Quaternary Science Reviews 112 (2015) 128–137

Contents lists available at ScienceDirect

Quaternary Science Reviews

journal homepage: www.elsevier.com/locate/quascirev

Uncertainties in dating constrain model choice for inferring extinction time from fossil records

Frédérik Saltré [a, *], Barry W. Brook [a, 1], Marta Rodríguez-Rey [a], Alan Cooper [a, b], Christopher N. Johnson [c], Chris S.M. Turney [d], Corey J.A. Bradshaw [a]

[a] The Environment Institute, School of Biological Sciences, The University of Adelaide, Adelaide, South Australia 5005, Australia
[b] Australian Centre for Ancient DNA, School of Biological Sciences, The University of Adelaide, Adelaide, South Australia 5005, Australia
[c] School of Zoology, Private Bag 5, University of Tasmania, Hobart, Tasmania 7001, Australia
[d] School of Biological, Earth and Environmental Sciences, University of New South Wales, Australia

ARTICLE INFO

Article history:
Received 2 August 2014
Received in revised form
23 January 2015
Accepted 24 January 2015
Available online

Keywords:
Model selection key
Extinction
Time series
Sensitivity analysis
Dating

ABSTRACT

Accurate estimates of the timing of extinctions (θ) are critical for understanding the causes of major die-off events and for identifying evolutionary or environmental transitions. Yet many studies have demonstrated that sampling biases and underlying statistical assumptions affect the accuracy of model-based estimates of extinction times ($\hat{\theta}$), and the added uncertainty contributed by inherent (laboratory) dating errors has largely been neglected. Here we provide a general guide (model-selection key) for choosing from among eight alternative 'frequentist sampling' (i.e., non-Bayesian) methods, differentiated by their treatment of both the probability of record occurrence and uncertainties in record dates, the most appropriate for a given record. We first provide a methodological framework to characterize time series of dated records as a function of the number of records, the size of the interval between successive records, and laboratory dating errors. Using both simulated data and dated Australian megafauna remains, we then assess how the characteristic of a dataset's time series dictates model performance and the probability of misclassification (false extant *vs*. false extinct). Among the four classic frequentist methods providing highest model performance, Marshall's (1997) and McCarthy's (1998) methods have the highest model precision. However, high model performance did not prevent misclassification errors, such that the Gaussian-resampled inverse-weighted McInerny (GRIWM) approach is the only method providing both high model accuracy and no misclassification issues, because of its unique down-weighting interval procedure and its ability to account for uncertainties in record dates. Applying the guideline to three time series of extinct Australian species, we recommend using Marshall's, McCarthy's and/or GRIWM methods to infer θ of both *Thylacinus* sp. and *Genyornis* sp., because each dataset is characterized by many sightings and a low variance of the interval between records, whereas McInerny's method better suits *Diprotodon* sp. due to an even lower interval variance.

1. Introduction

Mass extinction events, characterized by palaeontologists as high, planetary-wide species loss within a short geological time frame (e.g., over 75% of species within less than two million years, Barnosky et al., 2011), completely changed the global pattern of species distribution by both removing lineages and triggering evolutionary opportunities (Jablonski, 2001). However, the causes and mechanisms of mass extinctions, such as the end-Permian mass extinction (Grice et al., 2005; Payne and Clapham, 2012; Sun et al., 2012; Wang et al., 2014) or the late Quaternary megafauna extinction, are still debated by scientists from disciplines spanning palaeontology to archaeology and ecology (Alroy, 2001; Brook and Bowman, 2002; Barnosky et al., 2004; Lorenzen et al., 2011), in large part because of inaccuracy of inference of the timing of a species' extinctions (θ) (Flannery, 2002). Robust and accurate inferences are essential to test, for example, the evidence that the end-Permian transition was abrupt versus having multiple extinction phases (Jin et al., 2000; Song et al., 2013; Wang et al.,

* Corresponding author.
E-mail address: frederik.saltre@adelaide.edu.au (F. Saltré).
[1] Present address: School of Biological Sciences, Faculty of Science, Engineering & Technology, University of Tasmania, Hobart, Tasmania 7001, Australia.

http://dx.doi.org/10.1016/j.quascirev.2015.01.022
0277-3791/

Table 1
Description of the eight methods tested and categorized into five categories as a function of the kind of assumptions they make about sampling intensity over time (p-sampling assumption) and summary dataset characteristics (n, $\bar{i}$, $\sigma^2 i$, $\bar{\varepsilon}$, $\sigma^2\varepsilon$; see Table 2 for complete description). For each method, we indicated model constraints (high performance constraints) leading to its best performance from the sensitivity analysis (see Fig. 3 and Fig. A.6). For example, $\uparrow x$ means that a high value of the 'x' parameter leads to high model performance, considering that the number of arrows indicates the relative constraint intensity (i.e., $\uparrow\uparrow > \uparrow$ and $\downarrow\downarrow > \downarrow$).

Method	p-Sampling assumption	n	$\bar{i}$	$\sigma^2 i$	$\bar{\varepsilon}$	$\sigma^2\varepsilon$	High performance constraints
Strauss and Sadler (1989)							$\uparrow\uparrow n$
McInerny et al. (2006)		x	x	x	–	–	$\uparrow n$, $\downarrow\downarrow\sigma^2 i$
BRIWM	Poisson stationary process						$\uparrow n$, $\downarrow\bar{i}$
Solow et al. (2006)		x	x	x	x	–	$\downarrow\downarrow\bar{\varepsilon}$, $\downarrow\sigma^2 i$
McCarthy (1998) Marshall (1997)	Recovery potential	x	x	x	–	–	$\uparrow n$, $\downarrow\sigma^2 i$
Roberts and Solow (2003)	No assumptions	x	x	x	–	–	$\uparrow\uparrow n$, $\downarrow\bar{i}$
GRIWM (Bradshaw et al., 2012a)		x	x	x	x	x	$\uparrow n$, $\downarrow\bar{i}$, $\downarrow\bar{\varepsilon}$

2014), or that megafauna extinction was primarily climate- or human-driven in South America (Johnson et al., 2013; Lima-Ribeiro and Felizola Diniz-Filho, 2013) and Australia (Brook and Bowman, 2002; Wroe et al., 2013).

The megafauna extinction stalemate in particular persists primarily because the estimated timing of these species' extinctions ($\widehat{\theta}$) is uncertain due to the variable quality of the dated precursor fossil specimens, meaning that debates digress to matters of opinion rather than accurately measured phenomena and scientific hypothesis testing (Brook et al., 2013). Although quality fossil data are essential to improve our inferences of past extinctions, palaeoecological archives are inherently incomplete and geochronological dating methods are characterized by errors of centuries to millennia, so the reliability of θ inference based only on their scant information remains a major challenge. The absence of a species in a particular site or temporal window does not necessarily mean it was not present, so apparent declines of taxa in these records might simply reflect sampling artefacts rather than real trends in diversity (Prideaux et al., 2007). Such absences might also arise for taphonomic reasons (i.e., type of facies and sedimentary environments that can prevent the preservation of remains), life-history traits (e.g., taxa from lower trophic levels, because they are more abundant, have a relatively higher potential for fossilization) and ecological specialization (i.e., specialists living in a specific habitat will have their remains fossilised only there, whereas generalists will have an overall higher probability of being recorded). Evidence from extinctions observed in modern times suggests that as a doomed species approaches its final extinction date, population size tends to decrease exponentially due to the synergistic feedbacks (Brook et al., 2008) that lead to the extinction vortex (Fagan and Holmes, 2006), which reduces the probability of discovering fossil records near the terminal date and artificially truncates the true temporal range of a species' persistence window (Signor-Lipps effect; Signor and Lipps, 1982). Moreover, fossil records – retrieved from specific sites where the rare phenomenon of preservation was possible – only describe local losses of species such that the last date known cannot necessarily testify to a global extinction. Indeed, in some cases apparent disappearances can be followed by the subsequent reappearance of the species after further sampling (the 'Lazarus' effect; Keith and Burgman, 2004).

As population size tends to decline to incrementally lower densities prior to extinction (Fagan and Holmes, 2006), it is logical to assume that the last dated record of a species occurs sometime before its true extinction (i.e., the death of the last individual). Based on this assumption, many probabilistic methods (also called "classical frequentist methods", Alroy, 2014) have been developed to provide a confidence interval around $\widehat{\theta}$ given a particular time series of occurrence records, but uncertainties in dating techniques (e.g., inherent laboratory errors in radiometric dating), and the probability of sampling reliably dated specimens (i.e., sampling rate and location) make inference complex. For example, Roberts and Solow (2003) applied an optimal linear estimation method based on a record of historical sightings of the dodo (*Raphus cucullatus*) to determine the confidence interval surrounding its true extinction year. That method was extended to account explicitly for error in estimates of the record date for fossils (Solow et al., 2006), but comparisons within and among species were still difficult due to variation in sampling rates that can affect model performance (Rivadeneira et al., 2009). McInerny et al. (2006) proposed another frequentist-probabilistic method that incorporates sampling rate, which was further modified by Bradshaw et al. (2012a) to take into account the number and uncertainty of dates in the time series.

