Received: 18 March 2023 | Accepted: 15 April 2023

DOI: 10.1111/gcb.16841

Global Change Biology WILEY

RESEARCH ARTICLE

Quantifying reliability and data deficiency in global vertebrate population trends using the Living Planet Index

Shawn Dove[1,2] | Monika Böhm[2,3] | Robin Freeman[2] | Louise McRae[2] | David J. Murrell[1]

[1]Centre for Biodiversity and Environment Research, University College London, London, UK

[2]Institute of Zoology, Zoological Society of London, London, UK

[3]Global Center for Species Survival, Indianapolis Zoo, Indianapolis, Indiana, USA

Correspondence
Shawn Dove and David J. Murrell, Centre for Biodiversity and Environment Research, University College London, Gower Street, London WC1E 6BT, UK.
Email: s.dove@ucl.ac.uk; d.murrell@ucl.ac.uk

Funding information
H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: 766417

Abstract

Global biodiversity is facing a crisis, which must be solved through effective policies and on-the-ground conservation. But governments, NGOs, and scientists need reliable indicators to guide research, conservation actions, and policy decisions. Developing reliable indicators is challenging because the data underlying those tools is incomplete and biased. For example, the Living Planet Index tracks the changing status of global vertebrate biodiversity, but taxonomic, geographic and temporal gaps and biases are present in the aggregated data used to calculate trends. However, without a basis for real-world comparison, there is no way to directly assess an indicator's accuracy or reliability. Instead, a modelling approach can be used. We developed a model of trend reliability, using simulated datasets as stand-ins for the "real world", degraded samples as stand-ins for indicator datasets (e.g., the Living Planet Database), and a distance measure to quantify reliability by comparing partially sampled to fully sampled trends. The model revealed that the proportion of species represented in the database is not always indicative of trend reliability. Important factors are the number and length of time series, as well as their mean growth rates and variance in their growth rates, both within and between time series. We found that many trends in the Living Planet Index need more data to be considered reliable, particularly trends across the global south. In general, bird trends are the most reliable, while reptile and amphibian trends are most in need of additional data. We simulated three different solutions for reducing data deficiency, and found that collating existing data (where available) is the most efficient way to improve trend reliability, whereas revisiting previously studied populations is a quick and efficient way to improve trend reliability until new long-term studies can be completed and made available.

KEYWORDS
biodiversity data, biodiversity indicators, biodiversity trends, data deficiency, global biodiversity monitoring, indicator accuracy, indicator reliability, indicator testing, Living Planet Index

1 | INTRODUCTION

An urgent data crisis complicates the global biodiversity crisis (Turak et al., 2017). Attempts to assess global biodiversity (e.g., the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, IPBES) and to set global policies and goals that will halt or reverse its loss (e.g., the Convention on Biological Diversity, CBD, and Sustainable Development Goals, SDGs) need reliable and up-to-date scientific information (Jetz et al., 2019). Yet most studies and tracking programs are either species- or region-focused, temporally limited, and inherently biased, all of which results in large geographic and taxonomic knowledge gaps (Hortal et al., 2015; Jetz et al., 2019; Meyer et al., 2015; Proença et al., 2017; Turak et al., 2017). Advances in technologies such as camera tracking, satellite sensors, digital image recognition, network speed and capacity, data access, and mobile devices are improving our ability to track and count populations of birds and mammals (Lausch et al., 2016; Nichols et al., 2011; Rose et al., 2015) but our datasets are far from complete. The situation is worse for amphibians, reptiles, insects, and other groups, for which many species have yet to even be described (Mora et al., 2011).

We need tools to improve our understanding of global biodiversity within the limitations imposed by biased and incomplete datasets. Mace and Baillie (2007) suggested a solution: develop indicators based on existing data, understand data biases, and develop methods to reduce the bias. Biodiversity indicators summarize complex scientific information in a simple way, often serving as a bridge between science and policy (Secretariat of the Convention on Biological Diversity, 2006). But what can we expect from indicators that summarise only a fraction of the biodiversity they represent? To what extent can we rely on them to present a true picture of the state of global biodiversity?

Two of the best-known biodiversity indicators are the Living Planet Index (LPI), which tracks vertebrate population trends (McRae et al., 2017), and the Red List Index (RLI), which tracks extinction risk trends (Butchart et al., 2005). The RLI is based on extinction risk classifications at the species-level, created by expert assessment using an objective set of criteria (IUCN, 2012). By contrast, the LPI uses continuous population data collected by scientific surveys. However, as intensive global long-term studies do not exist for most species, the LPI calculates trends from data compiled from a variety of sources, including grey literature (McRae et al., 2017). This means a lack of standardization in study design (individual population time series are standardized, but there is no standardization between populations), monitoring strategy, frequency of assessment, monitoring intensity and effort, even data type (densities, counts of individuals or breeding pairs or even nests, and population size estimates are mixed together). The LPI has taxonomic and geographical imbalances (Collen et al., 2009; McRae et al., 2017), a problem found also in other global biodiversity datasets (Boakes et al., 2010; Collen et al., 2008; Yesson et al., 2007). Further, many included time series are short (McRae et al., 2016; Proença et al., 2017; Saha et al., 2018), and shorter trends tend to be less accurate than longer ones (Arkilanian et al., 2020; Wauchope et al., 2019). Recognizing these limitations, the LPI employs statistical techniques to increase the accuracy and precision of trends. Generalized additive modelling or log-linear interpolation are used (depending on the length of a given time series) to fill in missing values in time series, bootstrapping is used to generate confidence intervals (Collen et al., 2009), and a hierarchical weighting system is applied to account for geographical and taxonomic bias (Collen et al., 2009; McRae et al., 2017).

Nonetheless, the LPI's conclusions on biodiversity change have been questioned, with Buschke et al. (2021) finding an inherent negative bias in the calculation of LPI trends due to random population fluctuations and Leung et al. (2020) finding that the LPI is biased by clusters of extreme declining populations. Further, Leung et al. (2020) used the Living Planet Database (LPD) the LPI is based on to show that global biodiversity is not declining. While the analysis of Leung et al. (2020) has been contradicted by others (Loreau et al., 2022; Murali et al., 2022; Puurtinen et al., 2022), the controversy has placed a spotlight on the LPI and other global biodiversity indicators and increased the urgency of understanding how well we can rely on them.

Without a basis for real-world comparison, there is no way to directly assess an indicator's accuracy or reliability. However, there are ways to address this question indirectly. One solution was employed by the sampled approach to the Red List Index (sRLI), which uses the minimum representative sample size (sample size being the number of species represented in the index for a particular taxonomic group) needed to achieve less than a 5% probability of falsely detecting a positive slope when the Red List Index trend is negative (Baillie et al., 2008; Henriques et al., 2020). Minimum representative sample size was determined through sub-sampling of comprehensively assessed species groups on the IUCN Red List (e.g., mammals, birds etc.; Baillie et al., 2008; Henriques et al., 2020).

Two challenges presented by the LPI require a different approach than that taken for the sRLI. First, LPI trends are based on population time series that are often short and/or infrequently measured, and there are no regional or taxonomic groups within the LPI where the data is comprehensive enough to be certain of the real-world trend. Therefore, comparing sampled trends to LPI trends would tell us little about how the sampled trends might compare to reality. Second, the LPI uses non-linear trends that change slope and direction over time, so trends should be compared in a way that reflects this. Here, we use a modelling approach to overcome these challenges, based on thousands of datasets of synthetic population time series with variations in the underlying properties of the data to represent regional taxonomic groups in the real world and sampling from those datasets. We degraded the samples by randomly removing observations and adding observation error to resemble regional taxonomic groups in the Living Planet Database (LPD, the database underlying the LPI). We then compared the trends calculated from the samples with those from the complete datasets using the Jaccard distance metric (chosen using the distance measure selection method described in Dove et al., 2022) and constructed a multiple regression

model to understand how the distance values are influenced by variations in properties of the data. Here, distance metrics can be thought of as a measure of trend accuracy. By selecting a threshold value for accuracy and applying the model to the LPI, we were able to quantify the reliability of disaggregated LPI trends and determine the number of additional time series needed to meet the threshold. Finally, we modelled and compared three different solutions for reducing data deficiency: (a) tracking unstudied populations for a decade to generate new time series for the LPD, (b) resampling previously-studied populations to update old time series in the LPD, and (c) gathering more time series from existing studies to add to the LPD. The results from this study can be used to focus data-gathering and data-collation efforts on the regions, taxa, and populations that would be of greatest benefit to improving our understanding of the state of global vertebrate biodiversity.

2 | MATERIALS AND METHODS

Figure 1 shows an overview of our methods, with each numbered step corresponding to a numbered subheading in the text.

2.1 | Synthetic data generation

We first created simulated datasets to represent "real-world" regional vertebrate groups for which the LPI calculates biodiversity trends. The LPI is often represented as a single global index trend, but can also be disaggregated into hierarchical groups: first into systems (terrestrial, marine, freshwater), then geographical realms within each system, and finally taxonomic groups within each realm. It is this lowest level of the hierarchy, the regional taxonomic groups, which we simulated. From here on simulated regional taxonomic groups will be referred to as datasets. The base units of the LPI, and of our synthetic datasets, are population time series, which we will refer to simply as populations. These populations are grouped into species, and species are grouped into datasets.

Our procedure to simulate a dataset requires six parameters: (1) the total number of populations to simulate (set to 10,000), (2) the mean number of populations assigned to each species (set to 10), (3) the number of years (length of trend) to simulate (set to 50), (4) the mean of the population mean growth rates (μ_{ds}), (5) the standard deviation of the population mean growth rates (variation among populations, σ_{ds}), and (6) the mean of the population standard deviations of the growth rate (process error, μ_η, that determines annual variation in growth rates within time series). Before determining parameter settings, we tested each parameter individually for effects on trend accuracy. We did this by generating test datasets with a range of settings for the parameter being tested and keeping all other parameters fixed (see supplementary figures for details), then followed the methods described in Sections 2.2–2.7 (below) to determine if and how trend accuracy was affected. The first parameter, total populations, affected trend accuracy only when greater than half of all populations in a dataset were sampled (see Figure S1), a situation that is unlikely for regional taxonomic groups in the LPD, as it is rare even at the species level (see taxonomic representativeness in McRae et al., 2017). The second parameter, the mean number of populations per species, had no effect on trend accuracy within the wide range of values we tested (see Figure S2). The third, trend length, did affect trend accuracy (see Figure S3) and would therefore need to be set appropriately if adapting the model for a different indicator. However, it is relatively constant across regional taxonomic groups in the LPD (all trends begin at 1970 and end at the most recent year for which there are observations in the database, e.g., 2020). Therefore, we set the first three parameters at fixed values for the "real-world" simulations. Parameters four through six are variable in the LPD and did affect trend accuracy in our test results, and were therefore set to vary in the simulations.

