Assessing adequacy of models of phyletic evolution in the fossil record

Kjetil Lysne Voje^{1,2,*}

¹Centre for Ecological and Evolutionary Synthesis (CEES), Department of
Biosciences, University of Oslo, Oslo, Norway

²Department of Earth Sciences, Uppsala University, Villavägen 16, 75236 Uppsala,
Sweden.

*E-mail: k.l.voje@ibv.uio.no

Running headline: Absolute fit of phyletic models

The approach is implemented in the new R package adePEM, which is available on
github (<https://github.com/klvoje/adePEM>), DOI:10.5281/zenodo.1400988

Data and r code to reproduce all analyses are available at
https://github.com/klvoje/Phyletic_model_adequacy (DOI:10.5281/zenodo.1405131).

3 Tables

6 Figures

Abstract: 242

Main text: 5671

References: 1059

Keywords: fossil record, paleoTS, random walk, stasis, trend, time series, adePEM

Abstract

1. Comparing relative fit of different models of evolutionary dynamics to time series of phyletic change is a common tool when interpreting the fossil record. However, a measure of relative fit is no guarantee the preferred model describes the data well. Selecting a good model is essential for robust inferences, but we are currently lacking tools to investigate if a model of phyletic evolution represents an adequate description of trait dynamics in fossil data.
2. This study develops a general statistical framework implemented in R for assessing the adequacy of the three most commonly used models of evolution in the fossil record; stasis, directional change and random walk. The statistical framework is applied to 300 fossil time series in order to assess how often the three models represent adequate descriptions of evolutionary dynamics in the fossil record.
3. The model that showed the best relative fit to a particular fossil time series (using AICc) passed all adequacy tests in 219 out of 300 cases (73%, directional trend = 76%, stasis = 64%, random walk = 81%). It is therefore not uncommon that the best model according to AICc does not adequately describe the trait dynamics in a fossil time series.
4. Statistical tests of model adequacy ease evaluation of whether a particular model is a good descriptor of phyletic evolution and can assist in making meaningful inferences of model parameters (e.g. as rates of evolution) and interpretations of evolution in the fossil record.

56 Introduction

57 The fossil record is our only direct source of information on how past life forms have
58 evolved. How we interpret the fossil record is accordingly fundamental for our
59 understanding of evolution on timescales beyond a few centuries. A tight association
60 between pattern (mode) and process has often been argued for when interpreting the
61 fossil record (Eldredge & Gould 1972; Gould & Eldredge 1977; Stanley 1975; 1979),
62 but recognizing distinct patterns of morphological change in fossil time series were
63 for a long time a highly subjective exercise due to lack of statistical tools for
64 comparing competing interpretations. This changed when Hunt (Hunt 2006; Hunt,
65 Bell, & Travis 2008; Hunt 2008; Hunt, Wicaksono, Brown, & Macleod 2010; Hunt,
66 Hopkins, & Lidgard 2015) developed a model framework that allowed an objective
67 evaluation of relative fit of different models to fossil time series based on their AIC
68 scores. Hunt's models of the canonical modes stasis, directional change and random
69 walk are widely used when analyzing phyletic time series and his model framework
70 has greatly advanced our ability to interpret the fossil record (e.g., Hunt 2007;
71 Hopkins & Lidgard 2012; Pearson & Ezard 2014; Hunt *et al.* 2015; Voje 2016;
72 Brombacher, Wilson, Bailey, & Ezard 2017; Spanbauer, Fritz, & Baker 2018).

73 A potential shortcoming when relying only on relative model fit is the fact that
74 the best model among the candidates may not describe the data particularly well. This
75 is true because any list of candidate models will only reflect a subset of ways to
76 describe the data and Akaike information criterion and likelihood ratio tests lack the
77 ability to reject all candidate models if they do not provide an adequate fit to the data.
78 If we fit different models of evolutionary dynamics to a particular data set, one of
79 these model will show a better relative fit to the data, irrespective of whether this
80 model matches the observed evolutionary dynamics in the data or not. It is therefore

important to evaluate to what extent a best-fitting model among a set of candidates in fact represents an adequate description of the data. Blindly interpreting the top-ranked model based on relative fit alone may in fact prevent sensible interpretations of the analyzed data and hinder investigation of the research questions we seek to explore. A close match between model and data is especially important when the goal is to make meaningful inferences from parameters in the model (Hunt 2012; Pennell, FitzJohn, Cornwell, & Harmon 2015). For example, Hunt (2012) showed, using simulations, that evolutionary rates in the fossil record could be estimated as parameters in his models of stasis, directional change and random walk as long as these models accurately described the trait dynamics. While estimating rates of evolution as model parameters is a relatively new approach when analyzing fossil time series, model parameters are commonly interpreted as rates of change in phylogenetic comparative methods (e.g., Hansen 1997; Hansen, Pienaar, & Orzack 2008; Ackerly 2009; Harmon et al. 2010; Adams 2013; Slater 2013; 2015). Statistical procedures have also been developed to investigate the absolute fit of various models of evolution along a phylogeny to ensure meaningful interpretations of model parameters (e.g., Garland, Harvey, & Ives 1992; Boettiger, Coop, & Ralph 2012; Beaulieu, O'Meara, & Donoghue 2013; Slater and Pennell 2013; Pennell et al. 2015). However, we are currently short on tools to assess whether Hunt's (2006) models of phyletic evolution are actually capturing trait dynamics in fossil data in an adequate way.

Voje, Starrfelt, and Liow (2018) assessed the absolute fit of Hunt's (2006) stasis model to a large data set of fossil time series by applying adequacy tests. Here, I build on the work by Voje et al. (2018) and develop a statistical framework for investigating the adequacy of the three most popular models developed by Hunt (2006) when investigating evolution in fossil time series: directional change, random

walk and stasis. In short, the goal of the approach is to evaluate how likely it is that a particular model X with parameters Y can produce trait dynamics similar to what is observed in a data set Z. This is assessed by a parametric bootstrap approach: A model of evolution is judged as an adequate statistical representation of the trait dynamics in a fossil time series if the results of statistical tests on the observed data are similar to the same test statistics calculated on simulated data generated using the investigated model. Our confidence in a particular model increases if it is able to reproduce properties of the observed trait dynamics in the fossil data. This way of assessing model adequacy is similar to the approach developed by Pennell et al. (2015) for investigating the adequacy of various phylogenetic trait models.

