

# Occurrence records recover moth flight timing but compress interannual phenological variation

## Authors

Owen R. Jones<sup>\*1,2</sup>

## Affiliations

<sup>1</sup> Department of Biology, University of Southern Denmark, Odense, Denmark

<sup>2</sup> Institute of Ecology and Evolution, School of Biological Sciences, University of Edinburgh,  
Edinburgh, UK

## Corresponding author

\* Owen R. Jones, [jones@biology.sdu.dk](mailto:jones@biology.sdu.dk)

## Abstract

Phenological shifts are important indicators of climate-change responses, but large-scale inference requires coverage that structured monitoring rarely provides. Occurrence databases such as GBIF offer an alternative, although the reliability of different phenological metrics remains uncertain. I compared five metrics (Q10, Q50 and Q90 timing quantiles, weighted mean flight date and season width) derived from GBIF records for 34 UK moth species against the structured Rothamsted Insect Survey (RIS), using yearly and spatial analyses, and assessed agreement by absolute error, correlation and regression slope. Most GBIF records (97.1%) came from formal citizen-science recording schemes. In yearly analyses, timing quantiles and weighted mean flight date showed reasonable agreement with RIS, but substantial variation among species. GBIF consistently compressed interannual variation, recovering only about 60% of its amplitude even after correcting for benchmark measurement error. Season width tracked yearly patterns but with errors exceeding its natural variation, while peak timing was poorly recovered. Across latitude bands, central timing was most reliable, whereas Q90, season width and latitudinal contrasts showed weaker agreement. Temporal trends were generally poorly recovered; even Q50 trends, which showed moderate agreement, were strongly compressed. GBIF can provide moderately reliable estimates of seasonal timing, but season width, peak timing and temporal trends are less dependable. Compression of interannual variation also suggests that occurrence-derived estimates of phenological sensitivity to temperature will be biased towards zero. These findings provide a basis for metric selection, although their generality across regions and data sources requires further validation.

## Introduction

Phenology, the seasonal timing of biological activity, is one of the most sensitive biological indicators of environmental change. Phenological shifts have been documented across taxa worldwide (Parmesan and Yohe 2003; Menzel et al. 2006; Poloczanska et al. 2013; Root et al. 2003; Visser and Both 2005), with consequences for species interactions, population dynamics and ecosystem function. The evidence ranges from centuries-long records of cherry blossom flowering in Kyoto (Aono and Kazui 2008) to birds and marine systems (e.g. Sanz 2002; Crick 2004; Both et al. 2004; Edwards and Richardson 2004; Van Buskirk et al. 2009; Goodenough et al. 2010). Temperature is the main driver, because it sets the pace of development and the timing of resources. Warming generally advances spring activity, although the strength, and occasionally the direction, of the response varies among species and trophic levels.

Spring events such as budburst, insect emergence and bird breeding have advanced by days to weeks across temperate regions in recent decades, closely tracking rising temperatures. This makes phenology one of the clearest biological signals of climate change (e.g. Parmesan and Yohe 2003; Root et al. 2003). In the UK and elsewhere in northern Europe, many butterflies and moths now fly earlier (Roy and Sparks 2000; Hällfors et al. 2021; Hickinbotham et al. 2024), migratory birds arrive sooner (Crick and Sparks 1999; Thorup et al. 2007; Gordo 2007), and many plants flower earlier (Fitter and Fitter 2002). The ecological consequences of these shifts may matter even more than the shifts themselves. When interacting species respond to climate at different rates, their timing can become mismatched, so that peak food supply no longer coincides with peak demand (Donnelly et al. 2011; Thackeray et al. 2010; Both et al. 2009). This is clearest in well-studied systems such as plants and their pollinators (Memmott et al. 2007), and caterpillars and the birds that feed them to their chicks (Burgess et al. 2018). Such mismatches have been linked to lower breeding success in long-distance migrants (e.g. Reséndiz-Infante and

Gauthier 2020; Jones and Cresswell 2010) and to disrupted pollination networks (Memmott et al. 2007).