We modelled population time series using the stochastic exponential model with process error:

$$N_{t+1} = (1 + r_t)N_t, \quad (1)$$

where N_t is population size at year t, $1+r_t$ is annual growth rate (often referred to as lambda, or λ) at year t, and $r_t \sim N(\mu_{pop}, \sigma^2_{pop})$ models uncorrelated process error (i.e., temporal variation in the growth rate that could be caused by, for example, uncorrelated environmental variation) by sampling each annual growth rate from a normal distribution. The population process error, η, is also sampled from a distribution (so different populations have different, but similar levels of η), with $\sigma_{pop} \sim \text{Exp}\left(\frac{1}{\mu_\eta}\right)$, where $\frac{1}{\mu_\eta}$ is the rate parameter. Consequently, there is a tendency towards larger values for σ_{pop}, and therefore higher levels of process error, as μ_η, the expected value of the distribution increases.

The mean of the normal distribution of population growth rates was itself drawn from a normal distribution, $\mu_{pop} \sim N(\mu_{spec}, \sigma^2_{spec})$. Thus, populations from a species will have similar but not identical underlying mean population growth rates representing perhaps differences in environmental conditions between geographically isolated populations of a given species. In turn, similar species were grouped together into datasets, and we assumed that species within taxonomic groups had underlying population growth rates that were drawn from an identical distribution, $\mu_{spec} \sim N(\mu_{ds}, \sigma^2_{ds})$. Here, larger values for σ_{ds} lead to a broader range of underlying species growth rates, perhaps signifying broader species-specific variation in responses to drivers such as habitat change within a taxonomic group. Using this hierarchical approach therefore captures the similarity of time series within a species, and the similarity of time series between species within a taxonomic group.

Growth for each population was modelled for 50 years, starting at a population size of 100. Populations were assigned to species by randomly sampling from a pool of 1000 species labels, with replacement, resulting in a normal distribution of populations per species, $pps \sim N(\mu_{pps}, \sigma^2_{pps})$, with $\mu_{pps} = 10$ and $\sigma_{pps} = 4.5$. While

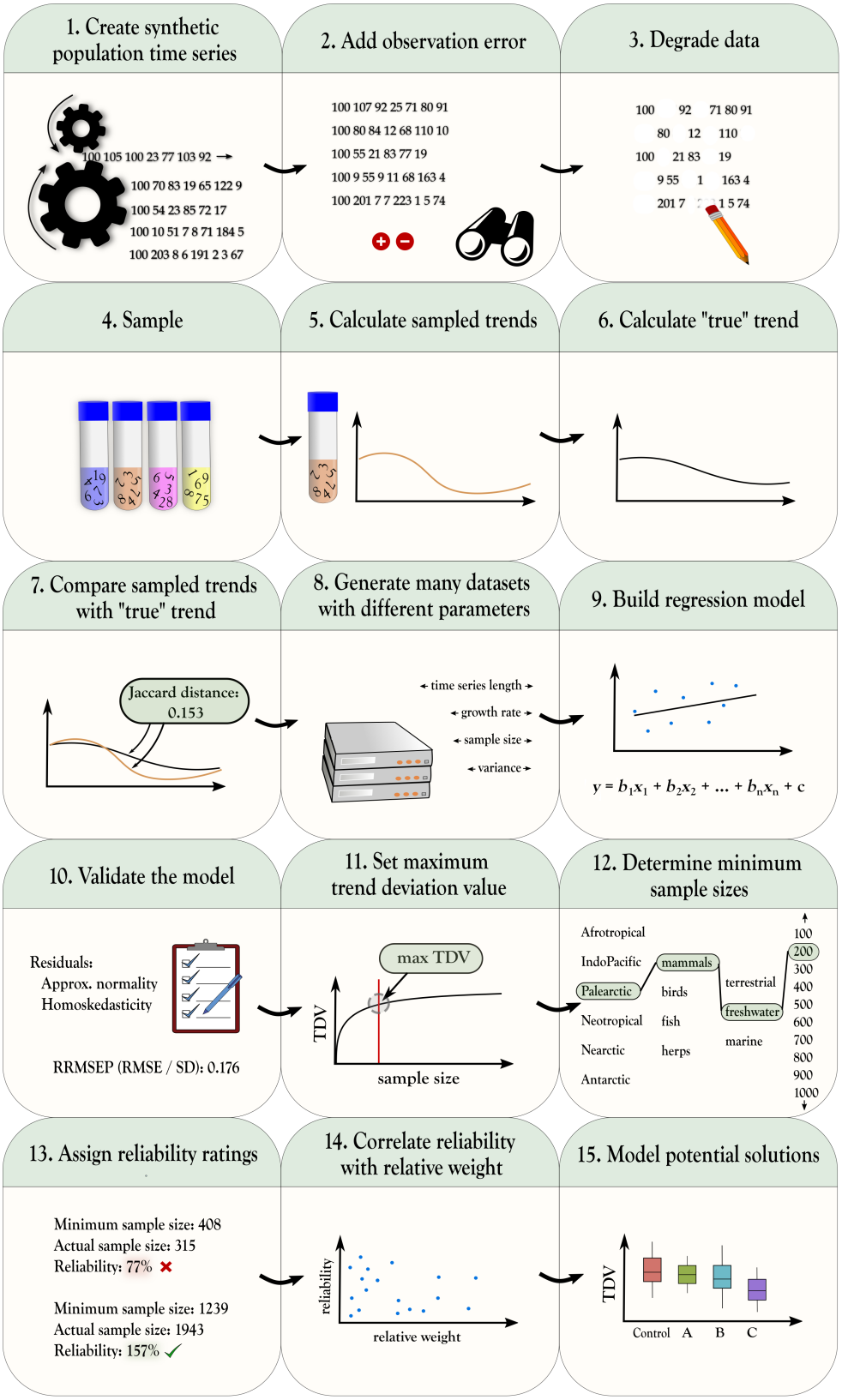

FIGURE 1 Modelling trend accuracy in the LPI: an overview. This figure illustrates the methodology used in this study. The numbered boxes correspond with the numbered steps in the methods section below. Values given in the figure are for illustration and are not intended to represent actual inputs or results.

populations are unlikely to be normally distributed across species in the real world (one would expect more rare species than common species), simulations confirmed that our modelling approach is robust against distributional assumptions for this parameter (see Figure S2).

2.2 | Observation error

The variation in population growth rates modelled above assumes all variation is due to process error. However, time series in the LPD are based on population estimates, which can be assumed to include some level of observation error due to for example, species misidentification, non-detection, and counting errors. This observation error is not accounted for in the LPI, but may affect trend reliability. Observation error, ε, can be calculated using the coefficient of variation (cv), defined as

$$\mathrm{cv}_{\varepsilon} = \frac{\sigma_{\mathrm{ab}}}{\mu_{\mathrm{ab}}}, \tag{2}$$

where μ_{ab} and σ_{ab} are the mean and standard deviation (respectively) of the abundance values. Since data in the LPD were collected using a variety of methods, and ε is not recorded in the database, we chose a range of ε consistent with values reported for other vertebrate surveys (Fryxell et al., 2014; Westcott et al., 2012; Zylstra et al., 2010). We determined through simulations that there is no effect of increasing observation error on trend accuracy (Figure S4), therefore an approximate range of ε should suffice. For each simulated population time series, ε was randomly selected from a normal distribution with $\mu_{\varepsilon}=0.15$ and $\sigma_{\varepsilon}=0.1$. We modelled observed abundances of a population at a time point, Z_t, as

$$Z_t = \begin{cases} N_t + \phi_t, \text{if } N_t + \phi_t \geq 0; \\ 0, \text{otherwise}, \end{cases} \tag{3}$$

where N_t is the population size at time t taken from Equation (1) and ϕ_t is a normally distributed variable, $\phi_t \sim N(0, \sigma^2_{\mathrm{obs}})$, with

$$\sigma_{\mathrm{obs}} = N_t \times \mu_{\varepsilon}, \tag{4}$$

where σ_{obs} is the standard deviation of ϕ_t. A value for μ_{ε} of 0.1 (10%) would result in approximately 68.2% of observations falling within 10% of their corresponding simulated 'true' values and 99.7% of simulated observations falling within 30%.

We chose a normal distribution for ϕ_t because we assumed there to be an equal chance of underestimating or overestimating population abundance. The LPI uses a mix of survey types and estimation methods and there does not seem to be conclusive evidence of whether one type of bias is dominant (e.g., Johansson et al., 2020; Kiffner et al., 2022; Lubow & Ransom, 2016; Manning & Goldberg, 2010).

2.3 | Data degradation

Observed versions of the datasets were then randomly degraded to resemble the varied quality of sampled real-world data present within the LPD. The length (number of years from first to final observation) for each degraded time series within a dataset was randomly chosen by sampling from a Poisson distribution. We determined through simulations that varying the number of observations does not affect trend accuracy at a given time series length, so we fixed the mean number of observations at half of the mean time series length (rounded up to the nearest integer). The starting years for each time series were assigned randomly. Time series were then cut to their assigned length, and half of the remaining observations were randomly removed.

2.4 | Sampling

Populations were randomly sampled from each dataset, without replacement until the desired sample size n was reached. This was repeated to obtain 20 random samples of the same size for each dataset. Values for four of the six dataset parameters described in Section 2.1 may be different for samples than for the dataset they are selected from, and may also vary between samples: the mean number of populations per species, the mean and standard deviation of population mean growth rates, μ_{ds} and σ_{ds}, and the mean of population standard deviations of the growth rate, μ_{η}.

