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]

Authors details

[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, Univ Montpellier, CNRS, EPHE, IRD, Univ Paul Valéry Montpellier 3, 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.

Correspondence 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 of biology, from evolutionary to molecular biology. The wide adoption of the concept of phenology has been associated with the publication of a large number of metrics, with sometimes confusion about what they measure. Here, we aimed to link these metrics to four characteristics of phenology: timing, synchrony, rhythmicity and regularity. From 52 published metrics previously used to quantify phenology of births in large herbivores, we conducted a quantitative analysis based on simulations of normal and non-normal distributions of births (both intra and inter-annual variations). We first compared and classified the metrics using a correlation matrix. Then, we evaluated the ability of each metric to capture the variation of the four phenology characteristics via a sensitivity analysis. We finally scored each metric according to eight criteria we thought were important to describe phenology correctly. The high correlation we found among the metrics corresponding to each of the four characteristics of phenology suggests that such diversity of metrics is all but useful. Additionally, we show that the best metrics are not necessarily the most commonly used, and that simpler is often better. Circular statistics seem particularly suitable to describe the timing and synchrony of births in a wide range of distribution patterns. The level of repeatability of both timing and synchrony is still poorly described, but general non-parametric tests such as Mood and Kolmogorov tests allow a first and easy quantification of this variability across the years. We provide some guidelines and advice to facilitate repeatability and comparison of phenology between studies. We trust our study can facilitate the use of well-defined and robust concepts and metrics in studies of phenology, in the context of climate change for instance.

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 Demarée 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 the term phenology is commonly employed to describe the temporal occurrence of many aspects of a species' biology (*e.g.* molting, migration, diapause in animals), but the phenology of reproduction (*e.g.* Sinclair et la. 2000, Rubenstein and Wikelski 2003, van den Hoff 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 is a key factor of 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 anti-predator strategies to reduce the mortality associated with predation (Darling 1938, Gosling 1969), but also many other social and biological mechanisms (Ims 1990), such as the avoidance of male harassment undergone by the females (Boness et al. 1995) or intra-specific competition between offspring (Hodge et al. 2011).

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 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 (at the population scale) of the events from one reproductive season 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 coined for tropical plants (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 remaining 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 phenology patterns within a species living in contrasting environments, or across species (Clauss et

al. 2020) is of major importance to assess the role of explanatory factors accounting for the often marked variability in phenology reported in empirical studies (Rutberg 1987). 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 of metrics 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, and a wide diversity of metrics associated to describe them in literature. As hinted above, we first clarified and formally defined four main terms describing phenology, built on our knowledge of the literature: timing, synchrony, rhythmicity and regularity. We then conducted a comparative analysis of 52 metrics that have been used to quantify the different characteristics of phenology of births in large herbivores, highlighting their strengths and weaknesses, and offering some guidance on which ones to use preferentially.

Materials and methods

We achieved a quantitative comparison of a wide range of metrics used to analyse phenology 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 by varying independently the four parameters that determine

timing, synchrony, rhythmicity and regularity of phenology of births (see details below). 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 that we considered important to identify robust and efficient metrics, but also meaningful from an ecological point of view (see Table 1 for a description of each criteria).

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 the 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. 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 initially used to describe the cyclical nature of the environment in a wider range than the study of birth phenology

(Visser et al. 2010). Thus, it should be used to describe organisms' phenology only when a direct relationship between periodic environmental phenomena and the cycle of the organism at stake has been demonstrated, which is not always the case in phenology studies. That is why we suggest to use the term "seasonality" only to describe the cyclicity of the environment and to prefer neutral terms such as those we introduced in this paper to describe phenology of births: rhythmicity and regularity.

Forty-seven articles (Supporting information 1) presented at least one mathematically-defined 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 tweak some of them (Supporting information 2).

