How to describe and measure phenology? An investigation on the diversity of metrics using phenology of births in large herbivores

Lucie Thel[1], Simon Chamaillé-Jammes[2, 3, 4] & Christophe Bonenfant[1, 4]

[1] Laboratoire de Biométrie et Biologie Évolutive, Unité Mixte de Recherche 5558, Bâtiment 711, Université Lyon I, 43 Boulevard du 11 novembre 1918, F-69622 Villeurbanne Cedex, France.

[2] CEFE, Université Montpellier, CNRS, EPHE, IRD, Montpellier, France.

[3] Mammal Research Institute, Department of Zoology & Entomology, University of Pretoria, Pretoria, South Africa.

[4] LTSER France, Zone Atelier "Hwange", Hwange National Park, Bag 62, Dete, Zimbabwe.

Corresponding author: Lucie Thel, lucie.mc.thel@gmail.com.

Abstract

Proposed in 1849 by Charles Morren to depict periodical phenomena governed by seasons, the term "phenology" has spread in many fields, from evolutionary to molecular biology. The adoption of the concept of phenology in such diverse fields has been associated with the creation of a wide diversity of metrics, and we believe, some confusion about what they measure. Here, as an example and to offer some perspective, we studied phenology of births in large herbivores and we identified 52 metrics that were previously used to quantify phenology. We conducted a quantitative analysis of these metrics based on simulations. We compared and classified the metrics using a correlation matrix, and we evaluated the ability of each metric to capture the variation of a set of phenological characteristics via a sensitivity analysis. We finally scored each metric according to eight criteria we thought were important to describe phenology well. We highlight the fact that four characteristics of phenology are important and can be clearly defined: timing, synchrony, rhythmicity and regularity. We also report on the high correlations that can usually be observed among the metrics corresponding to each of the four characteristics of phenology, suggesting that such diversity of metrics is not useful. Additionally, we show that the best metrics are not necessarily the most commonly used and that simpler is often better: easy to implement metrics often got the best scores. We give some guidelines and advice, in order to facilitate repeatability and comparison of phenology between studies. We also discuss some pitfalls to avoid when choosing a metric. We trust our study can facilitate the use of well-defined and robust concepts and metrics in studies of phenology, for instance in the context of global change effects on populations.

Keywords: regularity, rhythmicity, seasonality, synchrony, timing, ungulate

Introduction

In 1849, Charles Morren coined the term "phenology" to describe how periodical phenomena such as plant growth and reproduction are governed by the course of seasons (Morren 1849, see also Demaree 2011). With his observations, he opened a new field of research, and almost two centuries later the concept of phenology has become a cornerstone of ecology (Begon et al. 1986), used in plant and animal ecology indifferently (Forrest and Miller-Rushing 2010). By describing when particular life-history events (*e.g.* flowering, parturition) occur in relation to the characteristics or states of the individual (*e.g.* size, age) as well as to external factors (*e.g.* photoperiod, predation risk) the concept of phenology is key to understanding the temporal cycles in the life cycle of species (Forrest and Miller-Rushing 2010). Nowadays with animals, the term phenology is commonly employed to describe the temporal occurrence of many aspects of a species' biology (*e.g.* moulting, migration, hibernation), but phenology of reproduction (*e.g.* Sinclair et la. 2000, Rubenstein and Wikelski 2003, VanDenHoff 2020) has attracted most interest. Reproductive phenology is an integral part of life history theory as it is at the heart of inter-generational trade-offs (*i.e.* between parents and offspring) and participates in the reproductive success and fitness of the individuals (Stearns 1989, Forrest and Miller-Rushing 2010). On the one hand, the time of the year when most births occur is often linked to seasonal variations in food resources so that the flush of food resources matches the energetic needs of breeding, which ultimately improves the reproductive success of parents and the fitness of offspring (Plard et al. 2015). On the other hand, the spread of birth dates in a year is supposed to reflect an anti-predator strategy to reduce the mortality associated with predation (Darling 1938, Gosling 1969), but also many other social and biological mechanisms (Ims 1990).

In most ecological study cases, measurements and observations of phenology are frequently performed at the population level by characterizing the temporal distribution of biological events (Visser et al. 2010). Those rather complex and variable patterns are then reduced to two main components: "timing", which is the date at which the event of interest occurs, and "synchrony",

which is the spread of the dates at which the event occurs, *i.e.* the variability between individuals (Fig. 1). Besides, stimulated by research on the effects of climate change on biodiversity (*e.g.* Crick and Sparks 1999, Parmesan 2007, Sarkar et al. 2019), the question of whether phenology is consistent or varies in time both at individual and population levels, in particular with climate and climate-driven resources, has received increased interest in recent years (*e.g.* Renaud et al. 2019). We therefore need to quantify underappreciated properties of phenology: the consistencies of the timing and synchrony of the events from one reproductive event to the next. As those characteristics of phenology are not described by specific words yet, we suggest to use "rhythmicity" and "regularity" to describe the consistency of timing and synchrony respectively (Fig. 1), in line with Newstrom's terminology (Newstrom et al. 1994).

Even if it is simple to grasp at first sight, the concept of phenology nevertheless carries a lot of confusion in literature, both from a semantic and a descriptive point of view (Visser et al. 2010). Past studies have explored phenology through the analysis of a huge diversity of mathematical descriptors, many of which remain specific to one single study. This is problematic because well-defined, comparable and reliable descriptors of the temporal distribution of biological events are key to achieving meaningful comparisons of phenology patterns within or across species. For instance, English and colleagues reassessed the most influential factors of reproductive synchrony in large herbivores from the existing literature, but they had to narrow their original data set because there was no standardized way of measuring and comparing synchrony across the studies (English et al. 2012).

This large diversity of metrics is further associated with a lack of generally accepted definitions of them or even divergent definitions of the same word (see "seasonality" *sensu* Skinner et al. 2002 and Heideman and Utzurrum 2003), which really hampers our ability to make meaningful comparisons (*e.g.* Ryan et al. 2007, Heldstab et al. 2018). Because experimental works are logistically challenging or virtually impossible to conduct with large species, the comparison of phenological patterns within a species living in contrasting environments, or across species (Clauss

et al. 2020) is therefore of major importance to assess the role of explanatory factors accounting for the marked variability in phenology. Such comparative approaches (*sensu* Felsenstein 1985) shed light on the ecological and evolutionary causes shaping the main stages of the life cycle of organisms (Bronson 1989).