2.5 | Calculation of sampled trends

Non-linear index trends were calculated from each sample, following the LPI method described in McRae et al. (2017). Time series with six or more data points were modelled using a generalized additive model (GAM), as described in Collen et al. (2009), with a Gaussian (normal) distribution, smoothed by a thin plate regression spline, with the number of knots set to half the number of observations (rounded down to the nearest integer). Time series that had fewer than six data points were interpolated using the chain method (Loh et al., 2005), as described in Collen et al. (2009). The chain method imputes missing values using log-linear interpolation by

$$N_i = N_p \left(\frac{N_s}{N_p} \right)^{[(i-p)/(s-p)]}, \tag{5}$$

where N is the population estimate, i is the year for which the value is to be interpolated, p is the preceding year with an observed value, and s is the subsequent year with an observed value. For all populations, whether interpolated or modelled by a GAM, species indices were formed by a three-step process. First, population sizes were converted to annual rates of change by

$$d_t = \log_{10} \left(\frac{N_t}{N_{t-1}} \right), \tag{6}$$

where N is the population estimate and t is the year. Second, average growth rates were calculated for each species by

$$\overline{d_t} = \frac{1}{n}\sum_{i=1}^{n_t} d_{it}, \quad (7)$$

where n_t is the number of populations in a given species, d_{it} is the growth rate for population i at year t, and $\overline{d_t}$ is the average growth rate at year t. Growth rates were capped at [−1:1]. Finally, index values were calculated by.

$$I_t = I_{t-1} \times 10^{\overline{d_t}},\ I_0 = 1, \quad (8)$$

where I is the index value and t is the year. Equations (5–8) are from Collen et al. (2009).

2.6 | Calculation of the 'true' trend

A non-linear index trend was calculated for each complete, undegraded dataset (without observation error), following McRae et al. (2017), as for the sampled trends. However, the undegraded datasets had no missing values, therefore modelling each time series using the chain method or a GAM was unnecessary, and that step was skipped.

2.7 | Comparison of trends

We selected an appropriate distance measure to compare sampled trends with 'true trends' using the selection process described in Dove et al. (2022). Of the distance measures deemed appropriate, we chose the Jaccard distance because it uses a 0–1 scale, making it easier to interpret. The Jaccard distance is calculated as

$$d_{\text{Jaccard}} = \frac{\sum_{t=1}^{n}(P_t - Q_t)^2}{\sum_{t=1}^{n} P_t^2 + \sum_{t=1}^{n} Q_t^2 - \sum_{t=1}^{n} P_t Q_t} \quad (9)$$

(from Cha, 2007), where P_t and Q_t are index values from two trends P and Q at time point t, and n is the number of time points. From here on, any value calculated by applying the Jaccard distance to compare sampled versus 'true' trends will be referred to as a trend deviation value, or TDV.

We used TDV here as a measure of trend accuracy, but it is in fact the complement of accuracy (a perfectly accurate trend would yield a TDV of zero); lower TDV means higher accuracy. Furthermore, when referring to TDVs of simulated trends, we used the term 'trend accuracy,' but when referring to TDVs of LPI trends, we used the term 'trend reliability.' This is because TDVs for simulated trends were measured, while TDVs for LPI trends were estimated based on a model. Trend reliability is thus a measure of *expected* accuracy based on underlying data sufficiency or deficiency, but should not be considered a proxy for accuracy. In other words, a data deficient trend may be accurate but we cannot rely on it to be so.

2.8 | Generation of datasets

We generated 3000 datasets (each consisting of 1000 species and 10,000 populations), with each dataset sampled 20 times, resulting in 60,000 samples. Values for mean time series length, μ_{ds}, σ_{ds}, and μ_{η} were randomly selected from uniform distributions, while sample size was randomly selected from a log-uniform distribution, $\ln(SS) \sim U(\ln[a], \ln[b])$, where SS is sample size and a and b are the minimum and maximum values, respectively (log-uniform was chosen to ensure the model would be robust at small sample sizes, as most datasets in the LPD are small). Ranges for the distributions were chosen to ensure that parameter ranges in the samples would be broader than the ranges present in the LPD (Table 1). Regional taxonomic groups from the LPD with fewer than 20 populations were excluded from parameter range calculations to avoid extreme outliers. We set the minimum sample size to 50 because smaller samples rarely generated a complete trend, and the maximum to 10,000 to improve predictions of the effects of sample size increases.

2.9 | Multiple regression model

We built a multiple linear regression model to understand how variables in the simulated data determine trend accuracy (TDV). First, we removed all simulated datasets in which the mean of the sample parameter values fell outside of LPD parameter ranges (individual replicates were allowed to fall outside of LPD ranges), leaving 2361 datasets, or 47,220 samples. We then randomly selected 67% of the remaining datasets (1581 datasets) to train the

TABLE 1 Parameters with value ranges for simulated datasets, degraded samples and the LPD.

Independent variable	Range in datasets	Range in samples	Range in LPD
Sample size	—	50–9975	2–3000
Mean length of time series	6.0–38	5.5–39	6.0–39
Mean of Pop. mean growth rates, μ_{ds}	−0.13–0.12	−0.25–0.31	−0.19–0.16
St. dev. of Pop. mean growth rates, σ_{ds}	0.074–0.59	0.097–0.83	0.12–0.63
Mean of Pop. growth rate St. Dev., μ_{η} [a]	0.049–1.17	0.13–1.06	0.16–0.89

[a]This parameter is modelled as process error in the simulated datasets, but in the degraded samples it represents process error and observation error combined.

model. The other 33% (780 datasets) we set aside for testing the model.

2.10 | Model validation

The residuals of the combined data used to train the model were approximately normally distributed. Likewise, the residuals appeared homoscedastic when plotted against fitted values. We compared the actual TDV of each sample in the testing datasets to the predicted TDV for that sample calculated by the model, then calculated the RRMSEP (relative root mean squared error of prediction), defined as

$$\text{RRMSEP} = \text{RMSE} / \text{SD}, \tag{10}$$

where RMSE is the root mean squared error and SD is the standard deviation of the actual TDVs, and

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}\right)^2}{n}}, \tag{11}$$

where y_i is the ith actual TDV, $\hat{y}$ is the predicted TDV, and n is the number of samples.

2.11 | Maximum trend deviation value

We set a maximum predicted TDV as a threshold that regional taxonomic group trends within the LPI should not exceed to be considered reliable. First, we built a linear regression model of the square root of TDV from our training datasets, with the natural log of sample size as the predictor variable, since sample size is the only user-controlled variable within the LPD. Every regional taxonomic group within the LPD represents a single sample from the real world; therefore, we were not interested in the mean TDV achieved by each dataset, but in the range of possible TDV values, especially the upper part of the range (the least accurate sample trends from each dataset).

We used 10,000 bootstrap estimations of the mean of the TDV from each dataset to calculate the 90% confidence intervals using the bias corrected and accelerated bootstrap interval (BCa) method, also known as the adjusted bootstrap percentile method. The BCa method is a non-parametric method that does not assume the data is normally distributed (the TDV values have a beta distribution) and corrects for bias and skewness in the distribution of the mean estimates. We plotted the curve of the square root-log model of the upper 90% confidence interval of TDV in relation to sample size on a (non-log) graph of TDV versus sample size (Figure 2).

Increasing sample size should naturally lead to more desirable TDV but it is costly in terms of time and money to increase sample size, and it may also be prudent to put the resources elsewhere. It was therefore important to choose a maximum TDV that reflects these trade-offs. To choose a maximum TDV, we used a method called the concordance probability method (CZ) (Liu, 2012). We adapted CZ from the field of biomedical research, where it is often necessary to specify a cut-off value to discriminate between positive and negative results from screening or diagnostic tests (Liu, 2012). First, a receiver operating characteristic (ROC) curve is built, plotting the rate of true positives (sensitivity) against the rate of false positives (1 − specificity). The idea is to find the point on the curve that maximizes both sensitivity and specificity. The CZ method simply finds the point where their product is maximized.

By considering the square root-log model of the upper 90% confidence interval of TDV versus sample size (Figure 2) as equivalent to an ROC curve, we applied the CZ method to find the point on the curve where TDV and sample size are minimized. This is the point where we should achieve maximum value from the data. Further right along the curve, increasing the sample size would give a smaller improvement in trend reliability and is therefore not cost- or resource-effective. Since an ROC curve is intended for binary classification, the CZ method assumes that both sensitivity and specificity are on a 0–1 scale. TDV already ranges from 0 to 1, so we set sensitivity as 1–TDV. We normalized sample size to a 0–1 scale by converting it to a proportion of the complete dataset (dividing by the total number of

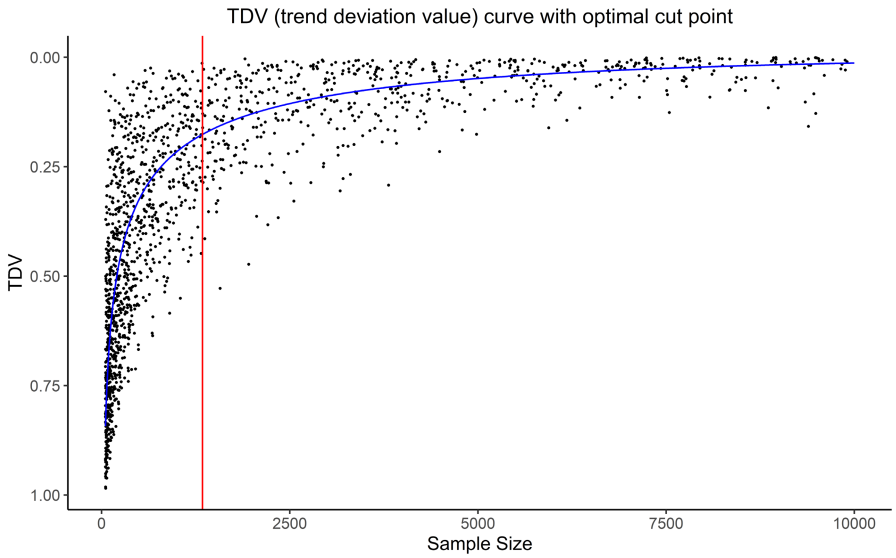

FIGURE 2 Trend deviation value (TDV) versus sample size. This plot includes only the upper 90% confidence interval of TDV from each simulated dataset. The curved blue line is the square root-log model of the plotted values. The vertical red line intersects the square root-log curve at the optimal cut-point.

time series in the dataset). Since all datasets were the same size, the relationship between TDV and sample size was not altered by the conversion to a proportion. Specificity was then 1–sample size. The optimal cut-point on the curve is defined as

$$\max(CZ), \quad CZ(c) = Se(c) \times Sp(c), \tag{12}$$

where Se is sensitivity, Sp is specificity, and c is any cut-point.

