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

Hannah Wauchope[1*], Alison Johnston[1, 2], Tatsuya Amano[1, 3], William Sutherland[1]
[1]Conservation Science Group, Dpt Zoology, University of Cambridge, David Attenborough Building, Cambridge, UK
[2]Cornell Lab of Ornithology, Cornell University, Ithaca, NY 14850, USA
[3]Centre for the Study of Existential Risk, University of Cambridge, 16 Mill Lane, Cambridge, UK
*Corresponding author, hannah.wauchope@uqconnect.edu.au

Running headline: Quantifying confidence in population trends

Abstract

1. Species' population trends are fundamental to conservation, underpinning IUCN red-list classifications, many national lists of threatened species and are used globally to convey to policy makers the state of nature. Clearly, it's crucial to quantify how much we can trust population trend data, yet many studies analyzing large numbers of population time series lack a straightforward way to do so.

2. Here we present a novel method that artificially degrades high quality population trend data to see how often subsets of the data, taken according to different sampling methods, correctly estimate the direction and magnitude of each population's complete trend. We test our method on a dataset of 27,930 waterbird population time series from across North America.

3. We find that if a significant trend is detected, even from only a small subset of years, it is likely to be representative of the complete trend. But you need to sample for many years to be confident that an insignificant trend (indicating a stable population) is correct, and it is more likely to be missing a decline than an increase in a population.

4. Our methods provide a clear and quantitative way to assign confidence to species trends, which will enable rigor to be added to large-scale population analyses, and can facilitate planning of future monitoring schemes. Our methods are readily applicable to other taxa and we provide the tools to do so.

Keywords: Population, trends, confidence, large-scale analyses, survey, monitoring, time-series, generation length

Introduction

Many crucial conservation decisions rely on knowing the general trend of a species or population. This information underpins IUCN red-list classification (Loh *et al.*, 2005; Rodrigues *et al.*, 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 (Gärdenfors, 2001; Collen *et al.*, 2009). It is important that conservationists understand how much 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. We have a poor understanding of how confident we can be in the trend of a population, especially when data on variability and measurement error are not available. In addition, many large-scale analyses and policy recommendations (e.g. Collen *et al.*, 2009; Living & Index, 2016) rely on aggregating trends on many populations and there is little guidance on how to weight trends by their likely veracity.

Studies estimating the trend of a population generally use a series of counts over time (typically years, as considered here). Linear or non-linear models are fit and regressed to estimate yearly change (e.g. 1% decline per year), or modelled counts are compared between two years to highlight change over a period (e.g. a 10% increase over 10 years). The number of years of data, sampled frequency, degree of measurement error and population variability all affect confidence in the derived trend. When data is available on measurement and population variability, power analyses are recommended to estimate levels of confidence in trend estimates (Hayes & Steidl, 2002; Magurran *et al.*, 2010; Johnson *et al.*, 2014). Although power analyses are worthwhile to carry out for the occasions where sufficient data are available, and in such cases, we would urge their use, this is often not the case, especially when assessing many populations, or using already gathered data.

Other studies have therefore attempted to quantify confidence in trends using simulated data. These conclude that longer time scales are needed for better trend estimates, and that there are high margins of error when detecting small declines (Wilson *et al.*, 2011; Prozt *et al.*, 2012; Tománková *et al.*, 2013; Connors *et al.*, 2014; Fox *et al.*, 2018). 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). The number of years of sampling required to reliably detect a trend has been estimated at both 10 (Rueda-Cediel *et al.*, 2015 - a snail; White, 2018 - various vertebrates) and 21 years (Reynolds *et al.*, 2011 - brown bears).

These studies are useful for gaining an approximate idea of confidence, but do not provide a straightforward way for a study to assign a value of confidence to a population time series. In the absence of better guidance, studies based on many population trends often lack the data or guidance to make any quantification of uncertainty (e.g. Craigie *et al.*, 2010, Loh *et al.*, 2005). Further, studies thus far have assumed that there is a 'true' trend exhibited by each species, but given most populations fluctuate over time in response to the positive and negative pressures affecting them, it is incorrect to maintain a static length of time from which to define a true trend for most species.

We propose a 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 (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 confidence of how likely that sample is to represent the complete trend, regardless of sample or complete trend length.

We ask two questions: 1) How confident can one be in a trend derived from a certain number of years of data, based on the time over which a trend estimate is desired? (e.g. how well do 8 consecutive years of survey data represent the trend of a population over 10 years?); and 2) How confident can one be in 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 quantify confidence both in terms of accuracy of trend direction and magnitude of change. Finally, we investigate two factors that we expect to impact confidence: species generation time and shape of the 'true' 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 while older individuals are still living. We also expect that samples from more complex trends will be less accurate than samples from linear trends.

As a case study, we use an empirical dataset of yearly counts of 122 waterbird species at 1104 sites in North America (a total of 27,930 site by species combinations). Providing these estimates for waterbird data is particularly beneficial as data on waterbirds are available at large spatio-temporal scales and waterbird studies can provide insights into broader conservation goals (Piersma & Lindström, 2004; Amat & Green, 2010; Amano *et al.*, 2018). However, our methods are general, and we provide code and instructions to generalize to other taxa. Our work provides an explicit measure of the confidence in a trend and allowing an evidence-based justification for excluding samples below a certain length, according to the degree of confidence desired for a study. Finally, these results could be used to plan multi-species monitoring programs, to give the highest likelihood of capturing representative trends for the most species.

Methods:

Data

We obtained an initial dataset of yearly count data for 208 waterbird species in North America from the Christmas Bird Count (CBC; Dunn *et al.*, 2005) at 1,125 sites spanning the years 1966 to 2013, from which 30 years of consecutive counts were taken for each site by species combination. We selected 30 years because it represented a good balance between a long-term survey period, but one for which adequate data was still available. In cases where a site was sampled for more than 30 years, the most recent 30 years were taken. The supporting material provides a list of species and a map of spatial coverage. The CBC has been conducted yearly since 1900, primarily in the US and Canada, and involves a count of all bird individuals within a 'site', defined as a circle 24.1km in diameter. Sites are accessed by a variety of means including foot, car, boat, and snow mobile. Effort varies between sites, but the number of survey hours spent per count is documented, allowing effort to be accounted for (see below). 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. This gave us an initial dataset of 48,653 populations.