Each method is characterized by a set of statistical assumptions conditioning its adequate application to a given time series (e.g., sampling probability uniformly distributed and independent, or dating error being constant; Table 1 and Solow et al., 2006), which if violated, can lead to the misclassification of a species as extinct or extant (so-called Type I and II statistical inference errors, respectively; Brosi and Biber, 2008; Jarić and Ebenhard, 2010; Fisher and Blomberg, 2012). In addition to methodological issues, the quality (number of records, record interval, variation in dating error over time) and the reliability of the datasets used to infer θ (e.g., species misidentification – Rasmussen and Prys-Jones, 2003; an erroneous ceiling on apparent dates due the time limit of radiocarbon [^{14}C] dating validity – Walker, 2005) also strongly affect model performance (Rivadeneira et al., 2009; Solow et al., 2011; Bradshaw et al., 2012a; Lee et al., 2014). Various classical frequentist methods have been tested and validated as a function of both the number of records and sampling intensity (Rivadeneira et al., 2009; Fisher and Blomberg, 2012), highlighting performance problems specifically when sampling probabilities decrease through time (Rivandeneira et al., 2009). Newly emerging Bayesian methods can, if used appropriately, reduce such performance issues and improve species classification (endangered or about to go extinct; Alroy, 2014), but the effect of inherent dating error and their variation over time on model performance have barely been assessed (Bradshaw et al., 2012a). As dating errors typically increase as sampling reaches deeper back in time (such as in palaeontological time series; Walker, 2005), providing rigorous measures of the biases generated by dating errors on $\widehat{\theta}$ is therefore essential.

Here we explore how the characteristics of time series of dated records, such as the number of occurrences, time gaps between records, and uncertainties in measured dates, act and interact to constrain different frequentist models used commonly to infer θ. More specifically, we provide both quantitative and qualitative criteria for: (*i*) maximizing the inferential capability of eight classical methods used to generate confidence intervals for θ; and (*ii*) provide a general guideline for selecting the most appropriate method to infer θ from a given time series of dated records. We first describe these eight frequentist methods focussing on their conceptual assumptions with respect to five summary variables characterizing the types of time series usually available (henceforth,

Table 2
Input and output variables involved in the sensitivity analysis.

Variable name	Short description	Range
n	Number of records	[3–100]
D	Record date (or age)	[−1000 to −15,000] in years
ε	Dating error as a function of the record's date/age	$\boldsymbol{\varepsilon} = 0.1203^{*}\mathbf{D} - 1.298$. Equation fitted using data from the Sahul fossil database (unpublished). $\boldsymbol{\varepsilon}$ is modified as being selected from within a window of the initial $\boldsymbol{\varepsilon} \pm 30$ years to include dating error variability
i	Interval between two records	
θ_t	Theoretical (set) final extinction date	−1000
$\bar{\boldsymbol{i}}$	Average i over time series expressed as a percentage of the entire observation period.	[40 to 6000], depends on both $\boldsymbol{i}$ and $\boldsymbol{n}$; expressed in "years"
$\sigma^2 i$	Variance of i over time series expressed as a percentage of the whole observation period	depends on $\boldsymbol{i}$ [3000 to 20×10^7]; expressed in "years"
$\bar{\varepsilon}$	Average $\boldsymbol{\varepsilon}$ over time series expressed as a percentage of the whole observation period.	[160 to 7300], depends on both $\boldsymbol{\varepsilon}$ and $\boldsymbol{n}$; expressed in "years"
$\sigma^2\varepsilon$	Variance of $\boldsymbol{\varepsilon}$ over time series expressed as a percentage of the whole observation period	depends on $\boldsymbol{\varepsilon}$ [400 to 15×10^7]; expressed in "years"

'time series' characteristics'; Table 2). As six of these methods have already been reviewed extensively (Rivadeneira et al., 2009; Bradshaw et al., 2012a; Alroy, 2014), we mainly describe the recently developed Gaussian-resampled inverse-weighted McInerny approach (GRIWM, Bradshaw et al., 2012a), and we introduce a new variant of GRIWM, called BRIWM (see description below). Second, based on both simulated times-series data and sensitivity analyses, we develop an index of model performance accounting for: (*i*) the probability that θ falls within the model's estimated confidence interval (i.e., the model's coverage probability); (*ii*) the bias in model estimates; and (*iii*) the width of model's estimated confidence interval to identify causes of variation in method performance and to highlight the range of values of time series' characteristics for each model that lead to its best performance. Third, we apply each of the eight models to extant and extinct, quality-controlled (i.e., dating quality checked) Australian mammal time series, to assess each model's ability to minimize Type I and II errors. Based on these results, we used a real-world demonstration by creating and testing a model-selection key to select the most appropriate model for inferring extinction from time series' characteristics of three extinct, Australian late Quaternary 'megafauna' species (*Thylacinus* sp., *Genyornis* sp., and *Diprotodon* sp.).

2. Materials and methods

2.1. Model descriptions and time series' characteristics

Many studies have described and contrasted different frequentist inference methods as a function of the data-sampling regime (Rivadeneira et al., 2009; Bradshaw et al., 2012a). Such time-series characterizations are typically designed to be easy to implement and convenient for theoretical analyses, by assuming that dated records follow well-established mathematical distributions (e.g., uniform or exponential; Bradshaw et al., 2012a), or mimicking theoretical sampling intensity (Rivadeneira et al., 2009). However, the 'true' distribution of sampling is rarely known, so we first propose a statistical framework that can be easily applied to characterize any given time series (Table 2). Here we propose to characterize time series of dated records as a function of five variables we refer to as 'times series' characteristics': (1) number of records (n); (2) average and (3) variance of the interval between successive records ($\bar{i}$ and $\sigma^2 i$, respectively); and (4) average and (5) variance of dating error ($\bar{\varepsilon}$ and $\sigma^2\varepsilon$, respectively) covering the time-series range of the dated specimens.

We then compared eight different methods to infer θ: (1) Strauss and Sadler's (1989), (2) Robert and Solow's (2003), (3) McCarthy's (1998), (4) Marshall's (1997), (5) Solow's (2006), (6) McInerny's (2006) methods, (7) GRIWM (Bradshaw et al., 2012a) and (8) the new bootstrap-resampled inverse-weighted McInerny (BRIWM) method. BRIWM has specifically been developed for this study in an attempt to account for data reliability. All methods except Solow's and GRIWM first assume that the species to which they are applied are actually extinct, and that θ lies somewhere between the last record and the present. Because they account for dating error, Solow's and GRIWM can allow extinction preceding the last record if the error on that estimate is high. Each method describes θ in terms of time since the last record and, depending on the total temporal span of all records, also estimates a desired level of confidence (usually expressed as $\alpha = 0.05$) for $\widehat{\theta}$. Such approaches are criticized because they assume a constant or declining sampling rate, but alternative Bayesian approaches either require *a priori* information about population dynamics weakly supported from records (Caley and Barry, 2014) or provide output not directly comparable to classical frequentist methods (i.e., return a probability that a species is extinct because it was not sampled, instead of the probability that the species was unsampled because it was extinct; Alroy, 2014). We therefore do not consider Bayesian approaches in this paper. Finally, classical frequentist approaches all take account of the total number of records, as well as the probability of presence of a species decreasing over time after the last record. The methods can be further categorized into those assuming a (*i*) uniform probability of record occurrence over time, and those (*ii*) accounting for uncertainties in record dates. None accounts (explicitly) for uneven sampling in space and related potential biases due to site selection.

Strauss and Sadler's, Solow's and McInerny's methods assume a uniform probability of record occurrence, but other methods relax this assumption either by integrating some temporal variation [so-called 'recovery function' – see Marshall (1997) and McCarthy (1998) – calculated here as a function of a probability of sampling fitted to each given time series following the Rivadeneira et al. (2009) approach] or making no distributional assumptions about the probability of sampling (Roberts and Solow, 2003; GRIWM: Bradshaw et al., 2012a, and BRIWM], although independence among records is still required. GRIWM and Solow's models are the only ones we tested here that take into account the uncertainties in record dates. While other methods only depend on $n, \bar{i}, \sigma^2 i$ (Table 1), both GRIWM and Solow's include $\bar{\varepsilon}$, but as Solow's assumes constant dating uncertainties across samples, GRIWM assumes variation in these uncertainties (considering $\sigma^2\varepsilon$) by 10,000 (or more) resamples of the standard deviation of each date from a Gaussian distribution (Bradshaw et al., 2012a).