2.12 | Minimum sample size for regional taxonomic groups

Minimum sample size was calculated by rearranging the formula for the multiple regression model to solve for sample size and replacing the TDV variable in the formula with the cut-off value determined above. Values for the other variables in the formula were determined separately for each regional taxonomic group from the LPD, as follows: populations with fewer than two data points were removed, missing data was interpolated using the chain method (Collen et al., 2009), then the mean growth rate, μ_{pop}, was calculated for each population. Growth rates were capped at [−1:1] before taking the mean, as in the LPI (McRae et al., 2017). Next, μ_{ds}, σ_{ds}, and μ_{η} were calculated. The mean time series length was calculated by dividing the total number of observations (after interpolation) by the total number of populations (excluding those with fewer than two data points). The calculated values were then placed into the model formula to determine minimum sample size.

2.13 | Assigning reliability ratings to regional taxonomic groups

The actual number of populations in each regional taxonomic group was divided by the minimum sample size and multiplied by 100 to determine the percentage of the minimum sample size actually met by each group. Groups achieving 100% or greater were designated as reliable, those achieving between 50% and 100% were designated as data deficient, and those achieving <50% were designated as severely data deficient.

2.14 | Correlations between reliability rating and LPI relative weighting

The LPI uses a weighting system to account for the estimated number of species in each regional taxonomic group to reduce representational bias (McRae et al., 2017). Each regional taxonomic group has a relative weighting assigned to it, which is used in the calculation of aggregated indices. We used the Pearson's product moment correlation coefficient test to determine if there was any significant correlation between percentage of the minimum sample size achieved for each regional taxonomic group and the assigned relative weightings in the LPI for each group. The test was performed on the combined dataset as well as each individual system.

2.15 | Modelling potential solutions

We used the model to simulate three different methods of improving trend reliability in the LPD: (A) tracking unstudied populations for 10 years, (B) resampling previously-studied populations, and (C) gathering more time series from existing studies. First, we generated 50 control datasets (as described in Sections 2.1–2.7) with a sample size of 200 and mean time series length of 14 (similar to the median values for regional taxonomic groups in the LPI of 180 and 13, respectively). We set μ_{ds} to 0, σ_{ds} to 0.30, and μ_{η} to 0.40. Using the same parameters, we then generated groups of 50 datasets with each of the following changes: group A had an extra 200 populations (total sample size: 400), but with observations only for the final 10 years, to simulate tracking additional populations for 10 years; group B had the final observation revealed on every sampled, degraded time series (total sample size: 200) to simulate resampling previously-studied populations; group C had an extra 200 randomly sampled populations (total sample size: 400) to simulate adding existing data to the LPD.

2.16 | Coding and data

All trends for synthetic data were produced using original code designed to reproduce the functionality of the rlpi package (Freeman et al., 2021). All coding was done in R (R Core Team, 2021) using RStudio (RStudio Team, 2022). Figure 1 was produced using Inkscape (Inkscape Project, 2020). All other figures were produced in R (R Core Team, 2021) using the ggplot2 package (Wickham, 2016). Population time series used to evaluate reliability of LPI trends are from the LPD (McRae et al., 2016). All original code is available on GitHub at https://github.com/ShawnDove/DD_LPI.

3 | RESULTS

3.1 | Regression model

The regression model contains five independent variables (Tables 1 and 2). Together they describe 62% of the variation (adjusted r-squared: .62) in the TDV associated with sampled trends, and with $F(5, 29{,}385) = 9686$, $p < .001$. All independent variables are statistically significant predictors, with $p < .001$ (Table 2). Interaction terms are also statistically significant but do not increase the adjusted r-squared of the model, so we left them out. RRMSEP is 0.231. Sample size is the most important variable affecting trend accuracy. As expected, higher sampling leads to a lower TDV (higher accuracy). The other variables all have smaller

TABLE 2 Multiple regression model of ln(TDV).

Coefficient	Estimate	Standard error	Beta coefficient	*t* value	*p* value
(Intercept)	3.957	0.04406	—	89.81	<.001
ln (Sample size)	−0.8460	0.004441	−0.6860	−190.5	<.001
ln (St. dev. of mean growth rate, σ_{ds})	0.7569	0.01630	0.1672	46.42	<.001
Mean growth rate, μ_{ds}	8.057	0.1454	0.1989	55.42	<.001
Mean of population St. dev., μ_{η}	1.503	0.02224	0.2426	67.57	<.001
Mean time series length	−0.03890	0.0007336	−0.1917	−53.02	<.001

effects on the variation in TDV. We found a positive effect of μ_{ds}, the mean growth rate for a taxonomic group, on the TDV and this is because there are no upper bounds on population size, whereas all populations are bounded by zero. Hence declining populations tend to have less variation across time series compared to growing populations. Increasing standard deviation of the mean growth rate (σ_{ds}) leads to an increase in the TDV, because higher variation in growth rates among populations effectively lowers the signal-to-noise ratio. Process error leads to more variation in growth rates within populations; therefore increasing μ_{η}, the parameter that modulates the strength of process error, also leads to higher values for TDV. Finally, longer time series lead to generally lower TDV because they provide data for each population over a larger portion of the overall trend length.

Much of the unexplained variance from the model is due to random sampling. We confirmed this by remaking the model using the sample means, which resulted in an adjusted *r*-squared of .87. Using the square root of TDV instead of the log further increased the adjusted *r*-squared to .93. This was not the case for the model using the individual samples, where the log resulted in a higher adjusted *r*-squared than the square root.

3.2 | Maximum trend deviation value

Using the concordance probability method to select a cut point on the square root-log model of the 90% upper confidence interval of TDV versus sample size, we found a maximum TDV value of 0.176. After placing this value into the model equation and reorganizing to solve for sample size, we applied the model to the LPI to find the minimum number of populations needed for each regional taxonomic group.

3.3 | Minimum sample size

The number of populations needed to achieve the TDV threshold for a reliable trend varies across taxonomic groups and realms (Table 3), but only weakly across systems, with medians of 269, 341, and 263 for terrestrial, freshwater, and marine systems, respectively. Fewer populations are needed in the global north (median: 213) than in the global south (median: 354). Birds show the highest variability, having both the smallest number of populations needed for any group (freshwater Nearctic birds: 19), and the largest (freshwater Afrotropical birds: 9081; however, this value is an extreme outlier—see Discussion). Mammals have the smallest sample size requirements, with a median of 165, while fishes have the largest, with a median of 465. Reptiles and amphibians (combined) and birds fall in between, with medians of 274 and 286, respectively.

3.4 | Trend reliability

Reliability varies strongly across realms, taxonomic groups, and systems (Figures 3 and 4). Terrestrial trends are the most reliable and freshwater trends the least. Terrestrial and freshwater trends are more reliable in the global north than in the global south, except for terrestrial reptiles and amphibians. Marine bird trends are more reliable in temperate areas than the tropics, while marine fish trends are more reliable in tropical waters than polar. Globally, bird trends are the most reliable but are nonetheless poor in the tropics, especially Africa. Reptile and amphibian trends are data deficient everywhere except the terrestrial Neotropical realm, and marine and freshwater mammal trends are data deficient everywhere (although marine IndoPacific mammals are very close to the threshold at 97%).

The regional taxonomic groups with the greatest potential to affect the reliability of aggregated LPI trends are exclusively tropical (Figure 5), due to a combination of high relative weighting and low reliability scores. The eight groups of greatest concern include five freshwater and three terrestrial groups, but no marine groups. All are from the tropics. Fishes, birds and reptiles and amphibians are represented, with mammals absent. Overall, the reliability scores of regional taxonomic groups do not show a statistically significant correlation with their relative weightings in the LPI, $r(55) = .085$, $t = 0.64$, $p = .53$. Likewise, there are no statistically significant correlations for individual systems, with terrestrial $r(13) = -.40$, $t = -1.56$, $p = .14$; freshwater $r(18) = -.09$, $t = -0.39$, $p = .70$; and marine $r(20) = .40$, $t = 1.95$, $p = .066$.

TABLE 3 The trend deviation value, the current number of populations in the LPD, the minimum number of populations that would meet the reliability threshold, and the number of additional populations that must be added to achieve the reliability threshold for each regional taxonomic group in the LPD. Note that the trend deviation values here were calculated using the model formula and therefore occasionally fall outside of the 0–1 range of the Jaccard distance the TDV is based on.