Data preparation

Christmas Bird Count data is not standardized by sampling effort, meaning this must be accounted for in the modelling process. To account for effort, we must know the relationship between effort and detection. The most common expected relationship is a linear relationship between count and the log-transformed effort: the likelihood of detecting individuals decreases with increasing effort. Following Butcher & McCulloch (1988) and Xu *et al.* (2015), we chose to only retain species where a significant linear relationship between detection and log of effort (defined as number of survey hours) was shown, found by running a simple linear model of the relationship between log of effort and number of individuals counted, at all years and sites:

$$Count_i = \log(e_i)$$

Where $Count_i$ is all counts of species i and e_i is the number of survey hours for each count. Any species found to have a non-significant relationship were removed from analysis, as were those that had a negative significant relationship between count and effort (i.e. as effort increased, detection decreased). We then used an offset term in our models to account for this sampling effort (see below). 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: 27,930 populations, comprising 122 species at 1104 sites.

As species varied in the extent to which they occurred at sites, we also ran our analysis on a standardized subset of the data: 98 species with 50 randomly selected sites each, 4900 populations in total. Even though this dataset was less than 20% of the size of our full dataset, the results were highly congruent.

Modelling Specifications

We used Generalized Linear Models (GLMs) to estimate the population growth rate, r, with population counts as the response variable and years as the explanatory variable, 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. GLMs consist of three parts, a random component (in this case is the yearly population counts), a linear predictor (in this case is the year and log of effort) that predicts the expected value of the count, and a link

function that transforms the expected value to the linear predictor. This link function describes a frequency distribution of the response variable (in our case, counts). For count data, it is usual to use Poisson, quasi-Poisson or negative binomial distributions, with the latter two being more appropriate if there is over-dispersion, where the variance of the response variable exceeds the mean. Waterbird count data is often highly over-dispersed: 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 as it is the most appropriate for modelling highly over-dispersed data.

Mathematically, the above is expressed as the following:

$$E(Count_i) = g^{-1}(\alpha + rx_i + \log(e_i)) \quad (1)$$

$$var(Count_i) = v_{NB}(E(Count_i)) \quad (2)$$

The expected value of our count is calculated as the inverse of the link function g(·), in our case negative binomial, wrapped around the linear predictor, which consists of an intercept, α, the population growth rate, *r*, multiplied by the year, *x*, and the log of effort (in hours), *e* (Eq. 1). Because the relationship between effort and count is known (i.e. a log linear relationship), it does not need a coefficient. The variance of our count data is defined as negative binomial (Eq. 2).

For each model the population growth rate, *r*, and p-value of *r* were noted. For our main analysis, we followed the convention of setting a significance level of $p<0.05$. This is an arbitrary threshold, and circumstances may arise where the risk of making a missed detection is greater than the risk of a false alarm (e.g. a high risk group of species), in which case it is better to set a higher p-value (Taylor & Gerrodette, 1993; Field *et al.*, 2007), and *vice versa*. We provide functionality to adjust p-values in the provided code, so this can be adjusted by future users if desired.

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 visited 14th Dec, 2018).

Sampling Methods (Consecutive and Interval)

We consider two ways in which sampling could be conducted, and took samples from the complete data according to these. First, we used consecutive sampling (Figure 1a), where all possible 3 to n-1 year subsets would be taken from a complete trend of n years in length. Second, we used interval sampling (Figure 1b): we varied the interval length (the number of years between samples) from 0 years (i.e. consecutive years) to 14 years (i.e. samples taken 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, that were at least three years in length. This is best understood through an example, see Table 1.

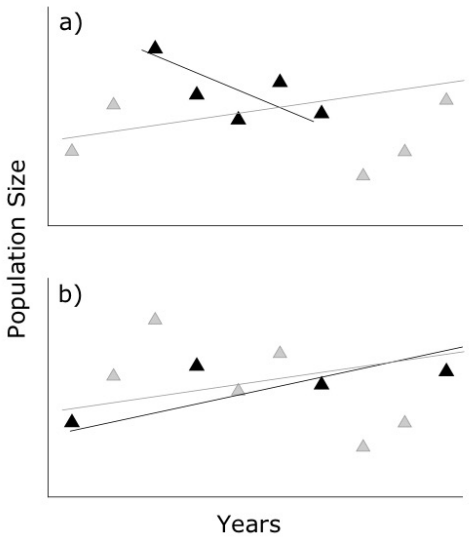

Figure 1. Schematic diagram showing sampling and modelling technique for comparing sample and complete trends. Triangles show counts for a hypothetical population, all counts are used to model the complete population trend (grey line), and only black triangles are used to model the sample population trend (black line). Shows examples of Consecutive sampling (a): a five-year chunk sampled from a ten-year period; and Interval sampling (b): four years sampled at three-year intervals from a ten-year period. We ran analysis on all possible Consecutive lengths, Interval lengths and complete trend lengths.

Table 1. All iterations used for interval samples taken every 8 years (i.e. interval length = 8). For lower interval lengths, higher numbers of years sampled were possible (up to interval lengths of 0, i.e. consecutive years, which could have up to 29 years sampled)

3 years sampled	4 years sampled
1, 9, 17	1, 9, 17, 25
2, 10, 18	2, 10, 18, 26
3, 11, 19	3, 11, 19, 27
4, 12, 20	4, 12, 20, 28
5, 13, 21	5, 13, 21, 29
6, 14, 22	6, 14, 22, 30
7, 15, 23	
8, 16, 24	
9, 17, 25	
10, 18, 26	
11, 19, 27	
12, 20, 28	
13, 21, 29	
14, 22, 30	

The complete trend was initially estimated from annual surveys over the full 30-year time-period, but what if one had a sample of 5 years of data, and wished to know how representative this was of the 20-year trend of a species? Reflecting this, we also conducted our consecutive analysis with differing lengths of complete years. For example, we might define the complete trend as the trend a population exhibited over a 10-year period. To do this, for each of our 27,930 populations, we took all possible 10 year segments (e.g. 1-10, 2-

11, 3-12 etc), and calculated the trend on all possible consecutive subsets of those ten years, which would range from 3 to 9 years in length. We would then compare these to trends calculated from the complete 10 years. We repeated this for Complete trend lengths ranging from 4 years to the full 30 years. Because interval sampling already had two dimensions (interval length and number of years sampled), we compared only to the complete trend length of 30 years.