BRIWM is a new variant of GRIWM we developed for this analysis to assess the importance of record reliability (i.e., the

confidence in the method used for dating) against the impact of dating uncertainties (i.e., the standard error of the estimated record's date). Like GRIWM, the model hypothesizes that the most-recent records (i.e., those closest to the last appearance date) are more useful in inferring $\widehat{\theta}$ as extinction than older dates, by down-weighting the contribution of each dated record to $\widehat{\theta}$ depending on its temporal distance from the most recent record. However, instead of accounting for all records and uncertainty in dates, by subsampling a large number of iterations (10,000 used here), BRIWM creates a new time series of records by resampling the original dataset (instead of resampling each date of the series into the standard deviation, as does GRIWM) using a bootstrap technique with replacement, and calculates $\widehat{\theta}$ for each iteration. This technique will modify both n (i.e., subsampling almost always decreases the number of unique records) and $\bar{i}$ (i.e., removing records from the original time series either enlarges or reduces the size of the intervals). From these 10,000 estimates of $\widehat{\theta}$, we can calculate a 95% confidence interval using percentiles.

2.2. Simulated time series for sensitivity analysis and model performance assessment

The aim of our sensitivity analysis (see below) is to assess how variation in time series' characteristics (Tables 1 and 2) affects model performance. Our analysis sets a theoretical θ of 1000 years before present ($\theta_t = -1000$), and we generated simulated time series stochastically by selecting single values for each time series' characteristic from within a specified range (Table 2) following a Latin hypercube sampling approach (Fig. A.2) to achieve a robust and efficient coverage of the parameter space (Saltelli et al., 2008). This approach also ensures that each variable is represented in a fully stratified design, without any prior knowledge of which variables will be most influential on the output. For each input combination, our models estimate the time of extinction (i.e., median estimates were obtained using $\alpha = 0.5$; Lima-Ribeiro and Diniz-Filho, 2014) and its 95% confidence interval.

2.3. Model performance index

Model performance is usually evaluated according to the coverage probability of θ_t by a model's estimated confidence interval (Rivadeneira et al., 2009), such that θ_t occurs in the interval defined by the last record and the upper bound of the 95% confidence interval. Such a metric favours methods that produce the widest confidence intervals, so we developed a specific index of model performance (ϕ) that we applied to each of the eight methods under the various simulated time series generated for the sensitivity analysis. We assumed that model performance depends on: (*i*) the coverage probability (o) of θ_t for each simulated time series; (*ii*) the distance ($\Delta_{\widehat{(\theta-\theta t)}}$) between the closest model's confidence bound (e.g., lower, upper boundary or the median value) that informs model accuracy; and (*iii*) the width of the model's confidence interval that informs model precision.

We defined the best-performing model (highest ϕ) as the one providing a high coverage rate, and having both a low $\Delta_{\widehat{(\theta-\theta t)}}$ and a narrow confidence interval under various scenarios:

$$\phi = [(o \times w_1) + (\beta \times w_2) + (\gamma \times w_3)]/\phi_{max}$$

where $o = sc_o/sc_{tot}$ with sc_o being the number of scenarios where the model's confidence interval successfully covered θ_t and sc_{tot} = the total number of time-series scenarios generated for the sensitivity analysis; β is a normalized measure of dispersion of $\Delta_{\widehat{(\theta-\theta t)}}$ in relation to θ_t such that: $\beta = \theta_t/(\theta_t + q_{\Delta_{\widehat{(\theta-\theta t)}}})$, with $q_{\Delta_{\widehat{(\theta-\theta t)}}}$ being the percentile at 0.975 of all $\Delta_{\widehat{(\theta-\theta t)}}$ calculated over all scenarios; γ is a normalized measure of dispersion of the width of the model's confidence interval in relation to a benchmark width (=1 to have both the lowest width of confidence interval as benchmark and to avoid technical issues due to a null numerator when the ratio is calculated) such that: $\gamma = 1/q_{CI}$, with q_{CI} being the percentile at 0.975 of all confidence intervals calculated over all scenarios. Thus, high values of ϕ indicate better method performance. Note that ϕ can be applied to a single scenario with $o = 0$ (no coverage) or 1 (coverage), $q_{\Delta_{\widehat{(\theta-\theta t)}}} = \Delta_{\widehat{(\theta-\theta t)}}$ and q_{CI} = confidence interval. As ϕ is unequally sensitive to variation in o, β and γ (see the detailed sensitivity analysis of ϕ; Fig. A.3), we integrated w_1, w_2 and w_3 as weighting coefficients to make ϕ equally sensitive to o, β and γ variation. Then, $\phi_{max} = 5.525$ (i.e., the maximum value of ϕ when $o = 1$, $\beta = 1$, and $\gamma = 1$) rescales ϕ between [0, 1].

Finally, we calculated the range of values for each time series' characteristic that maximizes ϕ for each model. For each time series' characteristic, the range of values is quantified using a coefficient of variation (CV) calculated as the variance of $\bar{i}$, $\sigma^2 i$, $\bar{\varepsilon}$, $\sigma^2 \varepsilon$ divided by the median date of the length of the entire time series (i.e., 7500 years for a time-series of length = 15,000 years in those generated for the sensitivity analysis). Because n does not refer to a time period, its CV is calculated following its variance divided by the median of the maximum number of records tested (i.e., 50 for a maximum number of 100 records used in the sensitivity analysis).

2.4. Sensitivity analysis

We evaluated the relative effect of each time series' characteristic for each simulated time series (n, $\bar{i}$, $\sigma^2 i$, $\bar{\varepsilon}$, $\sigma^2 \varepsilon$; see Table 2 for complete description) on each model's performance based on their relative effects on (*i*) the model's coverage probability, (*ii*) $\Delta_{\widehat{(\theta-\theta t)}}$ and (*iii*) the width of the confidence interval. We first constructed a series of generalized linear models (GLM) where the model's coverage probability (i.e., a binary response indicating whether or not θ_t fell within the model's estimated confidence interval), $\Delta_{\widehat{(\theta-\theta t)}}$ and the width of model's estimated confidence interval were the responses. The fixed effects represent the simulated times series' characteristics (used as explanatory variables in the GLM) as well as interactions between n and the four other variables to indicate how their combined effects modify performance. We then compared all GLM using the Bayesian information criterion (BIC) to identify the most influential predictors and to down-weight any tapering effects (Link and Barker, 2006). Finally, we calculated the standardized coefficients (ξ, described as α_n/SE_n in Bradshaw et al. 2012b) for each term of each GLM to indicate the relative influence of each of the five time series' characteristics on the model responses, which corrects for different scales of the predictors.

2.5. Australian datasets

For the real-world case study, we applied each of the eight models to infer θ and its confidence interval on six extant mammal species in Australia for which time series of dated fossil specimens exist (i.e., there are no true 'extinctions' for any of them): *Dasyurus maculatus*, *Lagostrophus fasciatus*, *Macropus rufogriseus*, *Perameles gunnii*, *Petrogale brachyotis*, and *Tachyglossus aculeatus* (Table B.1), as well as for three mammal species that went extinct in mainland Australia during the late Pleistocene or Holocene (*Thylacinus* sp., *Genyornis* sp., and *Diprotodon* sp.). We first calculated time series' characteristics for each of the nine mammal species (Table B.1) and we assessed a model's ability to deal successfully with statistical inference errors of Type I and II by comparing $\widehat{\theta}$ (in years) with the present date (0 before present, or 'BP'), where $\widehat{\theta}$ should be lower

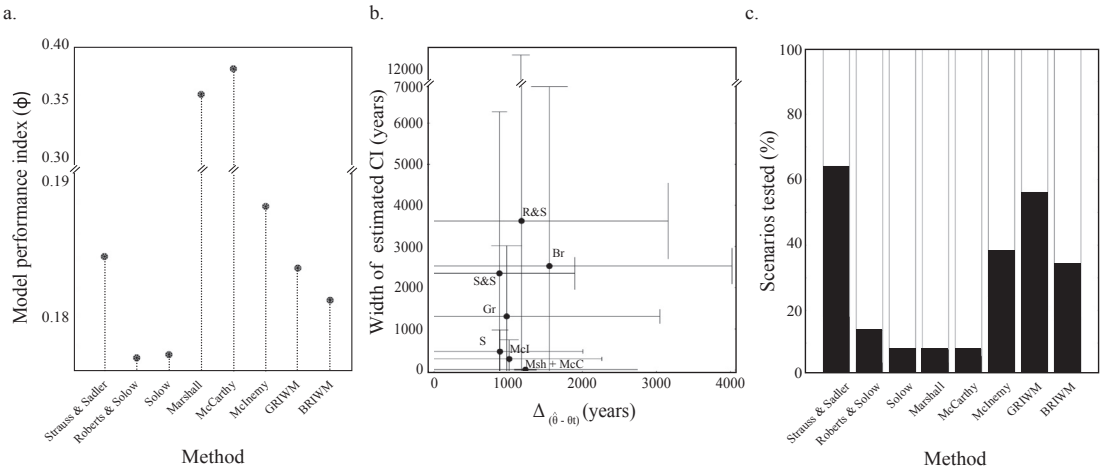

Fig. 1. Performance of eight models used to infer species extinction time (Strauss and Sadler's [S&S], Roberts and Solow's [R&S], McCarthy's [McC], Marshall's [Msh], McInerny's [McI], Solow's [S], GRIWM [Gr], and BRIWM [Br]), applied to various times series generated using a Latin hypercube approach from a theoretical true extinction date ($\theta_t = 1000$). (a) Model performance index (ϕ) is calculated for each method following the equations in Section 2.3, so that the higher ϕ, the better the performance. The equation assumes that ϕ depends on (b) the model's estimation bias ($\Delta_{\widehat{(\theta-\theta t)}}$ = the difference between the closest model's confidence bound to $\widehat{\theta}$) as a function of the width of the model's estimated confidence interval, (c) the method's coverage probability (i.e., proportion of times θ_t falls within the model's estimated confidence interval). Both $\Delta_{\widehat{(\theta-\theta t)}}$ and the confidence interval are expressed as averages (standard deviation associated) over all times series tested. Both *x*- and *y*-axes are expressed on a logarithmic scale.

than 0 BP for extant species, or higher than 0 BP for extinct species (i.e., Type I and II errors, respectively). We calibrated the 170 radiocarbon dates using the Southern Hemisphere Calibration curve (ShCal13, Hogg et al., 2013) to provide calendar-age estimates from the OxCal radiocarbon calibration tool Version 4.1 (Ramsey, 2010).