System	Realm	Taxon	TDV	Current sample size	Minimum sample size	Additional pops needed
Terrestrial	Afrotropical	Birds	0.444	330	983	653
		Mammals	0.036	794	122	0
		Reptiles & Amphibians	0.477	51	166	115
	IndoPacific	Birds	0.085	956	406	0
		Mammals	0.060	1581	441	0
		Reptiles & Amphibians	0.867	81	533	452
	Palearctic	Birds	0.030	988	122	0
		Mammals	0.019	2104	149	0
		Reptiles & Amphibians	0.373	55	134	79
	Neotropical	Birds	0.085	640	269	0
		Mammals	0.161	314	283	0
		Reptiles & Amphibians	0.161	238	214	0
	Nearctic	Birds	0.024	514	49	0
		Mammals	0.110	686	394	0
		Reptiles & Amphibians	0.564	129	511	382
Freshwater	Afrotropical	Birds	6.483	128	9081[a]	8953[a]
		Mammals	0.521	15	54	39
		Reptiles & Amphibians	0.729	18	96	78
		Fishes	0.925	149	1058	909
	IndoPacific	Birds	0.172	388	378	0
		Mammals	0.359	22	51	29
		Reptiles & Amphibians	0.269	114	188	74
		Fishes	0.334	367	781	414
	Palearctic	Birds	0.062	973	284	0
		Mammals	0.203	188	223	35
		Reptiles & Amphibians	0.383	90	225	135
		Fishes	0.183	580	607	27
	Neotropical	Birds	0.331	161	340	179
		Mammals	2.789	22	576	554
		Reptiles & Amphibians	0.824	103	638	535
		Fishes	0.125	2217	1482	0
	Nearctic	Birds	0.042	101	19	0
		Mammals	0.352	23	52	29
		Reptiles & Amphibians	0.280	280	484	204
		Fishes	0.090	752	342	0

TABLE 3 (Continued)

System	Realm	Taxon	TDV	Current sample size	Minimum sample size	Additional pops needed
Marine	Temperate Atlantic	Birds	0.044	581	112	0
		Mammals	0.379	138	342	204
		Reptiles & Amphibians	0.916	46	323	277
		Fishes	0.048	2170	465	0
	Tropical Atlantic	Birds	0.479	225	733	508
		Mammals	1.299	17	180	163
		Reptiles & Amphibians	0.937	90	649	559
		Fishes	0.040	3111	539	0
	Arctic	Birds	0.189	110	120	10
		Mammals	0.330	49	103	54
		Fishes	1.820	29	458	429
	South temperate	Birds	0.047	1361	288	0
		Mammals	0.426	30	85	55
		Fishes	0.181	230	238	8
	IndoPacific	Birds	0.038	5392	874	0
		Mammals	0.181	88	91	3
		Reptiles & Amphibians	0.623	81	361	280
		Fishes	0.069	1059	347	0
	Pacific temperate	Birds	0.127	146	100	0
		Mammals	0.292	117	213	96
		Reptiles & Amphibians	6.528	2	143	141
		Fishes	0.060	706	200	0

[a]Minimum sample size for freshwater Afrotropical birds is an extreme outlier. See Section 4 for explanation.

3.5 | Modelling potential solutions

Adding 200 additional time series to the sample with observations only in the final 10 years (solution A; equivalent to tracking 200 unstudied populations for 10 years) improved the mean TDV by 12%, while revealing the final year observation (solution B; equivalent to resampling previously-studied populations) for every population improved the median TDV by 11% (Figure 6). By contrast, simply doubling the sample size (solution C; equivalent to randomly adding 200 existing time series to the LPD) improved the median TDV by 50%. This solution shows a statistically significant improvement in trend accuracy compared to the control group ($p < .001$).

4 | DISCUSSION

Understanding the changing global state of biodiversity is crucial to making good policy and conservation decisions and 'bending the curve' of biodiversity loss (Mace et al., 2018). Acquiring accurate and comprehensive data is crucial, but the first step is to answer the question: what do we actually know? The present study quantifies the reliability of trends for each regional taxonomic group in the Living Planet Index and estimates the number of population time series needed to meet a standard of expected accuracy.

We used synthetic population time series datasets to construct a multiple regression model of trend accuracy by comparing trends of degraded samples with the trends of the full, undegraded datasets using a distance measure (Figure 1). We applied the model to regional taxonomic groups in the Living Planet Database to reveal that the majority need additional data for their trends to be considered reliable. Data deficiency is a problem globally but is more pronounced in the tropics. This is consistent with the analysis of geographical representativeness in McRae et al. (2017), which tested proportional representativeness of biodiversity compared to the global dataset and found that species groups in tropical realms are underrepresented. Bird trends are the most reliable and reptiles and amphibians the least. This is consistent with the picture of species representation in the LPD presented in McRae et al. (2017) and is unsurprising given that monitoring and data collection for birds is more extensive than for reptiles

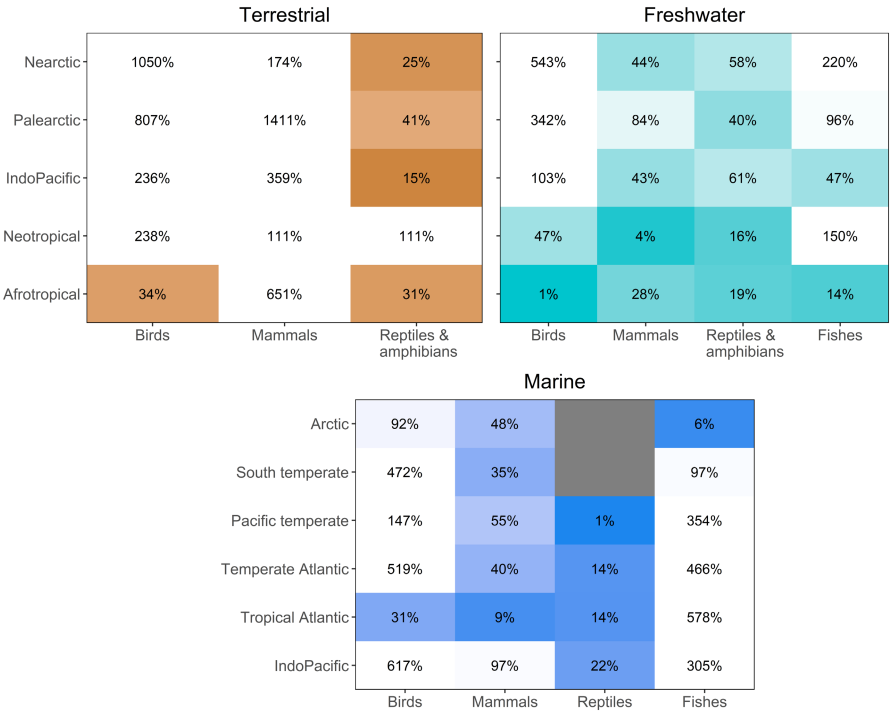

FIGURE 3 Proportion of the total amount of time series data needed to achieve the trend reliability threshold that each regional taxonomic group in the LPD currently contains. A score of 100% or greater means that group already has enough data to produce a reliable trend. A white box refers to a group that meets the reliability threshold, while a colored box means the threshold has not been met. The further the group is from meeting the threshold, the more intense the color. A grey box refers either to a group that could not be evaluated because there was too little data (South temperate marine reptiles) or due to an invalid realm-taxon combination (there are no marine reptiles in the Arctic).

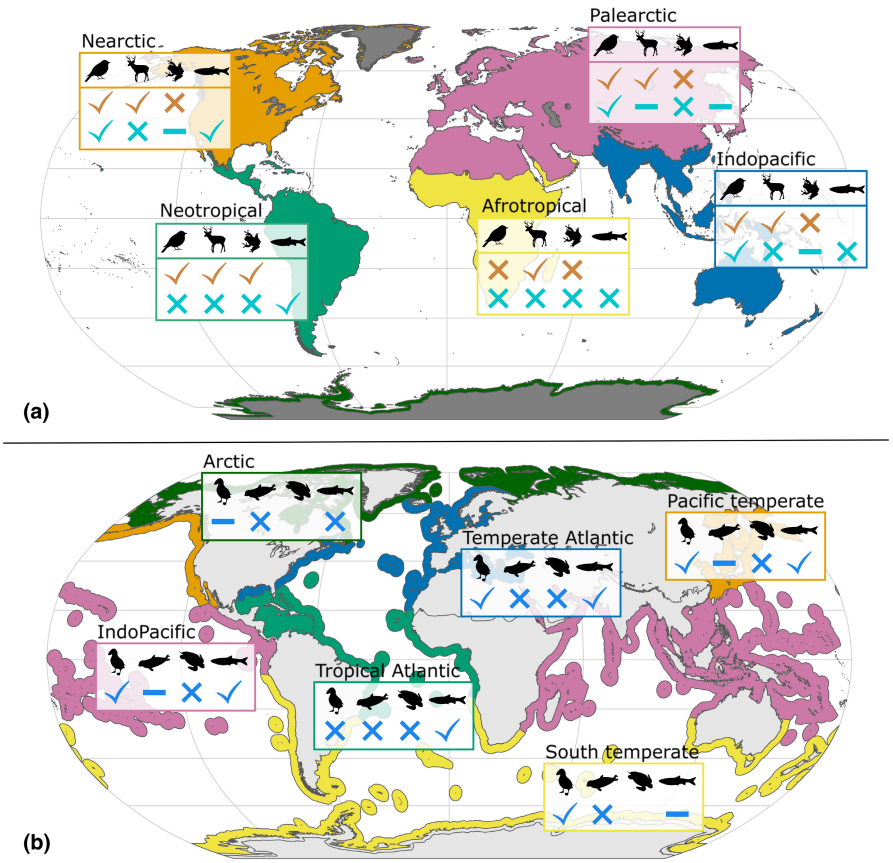

FIGURE 4 Reliability of regional taxonomic group trends in the LPI, grouped by system, realm, and taxon. Map (a) shows the terrestrial (top) and freshwater (bottom) results. Map (b) shows the marine results. The results box outlines are coloured to match their corresponding realms on each map. Reliability scores are binned into three categories, according to the number of time series in the LPD relative to the minimum sample size needed to achieve the TDV threshold. A check mark means that group has at least 100% of the minimum sample size and is considered reliable, a dash means it is data deficient (50%–99%), and an X mark means it is severely data deficient (<50%).

and amphibians (Oliver et al., 2021; Scheele et al., 2019), especially with the rise of citizen science (Oliver et al., 2021). However, many of our reliability scores differ from what would be expected given McRae et al.'s (2017) analysis of taxonomic representativeness. McRae et al. (2017) found that all Nearctic taxonomic groups are overrepresented, yet in our analysis Nearctic terrestrial and freshwater reptiles and amphibians as well as Nearctic freshwater mammals score as data deficient. The starkest differences occur in the marine system, where mammals and marine reptiles are overrepresented by species in all realms (except South temperate