Comparison Methods (Direction and Magnitude)

We used two ways to assess whether a sample trend (consecutive or interval) represented a complete trend. First, we took the simple direction of the trend, defining it as positive, negative or insignificant. Using this, a sample trend would be classified as correct if it was the same direction as the complete trend; opposing if it was the opposite direction; a false positive or negative if it was positive or negative, but the complete trend was insignificant; and a missed positive or negative if it was insignificant, but the complete trend was positive or negative (Table 2). We term this 'Direction' comparison.

Table 2. Categories used for Direction comparison. Columns show direction of trend from complete time series and rows direction of trend from sample time series. Cells show category, based on sample and complete trend direction.

		Complete Trend		
		Positive	Negative	Insignificant
Sample Trend	Positive	Correct	Opposing	False Positive
	Negative	Opposing	Correct	False Negative
	Insignificant	Missed Positive	Missed Negative	Correct

Second, for cases where a significant population growth rate *r* (i.e. the slope of a log-linear model) was obtained from *both* the sample and complete dataset (i.e. cases of 'Correct' or 'Opposing' from the Direction comparison method), we considered the absolute difference between *r* of the two; giving an idea of the degree of 'correctness'. We defined tolerance levels ranging from ±0.01 to ±0.5, specifying that if the sample *r* differed from the complete *r* by less than the tolerance level, the sample trend represented the complete trend and was correct, and was incorrect if it did not. We term this 'Magnitude' comparison (Fig. 2).

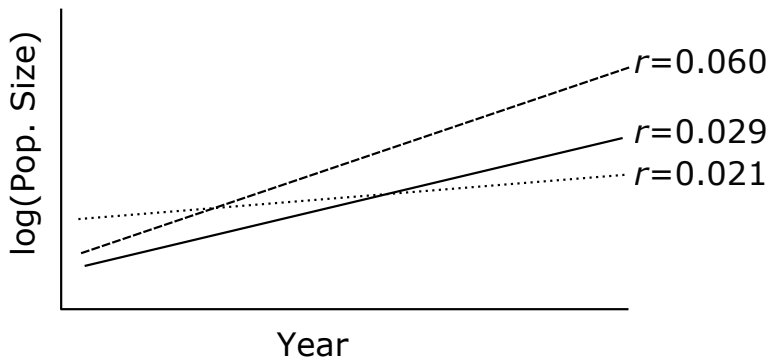

Figure 2. Example of Magnitude comparison, showing three populations with modelled log population size vs. time. If the solid line is the trend derived from the complete population,

then the dotted line sample would be correct at a tolerance of ±0.01 (because |0.021-0.029| < 0.01, but the dashed line would not (because |0.021-0.060| > 0.01).

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. We then found, for each combination of sample and complete dataset lengths and sampling types, the percentage of our 27,930 populations in each representativeness category.

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 and support population numbers 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 was obtained from birdlife.org species fact sheets (e.g. http://datazone.birdlife.org/species/factsheet/ruddy-turnstone-arenaria-interpres/details, accessed 26.07.2018). We then categorised our results using these three groups.

Trend Shape

To assess how our results are affected by trends of different shapes, we used Generalised Additive Models (GAMs) with the R package MASS (Venables & Ripley, 2002). These non-parametric models allow for non-linear relationships, i.e. they allow for a trend that 'wiggles' through time rather than following, say, a linear or exponential curve. A key aspect of these models is the degrees of freedom, which can be viewed as a rough proxy for the 'degree of wiggliness' of a trend: a trend with one degree of freedom will be linear, and one with many degrees of freedom will be highly wiggly. If the user does not specify a degree of freedom, the GAM will estimate it and this can be extracted to gain an idea of curve shape. We therefore ran GAMs on all complete 30 year trends, run 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 or quadratic (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 samples were more of less representative when sampled from complete trends of different shapes.

Results

See supporting information section 1 for summary statistics of the full dataset. All data used to produce plots is also provided, as well as code to reproduce full results on any set of population counts.

Direction Comparison of Consecutive Sampling

We found that large samples were required to be highly confident that the direction of a sample trend matched that of a complete trend. E.g. to have an 80-100% chance of a sample trend being correct based on trend direction, the sample needed to be just a few years fewer than the complete trend (Fig. 3a). However, sample trends opposed the complete sample trend less than 10% of the time. (Fig 3b).

The chance of a false positive or negative (i.e. the sample indicated a significant trend but the complete trend did not, Fig 3c, d) was low regardless of the length of the sample or complete trend. However, the chances of a missed positive or negative (i.e. the complete trend had a significant direction but the sample did not detect this) were higher, missed negatives were more likely than missed positives, and both were particularly likely when the sample trend length was considerably shorter than the complete trend length (Fig 3e, f). This suggests 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 of the complete trend.

