a straightforward way for a study to assign a value of reliability to population time series of varying lengths (i.e. numbers of years).

Therefore, in the absence of guidance, studies based on population trends often lack the data to make any quantification of uncertainty (e.g. Craigie et al., 2010; Loh et al., 2005). Furthermore, most studies assume that there is a 'true' trend exhibited by each population, but populations rarely demonstrate one linear trend continuously through time, rather fluctuating in response to the positive and negative pressures affecting them.

We propose a modified version of White (2019)'s method to quantify uncertainty in trend estimates. Our analyses hinge on the concept of comparing the trend derived from a 'sample' (a subset of the full set of counts for a population) to the 'complete' trend of that population, derived from the full set of counts (White, 2019). We have chosen to use the word 'complete' in this study rather than 'true' as even with yearly counts we cannot claim to know the true trend of a population. Normally, one would possess only the sample, and we therefore hope to provide an estimate of how likely that sample is to represent the complete trend, regardless of sample length or complete trend length. In our analysis we quantify reliability both in terms of trend sign and magnitude of change (Gelman & Carlin, 2014).

We ask two questions: (a) How reliable are trends derived from a certain number of years of data, based on the time over which a trend estimate is desired? For example how well do five consecutive years of survey data represent the trend of a population over 10 years?; and (b) How reliable are trends derived from data sampled at different intervals, such as samples taken every year over a 30-year period, compared to every 5 years over the same period? We also investigate two factors that we expect to impact reliability: species generation time and shape of the complete trend. We expect that species with longer generation times will require longer survey periods, as there will be a lag before populations show responses to changes in birth rate, as older individuals are still living (Kuussaari et al., 2009). We also expect that trends estimated from samples of populations with complex nonlinear complete trends will be less accurate than samples from populations with linear or near-linear complete trends.

As a case study, we use an empirical dataset of yearly counts of 129 waterbird species at 1,110 sites in North America (a total of 29,226 site by species combinations). Providing these estimates for waterbird data are particularly beneficial as data on waterbirds are available at large spatio-temporal scales and waterbird studies can provide insights into broader conservation goals (Amano et al., 2018; Amat & Green, 2010; Piersma & Lindström, 2004). However, our methods are general, and we provide code and instructions to generalize to other taxa. Our work provides an explicit measure of the reliability of a trend and gives an evidence-based justification for excluding samples below a certain length, according to the degree of confidence desired for a study. Finally, these results can be used to plan multi-species monitoring programs, to give the highest likelihood of capturing representative trends for the most species.

2 | MATERIALS AND METHODS

2.1 | Data preparation

We obtained an initial dataset of yearly count data for 174 waterbird species in North America from the Christmas Bird Count (CBC; Dunn et al., 2005, see Supporting Information Section 1 for details) at 1,123 sites spanning the years 1966 to 2013 (Amano et al., 2018), from which 30 years of consecutive counts were taken for each site by species combination. We selected 30 years because it was a long-term survey period, but sufficiently short that adequate data were still available. In cases where a site was sampled for over 30 years, the most recent 30 years were taken. We considered each species at each site as an independent population; as we were not attempting to estimate the trends of entire species, correlations between sites were irrelevant.

Christmas Bird Count data have variable sampling effort, which must be accounted for in the modelling process. The most common expected relationship between effort and detection is a linear relationship between log-transformed count and effort. Following Butcher and McCulloch (1988) and Xu, Barrett, Lank, and Ydenberg (2015), we chose to retain only those species where a significant linear relationship between detection and log of effort was shown, found by running a negative binomial generalized linear model (see modelling specifications, below) for each species, at all years and sites:

$$E\left(\text{Count}_i\right) = g^{-1}\left(\alpha + \beta \log\left(e_i\right)\right) \tag{1}$$

$$\text{var}\left(\text{Count}_i\right) = v_{NB}\left(E\left(\text{Count}_i\right)\right) \tag{2}$$

The link function $g(\cdot)$ is 'log', so the inverse is an exponential. The expected value of Count for species i is predicted by an intercept, α, the log of effort (in hours), e, and it's coefficient, β (Equation 1). The variance of our count data is defined as negative binomial (Equation 2). Any species found to have a non-significant β (i.e. no relationship between effort and detection) were removed from analysis, as were those with a significant, but negative β (i.e. as effort increased, detection decreased). We then included survey hours as an offset term in our models to account for this sampling effort.

We also removed any populations with a sum of less than 30 observations over the 30-year sampling period, to remove populations with mostly zero counts. This resulted in our final dataset of 29,226 populations, comprising 129 species at 1,110 sites (sees Supporting Information Section 2 for species list and site map).

As species varied in the extent to which they occurred at sites, we also ran our analysis on a standardized subset of the data: 99 species with 50 randomly selected sites each, 4,950 populations in total. Even though this dataset was less than 20% of the size of our full dataset, the results were highly congruent (Supporting Information Section 5).

2.2 | Modelling specifications

To estimate the population growth rate, r, with population counts as the response variable and years as the explanatory variable, we used generalized linear models (GLMs) run with the R package stats (R Core Team, 2017). We included effort using the 'offset' parameter, which allows a covariate with a known slope to be included in the model. For count data it is usual to use Poisson, quasi-Poisson or negative binomial distributions for the response, with the latter two being more appropriate if there is over-dispersion, where the variance of the response variable exceeds the mean. In our dataset 99.7% of samples were over-dispersed, with 77% of these by at least an order of magnitude. We therefore ran our models using the negative-binomial distribution, though our provided code allows specification of any of these three distributions.

Mathematically, the above model is expressed as the following:

$$E\left(\text{Count}_t\right) = g^{-1}\left(\alpha + \beta x_t + \log\left(e_t\right)\right) \tag{3}$$

$$\text{var}\left(\text{Count}_t\right) = v_{NB}\left(E\left(\text{Count}_t\right)\right) \tag{4}$$

As before, the link function $g(\cdot)$ was 'log', so the inverse is an exponential. The expected value of Count in year t is predicted by an intercept, α, the year coefficient, β, multiplied by the year value, x, and the log of effort (in hours), e (Equation 3). Because the relationship between effort and count is known (i.e. a log linear relationship), it does not need a coefficient. As before, the variance of our count data is defined as negative binomial (Equation 4).

For each model the population growth rate, r, and p-value of r were determined. r was obtained by raising e to the power of β (i.e. $r = e^{\beta}$). This value represents the population change per year: values above 1 indicate the population is increasing (e.g. 1.03 would indicate a 3% increase in the population each year) and values below 1 indicate the population is declining (e.g. 0.98 would indicate a 2% decrease in the population each year). p is obtained directly from the model output. For our main analysis, we followed the convention of setting a significance level of $p < .05$. This is an arbitrary threshold, and circumstances may arise where the risk of missing a trend is greater than the risk of erroneously concluding there is one (e.g. a high-risk group of species), in which case it is better to set a higher p-value (Field, O'Connor, Tyre, & Possingham, 2007; Taylor & Gerrodette, 1993), and *vice versa*. Our code provides the ability to adjust the p-value that defines significance if this is desired. In addition, Supporting Information Section 6 explores trend accuracy when no significance level cut-off is applied.

All models were run in R version 3.4.1 (R Core Team, 2017) using the Cambridge Service for Data Driven Discovery High Performance Computing service (https://www.hpc.cam.ac.uk, last accessed 26 September 2019).