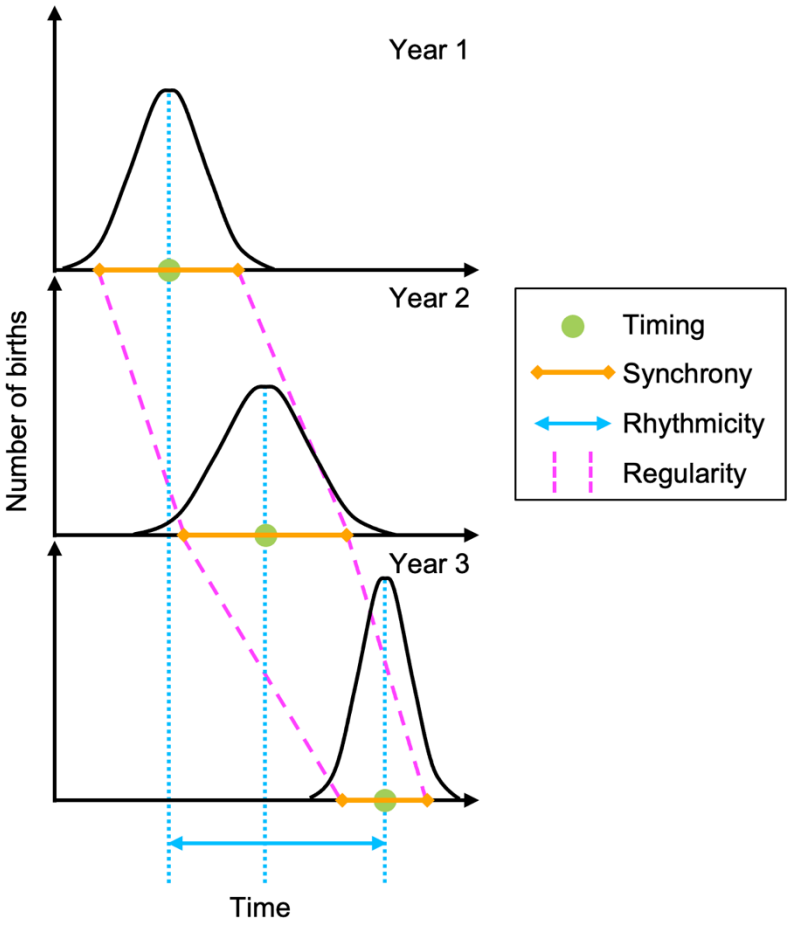

Figure 1: Four characteristics of phenology of births can be explored to fully describe phenology at the population scale: timing, synchrony, rhythmicity and regularity. Timing describes when within the year most births occur, synchrony illustrates whether females tend to give birth at the same time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

Step 2: Simulating phenology of births

We simulated phenology of births from statistical distributions with known parameters (Supporting information 3) to assess what characteristic of phenology (timing, synchrony, rhythmicity, regularity) each metric would capture, their sensitivity to changes into these four key characteristics of interest, and the correlation between the 52 metrics. We simulated the distributions of births over a year as most large herbivores breed once a year. This choice does not limit the generality of our results: for species breeding more than once per year (*e.g.* small species with short gestation length such as dikdik *Rynchotragus (Madoqua) kirki*, Sinclair et al. 2000), the same metrics may be applied on sub-periods of time, each displaying only one peak of births.

Each simulated phenology was generated by randomly distributing births in time, following a normal distribution. We distributed n = 1,000 births within a year of 365 days, repeated 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 the phenology of births respectively, and independently. Because the simulated phenology of births relied on random draws, the actual values of parameters in the simulated distribution of births could differ from the theoretical values used in the simulation algorithm. We used the realized values of the distribution parameters in the following analyses. Note that we replicated the same analyses using non-normal distributions of births (*i.e.* skewed normal, bimodal, Cauchy, and random distribution) to cover the variety of empirical distributions of births observed in *natura* (Supporting information 4).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 patterns of births

Among the 52 phenology metrics we coded, most applied to a single year, but others required two or more years of data to be computed (see the complete list in Supporting information 2). Because we wanted to compute all 52 metrics, we chose to simulate annual distributions of births over 10 consecutive years by default. For each simulation, we used data from the first year to compute metrics requiring only one year of data (n = 33 metrics), data from the first two years for metrics requiring two years of data (n = 9 metrics), and data from the whole simulation for the other metrics (n = 10 metrics).

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 groups of strongly correlated metrics from the pairwise correlation coefficients and assigned each metrics to one or several of the four characteristics of phenology it was 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 observed variation of each metric with a fixed variation 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 centering and scaling the proportions. We standardized the sensitivity of each metric individually to prevent the representation of highly varying metrics from flattening the representation of sparsely varying ones in the heat map representation.