Phenological shifts also have evolutionary consequences. Phenological traits are often heritable and can respond to selection, but observed shifts in populations are often driven mainly by plasticity rather than genetic change, and the two are hard to separate using observational data alone (e.g. Charmantier and Gienapp 2013; Gienapp et al. 2014). Occurrence databases such as GBIF cannot tell whether a detected shift reflects plasticity, evolution or a change in which individuals are active, and this ambiguity applies to all the metrics evaluated here.

Understanding these shifts requires consistent data, comparable across space and time and at meaningful biological scales, and such data are hard to obtain across many species and large areas. Structured monitoring programmes are the most reliable basis for phenological inference. Long-running schemes such as the UK Butterfly Monitoring Scheme (Roy et al. 2014) and the Rothamsted Insect Survey (Harrington and Woiwod 2007) produce abundance time series under standard protocols, from which timing metrics can be estimated with known uncertainty. But such schemes are expensive and cover limited areas. Remote sensing can sometimes fill the gap, but its resolution is limited (Klosterman and Richardson 2017; Richardson et al. 2018). With a few exceptions, such as the UKBMS, RIS and the PEP725 database (Templ et al. 2018), structured monitoring alone cannot provide the spatial resolution and species coverage that large-scale comparative phenology needs.

GBIF and similar databases appear to offer a solution. With hundreds of millions of records across all taxonomic groups, they make it possible, in principle, to estimate phenological metrics for thousands of species across large areas, a scale no structured scheme can match. Citizen-science platforms such as iNaturalist have greatly increased the volume and temporal coverage of GBIF records in recent decades (Tulloch et al. 2013; Di Cecco et al. 2021; Jarić et al. 2020), and large-scale research using these data has grown accordingly (Mason et al. 2025). However, GBIF

is not a single data source. It combines open citizen-science platforms, museum collections, formal citizen-science recording schemes, local records centres and research datasets, and the mix of sources differs strongly among taxa and regions. That mix is likely to matter for phenology. In the UK moth data analysed here, formal recording schemes and records centres supply 97.1% of GBIF records after filtering. This is therefore a test of scheme-dominated GBIF data (data mostly from organised citizen scientist recording schemes), not of open citizen-science platforms such as iNaturalist.

The difficulty is that these records were never designed for phenological monitoring. They reflect where and when people choose to record, not a systematic sample of seasonal activity. Effort varies across seasons, years and regions in ways that are hard to measure. Fewer recorders in early spring, for example, means fewer early-season records, which could make a species appear to emerge later than it really does. Differences in recorder density between regions add a further complication: regions with more active recorders produce more records regardless of the biology, which can exaggerate or hide real latitudinal gradients in timing. These biases are well known in species distribution modelling (Beck et al. 2014; Isaac and Pocock 2015), but their effects on phenology have received much less attention.

This matters because phenological studies extract several summaries from the same data. These include the median or mean date of activity, the start and end of the season (expressed here as timing quantiles, the dates by which a given fraction of the season's activity has occurred), season width, and spatial and temporal trends. These metrics make different demands on the data. Central timing should be fairly robust to moderate variation in effort, because most records fall near the seasonal peak. Tail quantiles are harder, because they depend on the edges of the season, where opportunistic records (those made without a standard sampling protocol) are sparsest and most prone to bias. Before examining the RIS–GBIF comparison, I therefore predicted that central timing would be recovered better than metrics that depend on the tails. A

database that gives reliable estimates of median flight date may still give misleading estimates of season onset, trends or north–south contrasts. The question is not whether GBIF contains phenological signal, which several studies have shown it does (e.g. Faltýnek Fric et al. 2020, but see Larsen and Shirey 2021; Cereghetti et al. 2026), including studies of herbarium specimens (e.g. Graves and Manzitto-Tripp 2026). The question is which metrics can be extracted reliably enough to support ecological inference, and which cannot.

Here I answer this question using the Rothamsted Insect Survey (RIS) as a structured benchmark. RIS runs a network of standardised light traps across the UK and has recorded moth abundance systematically for several decades (Harrington and Woiod 2007). I compare phenological metrics from RIS with those from GBIF records for 34 UK moth species. I test not only central timing but the range of metrics that ecologists routinely use: tail quantiles, season width, latitudinal differences and temporal trends. For these scheme-dominated UK data, GBIF recovers the year-to-year pattern in central timing and the timing quantiles reasonably well, though it understates its size. It recovers the late tail across space, season width, peak timing and most trends much less reliably. Scheme-dominated GBIF data are suitable for broad spatial analyses of central seasonal timing, but other uses need caution, and whether these results apply to platform-dominated data remains an open question.