In spite of the diversity of approaches to describe phenology, we found only a few attempts comparing phenology metrics to provide advice on which one should be used preferentially according to the context of the study (Moussus et al. 2010, Landler et al. 2018). Those initiatives are rare, and we currently lack a comprehensive comparison of metrics previously used to characterize phenology and assess the extent to which the different metrics actually capture the desired characteristics of the temporal distribution of events, nor do we have a proper assessment of the sensitivity of those metrics to actual changes in phenology.

Here, we propose such a comparison based on a literature survey of reproductive phenology in large herbivore species. We focus our overview on the taxonomic group of the large herbivores because it has been studied in many species and at many locations (Rutberg 1987). As a consequence, we expect to find a wide variety of patterns of births distributions, and a wide diversity of metrics associated to describe them in literature. As hinted above, we first clarify and formally define four main terms describing phenology, built on our knowledge of the literature: timing, synchrony, rhythmicity and regularity. We then conduct a comparative analysis of 52 metrics that have been used to quantify the different characteristics of phenology of births in large herbivores, highlight their strengths and weaknesses, and offer some guidance on which ones to use.

Materials and methods

We achieved a quantitative comparison of a wide range of different metrics used to analyse phenology by proceeding in six steps. In Step 1, we recorded all metrics employed to measure phenology in a selection of papers that we considered representative of the study of phenology of births in large herbivores. In Step 2, we simulated contrasting phenology from known parameters.

In step 3, we calculated all metrics on the simulated phenology to understand how they compare and what characteristics of phenology they actually measure. In Step 4, we explored the similarities between metrics from a correlation matrix, and identified categories of metrics capturing the same characteristic of phenology. In Step 5, we evaluated the sensitivity of each metric to changes in the estimated parameter. In Step 6, we ranked each metric based on eight criteria evaluating the quality of the relationship with the characteristic it measures (Table 1). Building on this work, we conclude by offering guidelines on the use of the existing metrics when studying phenology.

Step 1: Retrieving and coding the different phenology metrics

We opportunistically searched the literature for articles focusing on the distribution of births in large herbivores using keywords such as “phenology”, “timing”, “synchrony”, “seasonality”, “period” or “season”, and using various sources such as search engines and the references in previously found articles. From these articles, published between 1966 and 2019, we recorded all metrics used to describe phenology of births at the population level. We stopped our search once the rate at which we discovered new metrics with additional papers became negligible. We *a priori* classified each metric into one out of four categories, based on our understanding of the original description and formula of the metric (Fig. 1): (1) *timing* metrics, defining when within the year most births occur, (2) *synchrony* metrics, defining whether females tend to give birth at the same time in a population in a given year, (3) *rhythmicity* metrics, defining the consistency of timing between years, (4) *regularity* metrics, defining the consistency of synchrony between years.

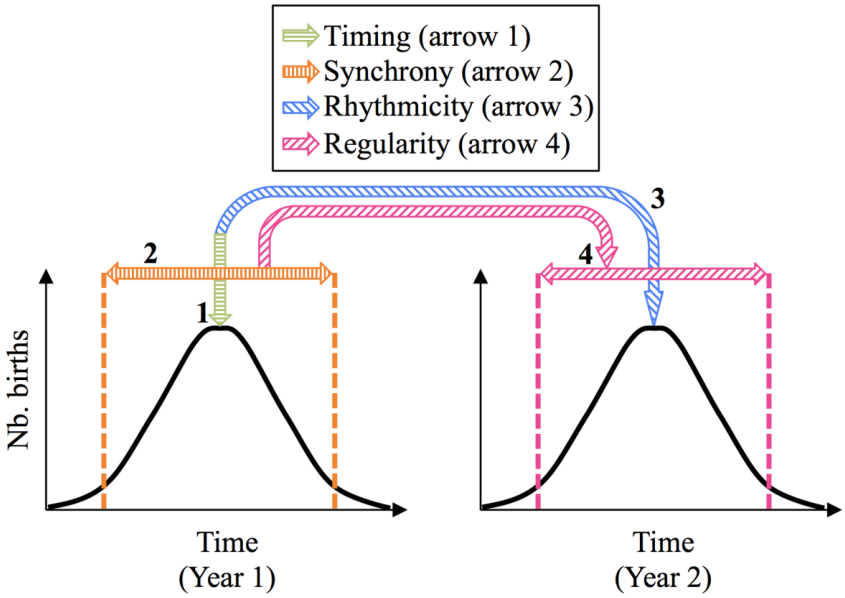

Figure 1: Four characteristics of phenology of births can be explored to fully describe phenology at the population scale: timing, synchrony, rhythmicity and regularity. Timing describes when within the year most births occur, synchrony illustrates whether females tend to give birth at the same time in a population in a given year, rhythmicity defines the consistency of timing between years, regularity refers to the consistency of synchrony between years. Green = timing metrics, orange = synchrony metrics, blue = rhythmicity metrics, pink = regularity metrics.

Forty-seven articles (Supporting information 1) presented at least one mathematically-described phenology metric yielding overall 52 different metrics. We could code all of them in R software (R Core Development Team 2019; except one in Perl, www.perl.org). In order to compare metrics quantitatively, we had to slightly adjust some of them (Supporting information 2).

Step 2: Simulating phenology of births

We simulated phenology of births with known characteristics to assess what characteristic of phenology (timing, synchrony, rhythmicity, regularity) each metric would capture, their sensitivity to change into these four key characteristics of interest, and the correlation between the 52 metrics. We simulated the distribution of births over a year as most large herbivores breed once a year. This choice does not limit the generality of our results because for species that would breed multiple times per year, the same metrics apply but on a more restricted time span.

Each simulated phenology was generated by randomly distributing births in time, following a normal distribution, n = 1,000 births approximately within a year of 365 days, replicated over 10 years (see why in Material and Methods - Step 3). We changed four parameters independently to modify the distribution of births: the mean day of birth for a given year (*mean*), the standard deviation of the distribution of births for a given year (*sd*), the range over which the mean birth date can vary across years (*Δmean*), and the range over which the standard deviation can vary across years (*Δsd*). Each parameter varied in a range from a minimum to a maximum value and was incremented with a constant step (Supporting information 3). By choosing the value of these parameters, we could simulate changes in the timing, synchrony, rhythmicity and regularity of phenology of births respectively, and independently. Because the simulated birth phenology relied on random draws (see above), the actual values of parameters in the simulated birth distribution could differ from the theoretical values used in the simulation algorithm. We used the realized values of the distribution parameters in the following analyses. We performed all simulations using the R software and made the code available on GitHub (see Data availability statement).