2.3 | Complete trend

As discussed in the Introduction, we wished to investigate sample reliability regardless of the length of the complete time series. We sampled from the complete time series in two ways: in consecutive years and in intervals of years; these are defined in detail below. For

Consecutive Sampling we varied the length of the complete time series from 4 to 30 years (the maximum number of years available in our data). We would take complete trends from all possible subsets of the full 30 years of data (Figure 1a), and then take samples from each of these (Figure 1b i). However, because Interval Sampling already has two dimensions (interval length and number of years sampled), we compared interval samples only to a complete time series length of 30 years (Figure 1b ii. Note that though this figure shows intervals being compared to a complete time series of 7 years, this is for simplicity only and all analysis compared intervals only to 30 year complete time series).

2.4 | Sampling methods (consecutive and interval)

For Consecutive Sampling (Figure 1b i), that is sampling from consecutive years, we took shorter adjacent subsets from a complete

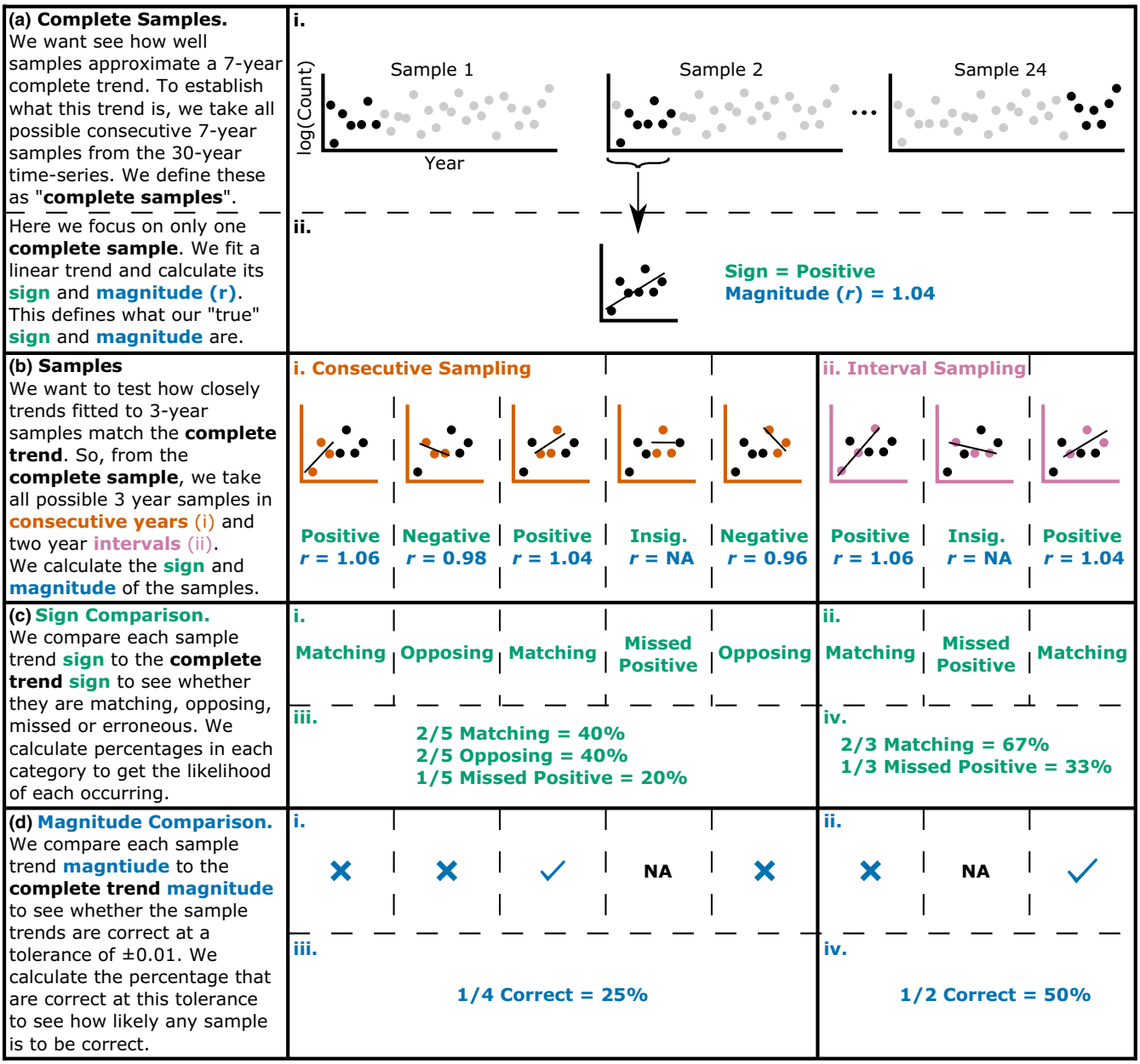

FIGURE 1 Schematic of methods with hypothetical data. Sections are explained in left hand column, focussing on one example from each step. A reminder that because magnitude represents proportion population change per year, values less than one indicate a negative sign. In (a) 7-year complete time series (bold black) are taken from the full 30 years of data, in actual analysis complete lengths range from 4 to 30 years. In (b i.) 3-year consecutive samples are taken (orange), in actual analysis these range from three years to complete time-series length minus one. In (b ii.) samples of three years at two year intervals are taken (pink), in actual analysis these range from 3 to 29 year samples taken in 1 to 14 year intervals. NOTE interval samples are only taken from complete time series of 30 years in analysis, shown here from 7 years for simplicity. In (c), Missed Negative or Erroneous categories are not shown, as they only occur when the complete trend is negative or insignificant respectively (see Table 1). In (d) only a tolerance of ±0.01 is shown, in actual analysis tolerances range from ±0.01 to ±0.5

dataset of *n* years in length. We sampled all possible consecutive subsamples from 3 years to *n* − 1 years (see also White, 2019). For Interval Sampling, we took samples at regular intervals from within the complete dataset (Figure 1b ii): we varied the interval length (i.e. samples taken every *x* years) from 1 year (i.e. consecutive years) to 14 years (i.e. samples taken every 14 years, either at years 1, 15 and 29, or 2, 16 and 30 years) and took all possible numbers of samples within these iterations to fill the 30-year period (e.g. 3 samples taken every 4 years could be samples taken at years 1, 5 & 9; 2, 6 & 10; 3, 7 & 11 etc).

2.5 | Comparison methods

We used two ways to assess whether a sample trend (Consecutive or Interval) was representative of a complete trend, as per Gelman & Carlin (2014). First, we took the sign (i.e. direction) of the trend, defining it as positive, negative or insignificant. Using this, a sample trend would be classified as matching if it was the same sign as the complete trend; opposing if it was the opposite sign; an erroneous positive or negative if it was positive or negative, respectively, but the complete trend was insignificant; and a missed positive or negative if it was insignificant, but the complete trend was positive or negative, respectively (Table 1). We term this 'Sign Comparison' (Figure 1c). Note that we conducted a final supplementary analysis considering how often insignificant sample trends still approximate the sign of significant complete trends (see Supporting Information Section 6).

Second, for cases where a significant trend was obtained from *both* the sample and complete time series (i.e. cases of 'Matching' or 'Opposing' from the Sign Comparison method), we considered the absolute difference between population growth rate *r* of the two; giving an idea of the degree of 'correctness'. That is, difference = $|r_{sample} - r_{complete}|$. We defined tolerance levels ranging from ±0.01 to ±0.5 and, if the difference was less than the tolerance level, the sample trend represented the complete trend and was correct and if it did not it was incorrect. We term this 'Magnitude Comparison' (Figure 1d).