Step 6: Scoring metrics

Finally, because there were too many different metrics, we could not discuss the pros and cons for all of them. We chose instead to provide guidance about the usefulness of the different metrics by scoring them according to a set of eight criteria that we considered as important behaviour for a metric to be relevant (Table 1). We ranked the metrics from 0 (not advised) 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 (Table 1) consisted in verifying if 1) the metric varied according to the phenology characteristic it was supposed to measure, 2) the variation of the metric according to the phenology characteristic was monotonous, 3) the relationship with the characteristic of phenology was strong (we evaluated the correlation visually as numerous relationships were not linear), 4) the metric did not saturate within a biologically realistic range of distributions of births. If those four criteria were satisfied, we then evaluated an additional set of four criteria (see Table 1 for details). We considered that suitable metrics were characterized by scores > 6, meaning the first four essential criteria are validated and at least two additional criteria are satisfied. Conversely metrics with scores ≤ 4 (for which the first four essential criteria are not validated) should not be advised.

Table 1: Ordered list of the criteria used to evaluate the relevance of each metric describing phenology of births. 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 possibly worthwhile and evaluate the remaining criteria. The sum of the value obtained for each criterion gives the relevance 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 phenology parameter it is expected to measure, *i.e.* the sign of the slope coefficient is constant	true = 1 false = 0
saturation	Does not saturate at the upper or lower boundary in a biological range of values (*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)	true = 1 false = 0
strength[1]	Is characterized by a strong relationship with the parameter it is expected to measure, *i.e.* is the scatter plot not too dispersed around the general trend of the relationship between the metric and the phenology characteristic, as an empirical approach of the predictive power	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, whatever the shape of the relationship (linear, but also logistic or quadratic for instance). [2] better for metrics that does not explicitly need normal distribution but more efficient if satisfied. [3] type 1 is a linear relationship, type 2 is a relationship with two inflexion points (i.e. sigmoid like function), type 3 is a relationship with one inflexion point (e.g. quadratic function), type 4 is a binary relationship.

Results

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 could be a rhythmicity or regularity metric *a priori*). Those metrics were based on descriptive statistics, circular statistics, statistical tests or statistical modeling such as general linear models. 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 phenology characteristics in addition to the relationship with the nominal phenology characteristic they were supposed to quantify (Supporting information 2 and 5).

The correlation matrix (Step 4) revealed obvious groups of metrics that were highly correlated and thus reflected the same characteristic of phenology (Fig. 2). Five groups were clearly identifiable, representing timing metrics (box 1), synchrony metrics (boxes 2 and 5), rhythmicity metrics (box 3), and 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. One metric, "*splcomp*" seemed misclassified *a priori* as it correlated better with synchrony than with regularity metrics. Another metric correlated well both with rhythmicity and regularity metrics ("*diffbgper*").