I start out by providing some background on Hunt's (2006) models directional change, random walk and stasis and briefly discuss how parameters in these models can be understood as measures of evolutionary rates, following Hunt (2012). I then introduce the rationale behind investigating model adequacy before I go through the process of assessing model adequacy using the statistical approach. I conduct a simulation study to investigate if the proposed adequacy tests behave as expected in relation to type 1 error. I then proceed and analyze 300 fossil times series to investigate how often models with a better relative fit according to AICc also represent an adequate statistical representation of the data. I analyze three fossil time series in more detail to exemplify how assessment of model adequacy can inform interpretations of phyletic evolution in the fossil record. These case studies highlight several aspects of evaluating model adequacy, including cases where interpretations of model parameters are troublesome and how a model with a higher (worse) AICc score may represent a better statistical representation of the data compared to a model with a lower (better) AICc score.

Material and Methods

Hunt's (2006) models of phyletic evolution and their rate parameters

This section provides some background info on Hunt's (2006) models of stasis, directional change and random walk and how parameters in these models can be interpreted as rates of evolution. This section can be skipped by readers familiar with Hunt's (2006) model framework.

Time occurs in discrete intervals in all three models (stasis, random walk and directional change) and the expected difference between sample means (step size) is represented by a normal distribution, with a given mean (μ) and variance (σ^2). The mean of the step distribution is zero in the random walk model, which means the expected difference between consecutive sample means is zero with a variance of $t \sigma^2$, where t is the number of time steps (e.g., generations) separating an ancestor and descendant population (i.e. two trait means). The directional trend model differs from the random walk model in that the mean of the normal distribution is different from zero. The mean of the distribution reflects the direction of evolution over time, while σ^2 represents the fluctuations around the directional trend. The random walk model is accordingly nested within the more general directional change model and contains one parameter less compared to the directional change model. The stasis model describes a trait with an optimal/fixed phenotype (θ) that the trait fluctuates around with a variance (ω). The stasis model therefore model trait variation in a lineage over time as a white noise process, with uncorrelated normally distributed trait values around a fixed mean through time. See Hunt (2006) for a detailed description of the three models and see Hunt and Carrano (2010) for an interpretation of similar models fitted to phylogenetic trees.

Hunt (2012) derived a rate metric for each of the three models. In the random walk model, the magnitude of evolutionary divergence over any specified time interval is determined solely by the variance of the estimated normal distribution from which evolutionary steps are drawn. The variance parameter is therefore an estimate of the evolutionary rate for this model (Hunt 2012). This interpretation is similar to how the variance parameter in a Brownian motion is interpreted as an evolutionary rate parameter in phylogenetic comparative studies (e.g., Ackerly 2009; Adams 2013).

As pointed out by Hunt (2012), a single rate metric for the directional change model is less straightforward to define as trait dynamics in this model both depends on the mean and the variance of the estimated normal distribution. If the variance parameter is zero or extremely small compared to the mean of the distribution (i.e. $\mu \gg \sigma^2$), the directional component (μ) dominates the trait dynamics and becomes the most direct measure of evolutionary rate. In the opposite situation, when the directional component is negligible compared to the variance (i.e. $\sigma^2 \gg \mu$), the model behaves close to a random walk, a case where the variance of the normal distribution becomes the measure of evolutionary rate. When both parameters are of a magnitude that substantially affects the trait dynamics, however, the expected trait divergence over a given time span is influenced by both parameters in a way that makes it difficult to precisely define the expected rate of change in a single parameter (Hunt 2012).

The variance (ω) around the optimal/fixed phenotype (θ) is the natural rate metric in the stasis model (Hunt 2012). Note, however, that the “rate” parameter in the stasis model has a different interpretation than the rates estimated in the two other models: while the directional change and random walk models are models of trait

change, the stasis model is a model of trait values around a fixed optimum. Time does not affect the expected trait divergence in the stasis model, which means a rate of change between discrete time units (e.g., generations) is not possible to estimate using this model. The variance (ω) parameter in the stasis mode is therefore more correctly interpreted as a measure of morphological disparity within a lineage (Roy & Foote 1997; Ciampaglio, Kemp, & McShea 2001; Hunt 2012), as deviations from a fixed fitness optimum on the adaptive landscape (Voje et al. 2018), or as permissible morphologies within an adaptive zone occupied by a lineage over time.

Adequacy tests for the models stasis, random walk and directional change

Voje et al. (2018) developed four adequacy tests to assess the absolute fit of Hunt's (2006) stasis model to fossil time series. The underlying evolutionary dynamics according to this stasis model is similar to a white noise process (see above) with uncorrelated normally distributed trait values around a fixed trait value through time. Three of four tests of model adequacy used in Voje et al. (2018) were designed to ensure that the data did not violate expectations of a white noise process (test of autocorrelation, a runs test and a fixed-variance test, see Table 1 for details on all adequacy tests). The fourth test used in Voje et al. (2018) investigates an essential part of the general (and verbal) definition of stasis, namely that lineages fitting stasis show little net change over time. All four tests have well-understood statistical properties and capture a range of possible model violations of a white noise-process.

All the tests listed in Table 1 (except the net evolution test) evaluate if the data behave as expected under a white noise process. However, a white noise process does not describe the expected trait dynamics predicted by the random walk or the directional change models. Applying the same tests on all three models ease

interpretation of how data may violate underlying model assumptions, as outcomes of the model adequacy tests would be directly comparable across the three models. Instead of creating model-specific tests, adequacy of the random walk or directional change models are therefore investigated by detrending the data to behave as a white noise process prior to applying the adequacy tests: Evolutionary change according to the random walk model is a stochastic process that consists of a series of random steps drawn from a normal distribution with $\mu = 0$ and a non-zero variance (see above). A random walk can accordingly be transformed to a white noise process simply by successively subtracting sample means in the time series. The directional trend model is a correlated (biased) random walk as evolutionary changes are drawn from a normal distribution where $\mu \neq 0$. The linear trend in the directional change model can accordingly be removed by subtracting a linear model from the data. To remove the linear trend and transform the data to fit a white noise process, the estimated mean of the step distribution is the slope parameter and the intercept is the first trait mean in the time series. In short, if either a random walk or a directional change model represents an adequate description of the observed trait dynamics in a given data set, the data is expected to behave as a white noise process after the data have been detrended as described above.