## Methods

The study had three stages (Table 1). First, I used simulation to test whether the phenology estimators could recover known values. Second, I tested the estimators on RIS data alone, to check that they worked on structured real-world data. Third, I compared RIS and GBIF directly, testing whether GBIF-derived metrics reproduced those from RIS. This design separates validation of the method from validation of the data source. It also avoids the circular argument that GBIF data are valid because a fitted model looks plausible.

*Table 1: Estimation methods used at each stage of the analysis. Group 1 compares cyclic GAMs and symmetric and asymmetric single- and two-scale kernels, each with and without effort correction, on simulated data. Groups 2 and 3 apply the selected two-scale asymmetric kernel to RIS and GBIF (the matched workflow), with raw GBIF summaries and a simple cyclic GAM as supplementary comparators. 'Effort aware?' shows how effort enters: in the likelihood, as an offset, or by dividing by observed or inferred effort afterwards. GAM = generalised additive model; nb = negative binomial; Q10, Q50, Q90 = 10th, 50th and 90th percentiles of the seasonal distribution.*

| Group                                             | Approach                                  | Data           | Effort aware?                          | Role                                                | Outputs                                                                    |
|---------------------------------------------------|-------------------------------------------|----------------|----------------------------------------|-----------------------------------------------------|----------------------------------------------------------------------------|
| <b>Group 1: Choose the estimator (simulation)</b> | Cyclic GAM                                | Simulated      | No and yes (paired benchmark)          | Comparator                                          | Q10, Q50, Q90, weighted mean, season width                                 |
|                                                   | Single-scale symmetric kernel             | Simulated      | No and yes (paired benchmark)          | Comparator                                          | Q10, Q50, Q90, weighted mean, season width                                 |
|                                                   | Single-scale asymmetric kernel            | Simulated      | No and yes (paired benchmark)          | Comparator                                          | Q10, Q50, Q90, weighted mean, season width                                 |
|                                                   | Two-scale symmetric kernel                | Simulated      | No and yes (paired benchmark)          | Comparator                                          | Q10, Q50, Q90, weighted mean, season width                                 |
|                                                   | Two-scale asymmetric kernel               | Simulated      | No and yes (paired benchmark)          | Candidate → selected                                | Q10, Q50, Q90, weighted mean, season width                                 |
| <b>Group 2: Validate it on RIS data</b>           | Two-scale asymmetric kernel               | RIS (observed) | Yes (exposure in likelihood)           | Chosen estimator; main yearly and spatial benchmark | Q10, Q50, Q90, weighted mean, season width, peak timing, curve shape       |
|                                                   | Empirical weekly aggregation              | RIS (observed) | No                                     | Supplementary raw benchmark                         | Q10, Q50, Q90, weighted mean, season width                                 |
|                                                   | RIS annual catch / trap-days              | RIS (observed) | Yes (observed effort divisor)          | Supplementary abundance benchmark                   | Annual abundance index                                                     |
| <b>Group 3: Compare GBIF with RIS</b>             | Two-scale asymmetric kernel               | GBIF           | Yes (exposure in likelihood; floor 50) | Main matched yearly and spatial workflow            | Q10, Q50, Q90, weighted mean, season width, peak timing, spatial contrasts |
|                                                   | Empirical weekly aggregation              | GBIF           | No                                     | Supplementary raw comparator                        | Q10, Q50, Q90, weighted mean, season width                                 |
|                                                   | Cyclic GAM (bs='cc', k=8, nb)             | GBIF           | Yes (effort offset term)               | Supplementary simple comparator                     | Q10, Q50, Q90, weighted mean, season width                                 |
|                                                   | GBIF annual individuals / inferred effort | GBIF           | Yes (inferred effort divisor)          | Supplementary abundance comparator                  | Annual abundance index                                                     |
|                                                   | GBIF annual raw individuals               | GBIF           | No                                     | Supplementary abundance comparator                  | Annual abundance index                                                     |