Step 3: Computing the phenology metrics from simulated phenology

Among the 52 phenology metrics coded, most applied to a single year, but others required two or more years of data to be computed. This was the case, for instance, for "*meanlin*", "*permean*" or "*watson*" metrics (see the complete list in Supporting information 2). Because we wanted to compute all 52 metrics for each individual simulation, we chose to simulate annual birth distributions over 10 consecutive years. For each simulation, we used data from the first year to compute metrics requiring only one year of data, data from the first two years for metrics requiring two years of data, and data from the whole simulation for the other metrics. Ten metrics were based on the analysis of several years, whereas nine compared phenology recorded for two years, and 33 were based on the analysis of one year of data.

Step 4: Comparing the metrics

From the results of Step 3, we computed the global correlation matrix between all pairs of metrics, using Pearson correlations. We then identified highly correlated metrics and to which of the four characteristics of phenology metrics were best related. We compared this categorization with our *a priori* classification of the metrics. This step enabled us to check our intuitive classification of the metrics in addition to reveal whether some metrics could incidentally capture several aspects of the distribution of births at once.

Step 5: Estimating the sensitivity of the metrics

For each metric, we performed a sensitivity analysis by quantifying the extent to which the metric varied with variations in the characteristic of phenology it was previously associated with in Step 4. We did so by computing, for each possible pair of simulations within the set of all simulations performed, the proportional difference between the realized values of the phenology parameter of interest of the two simulations, and the proportional difference between the values of the metric of interest of the same two simulations. In each case the proportional difference was calculated as $[(Value_{max} - Value_{min}) / Value_{min}] * 100$. This formulation allowed us to work with positive values only, as we were interested in the amplitude but not in the direction of the differences. We represented the scaled proportions of variation by centring and scaling the proportions. We did so for each metric individually, to prevent the representation of highly varying metrics from flattening the representation of sparsely varying ones in the heat map representation (Fig. 4).

Step 6: Scoring metrics

Finally, because there were too many different metrics, we could not discuss the pros and cons for all of them. Therefore, to provide guidance about the usefulness of the different metrics, we scored them according to a set of eight criteria that we considered as important behaviour for a metric to be optimal (Table 1). We ranked the metrics from 0 (not advisable) to 8 (strongly advised) according

to the number of criteria they fulfilled. Each criterion could be evaluated independently from the others, except from the first four that should all be met to consider that the metric describes the phenology characteristic *a minima*. The proposed criteria consisted in verifying if 1) the metric varied according to the phenology characteristic it was supposed to, 2) the variation of the metric according to the phenology characteristic was monotonous, 3) the relationship with the characteristic of phenology was strong (*i.e.* if the scatter plot was not too dispersed around the general trend of the relationship between the metric and the phenological characteristic, as an empirical approach of the predictive power), 4) the metric did not saturate within a biologically realistic range of distributions of births (*e.g.* if a synchrony metric returned the same value when all births occurred during a period of either one, two or three months for instance, it was considered to saturate within a biologically realistic range of birth distributions because such distributions of births can be found in the wild, such as in wildebeest, Estes 1976). If those four criteria were verified, we then evaluated an additional set of four criteria (normality, origin, linearity and unicity, see Table 1 for details). We considered that suitable metrics were characterized by scores > 6, and that metrics with scores ≤ 4 should be avoided.

Table 1: Ordered list of the criteria used to evaluate the quality of each metric. Each criterion can be individually fully (score of 1) or partially (score of 0.5) validated or no (score of 0) by each metric. The value for the first four criteria (in bold type) should be > 0 to consider a metric to be possibility worthwhile and evaluate the remaining criteria. The sum of the value obtained for each criterion gives the quality index of the metric (range between 0 and 8 points).

Criterion	Description	Score
goodness	Actually measures the parameter it is expected to measure	true = 1 false = 0
monotony	Varies monotonically with the value of the phenological parameter it is expected to measure	true = 1 false = 0
saturation	Does not saturate at the upper or lower boundary in a biological range of values	true = 1 false = 0
strength[1]	Is characterized by a strong relationship with the parameter it is expected to measure	strong relationship = 1 noisy relationship = 0.5 very noisy relationship = 0
normality[2]	Is not based on normality hypothesis	true = 1 better = 0.5 false = 0
origin	Does not depend on the temporal origin set by the investigator	true = 1 false = 0
linearity[3]	Is characterized by a linear relationship with the parameter it is expected to measure	type 1 = 1 type 2 and 3 = 0.5 type 4 = 0
unicity	Gives a unique result	true = 1 false = 0

[1] a strong relationship has a very small dispersion of points, a noisy relationship has a small dispersion of points that does not prevent from detecting a trend, a very noisy relationship has a dispersion of points large enough to prevent from detecting any trend. [2] better for metrics that does not explicitly need normal distribution but more efficient if verified. [3] type 1 is a linear relationship, type 2 is a relationship with two inflection points (i.e. sigmoid like function), type 3 is a relationship with one inflection point (e.g. square function), type 4 is a binary relationship.

Results

Overall we retrieved n = 47 papers from which we could extract 52 phenology metrics mathematically described by the authors. The mean number of metrics used in each paper was 3.8 ± 2.1 sd (*range* = 1-8). Eleven metrics were *a priori* associated with timing, 25 with synchrony, 10 with rhythmicity and five with regularity (we did not classify one metric because it was not possible to *a priori* discriminate between its affiliation to either rhythmicity or regularity metrics) (Fig. 2). Those metrics were based on descriptive statistics, circular statistics, statistical tests or statistical modelling such as linear regressions. The metrics were either date or duration, count information (*e.g.* a number of births), binary classification (*i.e.* if a given condition was satisfied or not), or unitless indices (Supporting information 2). We identified 48 metrics linked to the parameter they were supposed to measure according to our *a priori* classification. Among them, 8 metrics had a detectable relationship with at least one of the three remaining phenological parameters in addition to the relationship with the phenological parameter they were supposed to quantify (Supporting information 2 and 4).