Assessing model adequacy

The process of assessing whether a model shows an adequate fit to a particular data set follows these steps (Fig. 1): (1) The model we want to assess the adequacy of (stasis, random walk or directional change) is fitted to a fossil time series and model parameters are estimated by maximum likelihood using the paleoTS package version 0.5-1 (Hunt 2006; Hunt *et al.* 2008; Hunt 2008; Hunt *et al.* 2010; 2015). (2) Test

statistics are calculated on the observed time series. If it is the adequacy of the random walk or directional change models that is being assessed, the data are detrended prior to calculating the test statistics. (3) 1000 new time series are then simulated according to the model being evaluated using the parameter(s) estimated from the observed time-series (step 1). (4) Test statistics are estimated on each of the simulated time series. Again, the data are detrended prior to calculating the test statistics if it is the adequacy of the random walk or directional change model that are being assessed. (5) Lastly, each of the test statistics from the observed data are compared to the distribution of test statistics calculated on the 1000 simulated time series. The investigated model is judged unsuitable as a statistical description of a particular fossil time series if one or more test statistics calculated on the real data fall outside 95% of the calculated test statistics on the simulated time series.

Verifying the adequacy models using simulations

The performance of the statistical framework assessing model adequacy was investigated by conducting a simulation study. All the test statistics implemented in the framework have well-known statistical properties, but I used simulations to investigate the effects of variation in lengths of fossil time series (number of trait means) and sensitivity to varying parameters in the underlying models. The simulations follow the procedure described above and shown in Figure 1, except that step 1 is a simulated time series with known parameter values. Simulations of stasis time series were done using the *sim.Stasis* function while simulations of random walk and directional change were done using the *sim.GRW* in the paleoTS package version 0.5-1 (Hunt 2006). Four different variance parameters were investigated (0.01, 0.02, 0.04, 0.06) when simulating time series using the random walk and stasis models. The

theta (fixed/optimal trait value) was set to 1 for the stasis model. When simulating time series using the directional change model, the variance of the normal distribution was set to 0.01, while I tested different means of the normal distribution (0.01, 0.02, 0.04, 0.06). The sequence length was varied the same way for all three models and parameter combinations (number of sample means in time series =10, 20, 40, 80).

For each combination of parameters and sequence length for a particular model, one ‘observed’ time series was simulated (and detrended if the model generating the trait dynamics was either the random walk or directional change) before the test statistics were calculated. The estimated model parameters from the ‘observed’ time series were then used to simulate 1000 new time series. The test statistics were then applied on each of the simulated data sets in order to obtain distributions for each test statistic. Again, if either the random walk or directional trend models generated the simulated data, the simulated data were detrended before the test statistics were calculated. The distributions of test statistics were then used to investigate the frequency of type I error for the ‘observed’ data. This procedure was repeated 500 times for each combination of parameters and sequence length for each of the three models.

R code

The statistical framework for investigating model adequacy of Hunt’s (2006) models stasis, random walk and directional change has been implemented as a R package called adePEM (Assessing adequacy of phyletic-evolution models), available on github (<https://github.com/klvoje/adePEM>, DOI:10.5281/zenodo.1400988). The readme file contains info on how to install the package and examples of how to assess the adequacy of phyletic time series using the package. adePEM is compatible with

281 how fossil time series are analyzed using the package paleoTS version 0.5-1 (Hunt
282 2006; Hunt *et al.* 2008; Hunt 2008; Hunt *et al.* 2010; 2015), meaning paleoTS objects
283 can be analyzed directly using adePEM. Functions are provided so that the user can
284 run a given adequacy test for a specific model of interest (e.g. `auto.corr.test.RW`,
285 `runs.test.trend`, `net.change.test.stasis`). More useful to most users are probably
286 functions that run all adequacy tests simultaneously for a given model
287 (`fit3adequacy.RW`, `fit3adequacy.trend`, `fit4adequacy.stasis`). The user can define the
288 number of iterations in the bootstrap approach and can set the confidence level for
289 which a model is deemed suitable as a proper descriptor of a particular data set. The
290 functions running the adequacy tests automatically estimate model parameters
291 specifying joint parameterization, but the user has the option to define the model
292 parameters that should be investigated. The outcome of the adequacy tests is
293 presented both graphically (optional) and numerically.

294

295 *Applying the model adequacy framework to empirical data*

296 I apply the statistical framework for assessing model adequacy to 300 time series of
297 phyletic evolution in the fossil record in order to assess how often the models stasis,
298 random walk and directional change violates the implemented adequacy tests. The
299 majority of the data analyzed in this study overlap with the data analyzed in Voje et
300 al. 2018, Voje (2016), Hunt (2007), Hopkins and Lidgard (2012) and Hunt et al.
301 (2015), but were filtered to meet certain criteria: Each time series had to consist of at
302 least 10 sample means and time had to be on an absolute scale (instead of relative).
303 All traits analyzed were on a log-scale. The models directional change, random walk
304 and stasis were fit to each of the 300 time series by maximum likelihood using the
305 *fit3models* function (specifying joint parameterization) in the paleoTS package

version 0.5-1 (Hunt 2006; Hunt *et al.* 2008; Hunt 2008; Hunt *et al.* 2010; 2015) using R version 3.5.0. (R Core Team 2016). Based on the model that showed the best relative fit to the data according to its AICc score, I ran the adequacy tests for that model using the adePEM package. For example, if the random walk showed a better relative fit to a particular data set compared to the stasis and directional change models, I used the fit3adequacy.RW function to investigate if the random walk model also fitted the data in an absolute sense, i.e., I checked if the model passed all three adequacy tests, which would indicate that the model provides a good statistical explanation for the trait dynamics in the data.