3. Results

According to the three criteria used to build the model performance index (i.e., high coverage probability, with both a low $\Delta_{\widehat{(\theta-\theta t)}}$ and a narrow confidence interval), Marshall's and McCarthy's methods had the best performance ($\phi = 0.353$ and 0.376, respectively; Fig. 1a) followed by McInerny's ($\phi = 0.188$), Strauss & Sadler's and GRIWM methods ($\phi = 0.184$ and 0.183, respectively). None of the eight methods tested had maximum performance ($\phi = 1$) because they did not fulfil all of the three criteria required. Methods generated an estimated error ($\Delta_{\widehat{(\theta-\theta t)}}$) spanning 882 ± 1017 years (Strauss & Sadler's model, Fig. 1b) to 1557 ± 2463 years (BRIWM), and they provided confidence interval widths ranging from few years (7.84 ± 8.8 years and 7.50 ± 8.7 years for Marshall's and McCarthy's methods, respectively; Fig. 1b) to several millennia (3622 ± 9226 years for Roberts & Solow's method). Strauss & Sadler's and GRIWM provided the best coverage probability of θ_t (successfully covered 64 and 56% of the simulated time series, respectively; Fig. 1c) whereas Roberts & Solow's, Marshall's and McCarthy's methods had poorer coverage (successfully covered <15% of the simulated post-last-appearance-date time series).

High ϕ did not imply, however, that models guarded adequately against both Type I and II inference errors for the case studies (Fig. 2). Strauss & Sadler's, Roberts & Solow's and GRIWM were the only models that were able to predict accurately that the extant species were still alive, whereas the others wrongly predicted a premature extinction (Type 1 error; upper edge of boxplots <0; Fig. 2a, Table B.2). However, Strauss and Sadler's and Roberts & Solow's methods predicted the extinct *Thylacinus* sp. and *Diprotodon* sp. as extant (Type II error; upper edge of boxplots >0; Fig. 2b, Table B.2), whereas the other models predicted these species as extinct. GRIWM was the only model that avoided both Type I and II errors for these real-world datasets (Fig. 2).

Time series' characteristics (n, $\bar{i}$, $\sigma^2 i$, $\bar{\varepsilon}$ and $\sigma^2\varepsilon$; Table 2) affected ϕ for each method in different ways for coverage probability, the size of $\Delta_{\widehat{(\theta-\theta t)}}$ and the width of the estimated confidence interval. Each time series' characteristic had positive or negative effects on metrics, meaning that increasing the value of a characteristic either increased (positive effect: $\xi > 0$; Fig 3) or decreased (negative effect: $\xi < 0$) model outputs. Here we focused on Marshall's, McCarthy's, McInerny's and GRIWM methods (the four other methods are discussed in Appendix A, Fig. A.6). The average interval between records ($\bar{i}$) and the average dating error ($\bar{\varepsilon}$) positively affected both GRIWM $\Delta_{\widehat{(\theta-\theta t)}}$ ($\xi = +55$ and $+32$, respectively) and confidence interval ($\xi = +29$ and $+99$, respectively), meaning that high $\bar{i}$ increased the models' $\Delta_{\widehat{(\theta-\theta t)}}$. The number of records (n) negatively affected McInerny's, confidence interval ($\xi = -57$), meaning that lower n led to wider confidence intervals in the same way as the variance between records ($\sigma^2 i$) affected GRIWM's $\Delta_{\widehat{(\theta-\theta t)}}$ and confidence interval ($\xi = -95$ and -107, respectively). Some of the time series' characteristics had combined effects on model outputs. For example, $n \times \bar{i}$ negatively affected GRIWM's $\Delta_{\widehat{(\theta-\theta t)}}$ and confidence interval ($\xi = -83$ and -55, respectively; Fig. 3). Due to the negative relationship between n and $\bar{i}$ (i.e., $\bar{i}$ decreased as n increased, Fig. A.6), this combined effect means that increasing n reduced $\bar{i}$ and led to a decrease in both $\Delta_{\widehat{(\theta-\theta t)}}$ and confidence interval width. A similar relationship exists between n and $\sigma^2 i$ (Fig. A.5), such as $n \times \sigma^2 i$, negatively affecting McInerny's ($\xi = -48$), Marshall's and McCarthy's $\Delta_{\widehat{(\theta-\theta t)}}$ ($\xi = -51$ for the both methods), meaning that increasing n decreased $\sigma^2 i$ and led to a lower $\Delta_{\widehat{(\theta-\theta t)}}$.

We used coefficient of variation measures (CV, Table 3) to determine the optimal ranges of each time series' characteristic for which each model provided its highest ϕ (see detailed method in SI. 2). Roberts & Solow's and Solow's best performances occurred under high n (high ϕ: CV > 0.63 *vs.* low ϕ: CV = 0.5; Table 3), whereas Marshall's method performed better under lower n (high ϕ: CV = 0.55 *vs.* lower ϕ: CV = 0.59). A low variability in dating error ($\bar{\varepsilon}$) improved both Solow and GRIWM (high ϕ: CV = 9.04 and 11.68, respectively vs. low ϕ: CV = 755.25 and 615.26, respectively; Table 3). Both $\bar{i}$ and $\sigma^2 i$ drove performances of all models, but Strauss & Sadler's had no optimal range for any characteristic. For example, McInerny's, Solow's and GRIWM performed better under short-duration $\bar{i}$ (CV = 0.01 for both McInerny's and Solow's and

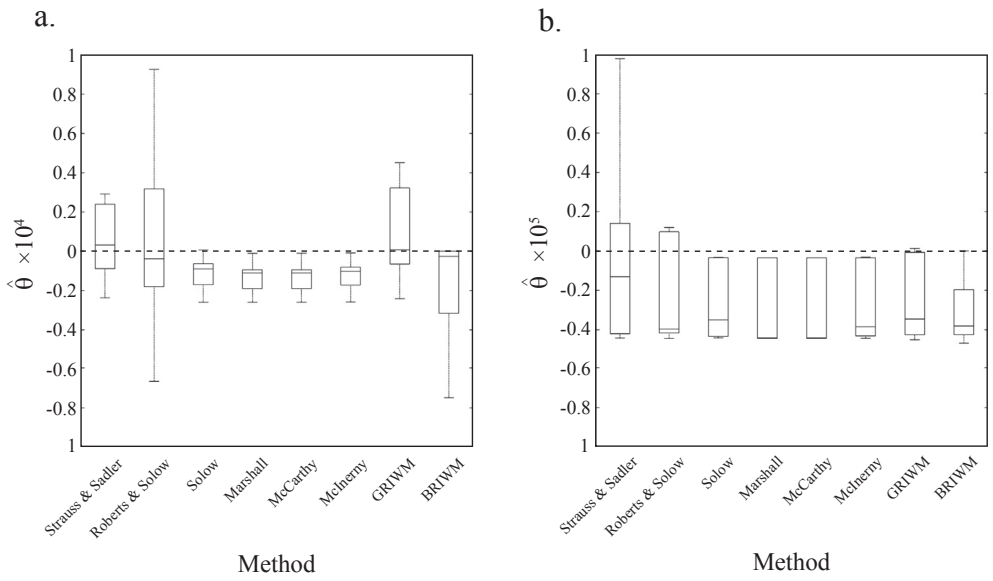

Fig. 2. Boxplots of model outputs (extinction time $\hat{\theta}$, in years for Strauss and Sadler's, Roberts and Solow's, McCarthy's, Marshall's, McInerny's, Solow's, GRIWM, and BRIWM) calculated for (a) six Australia extant species (*Dasyurus maculatus*, *Lagostrophus fasciatus*, *Macropus rufogriseus*, *Perameles gunnii*, *Petrogale brachyotis*, and *Tachyglossus aculeatus*) and (b) three extinct Australian species (*Diprotodon* sp., *Genyornis* sp., and *Thylacinus* sp.). Each boxplot is calculated on model estimates (i.e., confidence interval at 2.5%, 50% and 95%) from all six extant (a) and all three extinct (b) species (Table A.2) pooled together such that the central mark shows the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data not including outliers. Model estimates are compared with 1950 AD (0 years before present, dashed line), whereby positive values indicate that the model predicts species as 'extant' and negative values indicate the species as 'extinct'. For extant species, negative values indicate "Type I inference errors" (i.e., species wrongly predicted as extinct) and for extinct species, positive values indicate "Type II inference errors" (i.e., species wrongly predicted as extant).