## *Simulation study*

I began with simulation because it is the only part of the analysis in which the true seasonal distribution is known. This let me test whether the seasonal parameters could be recovered at all, and choose an estimator before fitting any RIS or GBIF data. I generated synthetic weekly phenology data under two regimes, symmetric and asymmetric. I included the asymmetric regime because moth flight curves are often skewed, and a method that works only for symmetric curves would be of little use. In the symmetric regime, the true daily abundance curve was Gaussian. Annual peak date was drawn uniformly from day 60 to 250, peak width from 10 to 20 days, and total annual abundance from 10,000 to 800,000 individuals. In the asymmetric regime, the curve was split-normal, rising and falling at different rates. Peak date was drawn from day 70 to 240, left-side width from 18 to 32 days, right-side width from 7 to 14 days, and abundance from the same range. In both regimes I simulated ten years across five latitudinal areas. I imposed fixed differences in timing and abundance among areas, so that the simulated data contained both year-to-year and spatial structure.

I then simulated observations. On top of each true abundance curve I placed a separate Gaussian curve for observer effort through the year. It had its own peak (day 150–200) and spread (20–60 days; see Supplementary Methods for the asymmetric-regime values), and was capped at 20 traps per day. I added log-scale noise so that effort varied randomly around this seasonal pattern. Observed counts were then drawn from a Poisson distribution. Its expected value was the product of true abundance, effort, and a fixed detection probability of 0.06, the small fraction of passing moths that a light trap catches. The observations were therefore noisy, and they were distorted by seasonal effort whose peak could differ from the true phenology. Overcoming this confounding is the central challenge for any estimator.

For each regime I compared five model families: a cyclic GAM, and single- and two-scale kernels in symmetric and asymmetric forms. Where possible I fitted each with and without effort

correction, so that I could assess model structure and effort treatment separately. I measured how well each recovered the known phenological summaries, focusing on Q50 but also on weighted mean day of year and season width. This stage showed that the phenology could be recovered and how the ranking of methods changed with the shape of the true curve. It also identified the two-scale asymmetric kernel as the best RIS estimator before any real data were used.

Finally, I tested recovery of variation in abundance between years and between areas. The target was relative, not absolute, abundance. Neither data source measures true population size on a comparable scale, and most ecological applications need only to know which years and places had high or low activity. I therefore standardised true and estimated seasonal totals across year  $\times$  area combinations, expressing each as a deviation from the species mean divided by its standard deviation. The comparison thus asked whether high- and low-abundance combinations were ranked correctly, not whether absolute counts were recovered.

## *Empirical study*

### *Rothamsted Insect Survey (RIS) data*

The benchmark data came from the Rothamsted Insect Survey (RIS) light-trap network, which runs traps under consistent protocols across the United Kingdom (Harrington and Woiwod 2007). The data were daily trap-level counts of macro-moths for 1999–2024, supplied by RIS as an extract in March 2026 (RIS 2021). RIS provides a repeated, structured record of seasonal activity, with known trap locations and controlled effort. It is not free of measurement error, but it is the strongest benchmark available for these moths and years, so agreement with RIS measures closeness to the best available benchmark, not absolute accuracy. I aggregated RIS records to weekly counts so that all comparisons used the same time step, excluded the two non-UK sites (Gibraltar and Lammi), and mapped RIS species names to the GBIF taxonomy. In the main comparisons, RIS metrics come from the two-scale asymmetric-kernel model, so that RIS

and GBIF are fitted with the same method. I kept empirical RIS summaries for supplementary benchmarks and RIS-only checks.

For RIS, effort is observed directly. I measured it as the number of traps operating in each week, counting a trap as operating in a week if it reported any records that week. I counted traps across the whole UK for the yearly comparisons, and within each latitude band for the spatial comparisons. Weeks in which no trap reported were left out of the fits. The number of operating traps entered the model as its exposure (the amount of trapping effort behind each weekly count), so the model estimates catch per operating trap rather than raw counts.