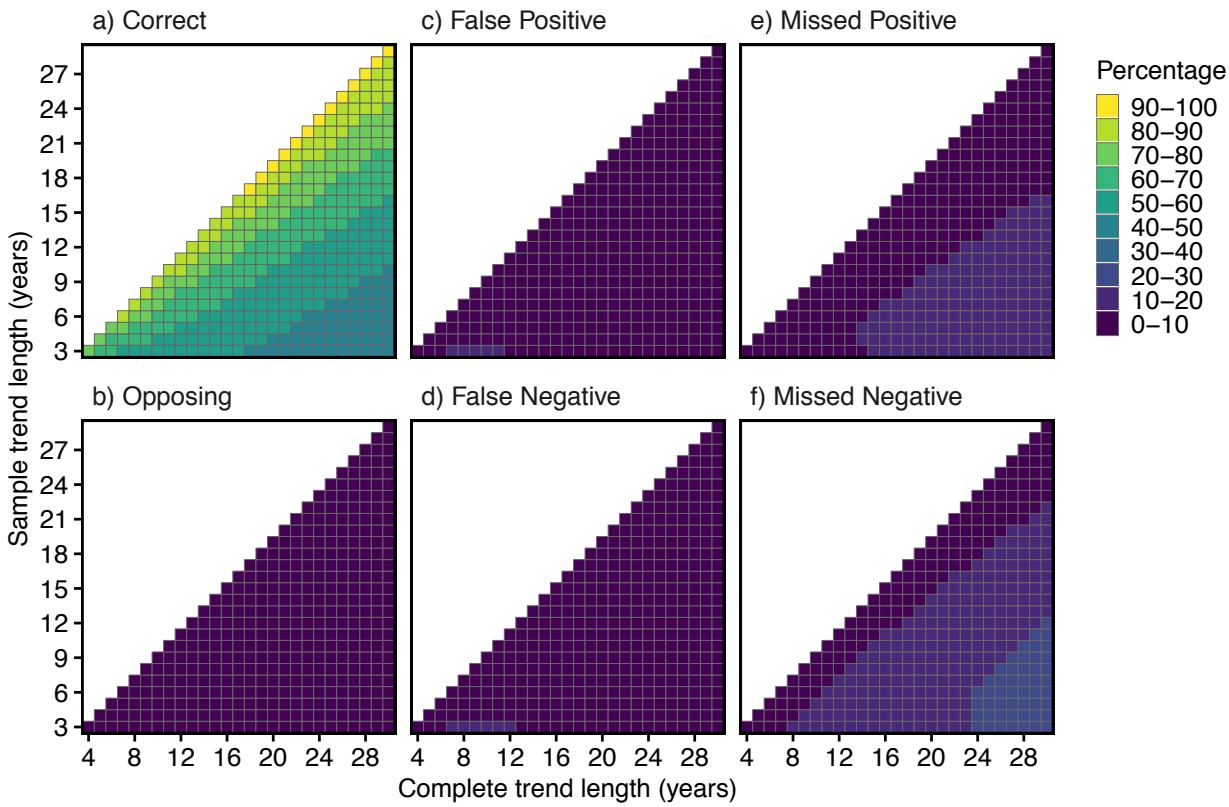

Figure 3. Direction comparison of sample population trends using consecutive sampling. Heat shows percentage of sample trends that, relative to the complete trend, were correct, opposing, a false positive/negative or a missed positive/negative (see Table 1). Shown for all combinations of sample lengths (y-axis), and complete lengths (x-axis).

Direction Comparison of Interval Sampling
Our results show that sampling in intervals can be more representative than sampling in consecutive years when considering trend direction. For example, sampling for 26 consecutive years (out of a 30-year complete trend) gave the correct result 80% of the time (Fig. 4a, bottom row, nb. this is equal to Fig. 3a rightmost column), but the same level of confidence could be obtained by sampling 13 times with one year between samples (Fig. 4a, second row). More strikingly, 4 years sampled every 8 years gave the same percentage correct (60-70%) as up to 20 years of consecutive sampling (Fig. 4a).

As with consecutive sampling, the percentage opposing for interval sampling was very low (Fig. 4b), the chance of making a false positive or negative was also very low for all sampling combinations (Fig. 4c, d) and though missed positives and negatives were slightly more likely, the chance never exceeded 40% (Fig. 4e, f). As before, missed negatives were more likely than positives.

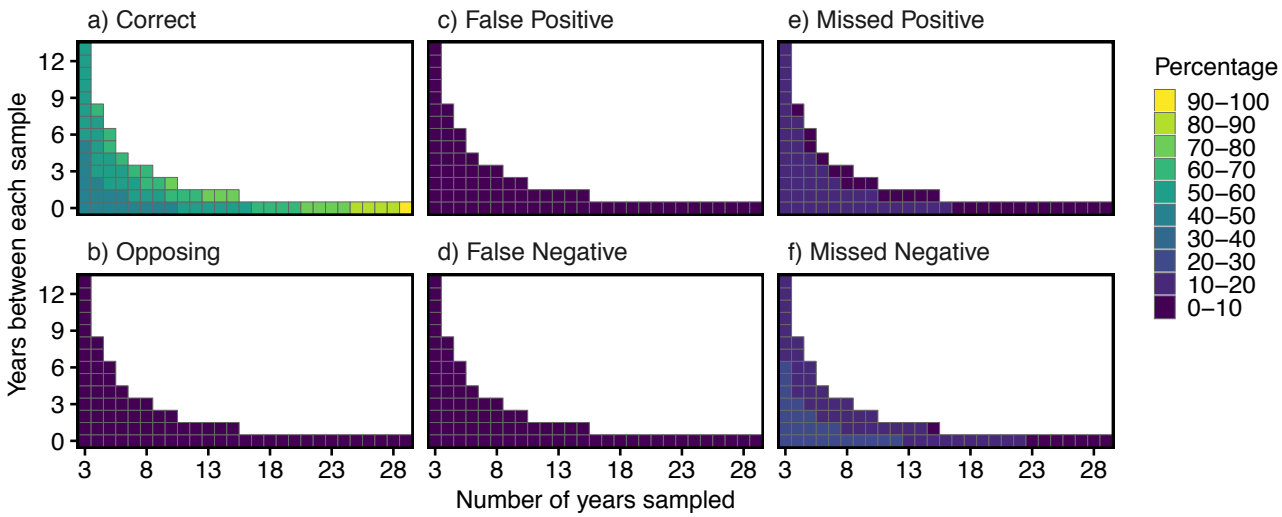

Figure 4. Direction comparison of sample trends using interval sampling. Heat shows percentage of sample trends that, relative to the complete trend, were correct, opposing, a false positive/negative or a missed positive/negative (see Table 1). Shown for all combinations of interval sampling, with number of years sampled (x-axis) and years between samples (y-axis). Thus 8 on the x-axis and 4 on the y-axis would mean 8 samples were taken, with 4 years between each one. The bottom row of each plot is equal to the right most column of the equivalent Figure 3 plot, but is included here to ease comparison. Complete trend length is always 30 years.