CV = 0.07 for GRIWM), whereas Roberts and Solow's required longer-duration (CV = 0.25). Low $\sigma^2 i$ promoted performance in Marshall's, McInerny's, McCarthy's and GRIWM's methods, whereas higher $\sigma^2 i$ led to better performance in Roberts & Solow's and BRIWM.

4. Discussion

Selecting the most appropriate method to infer species extinction time from dated fossil records is not straightforward, and decisions cannot be based on only one index. We argue that our selection process provides a balance between various constraints dictated by the time series' characteristics of each dataset. It could also be applied to evaluate new methods as they are developed. The 'best' method should demonstrate: (*i*) robustness and flexibility for successfully inferring extinction timings for various types of dated records; (*ii*) the ability to account explicitly for most time series' characteristics (i.e., high coverage probability); (*iii*) both high accuracy and precision of inference (i.e., a low $\Delta_{\widehat{(\theta-\theta t)}}$ and a narrow confidence interval); and (*iv*) an ability to deal successfully with Type I and II statistical inference errors. Our results showed that there is no 'best' method (maximum $\phi < 0.4$, Fig. 1a), but combined interpretations of (*i*) ϕ (Fig. 1), (*ii*) application to real datasets (Fig. 2) and (*iii*) the sensitivity analysis (Fig. 3) provide enough information to create a model-selection key to address this task.

4.1. Model assessments: importance of intervals between records and their dating errors

Although GRIWM had lower ϕ than Marshall's, McCarthy's, and McInerny's methods (Fig. 1a) due to the former's generally wider confidence intervals (Fig. 1b), GRIWM synthesized essential characteristics (i.e., robustness and flexibility) to be applied successfully to most time series: (*i*) it generated both better accuracy (low $\Delta_{\widehat{(\theta-\theta t)}}$) and a higher coverage probability (Fig. 1b and c); (*ii*) it is the only model among the eight we tested that deals adequately with both Type I and II inference errors (Fig. 2); and (*iii*) it accounts for the full set of time series' characteristics used to describe dated records (Table 1). GRIWM presents two main model-specific characteristics that improved its robustness and adaptive flexibility given various record uncertainties: a down-weighting interval procedure and a Gaussian resampling of the dating errors.

First, GRIWM weights later record intervals more strongly, thus increasing the importance of low-density specimens as the species approaches true extinction (Fagan and Holmes, 2006; Bradshaw et al., 2012a). This weighting procedure counters the unrealistic assumption of a stationary Poisson distribution (i.e., that records are uniformly distributed along the time series; Solow et al., 1993) made by Strauss & Stradler's, Solow's and McInerny's models. The stationary Poisson distribution usually produces high model precision (Fisher and Blomberg, 2012) as supported by Solow's and McInerny's results (Fig. 1b). However, narrow confidence intervals weakened model performance when it did not offset $\Delta_{\widehat{(\theta-\theta t)}}$, thus reducing Solow's coverage probability (e.g., Solow's, Fig. 1c) and decreasing ϕ (Fig. 1a). This ultimately leads these models to be prone to Type I errors (e.g., Solow's and McInerny's; Fig. 2a, Table B.2; Jarić and Ebenhard, 2010). Similarly, Marshall's and McCarthy's models failed to deal with Type I errors; they could potentially be construed as the 'best' models because of their higher ϕ (Fig. 1a) driven by their narrow confidence intervals and moderate $\Delta_{\widehat{(\theta-\theta t)}}$ (despite a low coverage rate, Fig. 1b and c).

Second, GRIWM accounts for variation in dating errors (Bradshaw et al., 2012a), thus necessarily widening its estimated confidence interval compared to Solow's (which assumes no variation in dating error; Table 1), or Marshall's, McCarthy's and McInerny's methods that do not account for dating error at all (Fig. 2b). A wider confidence interval reduces GRIWM's accuracy compared to Solow's (Fig. 1c) and improves its coverage probability (and so, its net performance ϕ; Fig. 1), as well as decreasing the risks of making

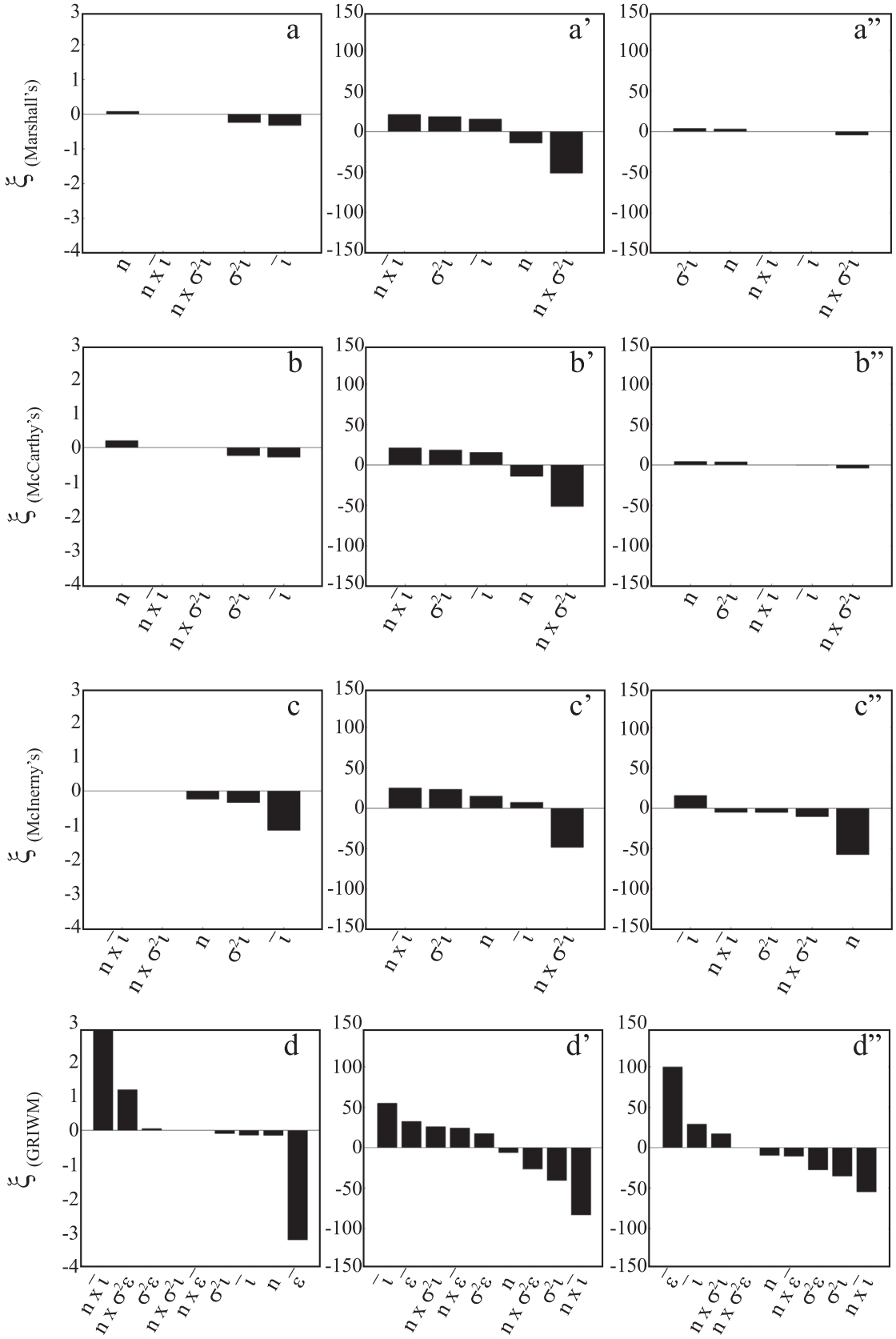

Fig. 3. Relative importance of the time series' characteristics (n, $\bar{i}$, $\sigma^2 i$, $\bar{\varepsilon}$, $\sigma^2 \varepsilon$; see Table 2 for complete description) on metrics used to calculate the model performance index (see equations in Section 2.3): the coverage probability of the theoretical timing of extinction (θ_t, panels a–d), size of model estimation biases ($\Delta_{\widehat{(\theta-\theta t)}}$ = the difference between the closest model's confidence bound to $\hat{\theta}$; panels a'–d') and the width of the estimated confidence interval (panels a"–d"). We displayed results of the four methods [(a, a' and a") McCarthy's, (b, b' and b") Marshall's, (c, c' and c") McInerny's, and (d, d' and d") GRIWM] among the eight methods tested, because they showed either the best model performance or they successfully dealt with both Type I and II errors (results from the four remaining models – Strauss and Sadler's, Roberts & Solow's, Solow's, and BRIWM – are shown in Appendices Fig. A.6). For each model and for each time series' characteristic (generated using a Latin hypercube within a range described in Table 2), we returned a standardized coefficient (ξ) calculated as the estimated coefficient of a generalized linear model fitted to (*i*) the model's ability to cover θ_t, (*ii*) $\Delta_{\widehat{(\theta-\theta t)}}$ and (*iii*) the confidence interval width divided by its standard deviation.