#### *GBIF data*

I took occurrence records from the Global Biodiversity Information Facility (GBIF), which combines data from museum collections, formal citizen-science recording schemes, research projects and open citizen-science platforms. Unlike RIS, these records are largely opportunistic: their density through time and space reflects where and when people record as much as it reflects the moths. Separating these two influences is the main challenge in extracting phenology from GBIF.

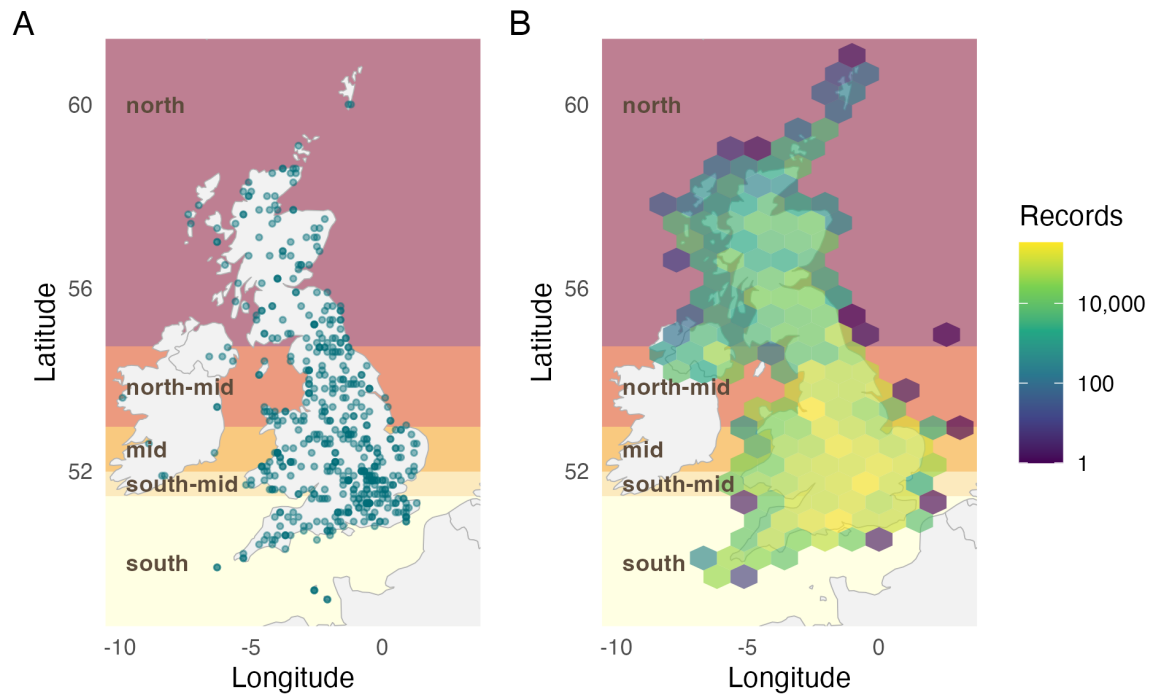
For the focal UK species, the GBIF records came mostly from formal citizen-science recording schemes rather than open citizen-science platforms. Of the filtered records, 97.1% came from formal recording schemes or monitoring programmes, 1.9% from open citizen-science platforms, 1.0% from research projects or surveys, and less than 0.1% from museums or other sources. This study is a test of scheme-dominated GBIF data for UK moths, not of data from open platforms such as iNaturalist. Both are citizen science, so the distinction is one of organisation (a common target group, standard record formats and expert verification), not of who records.

Some GBIF records come from the Rothamsted traps themselves: the National Moth Recording Scheme, which supplies most of the focal GBIF data, includes RIS catches, identifiable by the sampling protocol “Rothamsted Trap”. I removed these records (184,191; 2.6% of focal-species records) before any GBIF effort estimation, modelling or comparison, so that the benchmark and the data under test share no records. To check how much this mattered, I compared model-free agreement between RIS and GBIF in the spatial analysis for the same GBIF download processed with and without these records (Table S16). Leaving them in raised the mean correlation with RIS by 0.01 to 0.02 and lowered the mean absolute error by 0.13 to 0.25 days, depending on the metric. The effect was small and in the expected direction: records that come from the RIS traps make GBIF look slightly more like RIS. All results reported here exclude them.