I evaluate the relative and absolute fit of the stasis, random walk and directional change models to three fossil time series in more detail to exemplify how tests of model adequacy can aid interpretation of fossil data. These three data sets are the evolution of (log) dorsal fin ray number in *Gasterosteus doryssus* (Bell, Baumgartner, & Olson 1985), the (log) diameter of the proloculus in the foraminifer *Afrolivina afra* (Campbell & Reyment 1978) and (log) area of second cycle (proximal view) in the coccolithophore lineage *Chiasmolithus* (Bralower & Parrow 1996). Relative model fit was in each case investigated using the fit3models function in the paleoTS package specifying joint parameterization. R code and data to reproduce all analyses in this study are available at https://github.com/klvoje/Phyletic_model_adequacy (DOI:10.5281/zenodo.1405131).

Results

Simulations to evaluate adequacy tests

The four test statistics for evaluating model adequacy work as intended (Fig. 2, Fig1A

and 1B in supplement). Type 1 error rates for the test statistics for each of the three models are centered around 0.05. Length of the time series or variation in underlying parameters do not seem to have an effect on the type I error rate. Most of the simulated time series that are deemed non-adequate only violate one of the test statistics (Fig. 2, Fig1A and 1B in supplement).

Relative and absolute model fit

Table 2 summarizes the relative and absolute fit of the three models stasis, random walk and directional change to 300 fossil time series. The best model based on AICc (relative fit) passed all adequacy tests in 219 out of the 300 cases (73%); directional change model = 16/21 (76%); random walk model = 113/139 (81%); stasis model = 90/140 (64%).

It is not uncommon that more than one model show a similar relative fit to the same data set. The directional change and random walk models have similar AICc scores (difference in model fit ≤ 2 AICc units) for 58 fossil time series when one of these models shows the best relative fit to the data. Tests of model adequacy suggest that only one of the models represents an adequate description of the data in 16 of these data sets, i.e. one of the models (e.g. random walk) passed all adequacy tests while the alternative model (e.g. directional change) failed at least one adequacy test. The stasis and random walk models have similar relative fit (difference in model fit ≤ 2 AICc units) for 25 data sets. Investigating model adequacy shows that only one of the models represents an adequate description of the data in 12 of these 25 cases. The directional change and stasis models have never similar AICc scores (difference in model fit ≤ 2 AICc units) when one of these two models have the best relative fit to a data set.

More than one model can often adequately describe the same data. This is especially true for the random walk and the directional change models. In cases where the stasis model have the best relative fit, the random walk and the directional change models were found to be adequately describing the data in 53% and 33% of the cases respectively. Similarly, the directional change model describes the data adequately in 82% of the cases when the random walk model shows the best fit based on AICc, and the random walk model adequately describes data that show a better fit to the directional change model in 13 out of 16 cases. The stasis model is rarely an adequate description of the data when the two other models fit best. Stasis is only an adequate model in one of the 16 cases when the directional trend model has the lowest AICc score and is an adequate descriptor of the data in 6 out of 113 cases when the random walk model shows the best fit.

Case studies

The random walk model has the best fit to the data describing how (log) dorsal fin ray number evolve in *Gasterosteus doryssus* (Fig. 3, Table 3) and is 1.79 AICc units better than the next best model (directional change, which marginally failed the autocorrelation test). The random walk model also passes all three adequacy tests, which indicates the model represents an adequate description of the trait dynamics. The maximum likelihood estimate of the vstep parameter (0.14) can accordingly be meaningfully interpreted as a rate of evolution (per million years) in this lineage.

The stasis model shows a substantially better relative fit to the evolution of (log) diameter of the proloculus in the Late Cretaceous foraminifer *Afrobolivina afra* (Fig. 4, Table 3) compared to the random walk ($\Delta\text{AICc} = 10.92$) and directional change models ($\Delta\text{AICc} = 13.20$). However, the data show a much stronger

autocorrelation than expected if stasis was the true generating model. The model also fails the runs test. Failing these two tests suggests the stasis model does not provide a good statistical explanation for the trait dynamics in the data and we should accordingly be careful when drawing inferences from the estimated model parameters. Neither the random walk nor the directional trend models pass all the adequacy tests, which suggest none of the three models represent a good statistical description of the data.

The directional change model has the best relative fit to the data on the evolution of (log) area of second cycle (proximal view) in the coccolithophore lineage *Chiasmolithus* (Fig. 5, Table 3), but the random walk model has only a slightly worse fit ($\Delta\text{AIC } 0.12$). The results of the three adequacy tests for the directional change model strongly suggest that this model does not represent an adequate statistical explanation of the data. Inferences from the directional change model must therefore be conditioned on the known violation of its adequacy as a good descriptor of the data (Fig. 5). The data pass all three adequacy tests for the random walk model (Fig. 6).

Discussion

Time series of phyletic evolution in the fossil record represent important data for understanding evolutionary change spanning more than a few decades. Comparing how alternative models fit fossil time series has greatly increased our ability to understand evolution on macroevolutionary time-scales (e.g. Hunt 2007; 2008; Hopkins & Lidgard 2012; Voje 2016), but Akaike information criteria and likelihood ratio-tests cannot evaluate the adequacy of fit between the model and data.

Investigating the relative and absolute fit of Hunt's (2006) models stasis, random

walk and directional change to 300 fossil time series showed that most of the analyzed data on phyletic evolution in the fossil record are adequately described by one of the three models (73%), but more than a quarter is not. Investigating the adequacy of phyletic-evolution models is therefore important in order to meaningfully address questions of morphological change in the fossil record.