In all cases, we obtained a sample *r* and a complete *r* for each population, the significance level of each, and then compared them to give a category for representativeness (using either the Sign or Magnitude Comparison method). We then found, for each combination of sample and complete time-series lengths and sampling types, the percentage of our 29,226 populations in each representativeness category (Figure 1c iii, iv; Figure 1d iii, iv).

2.6 | Generation length

We considered generation length as a major factor that is likely to influence the duration of sampling required. This is because long-lived species often take longer to show responses to environmental pressures, as older individuals can continue to survive even if recruitment is falling (Kuussaari et al., 2009). To assess this, we divided our species into three groups based on generation length: short (1–5 years), medium (6–10 years) and long (11–15 years). Generation length data were obtained from birdlife.org species fact sheets (e.g. http://datazone.birdlife.org/species/factsheet/ruddy-turnstone-arenaria-interpres/details, last accessed 26 September 2019). We then organized our standard analysis according to these three categories.

2.7 | Trend shape

Thus far, our analysis has only considered that trends can be linear. To assess how our results are affected by trends of different shapes (i.e. nonlinear trends), we used generalized additive models (GAMs) with the R package MASS (Venables & Ripley, 2002). These nonparametric models allow for nonlinear relationships. We ran GAMs on all complete 30 year trends, model specification was the same as the GLMs but with a smoothing term on year, and took the estimated degrees of freedom (EDF) for each. EDFs ranged from 1 to 8.57, so we divided our trends into four shape groups, linear and quadratic up to cubic (EDF = 1–2.99), cubic or low degree polynomial (EDF = 3–4.99), mid-degree polynomial (EDF = 5–6.99) and high degree polynomial (EDF = 7–8.99). We then organized our standard analysis (using GLMs) according to these four categories, to see whether complete trend shape affects sample representativeness.

3 | RESULTS

See Supporting Information Section 2 for summary statistics of the full dataset. All data used to produce plots are also provided, as well as code to reproduce full results on any set of population counts (See Supporting Information Section 7).

3.1 | Sign Comparison of Consecutive Sampling

Reliable estimation of the sign of complete trends required many years of data for Consecutive Sampling. For example to have an

TABLE 1 Categories used for Sign Comparison. Columns show sign of trend derived from complete time series and rows show sign of trend derived from sample time series. Cells show category, based on sample and complete trend sign

	Complete trend sign		
Sample trend sign	**Positive**	**Negative**	**Insignificant ($p > .05$)**
Positive	Matching	Opposing	Erroneous Positive
Negative	Opposing	Matching	Erroneous Negative
Insignificant ($p > .05$)	Missed Positive	Missed Negative	Matching

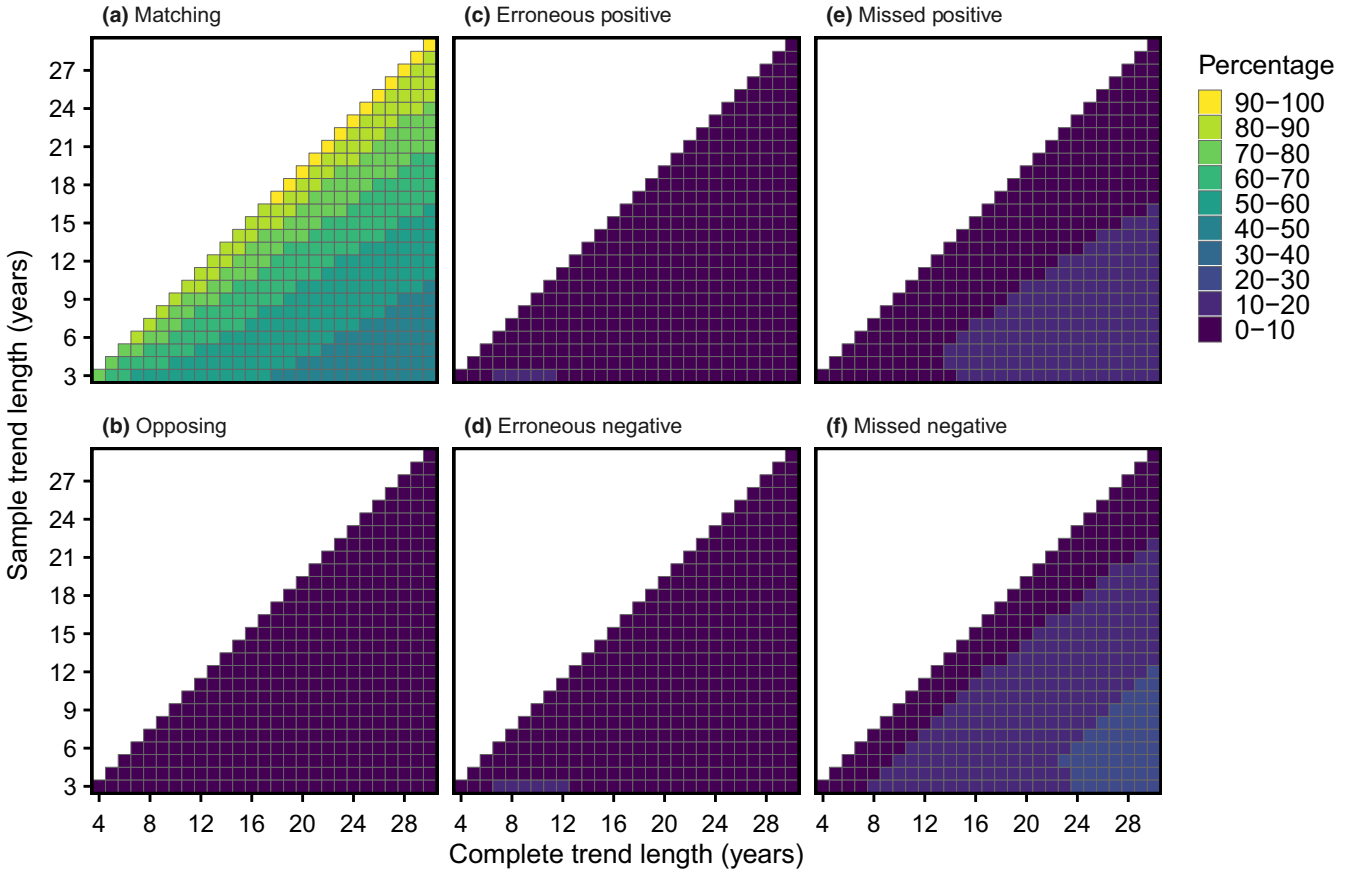

FIGURE 2 Sign comparison using Consecutive Sampling. Colour shows percentage of sample trends that, relative to the complete trend, were matching (a) opposing (b) an erroneous positive (c)/negative (d) or a missed positive (e)/negative (f) (see Table 1). Shown for all combinations of sample lengths (*y*-axis), and complete lengths (*x*-axis)

80%–100% chance of a sample trend having the same sign as a complete trend, the sample time series needed to be almost as long as the complete time series (Figure 2a). However, sample trends opposed the complete trend less than 10% of the time (Figure 2b).

The chance of an erroneous positive or negative (i.e. the sample indicated a significant trend but the complete trend did not reflect this, Figure 2c,d) was low regardless of the length of the sample or complete trend. However, the chances of a missed positive or negative trend (i.e. the complete trend had a significant sign but the sample did not detect this) were higher (see also Supporting Information Section 6); missed negatives were more likely than missed positives, and both were more likely when the sample time series was considerably shorter than the complete time series (Figure 2e,f). This implies that, particularly when trying to detect declines, shorter samples have low power to detect complete trends, but if they do detect a significant trend it is likely to be representative.