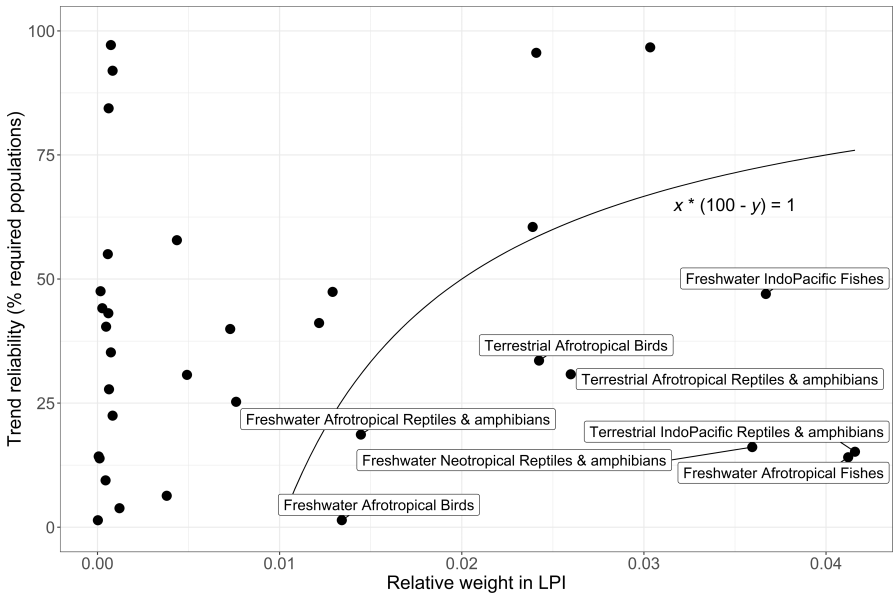

FIGURE 5 Trend reliability of regional taxonomic groups in the LPD (measured as the percentage of populations in the LPD relative to the number required to achieve the TDV threshold) versus the relative weighting applied to each group when calculating aggregated LPI trends. Only groups with reliability ratings below the threshold (<100%) are included here. To determine the groups having the strongest negative effect on the reliability of aggregated LPI trends, we calculated relative weight × (100 − reliability) and labelled the groups with a value higher than 1.

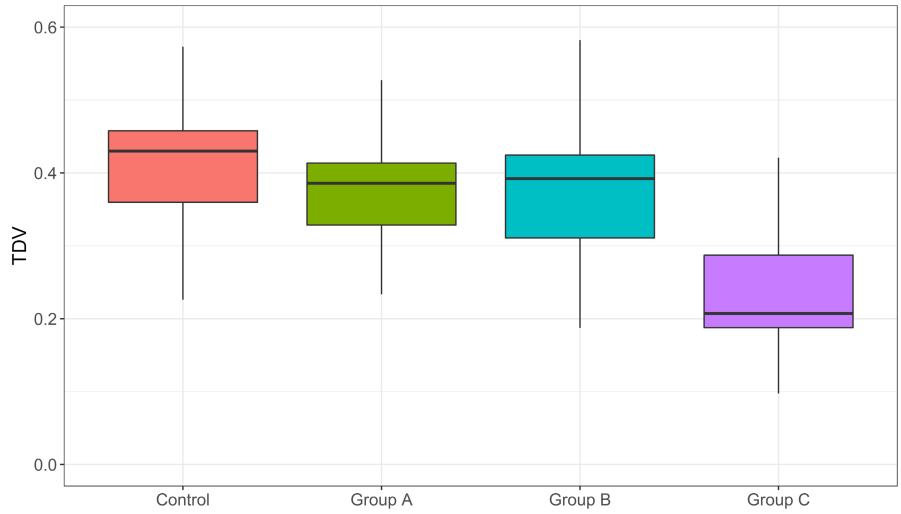

FIGURE 6 The effect on trend accuracy of potential solutions to data deficiency in LPI regional taxonomic groups. The control group has a sample size of 200 and mean time series length of 14. Group A has an additional 200 time series with observations only in the final 10 years of the index to simulate a 10-year data blitz. In group B, the final observation has been added back in for every time series to simulate resampling of previously-studied populations. Group C is like the control group, but the sample size has been doubled to 400 to simulate adding additional pre-existing studies to the LPI. All other parameters are fixed – μ_{ds}: 0; σ_{ds}: 0.3; μ_{η}: 0.4; populations per species: 10; trend length: 50; μ_{ε}: 0.15; σ_{ε}: 0.1. Each box represents the mean values of 20 datasets, with 20 samples per dataset.

reptiles, which are not represented at all) but which we found to be data deficient in all realms. In contrast, marine fishes are underrepresented by species numbers (McRae et al., 2017), but we found that in all except the Arctic realm marine fishes are data-rich enough to produce reliable trends. These differences strongly suggest that the percentage of species represented does not tell the whole story.

Geographical and taxonomic biases in the distribution of data in the LPI are well-known (McRae et al., 2017), and reflect underlying biases in the availability of data (Boakes et al., 2010; Collen et al., 2008; Yesson et al., 2007). McRae et al. (2017) introduced a weighting system to the LPI, which accounts for the estimated number of species in each regional taxonomic group to reduce representational bias. One problem with this is that most of the world's vertebrate species are located in the tropics (Collen et al., 2008; McRae et al., 2017), which are underrepresented in the LPD (McRae et al., 2017). Our concern was that if trends from these areas are the least reliable due to data deficiency, then the LPI could have simply replaced one problem, representation bias, with another: overreliance on data deficient trends. Indeed, our analysis shows that all regional taxonomic groups with a high relative weight and low reliability rating (bottom right of Figure 5) are tropical. Surprisingly, though, we did not find a statistically significant negative correlation between reliability of trends and their relative weights in the LPI. This also holds true for the terrestrial and freshwater systems when considered separately (the marine system actually shows a positive correlation) and is

consistent with Nori et al. (2020), who found that species richness and knowledge gaps are not always correlated.

According to our model, the size of a dataset, that is, the number of species or populations existing in the real world for any regional taxonomic group, is unimportant to the calculation of trend reliability for a given sample, as long as the sample represents less than half of the time series in the dataset (Figure S1). In other words, it is the absolute number of populations represented in the sample that matters, regardless of whether that sample represents 1% or 50% of the total populations in a regional taxonomic group. There are two principles working to cause this seemingly counterintuitive effect. First, the relationship between population size and the sample size needed to reach a desired level of precision is logarithmic and becomes more extreme at lower levels of precision (Israel, 1992). This means that a small sample size should be able to estimate a large population almost as well as it can estimate a small population. Second, there are limitations to the level of trend accuracy that can be achieved, regardless of sample size, because most time series in our simulated samples (and in the LPD) are much shorter than the length of the trend being estimated. Short time series tend to produce more extreme trends (Leung et al., 2020) and are less likely to accurately reflect long-term trends for individual populations (Wauchope et al., 2019). They also reduce the number of observations used for the calculation of group trends. For example, even if the mean time series length was 50% of the length of a trend (mean time series lengths for all regional taxonomic groups in the LPD are much shorter than that), if those time series were randomly distributed in time, only about 4% of them would begin at the first year and about 4% would end at the final year. Thus, the crucial early and final years of the trend would depend on only a fraction of the observations that the sample size indicates. This randomized distribution of time series across the trend results in less accurate trends than would be possible if observations were evenly distributed across time points (confirmed through simulations—see Figure S5). This issue is slightly complicated in the LPD. On one hand, the database begins 20 years earlier than the index, giving time for the number of observations to increase before measuring the trend. On the other hand, there is a delay in getting recent studies into the LPD (McRae et al., 2017), reducing the number of observations in the final years even more than a random distribution would suggest (see Figure S6).

This dramatic fall-off of observations suggests that more data is needed for the LPI to reliably reflect changes in the status of global vertebrate biodiversity over the past decade. While a reduction in the delay involved in getting new studies into the LPD might help, increasing the number of populations in the LPD is only possible to the extent that the necessary data exists. Therefore, we simulated two potential ways of generating new data to improve trend reliability: (A) a global data blitz, with researchers coordinating to track as many unstudied populations as possible for 10 years to generate new time series, and (B) resampling already-studied populations to uncover recent changes and lengthen existing time series (Figure 6). Both solutions had a slight but non-significant positive effect on trend accuracy but were far less effective than adding existing data (solution C; as is currently done for the LPD). It is likely that both solutions (data blitz and resampling) have a greater effect on the accuracy of the final portion of the trend than on the overall trend, but further study would be required to be certain. Either way, resampling would be more efficient than a data blitz, as the same improvement could be achieved in 1 year instead of 10. In the long term, tracking additional populations is essential to completing our picture of biodiversity change. Natural stochasticity means that short time series are of limited value in generating reliable trends (Wauchope et al., 2019), so tracking additional populations takes time to pay dividends. Nonetheless, overcoming indicator biases and data deficiencies will require a balanced global profile of populations, counted regularly to ensure changes can be detected quickly.

There is another limitation underlying the LPI, which cannot be solved by generating new data. All trends in the LPI begin in the year 1970, which is set as the base year for calculating the index values. Past trends can only be determined by existing data; therefore, while there may be some currently inaccessible data that either could be shared or made available for confidential storage in the LPD (Saha et al., 2018), there are likely to be severe limitations to relieving data deficiency for the early years of the LPI. However, other potential solutions could be examined in future studies. One would be to begin the index at a later year in which there is more data available (e.g., 1990). Another would be to change the base year for calculating the index to a more data-rich year, thus increasing the uncertainty around the early years of LPI trends (Gregory et al., 2019). The downside is that the interpretation of trends would be different. The LPI would no longer measure change in global vertebrate biodiversity relative to 1970, but relative to another year, and much of the change currently recorded in the index would have already occurred before the base year. A different approach would be to use other kinds of data, such as log books and catch records (e.g., Josephson et al., 2008), genetics (e.g., Beaumont, 2003), trade records (e.g., Collins et al., 2020), and land use/climate change modelling (e.g., Visconti et al., 2016) to infer historical abundance estimates for populations where no monitoring took place.