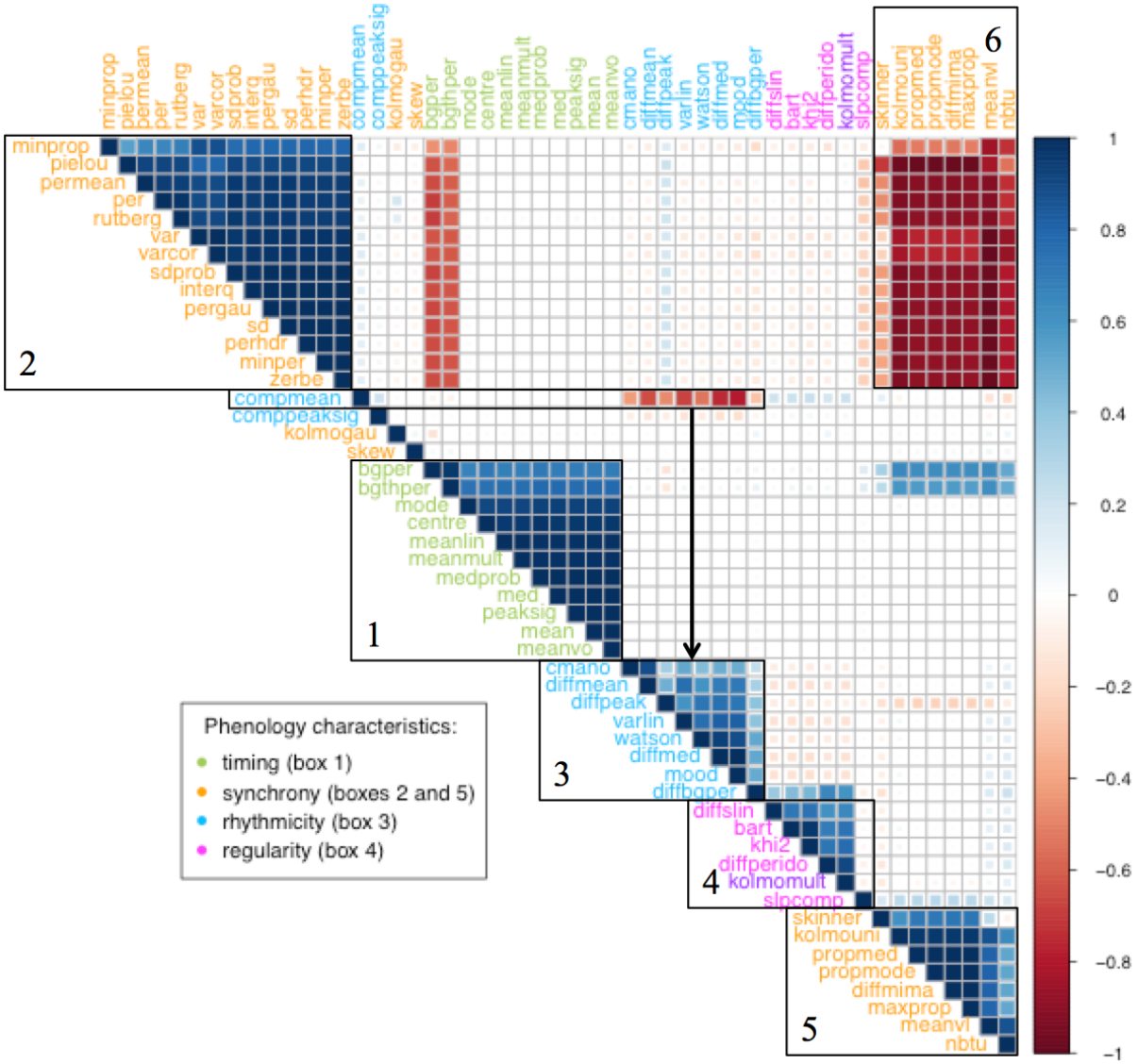

Figure 2: Correlation matrix between all pairs of metrics, using Pearson correlations (n = 51, "rayleigh" removed because no variation observed). It was not possible to classify "kolmomult" a priori in rhythmicity or regularity metrics, as it compares the complete distribution of births between two years. 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 characteristics (Step 5) differed markedly between metrics, especially in synchrony and regularity metrics (Fig. 3 and Supporting information 5). 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 almost all timing, rhythmicity and regularity metrics

according to variations of their associated parameter 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 nonexistent variations (*e.g.* "*skew*", "*compmean*").