3.2 | Sign Comparison of Interval Sampling

Our results show that sampling in intervals can be more representative than sampling in consecutive years, when considering trend sign. For example, sampling for 24 consecutive years (out of a 30-year complete time series) gave the matching result 70%–80% of the time (Figure 3a, bottom row, note this is equal to Figure 2a rightmost column), but the same level of reliability could be obtained by sampling 13 times every second year (Figure 3a, second row). More strikingly, four samples taken every 9 years (Figure 3a, second column, top cell) gave the same percentage matching (60%–70%) as up to 20 years of Consecutive Sampling (Figure 3a, bottom row).

As with Consecutive Sampling, the percentage opposing for Interval Sampling was very low (Figure 3b), the chance of making an erroneous positive or negative was also very low for all sampling combinations (Figure 3c,d) and, though missed positives and negatives were slightly more likely, the likelihood of missed trends never exceeded 40% (Figure 3e,f). As before, missed negatives were more likely than missed positives.

3.3 | Magnitude Comparison of Consecutive Sampling

When comparing the growth rate (*r*) of sample trends to complete trends, sample trends were regularly correct only at very high tolerances. Note that for ease of interpretation these results are displayed at four complete trend lengths: 5, 10, 20 and 30 years. In order for the sample trend *r* to be within ±0.1 (i.e. 10% population change per year) of the complete trend *r* 80% of the time, the

sample time series needed to be at least 9 years when compared to a complete time series of 10 years (Figure 4b); at least 15 years when compared to a 20-year complete time series (Figure 4c); at least 19 years when compared to a 30-year complete time series (Figure 4d); and could not be attained when the complete time series was 5 years long (Figure 4a). A sample of 29 years only estimated a trend within 1% (±0.01) of the 30-year trend in ~80% of situations (Figure 4d).

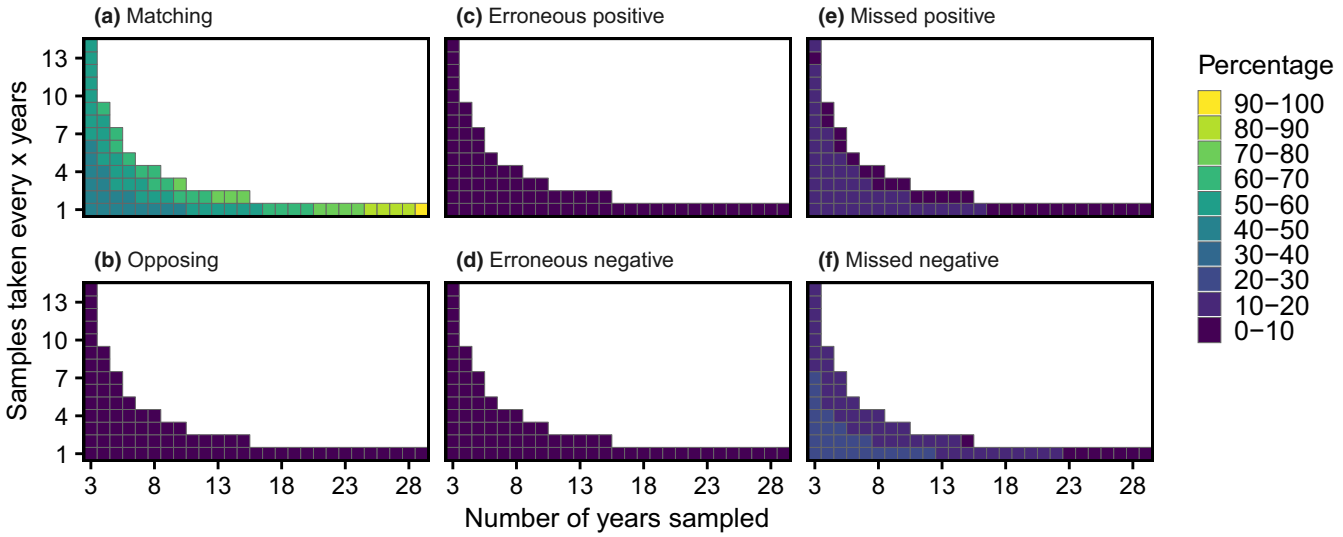

FIGURE 3 Sign comparison using Interval Sampling. Colour shows percentage of sample trends that, relative to the complete trend, were matching (a) opposing (b) an erroneous positive (c) or negative (d) or a missed positive (e) or negative (f) (see Table 1). Shown for all combinations of Interval Sampling, with number of years sampled (*x*-axis) and interval length (*y*-axis). Thus 8 on the *x*-axis and 4 on the *y*-axis would mean 8 samples were taken, one every 4 years. The bottom row of each plot is equal to the right most column of the equivalent Figure 2 plot, but is included here to ease comparison. Complete trend length is always 30 years

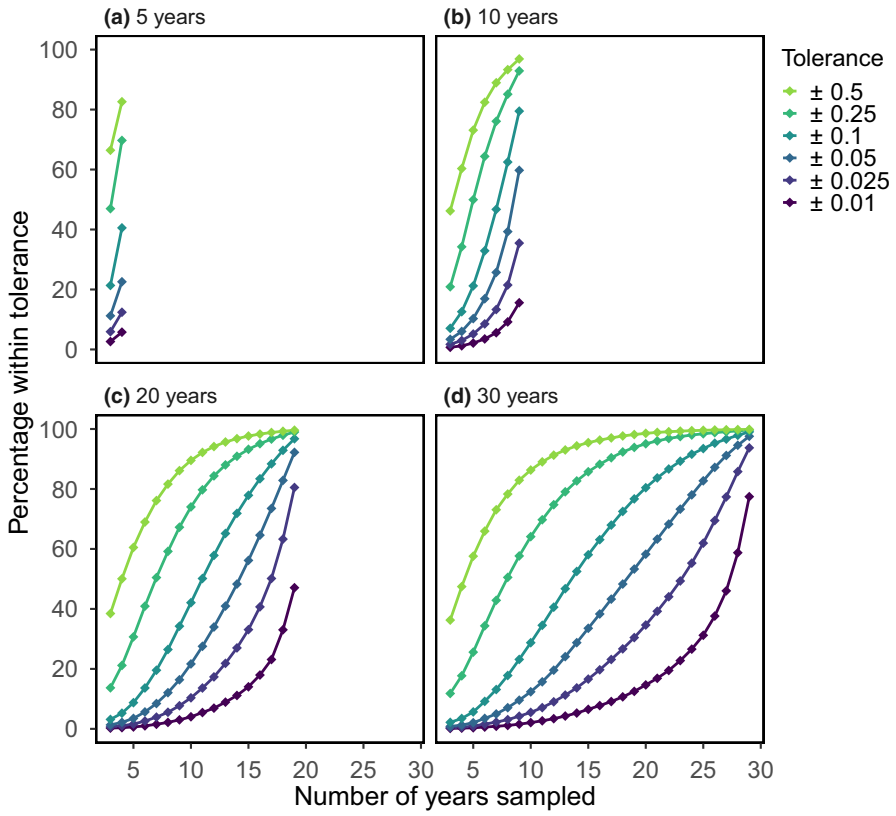

FIGURE 4 Magnitude comparison using Consecutive Sampling. Lines show percentage of sample trends that correctly estimate complete trends (*y*-axis), measured by whether the sample *r* matched the complete *r* within the tolerance (colours show different tolerances). Shown for four complete trend lengths, 5 (a), 10 (b), 20 (c) and 30 (d) years, and for all sample lengths (*x*-axis)

3.4 | Magnitude Comparison of Interval Sampling

Interval Sampling gave better results when comparing trend sign (Results 3.2), however did not perform as well for estimating magnitude of change. It was not possible for even 50% of sample trends to be correct at low thresholds (±0.01–0.025; Figure 5). The shape of curves indicates that high percentages could be attained with enough years of sampling at large intervals, but this would mean sampling over very long time-scales. Better reliability was achieved at higher tolerances, but only at ±0.5 and ±0.25, that is 25%–50% population change per year.