Table 3
Coefficient of variation (CV) for the set of summary characteristics (n, $\bar{i}$, $\sigma^2 i$, $\bar{\varepsilon}$, $\sigma^2 \varepsilon$; see Table 2 for complete description) for the proportion of times series whereby each model (Strauss and Sadler's, Roberts and Solow's, McCarthy's, Marshall's, McInerny's, Solow's, GRIWM, and BRIWM) provided both its high (+) and low (−) performance index. For each summary characteristic, the limit between (+) and (−) is defined as the summary characteristic value from which the model performance index decline precipitously (see detailed method in SI 3). Simulated time series were generated stochastically within a specified range (Table 1) following a Latin hypercube approach. For each model, CV is calculated as the variance of $\bar{i}$, $\sigma^2 i$, $\bar{\varepsilon}$, $\sigma^2 \varepsilon$ on (+) and (−) time series, divided by the median date of the width of the entire time series (i.e., 7000 years). Because n does not refer to a time period, we divided its variance by the median of maximum number of records tested (i.e., 50).

	n		$\bar{i}$		$\sigma^2 i$		$\bar{\varepsilon}$		$\sigma^2 \varepsilon$	
	(+)	(−)	(+)	(−)	(+)	(−)	(+)	(−)	(+)	(−)
Strauss and Sadler (1989)	0.59	0.59	0.21	0.18	0.40	0.31	–	–	–	–
Roberts and Solow (2003)	0.65	0.54	0.25	0.12	0.43	0.29	–	–	–	–
Solow et al. (2006)	0.63	0.55	0.01	0.19	0.03	0.39	9.04	755.27	–	–
Marshall (1997)	0.55	0.59	0.12	0.19	0.23	0.39	–	–	–	–
McCarthy (1998)	0.59	0.59	0.12	0.20	0.22	0.39	–	–	–	–
McInerny et al. (2006)	0.59	0.60	0.01	0.21	0.01	0.39	–	–	–	–
GRIWM (2012)	0.58	0.57	0.07	0.21	0.32	0.41	111.68	615.26	681.82	796.57
BRIWM	0.56	0.57	0.12	0.19	0.41	0.36	–	–	–	–

Type I errors (Fig. 2a). Nevertheless, although wider confidence intervals can reduce Type I errors, poor model precision (wide confidence intervals) such as those generated by Strauss & Sadler's and Roberts & Solow's methods (Fig. 1b), leads to Type II errors (Fig. 2b). Strauss and Sadler's method assumes a stationary Poisson distribution (which should theoretically narrow its confidence interval), but its high sensitivity to low numbers of records inflates its confidence interval (Strauss and Sadler, 1989; Rivadenera et al.,

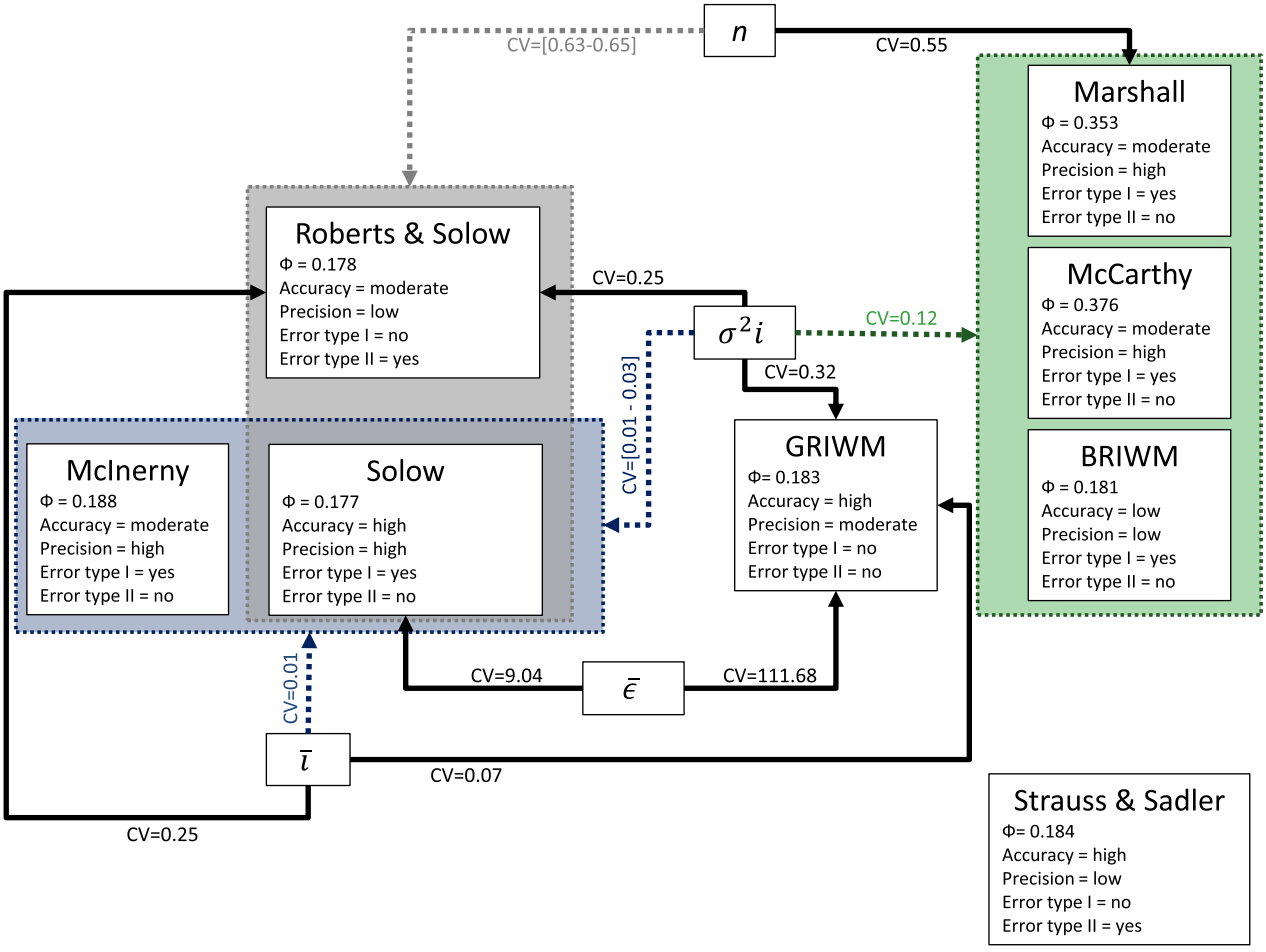

Fig. 4. Model selection guideline scheme. The most appropriate model is first selected as a function of the coefficient of variation (CV) of the time series' characteristics (n, $\bar{i}$, $\sigma^2 i$, $\bar{\varepsilon}$, $\sigma^2 \varepsilon$; see Table 2 for complete description) calculated from each given record compared with the closest value of benchmarked CVs (arrows under or above leading from the variable to the model). Solid arrows lead to only one model whereas dashed arrows (+coloured text) lead to one of the three groups of methods differentiated by coloured boxes (+dashed line colour). For each model, we specified (i) the performance index (ϕ; Fig. 1a), (ii) the model's ability to deal successfully (yes) or unsuccessfully (no) with Type I and II errors (extant vs. extinct misclassification; see Fig. 2) and model accuracy and precision (i.e., $\Delta_{\widehat{(\theta-\theta t)}}$ and confidence interval width, respectively; Fig. 1b). Benchmarked CVs are determined from the ϕ index calculated from simulated time series used for the sensitivity analyses (see details of methodology Fig. A.3). CVs are the variance of each time series characteristic divided by the median date of the width of the entire time series. Because n does not refer to a time period, we divided its variance by the same median of the maximum number of records tested in the sensitivity analysis (i.e., 50). Accuracy and precision are expressed as thresholds (a three-bin histogram on $\Delta_{\widehat{(\theta-\theta t)}}$ and confidence interval width): *high* ($\Delta_{\widehat{(\theta-\theta t)}} < 1000$ years; confidence interval width < 1200); *moderate* ($1000 \leq \Delta_{\widehat{(\theta-\theta t)}} < 1500$ years; $1200 \leq$ confidence interval width < 2000); *low* ($\Delta_{\widehat{(\theta-\theta t)}} \geq 1500$ years; confidence interval width ≥ 2000). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2009; Fig. A.4). Such a wide confidence interval (even wider using Roberts and Solow's models, Fig. 1b), makes models largely inefficient (Solow, 2005) because (*i*) it artificially improves coverage probability (confidence intervals offsets $\Delta_{\widehat{(\theta,\bar{\varepsilon}\theta t)}}$), biasing ϕ, and (*ii*) generates Type II error misclassification (Jarić and Ebenhard, 2010); Strauss & Sadler's and Roberts & Solow's models failed to predict the extinction of *Thylacinus* sp. and *Diprotodon* sp. (Fig. 2b and Table B.2).