Our regression analysis of the simulated data highlights some rather straightforward results—more data in terms of sampled populations and/or longer time series leads to higher reliability of trends, and more variation in population growth rates within and between populations leads to lower reliability. However, we also found that regional taxonomic groups that show positive trends might need more data (higher sampling, longer time series) than those that are declining. The corollary is that fewer samples might be required to obtain reliable trend estimates for declining groups, but this result also has implications for any biases in species selection. For example, monitoring efforts tend to focus on species at higher risk of extinction (Scheele et al., 2019). Many amphibian populations in the LPD were tracked because they were declining due to the devastating disease *chytridiomycosis*. This could negatively bias trends and falsely reduce variance in growth rates, leading the model to overestimate reliability because it assumes that tracked populations are randomly

selected. On the other hand, Murali et al. (2022) found that population coverage in the LPD is biased towards protected areas, where species are less likely to be threatened, therefore potentially causing a positive bias in LPI trends. Our results also imply that any biases towards or against species that have high/low process error, that is, have very variable annual growth rates, could potentially also bias our estimates of trend reliability. However, our analysis of the simulated data suggests overall sampling intensity far outweighs the other factors included in our model, not least because as sampling number increases so does the coverage of the variability in the taxonomic group.

Other biases in the LPD could also have important effects on our estimates of reliability. Time series are non-randomly distributed across time and/or space in the LPD. For example, while some biodiversity hotspots (e.g., tropical Africa) are poorly sampled, others, especially islands (e.g., Madagascar), are well-studied (Nori et al., 2020), and this may bias entire realms. In the Afrotropical realm, six (12%) of the 51 terrestrial reptile and amphibian time series in the LPD are from Round Island (a tiny uninhabited island near Mauritius) and more than half (57%; 29/51) are from a single study that took place at a reserve in Madagascar over a nine- year period; only seven (14%) are from mainland Africa, and of the seven, four are from a single study at a reserve in Nigeria. In this case, the model likely severely underestimates the amount of data needed to get a reliable trend. Valdez et al. (2023) found that a coarser sampling resolution increases the ability to detect global biodiversity change by reducing the effects of outlier population trends. Sampling resolution biases such as that in the Afrotropical realm will surely decrease trend reliability at a given sample size. While the Afrotropical realm may be an extreme example, it shows that there are important underlying aspects of the data that cannot be assessed by a model. Fortunately, these issues tend to diminish when more data is present, and thus should not have a large effect on trends assessed as reliable. Our model also assumes that adding additional time series to the LPD will maintain the parameters of the regional taxonomic group to which they are added (e.g., the mean time series length and the level of variance in population mean growth rates will not change). This can result in the model severely overestimating the numbers of populations required to achieve a reliable trend. For example, it suggested that 9087 populations of freshwater Afrotropical birds are needed. This likely occurred due to problems with the existing data. Although there are 128 freshwater Afrotropical bird populations in the LPD, most of them are short and/or sporadically observed (the mean number of observations is 4.0), and observations are clustered in the 1990s and 2000s, with only a single time series containing observations after 2009. However, this is an issue only for small or exceptionally poor quality samples (e.g., short time series, few studies, biased distribution in time and space), and if more and better time series are added to the LPD, the model should improve its estimates.

Another limitation of our modelling approach is that we could not correct for the sizes of the 'real-world' datasets (the number of populations that exist) that the LPD 'samples' are drawing from, and therefore may overestimate the sample size needed to achieve a reliable trend for very small datasets. Although there are estimates of the number of species for each regional taxonomic group, our model uses populations as the base unit to measure sample size. We chose to base sample size on populations rather than species for two reasons. First, we found that mean growth rates within the LPD vary almost as much between populations within a species as they do between species. Therefore, we cannot assume that the trend of a population represents the trend of the species it belongs to any better than it represents the trend of its entire regional taxonomic group. Second, localized threats such as land-use change and habitat destruction are likely to affect some populations within a species disproportionately. Population extinctions also occur much more frequently than species extinctions, and may serve as an early warning (Ceballos et al., 2017). However, a population is not a well-defined unit, and we do not have estimates of how many populations each species or regional taxonomic group is composed of. While our testing suggested we can assume the number of existing populations to be unimportant in determining trend reliability, this assumption breaks down when the sample comprises a large percentage of the dataset. It is unlikely that any regional taxonomic groups currently approach this level of representation within the LPD, but it is nonetheless an important caveat to be aware of.

Despite these caveats, the results of our study reveal the strengths and weaknesses in our understanding of global vertebrate biodiversity, highlighting the regional taxonomic groups for which we have enough data to make responsible decisions, as well as those on which future data gathering and collation efforts should focus. Some underlying aspects of the data create biases that are not taken into account by our modelling approach, and more fine-scale studies on gaps in population trends should be performed to better understand these biases and where to divert scientific resources. We show that revisiting previously-studied populations is a quick and efficient way to improve trend reliability for data deficient groups until more long-term studies can be completed and made available. The modelling approach we use to quantify trend reliability can also be generalized to assess other global and/or regional biodiversity indices that utilize population time series data. We are facing an urgent global biodiversity crisis made worse by biased and deficient data, but through careful study and cooperative global efforts we can solve the data problem and begin to 'bend the curve' of biodiversity toward a positive trend.

ACKNOWLEDGMENTS

We thank Sean Jellesmark, Gonzalo Albaladejo-Robles, and Bouwe Reijenga for their support. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 766417.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The simulated datasets used for this study are openly available in Dryad at https://doi.org/10.5061/dryad.mpg4f4r52 (Dove et al., 2023). Some of the Living Planet Database time series used in this study are not available to the public, but a limited version is available at https://www.livingplanetindex.org/data_portal. All R code used in this study is available at https://github.com/shawndove/DD_LPI.

ORCID

Shawn Dove https://orcid.org/0000-0001-9465-5638
Louise McRae https://orcid.org/0000-0003-1076-0874
David J. Murrell https://orcid.org/0000-0002-4830-8966

REFERENCES

Arkilanian, A. A., Clements, C. F., Ozgul, A., & Baruah, G. (2020). Effect of time series length and resolution on abundance- and trait-based early warning signals of population declines. *Ecology*, *101*(7), e03040. https://doi.org/10.1002/ecy.3040

Baillie, J. E. M., Collen, B., Amin, R., Akcakaya, H. R., Butchart, S. H. M., Brummitt, N., Meagher, T. R., Ram, M., Hilton-Taylor, C., & Mace, G. M. (2008). Toward monitoring global biodiversity. *Conservation Letters*, *1*(1), 18–26. https://doi.org/10.1111/j.1755-263X.2008.00009.x

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, *164*(3), 1139–1160. https://doi.org/10.1093/genetics/164.3.1139

Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-Qing, D., Clark, N. E., O'Connor, K., & Mace, G. M. (2010). Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology*, *8*(6), e1000385. https://doi.org/10.1371/journal.pbio.1000385

Buschke, F. T., Hagan, J. G., Santini, L., & Coetzee, B. W. T. (2021). Random population fluctuations bias the Living Planet Index. *Nature Ecology and Evolution*, *5*(8), 1145–1152. https://doi.org/10.1038/s41559-021-01494-0

Butchart, S. H. M., Stattersfield, A. J., Baillie, J., Bennun, L. A., Stuart, S. N., Akçakaya, H. R., Hilton-Taylor, C., & Mace, G. M. (2005). Using red list indices to measure progress towards the 2010 target and beyond. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1454), 255–268. https://doi.org/10.1098/rstb.2004.1583

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(30), E6089–E6096. https://doi.org/10.1073/pnas.1704949114

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, *4*(1), 300–307.

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*(2), 317–327. https://doi.org/10.1111/j.1523-1739.2008.01117.x

Collen, B., Ram, M., Zamin, T., & McRae, L. (2008). The tropical biodiversity data gap: Addressing disparity in global monitoring. *Tropical Conservation Science*, *1*(2), 75–88. https://doi.org/10.1177/194008290800100202

Collins, A. C., Böhm, M., & Collen, B. (2020). Choice of baseline affects historical population trends in hunted mammals of North America. *Biological Conservation*, *242*, 108421. https://doi.org/10.1016/j.biocon.2020.108421

Dove, S., Böhm, M., Freeman, R., Jellesmark, S., & Murrell, D. J. (2022). A user-friendly guide to using distance measures to compare time series in ecology. *bioRxiv*. https://doi.org/10.1101/2022.05.11.491333

Dove, S., Böhm, M., Freeman, R., McRae, L., & Murrell, D. J. (2023). Simulated population time series used to build and test a model of accuracy for population-based global biodiversity indicators. *Dryad*. Dataset. https://doi.org/10.5061/dryad.mpg4f4r52

Freeman, R., McRae, L., Deinet, S., Amin, R., & Collen, B. (2021). *rlpi: Tools for calculating indices using the Living Planet Index method*. R package version 0.1.0. https://github.com/Zoological-Society-of-London/living_planet_index

Fryxell, J. M., Sinclair, A. R. E., & Caughley, G. (2014). *Wildlife ecology, conservation, and management* (3rd ed.). Wiley Blackwell.

Gregory, R. D., Skorpilova, J., Vorisek, P., & Butler, S. (2019). An analysis of trends, uncertainty and species selection shows contrasting trends of widespread forest and farmland birds in Europe. *Ecological Indicators*, *103*, 676–687. https://doi.org/10.1016/j.ecolind.2019.04.064

Henriques, S., Böhm, M., Collen, B., Luedtke, J., Hoffmann, M., Hilton-Taylor, C., Cardoso, P., Butchart, S. H. M., & Freeman, R. (2020). Accelerating the monitoring of global biodiversity: Revisiting the sampled approach to generating red list indices. *Conservation Letters*, *13*(3), e12703. https://doi.org/10.1111/conl.12703

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, *46*, 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Inkscape Project. (2020). Inkscape. Retrieved from https://inkscape.org

Israel, G. D. (1992). Determining sample size. In *Fact sheet PEOD-6*. Florida cooperative extension service, Institute of Food and Agricultural Sciences, University of Florida.

IUCN. (2012). *IUCN Red List categories and criteria: Version 3.1* (2nd ed.). IUCN.