3.5 | Generation length and trend shape

When considering the Sign Comparison method, short samples of populations with long generation-lengths were less likely to match the complete trend (Figure 6, Supporting Figure S3). For example to have at least a 50% chance of the sample trend matching a 30-year complete trend, only a 4-year sample was required for short generation-length species (Figure 6a, 'Matching', bottom right corner), but an 11-year sample was required for long generation-length species (Figure 6c, 'Matching', last column, 11th cell). Erroneous trends and opposing trends were roughly equal among populations of different generation-lengths (Figure 6, Supporting Figure S3). Populations of different generation-lengths performed similarly according to the Magnitude Comparison method (Supporting Figures S4 and S5).

There was a surprising amount of consistency in results when considering trend shape (i.e. linear vs. nonlinear). When using the Sign Comparison method (Supporting Figures S6 and S7), percentages of matching, opposing and missed positives/negatives remained stable as trend complexity increased, though erroneous positives and negative were less common for simpler trend shapes (low estimated degrees of freedom). Similarly, trend shape did not seem to affect the percentage of samples that were correct according to the Magnitude Comparison method (Supporting Figures S8 and S9).

4 | DISCUSSION

In this paper, we provide and test a method that can estimate reliability of population trends of different lengths and sampling types, based on the total time over which a trend estimate is desired. Our

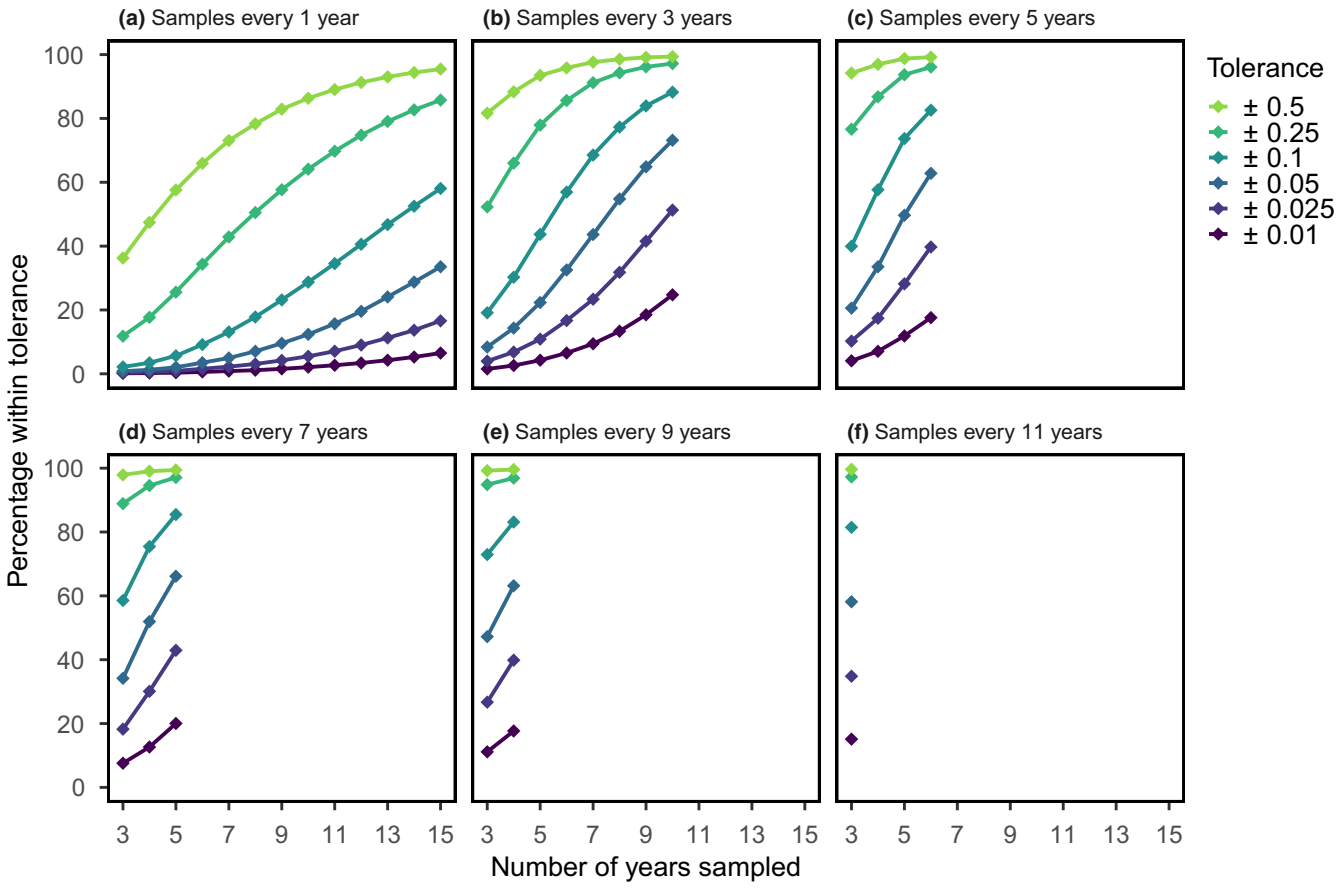

FIGURE 5 Magnitude comparison using Interval Sampling. Lines show percentage of sample trends that correctly estimate complete trends (*y*-axis), measured by whether the sample *r* matched the complete *r* within the tolerance (colours); samples taken using Interval Sampling. Shown for six stages of Interval Sampling: samples taken either every 1 (a), 3 (b), 5 (c), 7 (d), 9 (e) or 11 (f) years (facets), for all possible numbers of years sampled (*x*-axis). Complete trend length is always 30 years. Note that panel a is equal to Figure 4d (with a truncated *x* axis)