A large difference in AICc score among rival models is not an indication of whether the best model is a sufficiently good descriptor of the data. The stasis model was clearly better than the alternative models according to their AICc scores in describing evolutionary changes in the diameter of the proloculus in *Afrobolivina afra*. Still, tests of adequacy showed that the stasis model was not an adequate descriptor of the evolution of this trait. Visual inspection of data is invaluable when assessing model fit, but spotting violations of underlying assumptions of a statistical model can be difficult to detect by eyeballing alone. It is for example not obvious by looking at the data that the evolution of dorsal fin ray number in *Gasterosteus doryssus* shows acceptable levels of autocorrelation while changes in the diameter of the proloculus in *Afrobolivina afra* do not. Statistical tests of model adequacy can therefore be of help when evaluating whether a model is a sufficiently good descriptor of a particular data set.

An adequate model may in some cases be a valid alternative to a model identified by other model selection methods (Ripplinger & Sullivan 2010). For example, in cases where the relative model fit is very similar among candidate models, but where the model with the slightly better relative fit does not represent an adequate description of the trait dynamics, it may be advisable to also interpret the model that represents an adequate description of the data. The random walk model was commonly found to be an adequate descriptor of data even when it showed a

431 poorer relative fit compared to the alternative models. For example, the directional
432 change model shows a better fit to the evolution of log area of second cycle (proximal
433 view) of the coccolithophore lineage *Chiasmolithus* compared to the random walk
434 model, but the directional change model fails one of three adequacy tests while the
435 random walk model passes all three tests. The difference in how the directional
436 change and random walk models fit the *Chiasmolithus* data according to AICc is very
437 small (0.12 AICc units). It can therefore be argued that the variance parameter from
438 the random walk model in this example can be meaningfully interpreted, while the
439 rate parameters from the directional trend model are more difficult to trust.

440 Failure of a particular model to adequately describe a data set may have both
441 biological and non-biological causes (Pennell *et al.* 2015). For example, properties of
442 the fossil record make time series prone to variation in time resolution and time
443 averaging for estimated population samples, both factors that can affect the fit of the
444 data to simple process models (Hunt *et al.* 2015; Voje *et al.* 2018). Still, biological
445 explanations may also underlie failures to pass adequacy tests. Tests of model
446 adequacy can therefore guide our thinking regarding which models that may fit a
447 particular data set. For example, some of the data that are not adequately described by
448 the models stasis, random walk and directional change may be better explained by
449 more complex models, where different parts of the time series are allowed to be
450 described by different models. Certain pair-wise combinations of the three models can
451 already be fitted to times series using the paleoTS package (Hunt *et al.* 2015), and
452 there are several examples where such complex models have a better relative model
453 fit compared to fitting the single models to the data (Hunt *et al.* 2015; Saito-Kato,
454 Tanimura, Mori, & Julius 2015; Spanbauer *et al.* 2018). To what extent these more
455 complex trait models also represent adequate descriptions of the data can be

investigated by independently evaluating model adequacy of the different sections of the time series fitting different models. Failures of models to adequately describe the statistical properties of the data may also guide new model developments (Pennell et al. 2015). For example, the random walk model sometimes fails the fixed variance test, which indicates that the variance (i.e. the step size of the normal distribution from which evolutionary changes are drawn) either increase or decrease as a function of time. A decrease in the variance of the normal distribution with time resembles the early burst model of evolution developed for phylogenetic comparative data (Blomberg, Garland, & Ives 2003; Harmon *et al.* 2010). The early burst model is inspired by theory on adaptive radiations predicting the rate of evolution in a clade will slow down after an initial radiation due to ecological opportunity (Yoder *et al.* 2010). Similarly, single lineages can also be predicted to show a slowdown of evolution over time after entering a new habitat. Failures of the random walk model to adequately describe the trait dynamics in certain data sets may therefore suggest that an early burst model of phyletic evolution can be worthwhile to develop and fit to certain fossil time series.

I would like to stress that tests of model adequacy should not alone determine to what extent a particular model should be rejected or not. For some questions, it may not be very important that the model represents an adequate description of the data in every way. For example, a fossil time series may show a stationary behavior and at the same time not pass the adequacy tests for the stasis model. The data may therefore be judged as following a stationary dynamics at the same time as the variance (ω) parameter from the stasis model should be interpreted with care. It is also important to realize that model adequacy is not an either-or question. A model can represent everything from a very accurate to an extremely inaccurate description

of the trait dynamics in a time series. Again, how strict we demand a model to account for the trait dynamics in our data depends on the question we seek to explore. It is also important to remember that a model deemed an adequate descriptor of the evolution of a trait does not mean the trait evolved exactly as predicted by the model. An adequate model only means the data are not violating the specific model assumptions investigated by the applied adequacy tests. Furthermore, the tests listed in Table 1 only represent a subset of possible adequacy tests for the models stasis, directional change and random walk and additional tests should be applied if that is judged necessary for the particular biological question at hand. The observation that the random walk model represents an adequate description of a large number of time series, even for data where it has a much lower AICc score relative to the stasis and directional change models, is likely a reflection of the large range of trait dynamics this model can generate compared to the other models, and does not necessarily mean that most lineages evolve according to a random walk. It is also important to stress that even when a model is a good description of a fossil time series in both relative and absolute terms, careful interpretation of the parameters are still needed. For example, stasis is a label that for decades has been interpreted as low levels of evolution (e.g. Eldredge & Gould 1972; Gould & Eldredge 1977), but data fitting Hunt's (2006) stasis model show fluctuations that span a large range (Voje et al 2018) and may involve as much evolution through morphospace as data fitting alternative models (Voje 2016). Tests of model adequacy are important tools for meaningful interpretations of model parameters. Regardless, estimated parameters should always be carefully examined in the context of the investigated research question.

506 Acknowledgement

507 The manuscript benefited from comments from Jostein Starrfelt and two anonymous
508 reviewers. Thanks also to Jo Skeie Hermansen for testing the adePEM package. This
509 work would not have been possible without all the researchers who generated the
510 fossil time series data. The work was supported by grant No. 249961 from the
511 Norwegian Research Council.

512

513

514 Data Availability statement

515 The adePEM package (DOI:10.5281/zenodo.1400988) can be found on GitHub
516 (<https://github.com/klvoje/adePEM>). R code and data to reproduce all analyses in this
517 study are available at https://github.com/klvoje/Phyletic_model_adequacy
518 (DOI:10.5281/zenodo.1405131).