More comparisons with the bootstrap variant of GRIWM (BRIWM) support the idea that the temporal dependency of dating errors is essential to improve performance. BRIWM neglects dating errors and assumes that θ can be accurately inferred from a subsample of records of the original time series. It emphasises that some records are more important than others and accounts for dating error in θ inference. We suggest that this method could be potentially used to counter the assumption that all records are equally reliable in term of data quality. The methods described in this paper implicitly assume a high quality (reliability) of the underlying dates examined (Solow et al., 2011), but this assumption has to be checked carefully. Date reliability can be handled either by developing a quality rating based on both robust and objective criteria to select only highly reliable records, by rejecting obviously uncertain or dubious records prior to analysis, or by other down-weighting methods not described here (Solow et al., 2011; Thompson et al., 2013; Lee et al., 2014). However, GRIWM's better performance relative to BRIWM (i.e., better accuracy and precision; Fig. 1) suggests either that dealing with unreliable dates requires more complex methods (Thompson et al., 2013; Lee et al., 2014), or that accounting for dating error explicitly (GRIWM) prevails over record reliability (BRIWM).

4.2. Toward 'ideal' time series' characteristics for a given model

GRIWM's ability to handle both Type I and II errors successfully does not preclude the application of other models if they are cautiously applied to questions relating to definitively extinct species and when the time series have certain characteristics (Rivadeneira et al., 2009). Marshall's, McCarthy's and McInerny's methods performed better when the time series had both high n and low $\sigma^2 i$ (Fig. 3). These models' outputs are sensitive to $n \times \sigma^2 i$ (Fig. 3a–c) because the negative correlation (Fig. A.5) between these characteristics affects recovery potential (Marshall, 1997; McCarthy, 1998) or sampling rate (McInerny et al., 2006), which are central to calculating final extinction date. For example, using a similar method to that used to calculate the recovery potential of Marshall's and McCarthy's models (Marshall, 1997; McCarthy, 1998; Holland, 2003; Farnsworth and Ogurcak, 2006), we used a function of sampling probability that depended on the time series used to calculate extinction time (Rivadeneira et al., 2009). However, similar to McInerny's sampling rate (McInerny et al., 2006), such a function delays the final extinction date when n is low and $\sigma^2 i$ is high, which increases either $\Delta_{\widehat{(\theta-\theta t)}}$ (Marshall's and McCarthy's; Fig. 3a and b) or the width of the estimated confidence interval (McInerny's; Fig. 3c), and ultimately decreases their respective performance (Fig. 1a). As GRIWM accounts for dating error and down-weights the influence of intervals between consecutive records, both $\bar{\varepsilon}$ and $n \times \bar{\imath}$ most determine GRIWM's applicability (Fig. 3 and Fig. A.6). The more records are positioned near to the true extinction date (i.e., the "up sampling" scenario described by Rivadeneira et al., 2009) and the lower the average dating error ($\bar{\varepsilon}$), the better GRIWM performs. First, increasing the number of records decreases the average interval duration (Fig. A.5) and because of GRIWM's down-weighting procedure, the youngest date intervals are giving more influence. Second, $\bar{\varepsilon}$ characterises the precision of the dates and depends on the limits of dating (e.g., ^{14}C), where radiometric (^{14}C) dating in particular provides lower dating errors for the most recent samples (Walker, 2005).

4.3. Example of model-selection key applications

Ideal time series are rarely available, so our model-selection key helps to choose the most appropriate model for a given dataset (Fig. 4). For example, applied to dated fossil records of three Australian extinct species, we recommend using GRIWM or/and Marshall's method on *Thylacinus* sp., GRIWM or/and Marshall's or/and McCarthy's on *Genyornis* sp., and McInerny's method on *Diprotodon* sp. First, the coefficients of variation for the characteristics of the *Thylacynus* sp. time series (Table B.1) matched Marshall's (n), GRIWM ($\bar{\imath}$) and both Solow's and McInerny's ($\sigma^2 i$) requirements (Fig. 4). However, (*i*) Marshall's and GRIWM performed better (high precision and moderate accuracy, Figs. 1a–c and 4) than Solow's and McInerny's and (*ii*) GRIWM avoided misclassification (Type I and II errors; Fig. 2). Second, although many methods are appropriate to infer extinction timing for *Genyornis* sp. such as GRIWM ($\bar{\imath}$ and $\bar{\varepsilon}$), Marshall's, McCarthy's and BRIWM ($\sigma^2 i$), we recommend using Marshall's, McCarthy's and GRIWM for the same reason as described for *Thylacynus* sp. Finally, McInerny's and Solow's both suited *Diprotodon* sp ($\sigma^2 i$; Fig. 4 and Table A.2), but McInerny's performed better than Solow's mainly due to a better coverage probability and higher model precision (Fig. 1a–c).

5. Conclusion

Estimates of time of extinction depend highly on the sensitivity of the method used to a time series' characteristics. Choosing a suboptimal method can lead to misclassification of extinction events (i.e., extant or extinct) and thus lead to incorrect conclusions about ecological processes driving extinctions. However, the robustness of many frequentist (non-Bayesian) methods is highly sensitive to inherent (laboratory) dating errors. Among the four frequentist methods providing highest model performance, Marshall's (1997) and McCarthy's (1998) methods had the highest precision. However, the Gaussian-resampled inverse-weighted McInerny (GRIWM) approach is the only method providing model accuracy as well as no misclassification issues because of its inherent down-weighting interval procedure and because it accounts for uncertainties in record dates. With inference errors in mind, we suggest that GRIWM, Marshall's, McCarthy's & McInerny's methods can provide reasonably accurate estimates of a species' extinction time.

Acknowledgements

This paper emerged from the Sahul-Linnaeus workshop "Patterns of late Quaternary extinctions and their relationship to climate change" in Ballina, Australia, October 2013. We thank the Environment institute (The University of Adelaide, grant 2014-12) for financial support. Linnaeus Estate (linnaeus.com.au) provided meeting space and discounted accommodation. F.S. and M.R-R. were supported by an Australian Research Council (ARC) Discovery Grant (DP130103842), and C.J.A.B. & B.W.B. were supported by ARC Future Fellowships (FT110100306 and FT100100200, respectively).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.quascirev.2015.01.022.

References

Alroy, J., 2001. A multispecies overkill simulation of the End-Pleistocene megafaunal mass extinction. Science 292, 1893–1896.

Alroy, J., 2014. A simple Bayesian method of inferring extinction. Paleobiology 584–607.

Barnosky, A.D., Koch, P.L., Feranec, R.S., Wing, S.L., Shabel, A.B., 2004. Assessing the causes of Late Pleistocene extinctions on the continents. Science 306, 70–75.

Barnosky, A.D., Matzke, N., Tomiya, S., Wogan, G.O.U., Swartz, B., Quental, T.B., Marshall, C., McGuire, J.L., Lindsey, E.L., Maguire, K.C., Mersey, B., Ferrer, E.A., 2011. Has the Earth/'s sixth mass extinction already arrived? Nature 471, 51–57.

Bradshaw, C.J.A., Cooper, A., Turney, C.S.M., Brook, B.W., 2012a. Robust estimates of extinction time in the geological record. Quat. Sci. Rev. 33, 14–19.

Bradshaw, C.J.A., McMahon, C.R., Miller, P.S., Lacy, R.C., Watts, M.J., Verant, M.L., Pollak, J.P., Fordham, D.A., Prowse, T.A.A., Brook, B.W., 2012b. Novel coupling of individual-based epidemiological and demographic models predicts realistic dynamics of tuberculosis in alien buffalo. J. Appl. Ecol. 49, 268–277.

Brook, B.W., Bowman, D.M.J.S., 2002. Explaining the Pleistocene megafaunal extinctions: models, chronologies, and assumptions. Proc. Natl. Acad. Sci. 99, 14624–14627.

Brook, B.W., Bradshaw, C.J.A., Cooper, A., Johnson, C.N., Worthy, T.H., Bird, M., Gillespie, R., Roberts, R.G., 2013. Lack of chronological support for stepwise prehuman extinctions of Australian megafauna. Proc. Natl. Acad. Sci. 110, E3368.