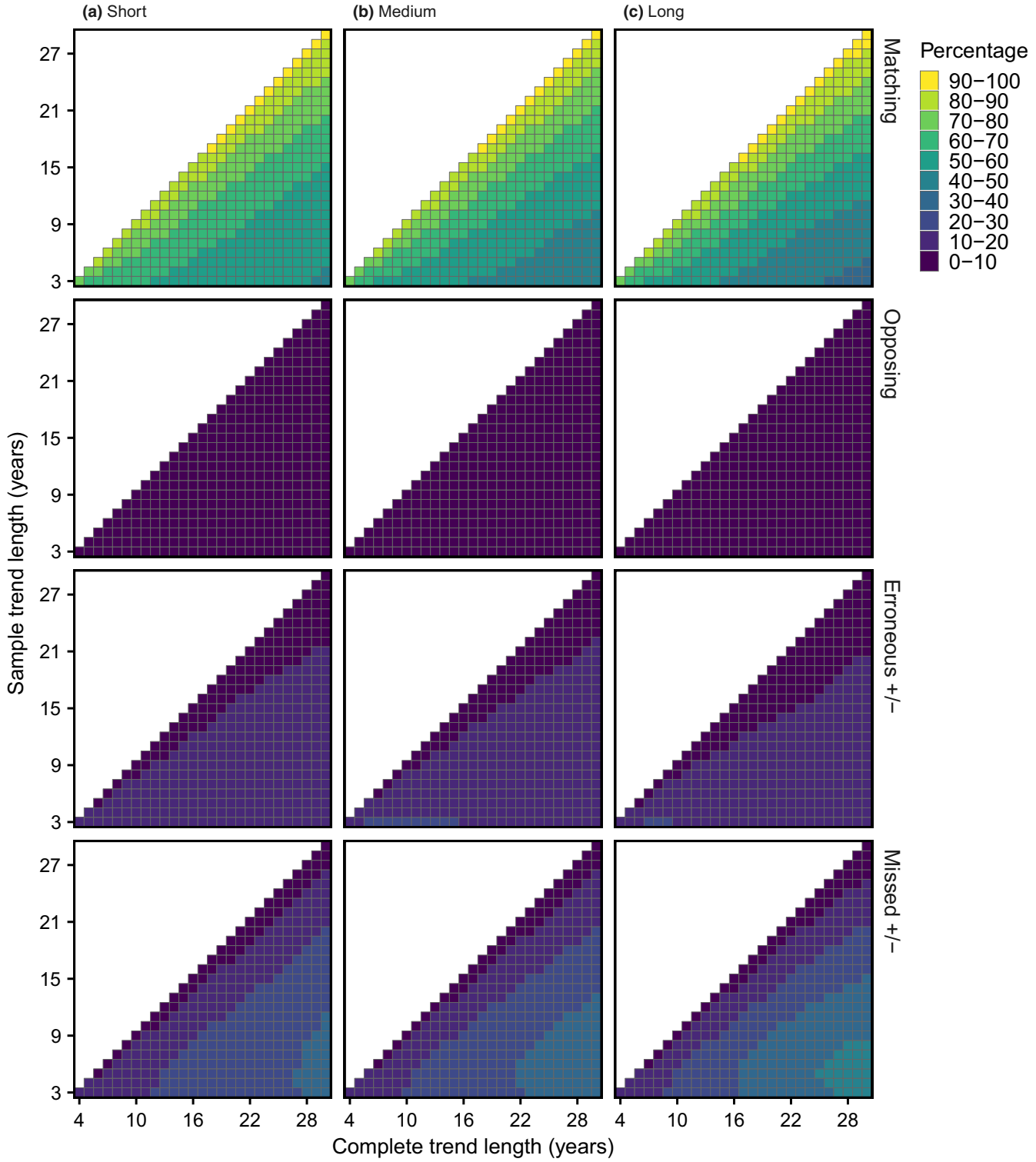

FIGURE 6 Sign comparison using Consecutive Sampling, for populations of different generation lengths. Colour shows percentage of sample trends that, relative to the complete trend, were matching, opposing, an erroneous positive/negative or a missed positive/negative (see Table 1). Shown for all combinations of sample lengths (*y*-axis), and complete lengths (*x*-axis). Divided by populations with either a) short (1–5 years), b) medium (6–10 years) or c) long (11–15 years) generation lengths

results are derived from an entirely empirical dataset with no simulations and they show a high amount of convergence (e.g. Figure 4), indicating that our sample sizes are large enough to give a reliable estimate of likelihood for each category. Our results were robust to standardization of the number of populations per species, subsetting of species into three separate groups based on generation length (see also White, 2019) and subsetting of populations into different trend shapes. We discuss the meaning of our results in the context

of trusting population trends, how our methods can be adapted to other taxa, and how results from our methods can be used to quantify reliability in large-scale studies of population trends and to design future monitoring schemes.

4.1 | Waterbird case study

Our results show that if a *significant* trend is obtained from a population time series, even when the data are from a few years, it is likely to reflect the longer term trend sign (direction) of the population, though not necessarily the magnitude. Keith et al. (2015) studying birds similarly found that past population trajectory was a good predictor of future population trajectory and Møller, Rubolini, and Lehikoinen (2008) and Sanderson, Donald, Pain, Burfield, and Bommel (2006) found a similar, though weak, correlation between trends of migratory birds between 1970–1990 and 1990–2000. However, we show that to be confident that an *insignificant* trend (implying a stable population) is representative, one must sample for many years. Further, if an insignificant trend is obtained, it is more likely to be missing a decline than an increase in a population, and we suggest caution with conclusions and decisions from insignificant trends. It should be noted that Keith et al. (2015) found that past trajectories were not a good predictor of future trajectories in mammals, salmon and other fish, meaning this method should be tested on other taxa using relevant data.

Sampling in intervals provided surprisingly accurate results that were better when compared to sampling the same number of years consecutively. When sampling a fixed number of times, accuracy increased with the distance between each sample. Presumably a limit exists at some point, but according to these results it is greater than 14 years. Other studies have found similar results (Starcevich, Irvine, & Heard, 2018; Urquhart, Paulsen, & Larsen, 1998), for example Reynolds et al. (2011) found that surveying brown bears every 10 years gave similar model performance to surveying in 3 out of every 5 years. Interval sampling could allow, say, a greater number of sites to be surveyed over a given area (Buckland & Johnston, 2017). However it is not always practical, especially for high-risk species where declines may need to be detected and acted on quickly. In addition, while interval sampling could be good for cheaply obtaining trend estimates, it is not a replacement for long-term monitoring that takes yearly samples, which can provide data for analyses considering drivers of population change.

In cases where analyses or management decisions depend not only on the sign of a population trajectory, but the actual rate of change, we find that sample trends are much less likely to be representative of complete trends. To be 80% reliable at a rate of population change of 1%–2.5% per year over 30 years, one must sample for at least 19 consecutive years. These results are roughly similar to White (2019) who found, with diverse taxa, that to detect a population change of 1% per year one would need to sample for 25–30 years; although a study of an invertebrates found diminishing returns on accuracy for samples greater than 10–15 years in length (Rueda-Cediel et al., 2015). In our study, sampling in intervals also struggled to produce accurate results: 80% reliability was never achieved if samples had to be correct at anything less than a rate of change of 10% per year (i.e. drastic population change).

Finally, we found that generation length appears to have some impact on results, with short samples from longer lived species less likely to be accurate. This is probably because longer generation times can create a lag in population responses to pressures (Kuussaari et al., 2009). The chance of making an erroneous conclusion, missing a trend, or falsely identifying a trend remained consistent regardless of generation length. These results were exploratory, not hypothesis testing, and should therefore not be considered conclusive. This would be an interesting area for further research.

This analysis was kept simple by restricting it to single location time series, and by restricting sampling to a minimum of once every year. Increasing the number of samples in each time period can improve confidence and accuracy in derived trends (Atkinson et al., 2006) and it is likely that percentage reliability in derived trends would increase if this was considered. Some studies have found that sampling at more locations improved trend detection better than longer time series (Sims, Wanless, Harris, Mitchell, & Elston, 2006, though see Schumann, Dann, Hoskins, & Arnould, 2013) and so modelling trends from multiple locations is also likely to improve our reliability estimates (Rhodes & Jonzén, 2011).

4.2 | Applications

For those working with large datasets who cannot conduct power analysis or quantify measurement error in their populations, reliability can be assigned to trends using the data from this study (see Supporting Information Section 7), or data produced using these methods with other taxa (using the provided code, see Data Accessibility). The user should define a target 'complete' trend length for the study, and extract the reliability estimates for this complete trend length. After this, two options are possible: (a) assign reliability estimates to each time series based on how long it is, and weight analysis according to these estimates or (b) select a threshold (e.g. time series must be at least 80% likely to represent the complete trend) and remove any time series that do not meet this criteria. We readily agree that our values are not infallible: they could vary with location, time period or taxa. However, this is an improvement on making an arbitrary cut-off point, or having no way of weighting population trends.