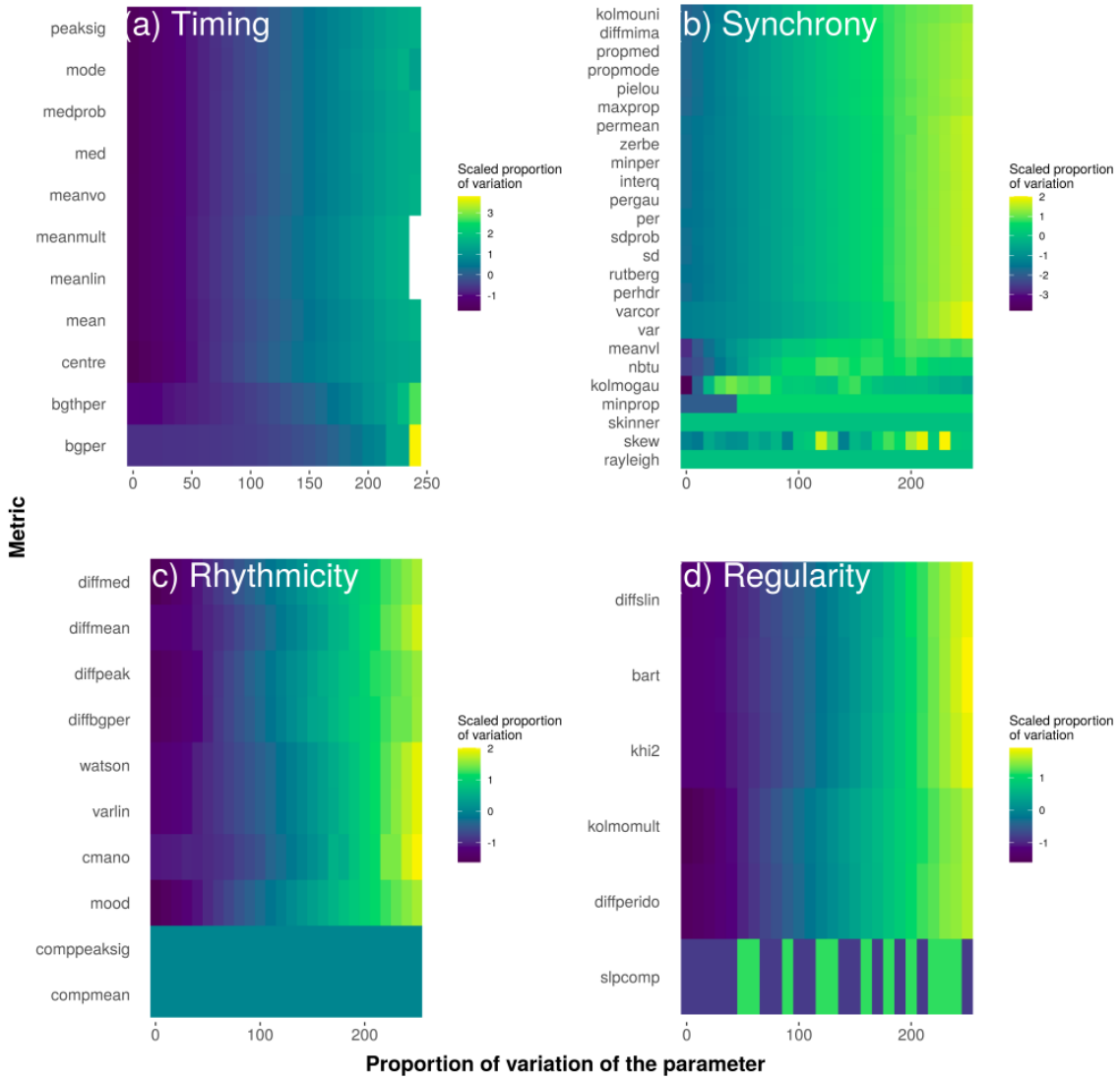

Figure 3: Heat maps representing the (scaled) proportion of variation of the metric in relation to the proportion of variation of the parameter of phenology associated: a) timing metrics according to the mean birth date for a given year (mean, n = 11), b) synchrony metrics according to the standard deviation 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). Colors 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 same analyses conducted on the basis of non-normal distributions led to similar observations in the case of asymmetric distributions (skewed normal, bimodal and Cauchy distributions). The correlation matrices showed similar patterns of correlations between the metrics, and the metrics varied analogously according to the variation of the *mean*, *sd*, *Δmean* and *Δsd* of the distributions for normal and asymmetric distributions either (see Supporting information 4 for a detailed analysis). Nevertheless, it is worth noting that a very limited number of metrics depending on the skewness of the distribution did not perform as well with the normal distribution than with asymmetric distributions. On the opposite, metrics depending on the presence of a period without any birth did not perform as well with non-normal distribution than with a normal distribution. In the case of a random distribution, no clear correlations between metrics nor relationships between the metrics and the four parameters of the distribution were detectable, except for some rare synchrony and timing metrics (Supporting information 4).