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., Fernandez, M., Geller, G. N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F. E., Pereira, H. M., Regan, E. C., Schmeller, D. S., & Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology and Evolution*, *3*(4), 539–551. https://doi.org/10.1038/s41559-019-0826-1

Johansson, Ö., Samelius, G., Wikberg, E., Chapron, G., Mishra, C., & Low, M. (2020). Identification errors in camera-trap studies result in systematic population overestimation. *Scientific Reports*, *10*(1), 6393. https://doi.org/10.1038/s41598-020-63367-z

Josephson, E., Smith, T. D., & Reeves, R. R. (2008). Historical distribution of right whales in the North Pacific. *Fish and Fisheries*, *9*(2), 155–168. https://doi.org/10.1111/j.1467-2979.2008.00275.x

Kiffner, C., Paciência, F. M., Henrich, G., Kaitila, R., Chuma, I. S., Mbaryo, P., Knauf, S., Kioko, J., & Zinner, D. (2022). Road-based line distance surveys overestimate densities of olive baboons. *PLoS One*, *17*(2), e0263314. https://doi.org/10.1371/journal.pone.0263314

Lausch, A., Bannehr, L., Beckmann, M., Boehm, C., Feilhauer, H., Hacker, J. M., Heurich, M., Jung, A., Klenke, R., Neumann, C., Pause, M., Rocchini, D., Schaepman, M. E., Schmidtlein, S., Schulz, K., Selsam, P., Settele, J., Skidmore, A. K., & Cord, A. F. (2016). Linking earth observation and taxonomic, structural and functional biodiversity: Local to ecosystem perspectives. *Ecological Indicators*, *70*, 317–339. https://doi.org/10.1016/j.ecolind.2016.06.022

Leung, B., Hargreaves, A. L., Greenberg, D. A., McGill, B., Dornelas, M., & Freeman, R. (2020). Clustered versus catastrophic global vertebrate declines. *Nature*, *588*(7837), 267–271. https://doi.org/10.1038/s41586-020-2920-6

Liu, X. (2012). Classification accuracy and cut pointselection. *Statistics in Medicine*, *31*(23), 2676–2686. https://doi.org/10.1002/sim.4509

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*(1454), 289–295. https://doi.org/10.1098/rstb.2004.1584

Loreau, M., Cardinale, B. J., Isbell, F., Newbold, T., O'Connor, M. I., & de Mazancourt, C. (2022). Do not downplay biodiversity loss. *Nature*, *601*, E27–E28. https://doi.org/10.1038/s41586-021-04179-7

Lubow, B. C., & Ransom, J. I. (2016). Practical bias correction in aerial surveys of large mammals: Validation of hybrid double-observer with sightability method against known abundance of feral horse (*Equus caballus*) populations. *PLoS One*, *11*(5), e0154902. https://doi.org/10.1371/journal.pone.0154902

Mace, G. M., & Baillie, J. E. M. (2007). The 2010 biodiversity indicators: Challenges for science and policy. *Conservation Biology*, *21*(6), 1406–1413. https://doi.org/10.1111/j.1523-1739.2007.00830.x

Mace, G. M., Barrett, M., Burgess, N. D., Cornell, S. E., Freeman, R., Grooten, M., & Purvis, A. (2018). Aiming higher to bend the curve of biodiversity loss. *Nature Sustainability*, *1*(9), 448–451. https://doi.org/10.1038/s41893-018-0130-0

Manning, J. A., & Goldberg, C. S. (2010). Estimating population size using capture–recapture encounter histories created from point-coordinate locations of animals. *Methods in Ecology and Evolution*, *1*(4), 389–397. https://doi.org/10.1111/j.2041-210X.2010.00041.x

McRae, L., Deinet, S., & Freeman, R. (2016). Data from: The diversity-weighted Living Planet Index: Controlling for taxonomic bias in a global biodiversity indicator. *Dryad*, Dataset.

McRae, L., Deinet, S., & Freeman, R. (2017). The diversity-weighted Living Planet Index: Controlling for taxonomic bias in a global biodiversity indicator. *PLoS One*, *12*(1), 1–20. https://doi.org/10.1371/journal.pone.0169156

Meyer, C., Kreft, H., Guralnick, R., & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*, *6*(8221). https://doi.org/10.1038/ncomms9221

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., & Worm, B. (2011). How many species are there on earth and in the ocean? *PLoS Biology*, *9*(8), e1001127. https://doi.org/10.1371/journal.pbio.1001127

Murali, G., de Oliveira Caetano, G. H., Barki, G., Meiri, S., & Roll, U. (2022). Emphasizing declining populations in the living planet report. *Nature*, *601*, E20–E22. https://doi.org/10.1038/s41586-021-04165-z

Nichols, J. D., O'Connell, A. F., & Karanth, K. U. (2011). Camera traps in animal ecology and conservation: What's next? In *Camera traps in animal ecology: Methods and analyses* (pp. 253–263). Springer. https://doi.org/10.1007/978-4-431-99495-4_14

Nori, J., Loyola, R., & Villalobos, F. (2020). Priority areas for conservation of and research focused on terrestrial vertebrates. *Conservation Biology*, *34*(5), 1281–1291. https://doi.org/10.1111/cobi.13476

Oliver, R. Y., Meyer, C., Ranipeta, A., Winner, K., & Jetz, W. (2021). Global and national trends, gaps, and opportunities in documenting and monitoring species distributions. *PLoS Biology*, *19*(8), e3001336. https://doi.org/10.1371/journal.pbio.3001336

Proença, V., Martin, L. J., Pereira, H. M., Fernandez, M., McRae, L., Belnap, J., Böhm, M., Brummitt, N., García-Moreno, J., Gregory, R. D., Honrado, J. P., Jürgens, N., Opige, M., Schmeller, D. S., Tiago, P., & van Swaay, C. A. M. (2017). Global biodiversity monitoring: From data sources to essential biodiversity variables. *Biological Conservation*, *213*, 256–263. https://doi.org/10.1016/j.biocon.2016.07.014

Puurtinen, M., Elo, M., & Kotiaho, J. S. (2022). The Living Planet Index does not measure abundance. *Nature*, *601*, E14–E15. https://doi.org/10.1038/s41586-021-03708-8

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rose, R. A., Byler, D., Eastman, J. R., Fleishman, E., Geller, G., Goetz, S., Guild, L., Hamilton, H., Hansen, M., Headley, R., Hewson, J., Horning, N., Kaplin, B. A., Laporte, N., Leidner, A., Leimgruber, P., Morisette, J., Musinsky, J., Pintea, L., … Wilson, C. (2015). Ten ways remote sensing can contribute to conservation. *Conservation Biology*, *29*(2), 350–359. https://doi.org/10.1111/cobi.12397

RStudio Team. (2022). *RStudio: Integrated development environment for R*. RStudio, PBC. http://www.rstudio.com

Saha, A., McRae, L., Dodd, C. K., Gadsden, H., Hare, K. M., Lukoschek, V., & Böhm, M. (2018). Tracking global population trends: Population time-series data and a Living Planet Index for reptiles. *Journal of Herpetology*, *52*(3), 259–268. https://doi.org/10.1670/17-076

Scheele, B. C., Legge, S., Blanchard, W., Garnett, S., Geyle, H., Gillespie, G., Harrison, P., Lindenmayer, D., Lintermans, M., Robinson, N., & Woinarski, J. (2019). Continental-scale assessment reveals inadequate monitoring for threatened vertebrates in a megadiverse country. *Biological Conservation*, *235*, 273–278. https://doi.org/10.1016/j.biocon.2019.04.023

Secretariat of the Convention on Biological Diversity. (2006). Global biodiversity outlook 2. https://www.cbd.int/gbo2/

Turak, E., Harrison, I., Dudgeon, D., Abell, R., Bush, A., Darwall, W., Finlayson, C. M., Ferrier, S., Freyhof, J., Hermoso, V., Juffe-Bignoli, D., Linke, S., Nel, J., Patricio, H. C., Pittock, J., Raghavan, R., Revenga, C., Simaika, J. P., & de Wever, A. (2017). Essential biodiversity variables for measuring change in global freshwater biodiversity. *Biological Conservation*, *213*, 272–279. https://doi.org/10.1016/j.biocon.2016.09.005

Valdez, J. W., Callaghan, C. T., Junker, J., Purvis, A., Hill, S. L. L., & Pereira, H. M. (2023). The undetectability of global biodiversity trends using local species richness. *Ecography*, *2023*(3), e06604. https://doi.org/10.1111/ecog.06604

Visconti, P., Bakkenes, M., Baisero, D., Brooks, T., Butchart, S. H., Joppa, L., Alkemade, R., Di Marco, M., Santini, L., Hoffmann, M., Maiorano, L., Pressey, R. L., Arponen, A., Boitani, L., Reside, A. E., van Vuuren, D. P., & Rondinini, C. (2016). Projecting global biodiversity indicators under future development scenarios. *Conservation Letters*, *9*(1), 5–13. https://doi.org/10.1111/conl.12159

Wauchope, H. S., Amano, T., Sutherland, W. J., & Johnston, A. (2019). When can we trust population trends? A method for quantifying the effects of sampling interval and duration. *Methods in Ecology and Evolution*, *10*(12), 2067–2078. https://doi.org/10.1111/2041-210X.13302

Westcott, D. A., Fletcher, C. S., Mckeown, A., & Murphy, H. T. (2012). Assessment of monitoring power for highly mobile vertebrates. *Ecological Applications*, *22*(1), 374–383.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., Gray, W. A., White, R. J., Jones, A. C., Bisby, F. A., & Culham, A. (2007). How global is the global biodiversity information facility? *PLoS One*, *2*(11), e1124. https://doi.org/10.1371/journal.pone.0001124

Zylstra, E. R., Steidl, R. J., & Swann, D. E. (2010). Evaluating survey methods for monitoring a rare vertebrate, the Sonoran Desert tortoise. *Journal of Wildlife Management*, *74*(6), 1311–1318. https://doi.org/10.2193/2009-331

SUPPORTING INFORMATION

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

How to cite this article: Dove, S., Böhm, M., Freeman, R., McRae, L., & Murrell, D. J. (2023). Quantifying reliability and data deficiency in global vertebrate population trends using the Living Planet Index. *Global Change Biology*, *00*, 1–17. https://doi.org/10.1111/gcb.16841