These results could also be used to help plan future monitoring schemes, but we would advise this be done cautiously. The goals of monitoring have been subject to much discussion (Hauser, Pople, & Possingham, 2006; Legg & Nagy, 2006; Lindenmayer & Likens, 2009; McDonald-Madden et al., 2010; Nichols & Williams, 2006), but in cases where programs are carried out with the goal of detecting trends (Marsh & Trenham, 2008) we can add our support to previous works (e.g. Rhodes & Jonzén, 2011; Rueda-Cediel et al., 2015; White, 2019) that suggest that many years of monitoring is essential to accurately capturing population trends. We also suggest that resources could be conserved and possibly allocated

to more locations or taxa if sampling is conducted in intervals rather than every year.

5 | CONCLUSIONS

In this age of increasing large-scale analyses, we believe the scientific community can do better at making informed decisions around uncertainty and reliability. Our methods and results provide a clear and quantitative way to add rigour to large-scale population analyses. We advocate an end to arbitrary cut-offs, and recommend that, where possible, users instead consider methods such as ours to quantify reliability and make decisions about their data accordingly. Our methods are fully transferable to other taxa, and the concepts can also be transferred to areas outside of population ecology.

ACKNOWLEDGEMENTS

CBC Data are provided by National Audubon Society and through the generous efforts of Bird Studies Canada and countless volunteers across the western hemisphere. H.S.W. was supported by a Cambridge Trust Cambridge-Australia Scholarship and the Cambridge Department of Zoology JS Gardiner Fellowship. W.J.S. is funded by Arcadia. We thank Richard Fuller, Benno Simmons and Martin Wauchope for helpful comments and discussion. Finally, we thank two anonymous reviewers for highly constructive feedback and the suggestion of Figure 1.

AUTHORS' CONTRIBUTIONS

H.S.W., T.A., W.J.S. and A.J. conceived the ideas and designed methodology. T.A. collated the data; H.S.W. analysed the data and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

All data used in our analysis are available from the Christmas Bird Count. Audubon, who manage the data, denied our request to archive the data in accordance with the BES Data Archiving Policy due to trademark issues, but are happy to provide it upon request at http://netapp.audubon.org/cbcobservation/. Code, which is written to be fully adaptable to other data, is available both in the Supporting Information and on GitHub, (https://github.com/hannahwauchope/TrustingTrends, https://doi.org/10.5281/zenodo.3417001; Wauchope, 2019) we recommend using the GitHub code as this will remain up to date.

ORCID

Hannah S. Wauchope https://orcid.org/0000-0001-5370-4616
Tatsuya Amano https://orcid.org/0000-0001-6576-3410
William J. Sutherland https://orcid.org/0000-0002-6498-0437
Alison Johnston https://orcid.org/0000-0001-8221-013X

REFERENCES

Amano, T., Székely, T., Sandel, B., Nagy, S., Mundkur, T., Langendoen, T., … Sutherland, W. J. (2018). Successful conservation of global waterbird populations depends on effective governance. *Nature*, *553*, 199. https://doi.org/10.1038/nature25139

Amat, J. A., & Green, A. J. (2010). Waterbirds as bioindicators of environmental conditions. In C. Hurford, M. Schneider, & I. Cowx (Eds.), *Conservation monitoring in freshwater habitats: A practical guide and case studies* (pp. 45–52). Dordrecht: Springer Netherlands.

Atkinson, P. W., Austin, G. E., Rehfisch, M. M., Baker, H., Cranswick, P., Kershaw, M., … Maclean, I. M. D. (2006). Identifying declines in waterbirds: The effects of missing data, population variability and count period on the interpretation of long-term survey data. *Biological Conservation*, *130*, 549–559. https://doi.org/10.1016/j.biocon.2006.01.018

Buckland, S. T., & Johnston, A. (2017). Monitoring the biodiversity of regions: Key principles and possible pitfalls. *Biological Conservation*, *214*, 23–34. https://doi.org/10.1016/j.biocon.2017.07.034

Butcher, G. S., & McCulloch, C. E. (1988). Influence of observer effort on the number of individual birds recorded on Christmas bird counts. *Biological Reports*, *90*, 120–129.

Collen, B., Loh, J., Whitmee, S., McRae, L., Amin, R., & Baillie, J. E. M. (2009). Monitoring change in vertebrate abundance: The living planet index. *Conservation Biology*, *23*, 317–327. https://doi.org/10.1111/j.1523-1739.2008.01117.x

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

Craigie, I. D., Baillie, J. E. M., Balmford, A., Carbone, C., Collen, B., Green, R. E., & Hutton, J. M. (2010). Large mammal population declines in Africa's protected areas. *Biological Conservation*, *143*, 2221–2228. https://doi.org/10.1016/j.biocon.2010.06.007

Dunn, E. H. (2002). Using decline in bird populations to identify needs for conservation action. *Conservation Biology*, *16*, 1632–1637. https://doi.org/10.1046/j.1523-1739.2002.01250.x

Dunn, E. H., Francis, C. M., Blancher, P. J., Drennan, S. R., Howe, M. A., Lepage, D., … Smith, K. G. (2005). Enhancing the scientific value of the Christmas Bird Count. *The Auk*, *122*, 338–346. https://doi.org/10.1093/auk/122.1.338

Farmer, C. J., Hussell, D. J. T., & Mizrahi, D. (2007). Detecting population trends in migratory birds of prey. *The Auk: Ornithological Advances*, *124*, 1047–1062. https://doi.org/10.1093/auk/124.3.1047

Field, S. A., O'Connor, P. J., Tyre, A. J., & Possingham, H. P. (2007). Making monitoring meaningful. *Austral Ecology*, *32*, 485–491. https://doi.org/10.1111/j.1442-9993.2007.01715.x

Fox, R., Harrower, C. A., Bell, J. R., Shortall, C. R., Middlebrook, I., & Wilson, R. J. (2018). Insect population trends and the IUCN Red List process. *Journal of Insect Conservation*, *23*, 269. https://doi.org/10.1007/s10841-018-0117-1

Gärdenfors, U. (2001). Classifying threatened species at national versus global levels. *Trends in Ecology & Evolution*, *16*, 511–516. https://doi.org/10.1016/S0169-5347(01)02214-5

Gelman, A., & Carlin, J. (2014). Beyond power calculations. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Hauser, C. E., Pople, A. R., & Possingham, H. P. (2006). Should managed populations be monitored every year? *Ecological Applications*, *16*, 807–819. https://doi.org/10.1890/1051-0761(2006)016[0807:SMPBME]2.0.CO;2

Johnson, P. C. D., Barry, S. J. E., Ferguson, H. M., & Müller, P. (2014). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution*, *6*, 133–142. https://doi.org/10.1111/2041-210X.12306

Keith, D., Akçakaya, H. R., Butchart, S. H. M., Collen, B., Dulvy, N. K., Holmes, E. E., … Waples, R. S. (2015). Temporal correlations in

population trends: Conservation implications from time-series analysis of diverse animal taxa. *Biological Conservation*, *192*, 247–257. https://doi.org/10.1016/j.biocon.2015.09.021