Brook, B.W., Sodhi, N.S., Bradshaw, C.J.A., 2008. Synergies among extinction drivers under global change. Trends Ecol. Evol. 23, 453–460.

Brosi, B.J., Biber, E.G., 2008. Statistical inference, Type II error, and decision making under the US Endangered Species Act. Front. Ecol. Environ. 7, 487–494.

Caley, P., Barry, S.C., 2014. Quantifying extinction probabilities from sighting records: inference and uncertainties. Plos One 9, e95857.

Fagan, W.F., Holmes, E.E., 2006. Quantifying the extinction vortex. Ecol. Lett. 9, 51–60.

Farnsworth, E.J., Ogurcak, D.E., 2006. Biogeography and decline of rare plants in New England: historical evidence and contemporary monitoring. Ecol. Appl. 16, 1327–1337.

Fisher, D.O., Blomberg, S.P., 2012. Inferring extinction of mammals from sighting records, threats, and biological traits. Conserv. Biol. 26, 57–67.

Flannery, T., 2002. The Future Eaters: an Ecological History of the Australasian Lands and People. Grove Press.

Grice, K., Cao, C., Love, G.D., Böttcher, M.E., Twitchett, R.J., Grosjean, E., Summons, R.E., Turgeon, S.C., Dunning, W., Jin, Y., 2005. Photic Zone Euxinia during the Permian-Triassic Superanoxic Event. Science 307, 706–709.

Hogg, A.G., Hua, Q., Blackwell, P.G., Niu, M., Buck, C.E., Guilderson, T.P., Heaton, T.J., Palmer, J.G., Reimer, P.J., Reimer, R.W., Turney, C.S.M., Zimmerman, S.R.J., 2013. SHCal13 southern hemisphere calibration, 0–50,000 cal BP. Radiocarbon 55.

Holland, S.M., 2003. Confidence limits on fossil ranges that account for facies changes. Paleobiology 29, 468–479.

Jablonski, D., 2001. Lessons from the past: evolutionary impacts of mass extinctions. Proc. Natl. Acad. Sci. 98, 5393–5398.

Jarić, I., Ebenhard, T., 2010. A method for inferring extinction based on sighting records that change in frequency over time. Wildl. Biol. 16, 267–275.

Jin, Y.G., Wang, Y., Wang, W., Shang, Q.H., Cao, C.Q., Erwin, D.H., 2000. Pattern of marine mass extinction near the Permian-Triassic Boundary in South China. Science 289, 432–436.

Johnson, C.N., Bradshaw, C.J.A., Cooper, A., Gillespie, R., Brook, B.W., 2013. Rapid megafaunal extinction following human arrival throughout the New World. Quat. Int. 308–309, 273–277.

Keith, D.A., Burgman, M.A., 2004. The Lazarus effect: can the dynamics of extinct species lists tell us anything about the status of biodiversity? Biol. Conserv. 117, 41–48.

Lee, T.E., McCarthy, M.A., Wintle, B.A., Bode, M., Roberts, D.L., Burgman, M.A., 2014. Inferring extinctions from sighting records of variable reliability. J. Appl. Ecol. 51, 251–258.

Lima-Ribeiro, M.S., Diniz-Filho, J.A.F., 2014. Obstinate overkill in Tasmania? The closest gaps do not probabilistically support human involvement in megafaunal extinctions. Earth-Sci. Rev. 135, 59–64.

Lima-Ribeiro, M.S., Diniz-Filho, J.A.F., 2013. American megafaunal extinctions and human arrival: improved evaluation using a meta-analytical approach. Quat. Int. 299, 38–52.

Link, W.A., Barker, R.J., 2006. Model weights and the foundations of multimodel inference. Ecology 87, 2626–2635.

Lorenzen, E.D., Nogués-Bravo, D., Orlando, L., Weinstock, J., Binladen, J., Marske, K.A., Ugan, A., Borregaard, M.K., Gilbert, M.T.P., Nielsen, R., Ho, S.Y.W., Goebel, T., Graf, K.E., Byers, D., Stenderup, J.T., Rasmussen, M., Campos, P.F., Leonard, J.A., Koepfli, K.-P., Froese, D., Zazula, G., Stafford Jr., T.W., Aaris-Sørensen, K., Batra, P., Haywood, A.M., Singarayer, J.S., Valdes, P.J., Boeskorov, G., Burns, J.A., Davydov, S.P., Haile, J., Jenkins, D.L., Kosintsev, P., Kuznetsova, T., Lai, X., Martin, L.D., McDonald, H.G., Mol, D., Meldgaard, M., Munch, K., Stephan, E., Sablin, M., Sommer, R.S., Sipko, T., Scott, E., Suchard, M.A., Tikhonov, A., Willerslev, R., Wayne, R.K., Cooper, A., Hofreiter, M., Sher, A., Shapiro, B., Rahbek, C., Willerslev, E., 2011. Species-specific responses of Late Quaternary megafauna to climate and humans. Nature 479, 359–364.

Marshall, C.R., 1997. Confidence intervals on stratigraphic ranges with nonrandom distributions of fossil horizons. Paleobiology 23, 165–173.

McCarthy, M.A., 1998. Identifying declining and threatened species with museum data. Biol. Conserv. 83, 9–17.

McInerny, G.J., Roberts, D.L., Davy, A.J., Cribb, P.J., 2006. Significance of sighting rate in inferring extinction and threat. Conserv. Biol. 20, 562–567.

Payne, J.L., Clapham, M.E., 2012. End-Permian mass extinction in the Oceans: an Ancient Analog for the Twenty-first Century? Annu. Rev. Earth Planet. Sci. 40, 89–111.

Prideaux, G.J., Roberts, R.G., Megirian, D., Westaway, K.E., Hellstrom, J.C., Olley, J.M., 2007. Mammalian responses to Pleistocene climate change in southeastern Australia. Geology 35, 33–36.

Ramsey, C.B., 2010. OxCal Version 4.1. Oxford Radiocarbon Accelaerator Unit, Oxford.

Rasmussen, P.C., Prys-Jones, R.P., 2003. History vs mystery: the reliability of museum specimen data. Bull. Br. Ornithol. Club 123A, 66–94.

Rivadeneira, M.M., Hunt, G., Roy, K., 2009. The use of sighting records to infer species extinctions: an evaluation of different methods. Ecology 90, 1291–1300.

Roberts, D.L., Solow, A.R., 2003. Flightless birds: when did the dodo become extinct? Nature 426, 245–245.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis: the Primer. Wiley.

Signor, P.W., Lipps, J.H., 1982. Sampling bias gradual extinction patterns and catastrophes in the fossil record. In: Silver, L.T., Schultz, P.H. (Eds.), In Geological Implications of Large Asteroids and Comets on the Earth. Geological Society of America, pp. 291–296.

Solow, A., Smith, W., Burgman, M., Rout, T., Wintle, B., Roberts, D., 2011. Uncertain sightings and the extinction of the Ivory-Billed woodpecker. Conserv. Biol. 26, 180–184.

Solow, A.R., 1993. Inferring extinction from sighting data. Ecology 74, 962–964.

Solow, A.R., Roberts, D.L., Robbirt, K.M., 2006. On the Pleistocene extinctions of Alaskan mammoths and horses. Proc. Natl. Acad. Sci. 103, 7351–7353.

Solow, A.R., 2005. Inferring extinction from a sighting record. Math. Biosci. 195, 47–55.

Song, H., Wignall, P.B., Tong, J., Yin, H., 2013. Two pulses of extinction during the Permian-Triassic crisis. Nat. Geosci. 6, 52–56.

Strauss, D., Sadler, P., 1989. Classical confidence intervals and Bayesian probability estimates for ends of local taxon ranges. Math. Geol. 21, 411–427.

Sun, Y., Joachimski, M.M., Wignall, P.B., Yan, C., Chen, Y., Jiang, H., Wang, L., Lai, X., 2012. Lethally Hot Temperatures during the Early Triassic Greenhouse. Science 338, 366–370.

Thompson, C.J., Lee, T.E., Stone, L., McCarthy, M.A., Burgman, M.A., 2013. Inferring extinction risks from sighting records. J. Theor. Biol. 338, 16–22.

Walker, M., 2005. Quaternary Dating Methods. Wiley.

Wang, Y., Sadler, P.M., Shen, S.Z., Erwin, D.H., Zhang, Y.C., Wang, X.D., Wang, W., Crowley, J.L., Henderson, C.M., 2014. Quantifying the process and abruptness of the end-Permian mass extinction. Paleobiology 40, 113–129.

Wroe, S., Field, J.H., Archer, M., Grayson, D.K., Price, G.J., Louys, J., Faith, J.T., Webb, G.E., Davidson, I., Mooney, S.D., 2013. Climate change frames debate over the extinction of megafauna in Sahul (Pleistocene Australia-New Guinea). Proc. Natl. Acad. Sci. U. S. A. 110, 8777–8781.