Magnitude Comparison of Consecutive Sampling
It is difficult to obtain a reasonably accurate growth rate estimate (r) of a complete trend. For ease of interpretation these results are displayed at four complete trend lengths: 5, 10, 20 and 30 years. If trying to estimate complete trends of 10 years or less (Fig 5a, b), samples do not reach very high levels of accuracy even at high tolerance levels (e.g. ±0.5, i.e. 50% population change per year). If trying to estimate complete trends over 10 years (Fig 5c, d), samples of over roughly half the complete length can reach high accuracy at very wide tolerances, but this is not particularly surprising and being correct to within 50% population change per year is not particularly useful. To be correct within more narrow tolerance thresholds it becomes increasingly difficult and more sample years are required, to the point

where at a tolerance of ±0.01 even a sample of 29 years can only correctly estimate a complete trend of 30 years ~80% of the time (Fig. 5d).

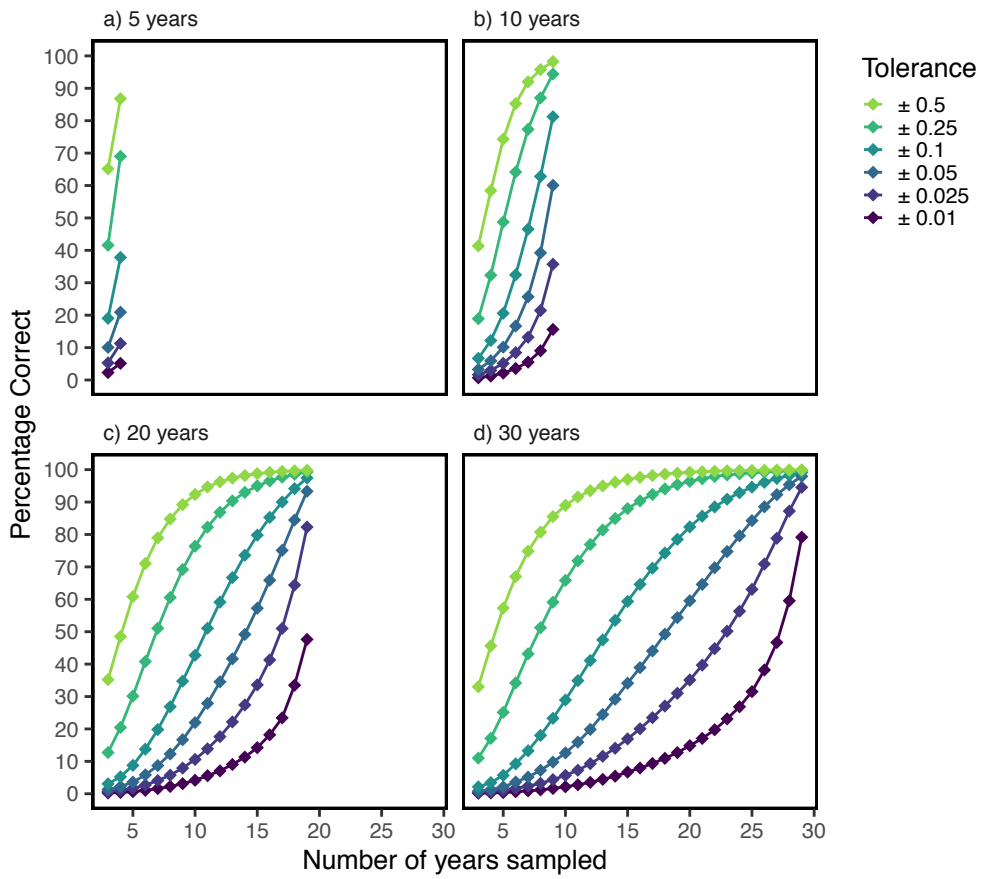

Figure 5. Percentage of sampled trends that correctly match the complete trend, measured by whether the sample *r* matches the complete *r* within the threshold (colours); samples taken using Consecutive sampling. Shown for four complete trend lengths, 5, 10, 20, and 30 years, and for all sample lengths (x-axis).

Magnitude Comparison of Interval Sampling

Though sampling at intervals gave better results when comparing trend direction, when comparing magnitude of change, interval sampling did not perform as well over the time scales we measured. It was not possible to achieve above even 65% of correct samples at low thresholds ($r = \pm 0.01$ - 0.025), i.e. population changes of 1-2.5% per year (Fig 6). The direction of curves indicates that high correct percentages could be attained with enough years of sampling at large intervals, but this would mean sampling over very long time scales. Better confidence was achieved at higher tolerances, but only at ±0.5 and ±-0.25, i.e. 25 - 50% population change per year.