The relevance score of the metrics (Step 6) varied between 0 and 7.5, while the possible range of variation was 0-8 (Supporting information 6). The best metrics to quantify timing *via* the analysis of several years reached a score of 7 and were based on the evaluation of the mean: "*meanlin*" which finds the mean birth date thank to a linear model with random effects, and "*meanmult*" which calculates the mean of the mean birth dates (Table 2). The metrics "*bgper*", "*bgthper*" (based on the detection of the beginning of the birth period), "*centre*", "*mean*", "*med*" and "*meanvo*" (respectively based on the estimation of the centre, mean, median and mean vector orientation of the distribution of births) were equally good to quantify timing of births when only one year of data was available, scoring 7 too (Table 2). All metrics used in the description of timing got a score of at least 6, meaning that they were all satisfying. Three metrics stood out from the others to measure synchrony from one year of data only, all of them getting a score of 7.5. Those metrics were "*pielou*" (corresponds to the evenness index), "*meanvl*" (which calculates the mean vector length), and "*kolmouni*" (which compares the distribution of births to a uniform distribution). The metrics "*interq*" (which evaluates the period length gathering a certain percentage of births),

"*sd*", "*var*", "*varcor*" (based on the calculation of the variance of the distribution of births) and "*rutberg*" (which evaluates the period length gathering a certain percentage of births from first birth) were also successfully measuring synchrony with a score of at least 6.5 (Table 2). The best metric to quantify rhythmicity was "*diffmed*", evaluating the period length between the median birth dates of two years (*score* = 7). The metrics "*diffbgper*" (which evaluates the period length between the first birth dates of two years), "*mood*" (which compares the median birth dates between two years) and "*varlin*" (which calculates the variance of the distribution of births of several years thanks to a linear model with random effects) also successfully measured rhythmicity with a score of at least 6.5 (Table 2). There was a very good metric quantifying regularity according to our criteria: "*kolmomult*", which compares two birth distributions (*score* = 7.5). The metric "*diffperiod*" (which evaluates the period length between two birth period durations), also successfully measured regularity with a score of 6.5 (Table 2). Our classification revealed what could be considered as ineffective (*score* = 0, n = 5) and poor metrics (*score* ≤ 4, n = 13). All criteria were attributed visually by one of us (LT) to each metric. However, the subjectivity was reduced by the use of systematic criteria described at length in Table 1. We invite the reader to consult Supporting information 2 and 5 to make his/her own opinion on the relevance of each metric for his/her own work.

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 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* linear model with random effects "*meanmult*": mean of 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	/
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

317

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, even when focusing on a specific taxonomic group. 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 phenology 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 confusing and unnecessary because we could identify a reduced set of simple metrics that work well to measure the different characteristics of phenology.

Many of the metrics we retrieved can be organised into four main categories, each one capturing a particular characteristic of phenology: timing, synchrony, rhythmicity and regularity. Of course, metrics belonging to the same category are not perfectly equivalent and interchangeable (Fig. 2, see also a discussion comparing "*zerbe*" and "*rutberg*" metrics in Zerbe et al. 2012). For instance, the correlations between timing metrics range between 0.68 and 1.00. The difference among metrics is more pronounced in the synchrony category with correlations ranging from 0.05 to 1.00 (excluding "*kolmogau*" and "*skew*" metrics that appear as singularities in the correlation matrix, Fig. 2). 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 between timing and synchrony through terms such as "period" or "season" of births. Indeed, several metrics we tested vary not only according to the phenology characteristic they were supposed to measure, but also according to other characteristics of the phenology (n = 8 metrics). For instance, we show 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 arises because the standard deviation of the distribution of births increases, all other parameters being equal, the first births occur earlier (Fig. 2).