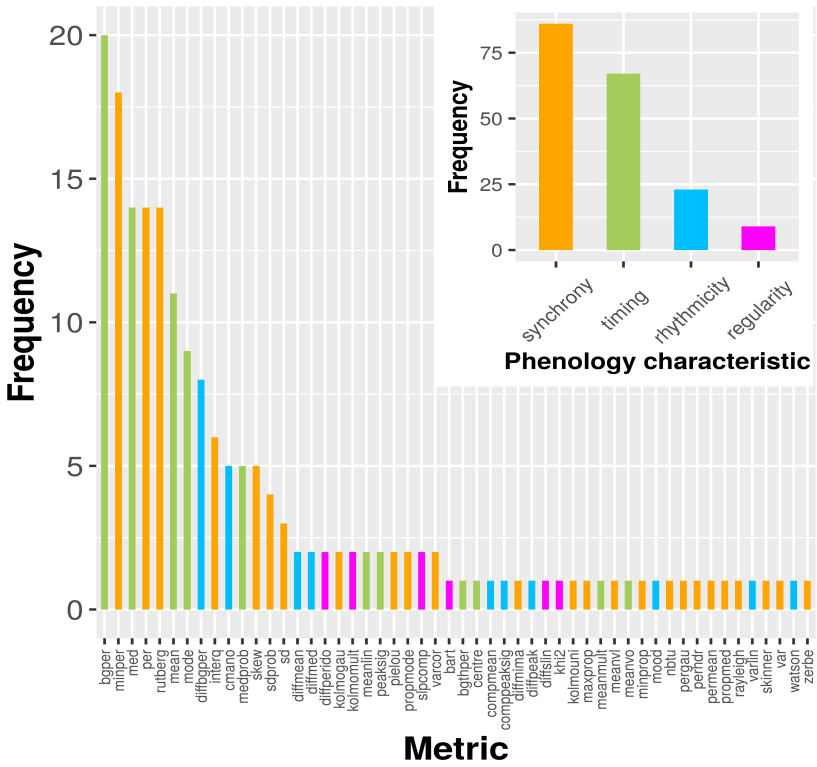

Figure 2: Frequency each metric and each characteristic of phenology was found to be studied in the reviewed articles (n = 47). Green = timing metrics, orange = synchrony metrics, blue = rhythmicity metrics, pink = regularity metrics.

The correlation matrix (Step 4) revealed obvious groups of metrics that were highly correlated and thus reflected the same characteristic of phenology (Fig. 3). Five groups were clearly identifiable, representing timing metrics (box 1), synchrony metrics (boxes 2 and 5), rhythmicity metrics (box 3) (note that the metric "*compmean*" metric is isolated on this representation, but negatively correlated with the other rhythmicity metrics), regularity metrics (box 4). The two groups of metrics measuring synchrony (boxes 2 and 5) had highly but negatively correlated values (box 6). This was indicative that all metrics of the two groups captured synchrony correctly, but in an opposite way. Three metrics had a particular behaviour and were isolated from the five groups, *i.e.* they were not clearly associated with any other metric ("*comppeaksig*", "*kolmogau*" and "*skew*" metrics). One metric seemed misclassified as it better correlated with a group of metrics measuring another phenology characteristic than the one it was supposed to measure *a priori* ("*slpcomp*" metric, which correlated better with synchrony than with regularity metrics). Another metric correlated well both with rhythmicity and regularity metrics ("*diffbgper*" metric).