519

521

522

523

524 **Figure 1. Schematic diagram of the parametric bootstrap approach to assess**

525 **model adequacy.** 1, Fit model to data and estimate model parameters. 2, Calculate

526 test statistics on data. If the fitted model in step 1 is either the random walk or

527 directional change model, the data is detrended before the test statistics are calculated.

528 3, Simulate a large number of data sets using the estimated parameters from step 1. 4,

529 Calculate test statistics on all the simulated data sets. If the generating model of the

530 simulated data is the random walk or directional trend model, each simulated data set
531 is detrended before the test statistics are calculated. 5, Investigate if the test statistics
532 from the observed data (step 2) fall within the distribution of test statistics calculated
533 on the simulated data. The model can be rejected as inadequate if an observed test
534 statistic is in one of the tails of the distribution of test statistics from the simulated
535 data. The circle arrow in the center illustrates that assessing model adequacy may
536 involve testing the adequacy of several models. For example, alternative models with
537 similar or poorer relative fit compared to the model with the best AICc score may
538 show good absolute fit and can in some cases be helpful when interpreting the data.
539

540

541 Figure 2 **Type 1 error rates of adequacy tests when the generating model is**
542 **random walk.** 500 ‘true’ random walk time-series were simulated for a given
543 sequence length (orange diamond = 10, red square= 20, blue cross = 40, black triangle
544 = 80) and a given size of the variance (vstep) parameter. For each of these ‘true’
545 random walk time series, 1000 time series were simulated to check if the test statistics
546 estimated on the ‘true’ data falls within the 95% of the observed test statistics
547 conducted on the simulated data. Type 1 error rates for the three test statistics are
548 around the expected 0.05 threshold. The simulation studies investigating type 1 error
549 rates in the directional change and stasis models show qualitatively identical results
550 (Fig 1A and 1B in the supplementary information). All = the proportion of simulated
551 time series that deviated from the 95% distribution in at least one of the four test

552 statistics. Fixed var. = test for a relationship between time and the size of deviations
553 from the fixed mean in the white noise process, Runs = test for non-random patterns
554 in the sign of deviations from the fixed mean in the white noise process, Autocor. =
555 test for autocorrelation in the data.
556

Figure 3. Assessing the adequacy of the random walk model. The random walk model has the best fit according to AICc to the evolutionary dynamics of (log) dorsal fin ray number in *Gasterosteus doryssus*. The random walk model passes all three adequacy tests since each of the three observed test statistics (red broken vertical bar) are always within the 2.5% tails of the distributions of test statistics calculated on the simulated data.

Figure 4. Assessing the adequacy of the stasis model. The evolution of the (log) diameter of the proloculus in *Afrobolivina afra* is best described by the stasis model according to AICc. However, the results of the autocorrelation and runs tests are falling outside the expected values for these test statistics based on the simulated data (i.e. they fall outside the 2.5% tails of the distributions), suggesting the stasis model is not an adequate description of the observed trait dynamics.

Figure 5. Assessing the adequacy of the directional change model. The directional change model outcompeted the alternative models according to their AICc scores in describing the evolution of the (log) area of second cycle (proximal view) of the coccolithophore lineage *Chiasmolithus*. Still, the directional change model does not represent an adequate statistical description of the observed dynamics in this trait since the model fails the runs test. The model is also close to fail the autocorrelation test. The data are adequately described by the random walk model (Fig. 6).

Figure 6. Random walk as an adequate model when the directional change model shows a better relative fit. The AICc score for the random walk model was slightly worse compared to the AICc score for the directional change model in describing the evolution of the log area of second cycle (proximal view) of the coccolithophore lineage *Chiasmolithus* (Table 3). Still, the random walk model is deemed an adequate description of the data as the model passes all three adequacy tests.

596 Table 1. Description of test statistics used to assess adequacy of models of phyletic
597 evolution.

598

Test	Description
<i>Autocorrelation</i>	The correlation of the first $n-1$ observations with the last $n-1$ observations in the time series. In a white noise process, observations (population means) represent random draws from a normal distribution and should exhibit low levels of autocorrelation.
<i>Runs test</i>	For a time series of length n , the number of <i>runs</i> (one run is a sequence of consecutive numbers with same sign), is approximately normal with mean $\mu = \frac{2(n_+n_-)}{n} + 1$ and variance $\frac{(\mu-1)(\mu-2)}{n-1}$, where n_+ and n_- are the number of residuals above and below the fixed mean/expected value respectively. The test statistic is the Z-score, which depends on the mean and variance. In a white noise process, there should be no tendency for the observations (population means) to successively deviate in the same direction.
<i>Fixed variance</i>	The slope of the least-squares regression of deviations (their absolute value) from the fixed phenotype/expected value as a function of time. Basically a test of heteroscedasticity of the absolute values of the residuals from the fixed mean in the white noise process. The slope indicates whether the variance of the white noise process is constant, increases or decreases as a function of time. A slope of zero is expected if the data follow a true white noise.

<i>Net evolution</i>	The absolute difference between the first and last sample mean in the time series. Low amounts of net evolution are an essential part of the general (verbal) definition of stasis. This test only applies to the stasis model.
----------------------	---

599
600

601 Table 2. Relative and absolute fit of the models stasis, random walk and directional
602 change to 300 fossil time series.

	Directional change	Random walk	Stasis
Best relative fit (based on AICc)	21	139	140
Absolute fit (passed all adequacy tests)	16 (76%)	113 (81%)	90 (64%)

603
604
605
606

607

608 Table 3. Relative model fit (AICc scores) and maximum likelihood parameter
609 estimates of the models stasis, random walk and directional change for the three case
610 study datasets. Bold AICc value indicates best model.