343 *We* attempted to identify what metrics could be the most suitable for measuring timing,
344 synchrony, regularity and rhythmicity by scoring them according to what we subjectively
345 considered as the main suitable properties. We considered for instance that a good metric should not
346 be restricted to one kind of pattern (*e.g.* unimodal), and we actually found that less than 25% of the
347 metrics (based on the metrics for which this criterion was evaluated, Supporting information 2)
348 theoretically require unimodal distribution to work well. A workaround for multimodal distributions
349 of births is to separate each reproductive period and conduct independent analyses (*e.g.* Heideman
350 and Utzurrum 2003 in bats). Besides, we illustrated that only a limited number of metrics really
351 changed their behaviour when applied to non-normal distributions (Supporting information 4). The
352 metric should also not be sensitive to the temporal origin set by the investigator, as each species
353 (and even each population) has different favorable periods for its reproduction cycle (*e.g.* mountain
354 sheep from desert and alpine habitat, Bunnell 1982). Using the calendar year would be biologically
355 meaningless and might create artificial patterns of births by splitting the distribution at the end of
356 the year. Metrics that do not take into account the temporal origin of the data prevent such
357 limitations and allow an objective comparison. To our knowledge, only five metrics ("*pielou*",
358 "*mode*", "*meanvl*", "*kolmouni*" and "*kolmomult*") meet this assumption. Circular statistics could be
359 prioritized to answer the difficulties linked to the selection of temporal origin, as it is frequently
360 done in primate literature (*e.g.* DiBitetti and Janson 2000, see also MacKay et al. 2018 in Eastern
361 grey kangaroo *Macropus giganteus*). Notwithstanding such limitations, we could find several
362 metrics that meet our expectations of a good metric for each phenology characteristic (Table 2).
363 Conversely, a few metrics should not be recommend to describe phenology of births like "*diffmean*"
364 (which evaluates the slope of the linear regression describing the evolution of the mean dates of
365 births of several years) or "*nbtu*" (which identifies the duration of the period gathering at least a
366 certain percent of births). Those latter metrics do not capture the changes in the phenology
367 parameter that they are supposed to measure, or present undesirable statistical properties. For
368 instance, "*nbtu*" varied non-monotonously with the level of synchrony of the birth phenology.

Another one like "*per*" (which identifies the duration between first and last birth) plateaued for a range of biologically realistic values limiting its use in a wide range of ecological conditions (Supporting information 5).

Overall, some phenology characteristics have been more consistently evaluated across studies, a fact illustrated by the number of metrics of each category used in more than two papers (Supporting information 7). If timing and synchrony of births are the easiest and most frequent characteristics of phenology estimated and compared, only a handful of metrics evaluates rhythmicity and regularity of the phenology of births across the years, and even less are used more than twice (n = 2 and 0 metrics respectively). This gap might reflect the difficulty in acquiring the adequate data. Indeed, robust analysis of rhythmicity and regularity requires many years of data, which may not be available as phenology data collection is costly and time-consuming (Kharouba and Wolkovich 2020). Moreover, scientists are less interested in timing and synchrony consistency *per se* than in the relationship between timing and synchrony, and ecological or environmental factors. Instead of measuring the consistency of timing and synchrony *per se* across the years, they assess how the timing and synchrony changes with temperature, rainfall or spring snow cover for instance (Paoli et al. 2018). In any case, our study makes it clear that the currently available metrics are seldom and only moderately correlated, particularly in the context on non-normal distributions of births,(Supporting information 4). Capturing the temporal variation of phenology across years appears difficult and requires thoughtful selection and interpretation of the used metric. In the context of climate change, there is an urgent need to develop statistical tools for measuring regularity and rythmicity, to test the hypothetical responses of phenology to changing climate and predation regimes, for instance. Cosinor and Fourier analyses can help detecting long-term trends and the phenology departure from such trends for example, but require a large amount of data (Cancho-Candela et al. 2007).

The past development of so many metrics to quantify phenology likely arose from the need to describe the large diversity of birth distribution documented across species and populations

395 (Bronson 1989). For instance, the wildebeest *Connochaetes taurinus* have very synchronous and
396 normal-like distribution of births (Estes 1976). At the other extreme, the distribution of births of
397 giraffe *Giraffa camelopardalis* is quite unpredictable and spread out over the course of the year
398 (Sinclair et al. 2000). Offspring's survival depends on numerous environmental, biotic and even
399 social factors, and so does the shape of the distribution of births at the population scale (Festa-
400 Bianchet 1988). Besides, reproductive females face their own constraints, such as failure to
401 conceive during the first estrous because of a degraded body condition (Guinness et al. 1978).
402 Conception failure may then disrupt the general trend of the population, and contributes to the
403 diversity of phenology of births encountered in the wild.