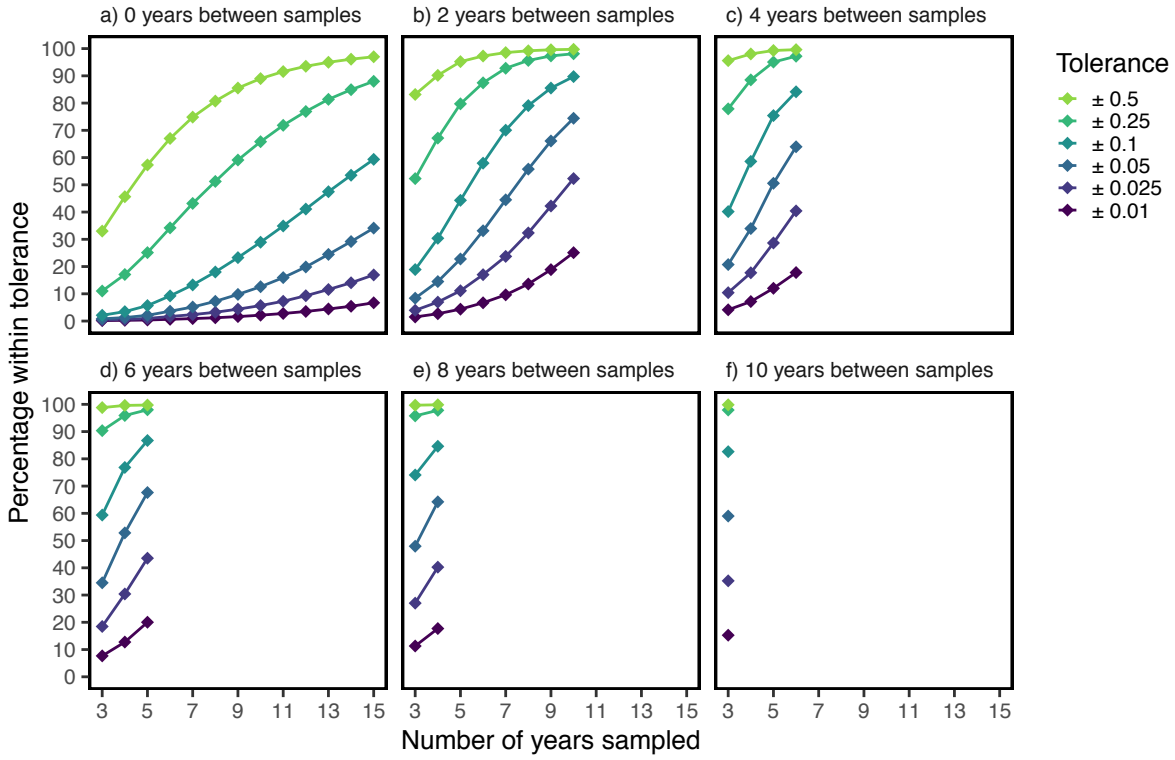

Figure 6. Percentage of samples (y-axis) that correctly match the complete trend, measured by whether the sample *r* matches the complete *r* within the threshold (colours); samples taken using Interval sampling. Shown for four stages of interval sampling: samples taken either every 0, 2,4, 6, 8, or 10 years (facets), for all possible number of years sampled (x-axis). Complete trend length is always 30 years.

Generation Length and Trend Shape

We found that, for populations with long generation lengths, short samples of long time series are less likely to be correct and missed positives/negatives are more likely (Fig. 7, Supporting Figures 2-4). False positives/negatives and opposing trends were roughly equal among populations of different generation lengths (Fig. 7, Supporting Figures 2-4). However, populations of different generation lengths performed similarly when comparing the magnitude of difference between sample and complete trends.

For populations showing different trend shapes, false positives/negatives were more likely with small samples of complete trends of high EDFs, but the percentage of samples that were correct remained reasonably constant across EDFs (Supporting Figure 5-8). Populations of trend shapes performed similarly when comparing the magnitude of difference between sample and complete trends.

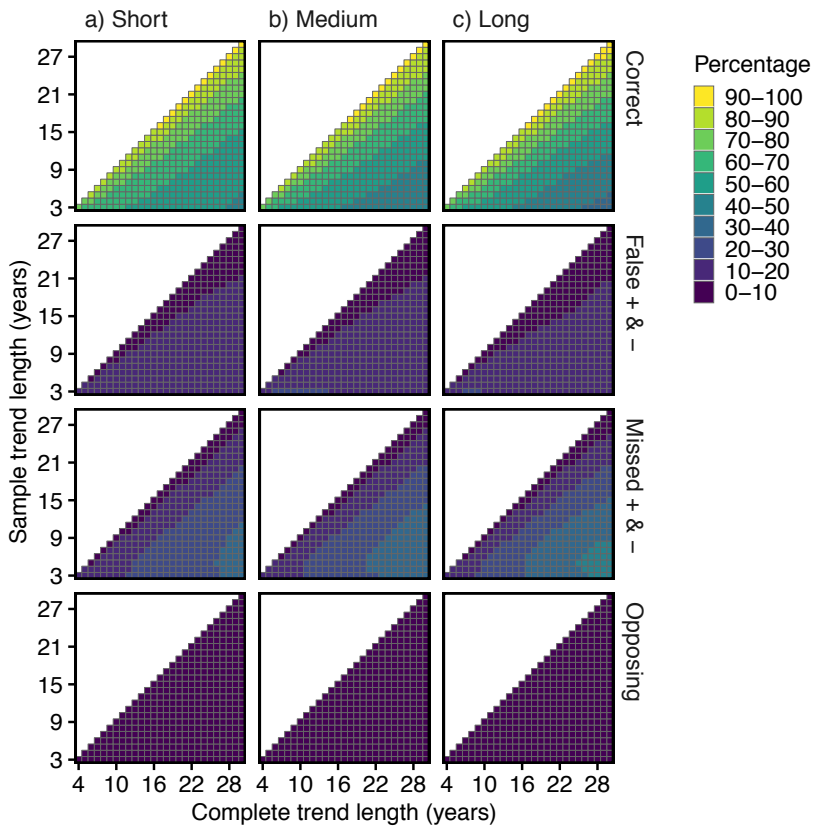

Figure 7. Direction comparison of sample population trends using consecutive sampling. Heat shows percentage of sample trends that, relative to the complete trend, were correct, opposing, a false positive/negative or a missed positive/negative (see Table 1). Shown for all combinations of sample lengths (y-axis), and complete lengths (x-axis). Divided by populations whose species are classified into three generation lengths: a) short (1-5 years), b) medium (6-10 years) and c) long (11-15 years).

Discussion

In this paper, we provide and test a method that can estimate confidence in population trends of different lengths and sampling types, based on the total time over which a trend estimate is desired. Our results show high levels of convergence and are robust to different subsets, indicating that our method is producing reasonable estimates for waterbirds. 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 add confidence to large scale studies of population trends and to design future monitoring schemes.

Waterbird case study

Our results show that if a significant trend is obtained from a population time series, even when the data is from a few years, it is likely to reflect the true trend direction of the population, though not necessarily the magnitude. Further, few samples at widely spaced intervals can produce similar levels of confidence in trend direction compared to double the number of samples taken in consecutive years. We also show that if an insignificant trend is obtained, it is more likely to be missing a decline than an increase in a population, and should be trusted with caution.