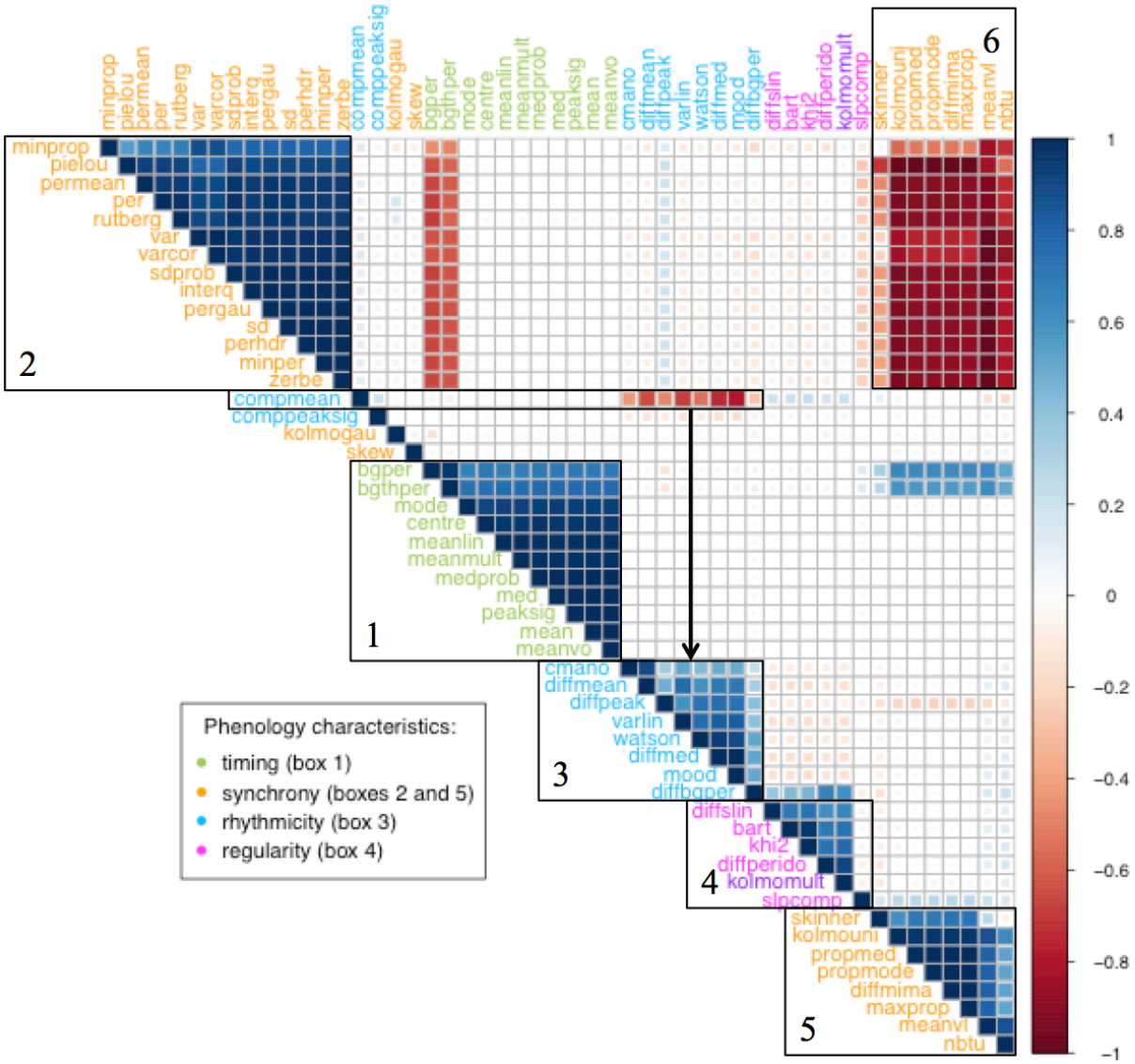

Figure 3: Correlation matrix between all pairs of metrics, using Pearson correlations (n = 51, "rayleigh" removed because no variation observed). It was not possible to classify "kolmomult" a priori in rhythmicity or regularity metrics, as it compares the complete distribution of births between two years. Box 6 highlights the high but negative correlation between the two groups of metrics measuring synchrony (boxes 2 and 5). Green = timing metrics, orange = synchrony metrics, blue = rhythmicity metrics, pink = regularity metrics.

The sensitivity of the metrics to the simulated variation of the phenology parameters (Step 5) differed markedly between metrics, especially in synchrony and regularity metrics (Fig. 4 and Supporting information 4). The scaled proportion of variation of the metrics according to the proportion of variation of the associated parameters ranged from -1.73 to 3.79 for timing metrics, from -3.81 to 2.01 for synchrony metrics, from -1.62 to 2.01 for rhythmicity metrics and from -1.60

to 1.91 for regularity metrics. The variation of all timing and rhythmicity metrics (except "*comppeaksig*" and "*compmean*" metrics), and most of the regularity metrics (except "*slpcomp*" metric) according to variations of the associated parameter (*mean*, *Δmean* and *Δsd* respectively) was highly homogeneous. Synchrony metrics were less homogeneous, certainly due to the fact that those metrics were the most numerous and based on more diverse methods (proportion of variation, integrative indexes or moments of the distribution of births, for instance). The metrics that were isolated in the correlation matrix were clearly visible in the heat maps, characterized by erratic or non-existent variations (*e.g.* "*skew*" metric).

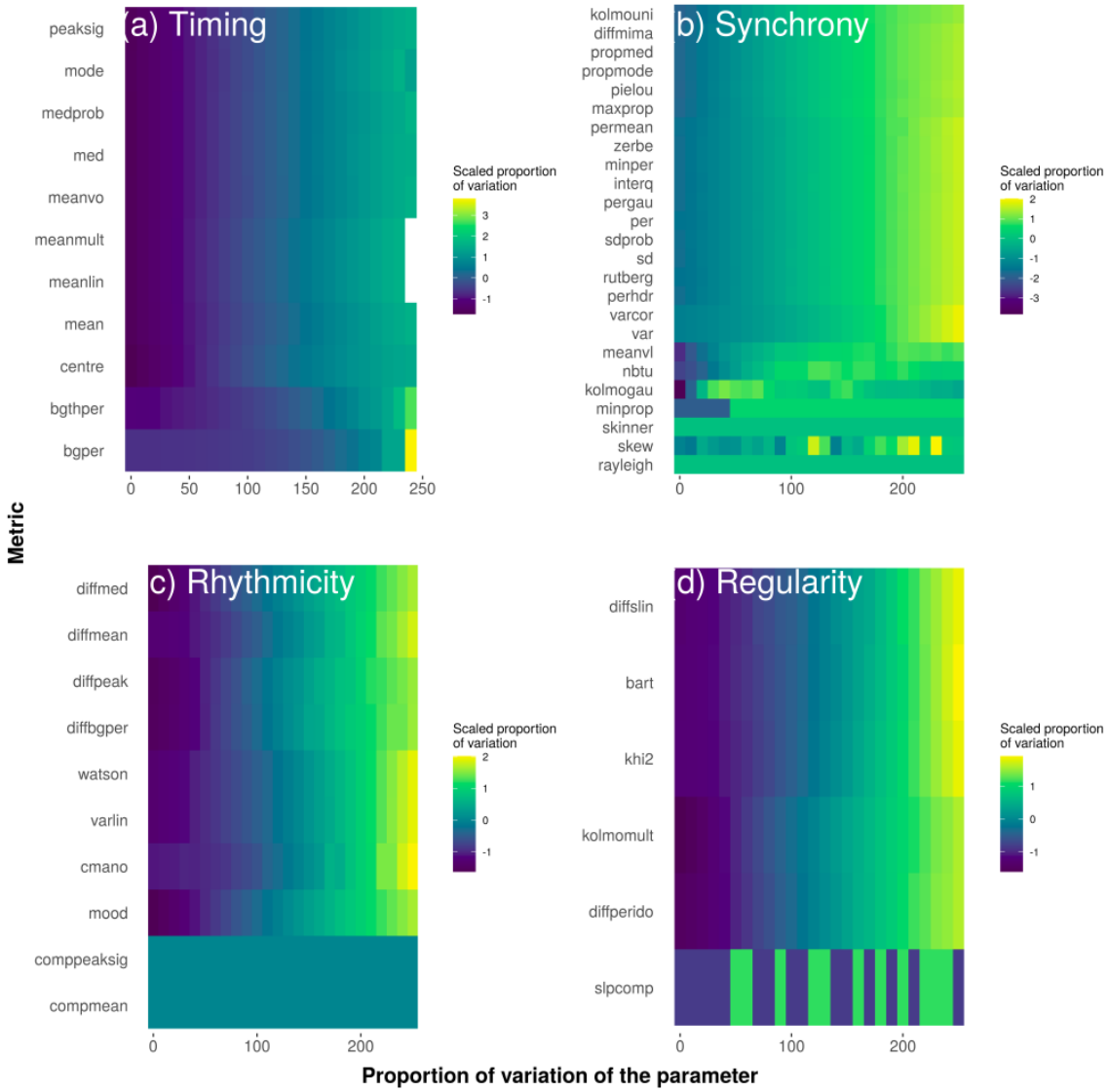

Figure 4: Heat maps representing the (scaled) proportion of variation of the metric in relation to the proportion of variation of the parameter of phenology associated: a) timing metrics according to the mean birth date for a given year (mean, n = 11), b) synchrony metrics according to the standard deviation of the distribution of births for a given year (sd, n = 25), c) rhythmicity metrics according to the range over which the mean birth date can vary across years (Δmean, n = 10), d) regularity metrics according to the range over which the standard deviation of the distribution of births can vary across years (Δsd, n = 6). Colours in the heat maps reflect the amplitude of variation of each metric, scaled within metric. This scaling was done to prevent the large variation of some metrics to obscure the smaller but meaningful variations of other metrics to be visible.