404 We show that the assumption of a normal distribution or another bell-shaped (asymmetric or
405 not) distribution mimicking those found in *natura* (*e.g.* skewed normal, bimodal or Cauchy
406 distribution) has little consequences on our conclusions (Supporting information 4). The conclusion
407 differs, however, when no clear structure in the timing of births, as often encountered in some
408 populations of large herbivores living in the southern hemisphere (Sinclair et al. 2000). In this case
409 most metrics give inconsistent and unreliable results (Supporting information 4) because beyond the
410 heterogeneity of results provided by metrics measuring the same phenology characteristic for such
411 patterns, describing them using common metrics is simply not appropriate. When births occur year
412 round, the timing and rhythmicity are meaningless as they cannot reduce to one or two summarizing
413 statistics. Using evenness indexes such as "*pielou*" could provide a quantification of at least the
414 heterogeneity of such phenology. In the context of a comparative study with mainly non random
415 distributions, we recommend using "*meanvo*" to describe timing and "*meanvl*" synchrony
416 respectively, because both are among the best metrics of their category, and are not influenced by
417 temporal origin. Finally we recommend using the underused "*mood*" (which compares the median
418 birth dates, a better descriptor than the first birth) to describe rhythmicity and "*kolmomult*" to
419 describe regularity (see Table 2 and Supporting information 2 for a description of those metrics).

Being non-parametric tests, they are applicable in wide range of distributions as frequently observed in large herbivores populations.

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 two anonymous reviewers and the subject editor form their comments on a previous version of the manuscript. We also thank Marion Valeix for her comments on early 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 the "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.

Boness, D. J. et al. 1995. Does male harassment of females contribute to reproductive synchrony in the grey seal by affecting maternal performance? - Behav. Ecol. Sociobiol. 36: 1–10.

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.

Cancho-Candela, R. et al. 2007. Decline and loss of birth seasonality in Spain: analysis of 33 421 731 births over 60 years. - J. Epidemiology Community Health 61: 713–718.

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.

Festa-Bianchet, M. 1988. Birthdate and survival in bighorn lambs (*Ovis canadensis*). - J. Zool. 214: 653–661.

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.

Guinness, F. et al. 1978. Calving times of red deer (*Cervus elaphus*) on Rhum. J. Zool. 185: 105–114.

Heideman, P. D. and Utzurrum, R. C. B. 2003. Seasonality and synchrony of reproduction in three species of nectarivorous Philippines bats. - BMC Ecol. 3: 11.

Heldstab, S. A. et al. 2018. Geographical Origin, Delayed Implantation, and Induced Ovulation Explain Reproductive Seasonality in the Carnivora. - J. Biol. Rhythms 33: 402–419.

Hodge, S. et al. 2011. Reproductive competition and the evolution of extreme birth synchrony in a cooperative mammal. - Biol. Lett. 7: 54-56.

Ims, R. A. 1990. The ecology and evolution of reproductive synchrony. - Trends Ecol. Evol. 5: 135–140.

Kharouba, H. M. and Wolkovich, E. M. 2020. Disconnects between ecological theory and data in phenological mismatch research. - Nat. Clim. Chang.: 1–10.

Landler, L. et al. 2018. Circular data in biology: advice for effectively implementing statistical procedures. - Behav. Ecol. Sociobiol. 72: 128.

MacKay, A. E. et al. 2018. Maternal resource allocation adjusts to timing of parturition in an asynchronous breeder. Behav. Ecol. Sociobiol. 72: 1-10.

Morren, C. 1849. Le globe, le temps et la vie. - Bulletins de l'Académie royale des Sciences, des Lettres et des Beaux-Arts de Belgique 16: 660–684.

Moussus, J.-P. et al. 2010. Featuring 10 phenological estimators using simulated data. - Methods Ecol. Evol. 1: 140–150.

Newstrom, L. E. et al. 1994. A new classification for plant phenology based on flowering patterns in lowland tropical rain forest trees at La Selva, Costa Rica. - Biotropica: 141–159.

Paoli, A. et al. 2018. Winter and spring climatic conditions influence timing and synchrony of calving in reindeer. - PLoS One 13: e0195603.

Parmesan, C. 2007. Influences of species, latitudes and methodologies on estimates of phenological response to global warming. - Glob. Chang. Biol. 13: 1860–1872.

Plard, F. et al. 2015. The influence of birth date via body mass on individual fitness in a long-lived mammal. - Ecology 96: 1516–1528.

Renaud, L.-A. et al. 2019. Phenotypic plasticity in bighorn sheep reproductive phenology: from individual to population. - Behav. Ecol. Sociobiol. 73: 50.

Rubenstein, D. R. and Wikelski, M. 2003. Seasonal changes in food quality: a proximate cue for 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.

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.