	Directional change	Random walk	Stasis
Trait: log dorsal fin ray number (<i>Gasterosteus doryssus</i>)			
AICc	-114.68	-116.47	-83.33
Parameters	mstep= 1.0059; vstep= 0.1372	vstep= 0.1432	theta = 2.2206; omega = 0.0020
Trait: log mean diameter of the proloculus (<i>Afrobolivina afra</i>)			
AICc	-25.54	-27.811	-38.73
Parameters	mstep= 0.7130; vstep= 6.3764	vstep= 6.3789	theta = 2.2157; omega = 0.0221
Trait: log area of second cycle (<i>Chiasmolithus</i>)			
AICc	-137.22	-137.10	-134.36
Parameters	mstep= -0.0034; vstep=0.0000	vstep= 0.0002	theta = -0.5300; omega = 0.0001

611

612 References

- 613 Ackerly, D. (2009). Conservatism and diversification of plant functional traits:
614 Evolutionary rates versus phylogenetic signal. *Proceedings of the National*
615 *Academy of Sciences*, 106 Suppl 2, 19699–19706.
616 <http://doi.org/10.1073/pnas.0901635106>
- 617 Adams, D.C. (2013). Comparing evolutionary rates for different phenotypic traits on a
618 phylogeny using likelihood. *Systematic biology*, 62, 181–192.
619 <http://doi.org/10.1093/sysbio/sys083>
- 620 Beaulieu, J. M., O'Meara, B. C., & Donoghue, M. J. (2013). Identifying hidden rate
621 changes in the evolution of a binary morphological character: the evolution of
622 plant habit in campanulid angiosperms. *Systematic Biology*, 62, 725–737.
623 <http://doi.org/10.1093/sysbio/syt034>
- 624 Bell, M.A., Baumgartner, J.V. & Olson, E.C. (1985). Patterns of temporal change in
625 single morphological characters of a Miocene stickleback fish. *Paleobiology*, 11,
626 258–271. <http://doi.org/10.1017/S0094837300011581>
- 627 Blomberg, S.P., Garland, T.Jr, & Ives, A.R. (2003). Testing for phylogenetic signal in
628 comparative data: behavioral traits are more labile. *Evolution*. 57:717-745.
629 [https://doi.org/10.1554/0014-3820\(2003\)057\[0717:TFPSIC\]2.0.CO;2](https://doi.org/10.1554/0014-3820(2003)057[0717:TFPSIC]2.0.CO;2)
- 630 Boettiger, C., Coop, G., & Ralph, P. (2012). Is your phylogeny informative?
631 Measuring the power of comparative methods. *Evolution*, 66, 2240–2251.
632 <http://doi.org/10.1111/j.1558-5646.2011.01574.x>
- 633 Bralower, T.J. & Parrow, M. (1996). Morphometrics of the Paleocene coccolith

genera *Crucioplacolithus*, *Chiasmolithus* and *Sullivaniana*: a complex evolutionary history. *Paleobiology*, 22, 352–385. <http://doi.org/10.1017/S009483730001633X>

Brombacher, A., Wilson, P.A., Bailey, I. & Ezard, T.H.G. (2017). The breakdown of static and evolutionary allometries during climatic upheaval. *The American Naturalist*, 190, 350–362. <http://doi.org/10.1086/692570>

Campbell, N.A., & Reymont R.M. (1978). Discriminant analysis of a Cretaceous foraminifer using shrunken estimators. *Journal of the International Association for Mathematical Geology*, 10, 347–359. <https://doi.org/10.1007/BF01031739>

Ciampaglio, C.N., Kemp, M. & McShea, D.W. (2001). Detecting changes in morphospace occupation patterns in the fossil record: characterization and analysis of measures of disparity. *Paleobiology*, 27, 695–715. [https://doi.org/10.1666/0094-8373\(2001\)027<0695:DCIMOP>2.0.CO;2](https://doi.org/10.1666/0094-8373(2001)027<0695:DCIMOP>2.0.CO;2)

Eldredge, N. & Gould, S.J. (1972). Punctuated equilibria: an alternative to phyletic gradualism. *Models in Paleobiology* (ed T. Schopf), pp. 82–115. Models in paleobiology, San Francisco.

Garland, T. Jr., Harvey, P. H., & Ives, A. R. (1992). Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Systematic Biology*, 41, 18–32. <http://doi.org/10.1093/sysbio/41.1.18>

Gould, S.J. & Eldredge, N. (1977). Punctuated Equilibria: The Tempo and Mode of Evolution Reconsidered. *Paleobiology*, 3, 115–151. <https://doi.org/10.1017/S0094837300005224>

Hansen, T. F. (1997). Stabilizing selection and the comparative analysis of adaptation.

656 Evolution, 51, 1341–1351. <https://doi.org/10.2307/2411186>

657 Hansen, T. F., Pienaar, J., & Orzack, S. H. (2008). A comparative method for
658 studying adaptation to a randomly evolving environment. *Evolution*, 62, 1965–
659 1977. <https://doi.org/10.1111/j.1558-5646.2008.00412.x>

660 Harmon, L.J., Losos, J.B., Davies, T.J., Gillespie, R.G., Gittleman, J.L., Jennings,
661 W.B., Kozak, K.H., McPeck, M.A., Moreno-Roark, F., Near, T.J., Purvis, A.,
662 Ricklefs, R.E., Schluter, D., Schulte, J.A.I., Seehausen, O., Sidlauskas, B.L.,
663 Torres-Carvajal, O., Weir, J.T. & Mooers, A.O. (2010). Early Bursts of Body
664 Size and Shape Evolution Are Rare in Comparative Data. *Evolution*, 64, 2385–
665 2396. <http://doi.org/10.1111/j.1558-5646.2010.01025.x>

666 Hopkins, M.J. & Lidgard, S. (2012). Evolutionary mode routinely varies among
667 morphological traits within fossil species lineages. *Proceedings of the National*
668 *Academy of Sciences*, 109, 20520–20525.
669 <https://doi.org/10.1073/pnas.1209901109>

670 Hunt, G. (2006). Fitting and Comparing Models of Phyletic Evolution: Random
671 Walks and beyond. *Paleobiology*, 32, 578–601. <https://doi.org/10.1666/05070.1>