The quality score of the metrics (Step 6) varied between 0 and 7.5, while the possible range of variation was 0-8 (Supporting information 5). The best metrics to quantify timing *via* the analysis

291 of several years were based on the evaluation of the mean: "*meanlin*", which finds the mean birth
292 date thank to a linear model with random effects, and "*meanmult*" that calculates the mean of the
293 mean birth dates, with a score of 7 for both (Table 2). The metrics "*bgper*", "*bgthper*" (based on the
294 detection of the beginning of the birth period), "*centre*", "*mean*", "*med*" and "*meanvo*" (respectively
295 based on the estimation of the centre, mean, median and mean vector orientation of the distribution
296 of births) were equally good to quantify timing of births when only one year of data was available,
297 with also a score of 7 (Table 2). All metrics used in the description of timing got a score of at least
298 6, meaning that they were all satisfying. Three metrics stood out from the others to measure
299 synchrony using the analysis of only one year of data, all of them getting a score of 7.5. Those
300 metrics were "*pielou*" (corresponds to the evenness index), "*meanvl*" (calculates the mean vector
301 length), and "*kolmouni*" (compares the birth distribution to a uniform distribution). The metrics
302 "*interq*" (evaluates the period length gathering a certain percentage of births), "*sd*", "*var*", "*varcor*"
303 (based on the calculation of the variance of the distribution of births) and "*rutberg*" (evaluates the
304 period length gathering a certain percentage of births from first birth) were also successfully
305 measuring synchrony with a score of at least 6.5 (Table 2). The best metric to quantify rhythmicity
306 was the one evaluating the period length between the median birth dates of two years (namely
307 "*diffmed*", *score* = 7). The metrics "*diffbgper*" (evaluates the period length between the first birth
308 dates of two years), "*mood*" (compares the median birth dates between two years) and "*varlin*"
309 (calculates the variance of the distribution of births of several years thanks to a linear model with
310 random effects) also successfully measured rhythmicity with a score of at least 6.5 (Table 2). There
311 was a very good metric quantifying regularity according to our criteria, "*kolmomult*" (compares two
312 birth distributions), with a score of 7.5. The metric "*diffperiod*" (evaluates the period length
313 between two birth period durations), also successfully measured regularity with a score of 6.5
314 (Table 2). Our classification revealed what could be considered as ineffective (*score* = 0, n = 5) and
315 poor metrics (*score* ≤ 4, n = 13). All criteria were attributed visually by us to each metric and are
316 somewhat subjective. Hence, some metrics with a score lower than 6 could seem more appropriate

than others with a better score (*e.g.* "*zerbe*" metric is an improvement of "*rutberg*" metric but got a smaller score, Zerbe et al. 2012). That is why we invite the reader to consult Supporting information 2 and 4 to make his/her own opinion on the quality of each metric.

Table 2: List of the best metrics (according to the criteria defined in Table 1) to describe each characteristic of phenology of births (timing, synchrony, rhythmicity, regularity). The preferred metric for each characteristic of phenology of births for comparative purpose is reported in bold type.

	Best metrics (when one year available)	Best metrics (when several years available)
Timing	"*bgper*": first birth date "*bgthper*": first birth date when at least a certain percentage of births have occurred "*centre*": central date between first and last birth dates "*mean*": mean birth date "*med*": median birth date "***meanvo***": mean vector orientation for circular data	"*meanlin*": mean birth date *via* a linear model with random effects "*meanmult*": mean of the mean birth dates
Synchrony	"*pielou*": evenness index of distribution of births "***meanvl***": mean vector length for circular data "*kolmouni*": compare birth distribution to a uniform distribution "*interq*": period gathering a certain percentage of births based on quantiles "*sd*": standard deviation of distribution of births "*var*": variance of distribution of births "*varcor*": variance of distribution of births corrected by the Sheppard method "*rutberg*": shortest period gathering at least a certain percentage of births since first birth	/

Table 2: Continued.

Rhythmicity	Not applicable	"*diffmed*": period length between median birth dates of two years "*diffbgper*": period length between first birth dates of two years "***mood***": compare median birth dates of two years "*varlin*": variance between years *via* a linear model with random effects
Regularity	Not applicable	"***kolmomult***": compare distributions of births of two years "*diffperiod*": period length between birth period durations (between first and last birth dates) of two years

Discussion

With more than fifty metrics used to describe and analyse the distribution of births in large herbivores since 1966, our survey of the literature clearly illustrates the diversity of approaches used to study phenology of births. Although the choice of a metric is most of the time justified to answer a specific ecological question or on statistical grounds, the lack of consensual methods to quantify phenological patterns makes comparisons across species or populations difficult at best, if possible at all. Our simulation study suggests that such a diversity of metrics might be unnecessary, and confusing.