In simple terms, these results indicate that to be confident in a *lack* of trend, i.e. an insignificant result, one must obtain a very high number of samples. In other words, the power of an insignificant trend is low, unless it is close to the length over which a complete trend is trying to be estimated. However, if a *significant* trend is detected, it is likely to be representative of the complete trend; we found that the chances of a sample opposing or falsely detecting a positive or negative trend were very low. Keith *et al.* (2015) studying birds similarly found that past population trajectory was a good predictor of future population trajectory. Møller *et al.* (2008) and Sanderson *et al.* (2006) also found a similar, though weak, correlation between trends of migratory birds between 1970-1990 and 1990-2000. However, 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 extended to other taxa using relevant data, rather than our results with water birds.

Sampling in intervals provided better, and surprisingly accurate, results than sampling in consecutive years. We found that, 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 13 years. This is surprising given the high amounts of variability in waterbird data, we would expect that sampling at such wide intervals would not be able to produce such accurate results. Other studies have found similar results (Urquhart *et al.*, 1998; Starcevich *et al.*, 2018), 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. Sampling at such wide intervals is not always practical however, especially for high risk species where declines may need to be detected and acted on quickly. In addition, whilst 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 direction of a population trajectory, but the actual rate of change, we find that samples are much less likely to be representative of complete trends. To be 80% correct at a rate of population change of 1-2.5% per year, one must sample for at least 19 years. These results are roughly in agreement with other studies on diverse taxa that find samples should be between 10 (Rueda-

Cediel *et al.*, 2015; White, 2018) and 21 years (Reynolds *et al.*, 2011). Sampling in intervals also struggled to produce accurate results, 80% confidence 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).

Applications

Given our results are derived from an entirely empirical dataset with no simulations, they show a surprising amount of convergence (e.g. Fig 5.), 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, sub-setting of species into three separate groups based on generation length (see also White, 2018) and sub-setting of populations into different trend shapes. This indicates that our method is a valid way to get broad-scale estimates of confidence in trends derived from data of different lengths. The method is general, enabling derivations of estimates of confidence for any taxa for which there is sufficient data. Code is available (see Data Availability) through which this can be done.

The data from this study (see supporting information), or data produced using these methods with other taxa, can be used to assign confidence to large scale analyses that use many population time-series, either by removing all populations below a certain confidence threshold, or by weighting analysis by confidence in each trend. The analysis should define a 'complete' trend length for the study (in most cases this would be the length of the longest time series possessed in the dataset), and extract the confidence estimates for this complete trend length. If the analysis intends to use all data available, confidence estimates can then be assigned to each time series based on how long each is, and any analyses can be weighted according to confidence in each time series. Or, a confidence threshold can be set, and any time series shorter that this can be removed, and analyses can proceed using this 'high quality' dataset.

We readily agree that such confidence levels are not infallible: we have not taken a universal survey of all species always, and therefore can never know exactly how representative any trend derived from a time-period can be. However, this seems is a significant improvement on making an arbitrary cutoff point to clean datasets, or on having no way of weighting populations. For those working with large datasets who cannot conduct power analysis or quantify measurement error in their populations, this is a way to assign at least some level of confidence to results.

These results could also be used to help plan future monitoring schemes, but we would advise this be done with a cautious eye. Monitoring is often best when tailored to the species in question (Ficetola *et al.*, 2017), should be adaptive and adjusted according to desired management outcomes (Hauser *et al.*, 2006; Legg & Nagy, 2006; Nichols & Williams, 2006; Lindenmayer & Likens, 2009; McDonald-Madden *et al.*, 2010). But in cases where programs are carried out with the goal of detecting trends (Marsh & Trenham, 2008) we can provide some very general guidance, namely that monitoring for many years is essential to accurately capturing population trends (unless significant trends are detected), and that resources can be conserved and possibly allocated to more locations if sampling is conducted in intervals rather than every year.

This analysis was kept deliberately 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 confidence in derived trends would increase if this was considered. Detecting trends in species, not populations, generally means combining analyses from a number of survey locations (Rhodes & Jonzén, 2011), which would also influence results, however this introduced too many variables to consider for our analysis. Some studies have found that sampling at more locations improved trend detection better than longer time series (Sims *et al.*, 2006, though see Schumann *et al.*, 2013) and so modelling trends from multiple locations is likely to influence our confidence estimates.

Acknowledgements
CBC Data is provided by National Audubon Society and through the generous efforts of Bird Studies Canada and countless volunteers across the western hemisphere. HW was supported by a Cambridge Trust Cambridge-Australia Scholarship and the Cambridge Department of Zoology JS Gardiner Fellowship. We thank Richard Fuller and Benno Simmons for helpful comments and discussion.

Author Contributions
HW, TA, AJ and WS conceived the ideas and designed methodology. TA collected the data, HW analyzed the data and HW led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data Accessibility
All data used in our analysis are available from the Christmas Bird Count (www.christmasbirdcount.org). Code, which is written to be fully adaptable to other data, is available both in the supporting materials and on GitHub (https://github.com/hannahchoppie/TrustingTrends), we recommend using the GitHub code as this will remain up to date.

Supporting Information
Supporting information contains two files, a PDF containing supporting figures and explanations, and a zip file containing modelled data from our methods, and code the reproduce methods.

References

Amano T, Székely T, Sandel B et al. (2018) Successful conservation of global waterbird populations depends on effective governance. *Nature*, **553**, 199.

Amat JA, Green AJ (2010) Waterbirds as Bioindicators of Environmental Conditions. In: *Conservation Monitoring in Freshwater Habitats: A Practical Guide and Case Studies*, pp. 45–52. Springer, Dordrecht, Heidelberg, New York, London.

Atkinson PW, Austin GE, Rehfisch MM et al. (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.

Butcher GS, McCulloch CE (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 JEM (2009) Monitoring Change in Vertebrate Abundance: the Living Planet Index. *Conservation Biology*, **23**, 317–327.