672 Hunt, G. (2007). Evolutionary divergence in directions of high phenotypic variance in
673 the ostracode genus *poseidonamicus*. *Evolution*, 61, 1560–1576.
674 <https://doi.org/10.1111/j.1558-5646.2007.00129.x>

675 Hunt, G. (2008). Gradual or Pulsed Evolution: When should punctuational
676 explanations be preferred? *Paleobiology*, 34, 360–377.
677 <https://doi.org/10.1666/07073.1>

678 Hunt, G. (2012). Measuring rates of phenotypic evolution and the inseparability of
679 tempo and mode. *Paleobiology*, 38, 351–373. <https://doi.org/10.1666/11047.1>

680 Hunt, G., Bell, M.A. & Travis, M.P. (2008). Evolution towards a new adaptive
681 optimum: phenotypic evolution in a fossil stickleback lineage. *Evolution*, **62**,
682 700–710. <https://doi.org/10.1111/j.1558-5646.2007.00310.x>

683 Hunt, G., and Carrano, M.T. (2010). Models and methods for analyzing phenotypic
684 evolution in lineages and clades. In J. Alroy & G. Hunt (Eds), *Short Course on*
685 *Quantitative Methods in Paleobiology* (pp. 245–269). Paleontological Society,
686 Denver, CO. <https://doi.org/10.1017/S1089332600001893>

687 Hunt, G., Wicaksono, S.A., Brown, J.E. & Macleod, K.G. (2010). Climate-driven
688 body-size trends in the ostracod fauna of the deep Indian Ocean. *Palaeontology*,
689 53, 1255–1268. <https://doi.org/10.1111/j.1475-4983.2010.01007.x>

690 Hunt, G., Hopkins, M.J. & Lidgard, S. (2015). Simple versus complex models of trait
691 evolution and stasis as a response to environmental change. *Proceedings of the*
692 *National Academy of Sciences*, 112, 4885–4890.
693 <https://doi.org/10.1073/pnas.1403662111>

694 Pearson, P.N. & Ezard, T.H.G. (2014). Evolution and speciation in the Eocene
695 planktonic foraminifer *Turborotalia*. *Paleobiology*, 40, 130–143.
696 <https://doi.org/10.1666/13004>

697 Pennell, M.W., FitzJohn, R.G., Cornwell, W.K. & Harmon, L.J. (2015). Model
698 adequacy and the macroevolution of angiosperm functional traits. *The American*
699 *Naturalist*, 186, E33–E50. <https://doi.org/10.1086/682022>

700 R Development Core Team. (2013). R: a language and environment for statistical
 701 computing. R Foundation for Statistical Computing, Vienna.

702 Ripplinger, J. & Sullivan, J. (2010). Assessment of substitution model adequacy using
 703 frequentist and bayesian methods. *Molecular Biology and Evolution*, 27, 2790–
 704 2803. <http://doi.org/10.1093/molbev/msq168>

705 Roy, K. & Foote, M. (1997). Morphological approaches to measuring biodiversity.
 706 *Trends in ecology & evolution*, 12, 277–281. [https://doi.org/10.1016/S0169-](https://doi.org/10.1016/S0169-5347(97)81026-9)
 707 [5347\(97\)81026-9](https://doi.org/10.1016/S0169-5347(97)81026-9)

708 Saito-Kato, M., Tanimura, Y., Mori, S. & Julius, M.L. (2015). Morphological
 709 evolution of *Stephanodiscus* (Bacillariophyta) in Lake Biwa from a 300 ka fossil
 710 record. *Journal of Micropalaeontology*, 34, 165–179.
 711 <http://doi.org/10.1144/jmpaleo2014-015>

712 Slater, G. J. (2013). Phylogenetic evidence for a shift in the mode of mammalian body
 713 size evolution at the Cretaceous - Palaeogene boundary. *Methods in Ecology and*
 714 *Evolution*, 4, 734–744. <http://doi.org/10.1111/2041-210X.12084>

715 Slater, G. J. (2015). Iterative adaptive radiations of fossil canids show no evidence for
 716 diversity-dependent trait evolution. *Proceedings of the National Academy of*
 717 *Sciences*, 112, 4897–4902. <http://doi.org/10.1073/pnas.1403666111>

718 Slater, G. J., & Pennell, M. W. (2013). Robust Regression and Posterior Predictive
 719 Simulation Increase Power to Detect Early Bursts of Trait Evolution. *Systematic*
 720 *Biology*, 63, syt066–308. <http://doi.org/10.1093/sysbio/syt066>

721 Spanbauer, T.L., Fritz, S.C. & Baker, P.A. (2018). Punctuated changes in the

722 morphology of an endemic diatom from Lake Titicaca. *Paleobiology*, 44, 89-100.
 723 <https://doi.org/10.1017/pab.2017.27>

724 Stanley, S.M. (1975). A theory of evolution above the species level. *Proceedings of*
 725 *the National Academy of Sciences of the United States of America*, 72, 646–650.
 726 <https://doi.org/10.1073/pnas.72.2.646>

727 Stanley, S.M. (1979). *Macroevolution—pattern and process*. WH Freeman. San
 728 Francisco.

729 Voje, K.L. (2016). Tempo does not correlate with mode in the fossil record.
 730 *Evolution*, 70, 2678–2689. <https://doi.org/10.1111/evo.13090>

731 Voje, K. L., Starrfelt, J., & Liow, L. H. (2018). Model adequacy and
 732 microevolutionary explanations for stasis in the fossil record. *The American*
 733 *Naturalist*, 191, 509-523. <http://doi.org/10.1086/696265>

734 Yoder, J.B., Clancey, E., des Roches, S., Eastman, J.M., Gentry, L., Godsoe, W.,
 735 Hagey, T.J., Jochimsen, D., Oswald, B.P., Robertson, J., Sarver, B.A.J., Schenk,
 736 J.J., Spear, S.F. & Harmon, L.J. (2010). Ecological opportunity and the origin of
 737 adaptive radiations. *Journal of Evolutionary Biology*, 23, 1581–1596.
 738 <http://doi.org/10.1111/j.1558-5646.2010.01025.x>

739