Many of the metrics we retrieved can be organized into four main categories, each one capturing a particular aspect of phenology: timing, synchrony, rhythmicity and regularity. Of course all metrics belonging to the same category are not perfectly equivalent and interchangeable (Fig. 3, see also a discussion comparing "*zerbe*" and "*rutberg*" metrics in Zerbe et al. 2012). For instance,

339 the correlations between timing metrics range between 0.68 and 1.00. The difference among metrics
340 is even more pronounced in the synchrony category with correlations ranging from 0.05 to 1.00
341 (when excluding "*kolmogau*" and "*skew*" metrics that are far from the other synchrony metrics in
342 the correlation matrix, Fig. 3). Despite the rather high average correlation within categories of
343 metrics visible in Fig. 3, some are definitely more closely associated with the simulated changes in
344 phenology of births and should be preferred whenever possible (Table 2). Overall, some phenology
345 characteristics have been more consistently measured across studies. This can be illustrated by the
346 number of metrics of each category used in more than two papers (Fig. 2): this was true for timing
347 (n = 5 metrics used more than two times) and synchrony (n = 7 metrics) metrics, but less true for
348 rhythmicity (n = 2 metrics) and even less for regularity (n = 0 metric) metrics. Also, we identified a
349 few metrics that we do not recommend to describe phenology of births because they are not
350 sensitive to changes in the phenology parameter that they are supposed to measure, or appear to
351 have undesirable statistical properties, such as a non-monotonous variation according to the
352 variation of the parameter they are supposed to measure, or a saturation in a range of biological
353 values.
354
355 *What is a good metric?*
356 The development of so many metrics to quantify phenology might have arisen from the need to
357 describe a large diversity of distributions of births across species and populations (Bronson 1989).
358 Despite some obvious redundancies between metrics revealed in this article, it could be difficult to
359 propose a set of universal metrics satisfying all the desired criteria and the sometimes varied goals
360 of studies. Here, we attempted to identify what metrics could be the most suitable for timing,
361 synchrony, regularity and rhythmicity by scoring them according to what we subjectively
362 considered as the main suitable properties. We consider for instance that a good metric should not
363 be restricted to one kind of pattern (*e.g.* unimodal), and we actually found that less than 25% of the
364 metrics (based on the metrics for which this criterion was evaluated, Supporting information 2)

theoretically require unimodal distribution to work well. This can be circumvented in the case of multimodal distributions by separating each reproductive period for independent analysis (*e.g.* Heideman and Utzurrum 2003 in bats). The metric should not be sensitive to the temporal origin set by the investigator, as each species (and even each population) has different favourable periods for its reproduction cycle (*e.g.* mountain sheep from desert and alpine habitat, Bunnell 1982). Using the calendar year would be biologically meaningless and might create artificial distribution patterns of births by splitting the distribution at the end of the year. Metrics that do not take into account the temporal origin of the data prevent such limitations and allow an objective comparison. To our knowledge, only five metrics ("*pielou*", "*mode*", "*meanvl*", "*kolmouni*" and "*kolmomult*") meet this assumption. Circular statistics could be prioritized to answer the difficulties linked to the selection of temporal origin, as it is frequently done in primate literature (*e.g.* DiBitetti and Janson 2000). In each category, we could find several metrics that meet the expectations of a good metric (Tab. 2).

Special considerations about consistency metrics, rhythmicity and regularity

In the literature, the term "seasonality" can be used to describe the position of births in the year (*i.e.* timing, *e.g.* in Sinclair et al. 2000), the duration of birth period (*i.e.* synchrony, *e.g.* in Zerbe et al. 2012), and even the fact that births occur at the same period of the year every year (*i.e.* rhythmicity and/or regularity, *e.g.* in Heideman and Utzurrum 2003). However, this term is also used to describe the cyclical nature of the environment (Wittemyer et al. 2007). That is why we suggest to use the term "seasonality" only to describe the cyclical nature of the environment and to prefer neutral terms such as those we introduced in this paper to describe phenology of births: rhythmicity and regularity.

If timing and synchrony of birth are the easiest components of phenology to quantify and compare, only a handful of metrics evaluates rhythmicity and regularity of the phenology of births across the years. It might reflect the difficulty in acquiring the needed data. Indeed, robust analysis of rhythmicity and regularity requires a large number of years of data, which may not be available

as phenological data collection is costly and time-consuming (Kharouba and Wolkovich 2020). Moreover, scientists are not interested in timing and synchrony consistency *per se*, but rather in the relationship between timing and synchrony and environmental factors. Instead of evaluating the consistency of timing and synchrony across the years, they rather evaluate the variation of timing and synchrony regarding the variation of their environmental drivers (*e.g.* Paoli et al. 2018). In any case, our study makes it clear that the currently available metrics are seldom and only moderately correlated, suggesting that capturing the temporal variation of phenology across years is difficult and requires thoughtful selection and interpretation of the metric used.

Biological dependence of the different characteristics of phenology

The way the different characteristics of phenology are measured can lead to dependency between one another, and this could explain the confusions noticed in literature. Indeed, several metrics we tested vary not only according to the distribution characteristic they are supposed to measure, but also according to other phenological characteristics (n = 8 metrics). For instance, we evidenced a strong correlation between "*bgper*" and "*bgthper*" metrics, which evaluate the start of the birth period (*i.e.* timing metrics), and the synchrony metrics. This unexpected association between different kinds of metrics is linked to the fact that, as the standard deviation of the distribution of births increases, all other parameters being equal, the first births occur earlier (Fig. 3).

Conclusion

In the context of a comparative study, we recommend using "*meanvo*" to describe timing as it is as efficient as any other timing metric with the strong advantage not to being influenced by temporal origin. We recommend using "*meanvl*" to describe synchrony because here again, this metric is among the best synchrony metrics and it is not influenced by temporal origin. We recommend using "*mood*" to describe rhythmicity even if it is not largely used, as it compares the median birth dates, a better descriptor of phenology of births than the first birth more commonly used. We recommend

using "*kolmomult*" to describe regularity, as it is the only regularity metric to meet almost perfectly all of our criteria and it is based on a non-parametric test. All those metrics are easy to implement and do not have strong prerequisites on the structure of the input data. It is worth noting that despite their good properties, these metrics are not the most commonly used in literature. We hope our work will facilitate considerations of these metrics for future studies, and more generally pave the way towards more robust and comparable analyses of phenology.

Declarations

Acknowledgments: this work was performed using the computing facilities of the CC LBBE/PRABI and the IFB Cloud. The authors thank Simon Penel for his help with Perl Programming Language, and Marcus Clauss for providing the code of one of the metrics. We thank Marion Valeix for comments on previous drafts of the manuscript, and Primaëlle Fusto for proofreading.

Funding: This work was supported by a grant from the "Ministère français de l'Enseignement supérieur, de la Recherche et de l'Innovation" through the "Ecole Doctorale E2M2" of "Université Claude Bernard Lyon 1".