Kuussaari, M., Bommarco, R., Heikkinen, R. K., Helm, A., Krauss, J., Lindborg, R., … Steffan-Dewenter, I. (2009). Extinction debt: A challenge for biodiversity conservation. *Trends in Ecology & Evolution*, *24*, 564–571. https://doi.org/10.1016/j.tree.2009.04.011

Legg, C. J., & Nagy, L. (2006). Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management*, *78*, 194–199. https://doi.org/10.1016/j.jenvman.2005.04.016

Lindenmayer, D. B., & Likens, G. E. (2009). Adaptive monitoring: A new paradigm for long-term research and monitoring. *Trends in Ecology & Evolution*, *24*, 482–486. https://doi.org/10.1016/j.tree.2009.03.005

Loh, J., Green, R. E., Ricketts, T., Lamoreux, J., Jenkins, M., Kapos, V., & Randers, J. (2005). The living planet index: Using species population time series to track trends in biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*, 289–295. https://doi.org/10.1098/rstb.2004.1584

Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. M. P., Elston, D. A., Scott, E. M., … Watt, A. D. (2010). Long-term datasets in biodiversity research and monitoring: Assessing change in ecological communities through time. *Trends in Ecology & Evolution*, *25*, 574–582. https://doi.org/10.1016/j.tree.2010.06.016

Marsh, D. M., & Trenham, P. C. (2008). Current trends in plant and animal population monitoring. *Conservation Biology*, *22*, 647–655. https://doi.org/10.1111/j.1523-1739.2008.00927.x

McCain, C., Szewczyk, T., & Knight, K. B. (2016). Population variability complicates the accurate detection of climate change responses. *Global Change Biology*, *22*, 2081–2093. https://doi.org/10.1111/gcb.13211

McDonald-Madden, E., Baxter, P. W. J., Fuller, R. A., Martin, T. G., Game, E. T., Montambault, J., & Possingham, H. P. (2010). Monitoring does not always count. *Trends in Ecology & Evolution*, *25*, 547–550. https://doi.org/10.1016/j.tree.2010.07.002

Møller, A. P., Rubolini, D., & Lehikoinen, E. (2008). Populations of migratory bird species that did not show a phenological response to climate change are declining. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 16195–16200. https://doi.org/10.1073/pnas.0803825105

NESP Threatened Species Recovery Hub. (2018). The threatened species index for australian birds. Retrieved from https://tsx.org.au/ (last accessed 19 July 2019).

Nichols, J. D., & Williams, B. K. (2006). Monitoring for conservation. *Trends in Ecology & Evolution*, *21*, 668–673. https://doi.org/10.1016/j.tree.2006.08.007

Piersma, T., & Lindström, Å. (2004). Migrating shorebirds as integrative sentinels of global environmental change. *Ibis*, *146*, 61–69. https://doi.org/10.1111/j.1474-919X.2004.00329.x

Prozt, E. J., Peterman, R. M., Dulvy, N. K., Cooper, A. B., & Irvine, J. R. (2012). Reliability of indicators of decline in abundance. *Conservation Biology*, *26*, 894–904. https://doi.org/10.1111/j.1523-1739.2012.01882.x

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reynolds, J. H., Thompson, W. L., & Russell, B. (2011). Planning for success: Identifying effective and efficient survey designs for monitoring. *Biological Conservation*, *144*, 1278–1284. https://doi.org/10.1016/j.biocon.2010.12.002

Rhodes, J. R., & Jonzén, N. (2011). Monitoring temporal trends in spatially structured populations: How should sampling effort be allocated between space and time? *Ecography*, *34*, 1040–1048. https://doi.org/10.1111/j.1600-0587.2011.06370.x

Rodrigues, A. S. L., Pilgrim, J. D., Lamoreux, J. F., Hoffmann, M., & Brooks, T. M. (2006). The value of the IUCN Red List for conservation. *Trends in Ecology & Evolution*, *21*, 71–76. https://doi.org/10.1016/j.tree.2005.10.010

Rueda-Cediel, P., Anderson, K. E., Regan, T. J., Franklin, J., & Regan, H. M. (2015). Combined influences of model choice, data quality, and data quantity when estimating population trends. *PLoS ONE*, *10*, e0132255. https://doi.org/10.1371/journal.pone.0132255

Sanderson, F. J., Donald, P. F., Pain, D. J., Burfield, I. J., & van Bommel, F. P. J. (2006). Long-term population declines in Afro-Palearctic migrant birds. *Biological Conservation*, *131*, 93–105. https://doi.org/10.1016/j.biocon.2006.02.008

Schumann, N., Dann, P., Hoskins, A. J., & Arnould, J. P. Y. (2013). Optimizing survey effort for burrow-nesting seabirds. *Journal of Field Ornithology*, *84*, 69–85. https://doi.org/10.1111/jofo.12007

Sims, M., Wanless, S., Harris, M. P., Mitchell, P. I., & Elston, D. A. (2006). Evaluating the power of monitoring plot designs for detecting long-term trends in the numbers of common guillemots. *Journal of Applied Ecology*, *43*, 537–546. https://doi.org/10.1111/j.1365-2664.2006.01163.x

Starcevich, L. A. H., Irvine, K. M., & Heard, A. M. (2018). Impacts of temporal revisit designs on the power to detect trend with a linear mixed model: An application to long-term monitoring of Sierra Nevada lakes. *Ecological Indicators*, *93*, 847–855. https://doi.org/10.1016/j.ecolind.2018.05.087

Taylor, B. L., & Gerrodette, T. (1993). The uses of statistical power in conservation biology: The vaquita and northern spotted owl. *Conservation Biology*, *7*, 489–500. https://doi.org/10.1046/j.1523-1739.1993.07030489.x

Tománková, I., Boland, H., Reid, N., & Fox, A. D. (2013). Assessing the extent to which temporal changes in waterbird community composition are driven by either local, regional or global factors. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *23*, 343–355. https://doi.org/10.1002/aqc.2303

U.S. Fish & Wildlife Service. (2018) Endangered Species. Retrieved from https://www.fws.gov/endangered/ (last accessed 19 July, 2019).

Urquhart, N. S., Paulsen, S. G., & Larsen, D. P. (1998). Monitoring for policy-relevant regional trends over time. *Ecological Applications*, *8*, 246–257. https://doi.org/10.2307/2641064

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer.

Wauchope, H. (2019). hannahwauchope/TrustingTrends: Trusting Trends MEE Release (Version Major_v.1). Zenodo. https://doi.org/10.5281/zenodo.3417001

White, E. R. (2019). Minimum time required to detect population trends: The need for long-term monitoring programs. *BioScience*, *69*, 40–46. https://doi.org/10.1093/biosci/biy144

Wilson, H. B., Kendall, B. E., & Possingham, H. P. (2011). Variability in population abundance and the classification of extinction risk. *Conservation Biology*, *25*, 747–757. https://doi.org/10.1111/j.1523-1739.2011.01671.x

WWF. (2016). *Living planet report 2016. Risk and resilience in a new era*. Gland, Switzerland: WWF International.

Xu, C., Barrett, J., Lank, D. B., & Ydenberg, R. C. (2015). Large and irregular population fluctuations in migratory Pacific (*Calidris alpina pacifica*) and Atlantic (*C. a. hudsonica*) dunlins are driven by density-dependence and climatic factors. *Population Ecology*, *57*, 551–567. https://doi.org/10.1007/s10144-015-0502-5

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Wauchope HS, Amano T, Sutherland WJ, Johnston A. When can we trust population trends? A method for quantifying the effects of sampling interval and duration. *Methods Ecol Evol*. 2019;00:1–12. https://doi.org/10.1111/2041-210X.13302