Connors BM, Cooper AB, Peterman RM, Dulvy NK (2014) The false classification of extinction risk in noisy environments. *Proceedings of the Royal Society B: Biological Sciences*, **281**.

Craigie ID, Baillie JEM, Balmford A, Carbone C, Collen B, Green RE, Hutton JM (2010) Large mammal population declines in Africa's protected areas. *Biological Conservation*, **143**, 2221–2228.

Dunn EH (2002) Using Decline in Bird Populations to Identify Needs for Conservation Action. *Conservation Biology*, **16**, 1632–1637.

Dunn EH, Francis CM, Blancher PJ et al. (2005) Enhancing the scientific value of the Christmas Bird Count. *The Auk*, **122**, 338–346.

Ficetola GF, Romano A, Salvidio S, Sindaco R (2017) Optimizing monitoring schemes to detect trends in abundance over broad scales. *Animal Conservation*, **21**, 221–231.

Field SA, O'Connor PJ, Tyre AJ, Possingham HP (2007) Making monitoring meaningful. *Austral Ecology*, **32**, 485–491.

Fox R, Harrower CA, Bell JR, Shortall CR, Middlebrook I, Wilson RJ (2018) Insect population trends and the IUCN Red List process. *Journal of Insect Conservation*.

Gärdenfors U (2001) Classifying threatened species at national versus global levels. *Trends in Ecology & Evolution*, **16**, 511–516.

Hauser CE, Pople AR, Possingham HP (2006) Should Managed Populations Be Monitored Every Year? *Ecological Applications*, **16**, 807–819.

Hayes JP, Steidl RJ (2002) Statistical Power Analysis and Amphibian Population Trends. *Conservation Biology*, **11**, 273–275.

Johnson PCD, Barry SJE, Ferguson HM, Müller P (2014) Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution*, **6**, 133–142.

Keith D, Akçakaya HR, Butchart SHM et al. (2015) Temporal correlations in population trends: Conservation implications from time-series analysis of diverse animal taxa. *Biological Conservation*, **192**, 247–257.

Kuussaari M, Bommarco R, Heikkinen RK et al. (2009) Extinction debt: a challenge for biodiversity conservation. *Trends in Ecology & Evolution*, **24**, 564–571.

Legg CJ, Nagy L (2006) Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management*, **78**, 194–199.

Lindenmayer DB, Likens GE (2009) Adaptive monitoring: a new paradigm for long-term research and monitoring. *Trends in Ecology & Evolution*, **24**, 482–486.

Loh J, Green RE, 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 LP-

295.
Magurran AE, Baillie SR, Buckland ST et al. (2010) Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in Ecology & Evolution*, **25**, 574–582.
Marsh DM, Trenham PC (2008) Current Trends in Plant and Animal Population Monitoring. *Conservation Biology*, **22**, 647–655.
McDonald-Madden E, Baxter PWJ, Fuller RA, Martin TG, Game ET, Montambault J, Possingham HP (2010) Monitoring does not always count. *Trends in Ecology & Evolution*, **25**, 547–550.
Møller AP, 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.
NESP Threatened Species Recovery Hub (2018) The Threatened Species Index for Australian Birds.
Nichols JD, Williams BK (2006) Monitoring for conservation. *Trends in Ecology & Evolution*, **21**, 668–673.
Piersma T, Lindström Å (2004) Migrating shorebirds as integrative sentinels of global environmental change. *Ibis*, **146**, 61–69.
Prozt EJ, Peterman RM, Dulvy NK, Cooper AB, Irvine JR (2012) Reliability of Indicators of Decline in Abundance. *Conservation Biology*, **26**, 894–904.
R Core Team (2017) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Reynolds JH, Thompson WL, Russell B (2011) Planning for success: Identifying effective and efficient survey designs for monitoring. *Biological Conservation*, **144**, 1278–1284.
Rhodes JR, Jonzén N (2011) Monitoring temporal trends in spatially structured populations: how should sampling effort be allocated between space and time? *Ecography*, **34**, 1040–1048.
Rodrigues ASL, Pilgrim JD, Lamoreux JF, Hoffmann M, Brooks TM (2006) The value of the IUCN Red List for conservation. *Trends in Ecology & Evolution*, **21**, 71–76.
Rueda-Cediel P, Anderson KE, Regan TJ, Franklin J, Regan HM (2015) Combined Influences of Model Choice, Data Quality, and Data Quantity When Estimating Population Trends. *PLOS ONE*, **10**, e0132255.
Sanderson FJ, Donald PF, Pain DJ, Burfield IJ, van Bommel FPJ (2006) Long-term population declines in Afro-Palearctic migrant birds. *Biological Conservation*, **131**, 93–105.
Schumann N, Dann P, Hoskins AJ, Arnould JPY (2013) Optimizing survey effort for burrow-nesting seabirds. *Journal of Field Ornithology*, **84**, 69–85.
Sims M, Wanless S, Harris MP, Mitchell PI, Elston DA (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.
Starcevich LAH, Irvine KM, Heard AM (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.
Taylor BL, Gerrodette T (1993) The Uses of Statistical Power in Conservation Biology: The Vaquita and Northern Spotted Owl. *Conservation Biology*, **7**, 489–500.
Tománková I, Boland H, Reid N, Fox AD (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.
U.S. Fish & Wildlife Service (2018) Endangered Species.
Urquhart NS, Paulsen SG, Larsen DP (1998) Monitoring for policy-relevant regional trends

over time. *Ecological Applications*, **8**, 246–257.
Venables WN, Ripley BD (2002) *Modern Applied Statistics with S*, 4th edn. Springer, New York.
White ER (2018) Minimum time required to detect population trends: the need for long-term monitoring programs. *PeerJ Preprints*, **6**, e3168v4.
Wilson HB, Kendall BE, Possingham HP (2011) Variability in Population Abundance and the Classification of Extinction Risk. *Conservation Biology*, **25**, 747–757.
WWF (2016) *Living Planet Report 2016. Risk and resilience in a new era*. WWF International, Gland, Switzerland.
Xu C, Barrett J, Lank DB, Ydenberg RC (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.