Authors' contributions: LT, CB and SCJ conceived the ideas and designed the methodology; LT performed data collection and analysed the data; LT wrote the first draft of the manuscript, which was then edited by all authors. All authors gave final approval for publication.

Data availability statement: code available on GitHub: https://github.com/LucieThel/phenology-metrics.

References

Begon, M. et al. 1986. Ecology. Individuals, populations and communities. - Blackwell scientific publications.

Bronson, F. H. 1989. Mammalian reproductive biology. - University of Chicago Press.

Bunnell, F. L. 1982. The lambing period of mountain sheep: synthesis, hypotheses, and tests. - Can. J. Zool. 60: 1–14.

Clauss, M. et al. 2020. Basic considerations on seasonal breeding in mammals including their testing by comparing natural habitats and zoos. - Mamm. Biol.: 1–14.

Crick, H. Q. P. and Sparks, T. H. 1999. Climate change related to egg-laying trends. - Nature 399: 423.

Darling, F. 1938. Bird flocks and the breeding cycle. - Cambridge University Press.

Demarée, G. R. 2011. From “Periodical Observations” to “Anthochronology” and “Phenology”- the scientific debate between Adolphe Quetelet and Charles Morren on the origin of the word “Phenology.” - Int. J. Biometeorol. 55: 753–761.

Di Bitetti, M. S. and Janson, C. H. 2000. When will the stork arrive? Patterns of birth seasonality in neotropical primates. - Am. J. Primatol. Off. J. Am. Soc. Primatol. 50: 109–130.

English, A. K. et al. 2012. Reassessing the determinants of breeding synchrony in ungulates. - PLoS One 7: e41444.

Estes, R. D. 1976. The significance of breeding synchrony in the wildebeest. - Afr. J. Ecol. 14: 135–152.

Felsenstein, J. 1985. Phylogenies and the comparative method. - Am. Nat. 125: 1–15.

Forrest, J. and Miller-Rushing, A. J. 2010. Toward a synthetic understanding of the role of phenology in ecology and evolution. - Philos. Trans. R. Soc. B Biol. Sci. 365: 3101–3112.

Gosling, L. M. 1969. Parturition and related behaviour in Coke’s hartebeest, Alcelaphus buselaphus cokei Günther. - J. Reprod. Fertil. Suppl. 6: 265–286.

463 Heideman, P. D. and Utzurrum, R. C. B. 2003. Seasonality and synchrony of reproduction in three
464 species of nectarivorous Philippines bats. - BMC Ecol. 3: 11.
465 Heldstab, S. A. et al. 2018. Geographical Origin, Delayed Implantation, and Induced Ovulation
466 Explain Reproductive Seasonality in the Carnivora. - J. Biol. Rhythms 33: 402–419.
467 Ims, R. A. 1990. The ecology and evolution of reproductive synchrony. - Trends Ecol. Evol. 5:
468 135–140.
469 Kharouba, H. M. and Wolkovich, E. M. 2020. Disconnects between ecological theory and data in
470 phenological mismatch research. - Nat. Clim. Chang.: 1–10.
471 Landler, L. et al. 2018. Circular data in biology: advice for effectively implementing statistical
472 procedures. - Behav. Ecol. Sociobiol. 72: 128.
473 Morren, C. 1849. Le globe, le temps et la vie. - Bull. l'Académie R. des Sci. des Lettres des B.-art.
474 Belgique 16: 660–684.
475 Moussus, J.-P. et al. 2010. Featuring 10 phenological estimators using simulated data. - Methods
476 Ecol. Evol. 1: 140–150.
477 Newstrom, L. E. et al. 1994. A new classification for plant phenology based on flowering patterns
478 in lowland tropical rain forest trees at La Selva, Costa Rica. - Biotropica: 141–159.
479 Paoli, A. et al. 2018. Winter and spring climatic conditions influence timing and synchrony of
480 calving in reindeer. - PLoS One 13: e0195603.
481 Parmesan, C. 2007. Influences of species, latitudes and methodologies on estimates of phenological
482 response to global warming. - Glob. Chang. Biol. 13: 1860–1872.
483 Plard, F. et al. 2015. The influence of birth date via body mass on individual fitness in a long-lived
484 mammal. - Ecology 96: 1516–1528.
485 Renaud, L.-A. et al. 2019. Phenotypic plasticity in bighorn sheep reproductive phenology: from
486 individual to population. - Behav. Ecol. Sociobiol. 73: 50.
487 Rubenstein, D. R. and Wikelski, M. 2003. Seasonal changes in food quality: a proximate cue for
488 reproductive timing in marine iguanas. - Ecology 84: 3013–3023.

Rutberg, A. T. 1987. Adaptive hypotheses of birth synchrony in ruminants: an interspecific test. - Am. Nat. 130: 692–710.

Ryan, S. J. et al. 2007. Ecological cues, gestation length, and birth timing in African buffalo (Syncerus caffer). - Behav. Ecol. 18: 635–644.

Sarkar, U. K. et al. 2019. Climato-environmental influence on breeding phenology of native catfishes in River Ganga and modeling species response to climatic variability for their conservation. - Int. J. Biometeorol. 63: 991–1004.

Sinclair, A. R. E. et al. 2000. What determines phenology and synchrony of ungulate breeding in Serengeti? - Ecology 81: 2100–2111.

Skinner, J. D. et al. 2002. Inherent seasonality in the breeding seasons of African mammals: evidence from captive breeding. - Trans. R. Soc. South Africa 57: 25–34.

Stearns, S. C. 1989. Trade-offs in life-history evolution. - Funct. Ecol. 3: 259–268.

van den Hoff, J. 2020. Environmental constraints on the breeding phenology of Giant Petrels Macronectes spp., with emphasis on Southern Giant Petrels M. giganteus. - Mar. Ornithol. 48: 33–40.

Visser, M. E. et al. 2010. Phenology, seasonal timing and circannual rhythms: towards a unified framework. - Philos. Trans. R. Soc. B Biol. Sci. 365: 3113–3127.

Wittemyer, G. et al. 2007. Breeding phenology in relation to NDVI variability in free-ranging African elephant. - Ecography (Cop.). 30: 42–50.

Zerbe, P. et al. 2012. Reproductive seasonality in captive wild ruminants: implications for biogeographical adaptation, photoperiodic control, and life history. - Biol. Rev. 87: 965–990.