I harmonised taxonomic names, selected records for the focal species, filtered them to the study region and period, and kept only records with usable dates and coordinates. As with RIS, I aggregated records to weekly counts. To account for variation in recording effort, I estimated effort separately and included it in the model as exposure. From the fitted seasonal curves I extracted five phenological summaries (Q10, Q50, Q90, weighted mean day of year and season width), for each year and for each latitude band, and compared each with the corresponding RIS value. The two datasets overlap widely in space (Figure 1).

Because GBIF effort is not observed, I estimated it from the records themselves. Records close together on the same date probably come from the same light trap, and separate clusters of records probably come from different traps. I did not require identical coordinates, because record coordinates are often rounded or imprecise. Instead, I applied DBSCAN density-based clustering (Hahsler et al. 2019) to each day’s coordinates, grouping records within 500 m of one another and requiring at least three records per cluster. I treated each cluster as one trap and summed the daily number of clusters to give weekly effort. I fixed both settings in advance

rather than tuning them against RIS: 500 m allows for imprecise coordinates, and the three-record minimum stops one or two isolated records from counting as a trap. The resulting weekly effort entered the kernel model as exposure and the cyclic GAM comparator as an offset, so in both the fitted curve described abundance per unit effort rather than raw record counts.



**Figure 1.** RIS and GBIF data overlap widely across the UK. Left: RIS light-trap locations. Right: density of GBIF records (hexagonal bins). Shaded horizontal bands show the five latitude bands used in the spatial analyses.

### *Species selection*

RIS and GBIF use different taxonomies, so I first mapped RIS names to current accepted names (for example, *Orthosia munda*, *Omphaloscelis lunosa*, *Spilosoma luteum* and *Hoplodrina alsines* became *Anorthoa munda*, *Anchoscelis lunosa*, *Spilarctia lutea* and *Hoplodrina octogenaria*). I excluded both *Mesapamea secalis* and *M. didyma*. The two can be told apart only by examining genitalia, so RIS

mostly records them as a combined category, and species-level RIS records exist for only 12 of the 26 years.

From this pool, I chose species to cover the whole flight season rather than only the most frequently recorded taxa. I chose species from across the season, so that early-, mid- and late-season species were all included. I kept only species with a single annual flight peak, and required each to have at least 20 years of RIS data and plentiful GBIF records. This stopped abundant mid- and late-season species from dominating the list, and meant the methods were tested across a wide range of phenological shapes.

I tested the single-peak rule on the RIS weekly catch curves rather than assuming it. A known multivoltine species, *Gymnoscelis rufifasciata*, showed a clear second peak and was excluded. Two retained species, *Lacanobia oleracea* and *Axyليا putris*, show only a weak, occasional late-season second peak, visible as a low tail rather than a second generation, and I kept them.

### *Spatial binning*

For the spatial analysis, I divided the study area into five latitudinal bands at the quintiles of RIS trap latitude, which balances spatial resolution against sample size within each band. I assigned RIS traps and GBIF records to bands by their coordinates (Figure 1). Within each band and year I built a seasonal activity curve, using trap counts and exposure for RIS and occurrence records for GBIF. From these curves I extracted band-specific metrics, which allowed comparison of timing within bands, latitudinal contrasts in Q50 and within-band trends.

### *Phenology model*

I modelled flight phenology with a two-scale asymmetric-kernel mixture model. I fitted it first to RIS, to create the benchmark, and then to GBIF, so that both datasets were analysed with the same method. Think of the model as adding together a few smooth bumps, some broad and some narrow, each allowed to be lopsided. A single smooth curve often cannot fit both the

gradual shoulders and the sharp central peak of a flight season. The model fits them separately: a bank of broad kernels describes the overall seasonal envelope, including long shoulders, and a bank of narrow kernels describes sharp peaks. Each kernel can rise and fall at different rates, so the model can fit seasons in which activity builds slowly and ends abruptly, or the reverse. Kernel weights were constrained to be non-negative, and a sparsity penalty stopped the model from placing activity in weeks with little data. Formally, weekly expected abundance for a given species, year and (where relevant) latitude band was the sum of the two banks of non-negative, asymmetric kernels, multiplied by exposure. The full model equations are given in the Supplementary Methods.

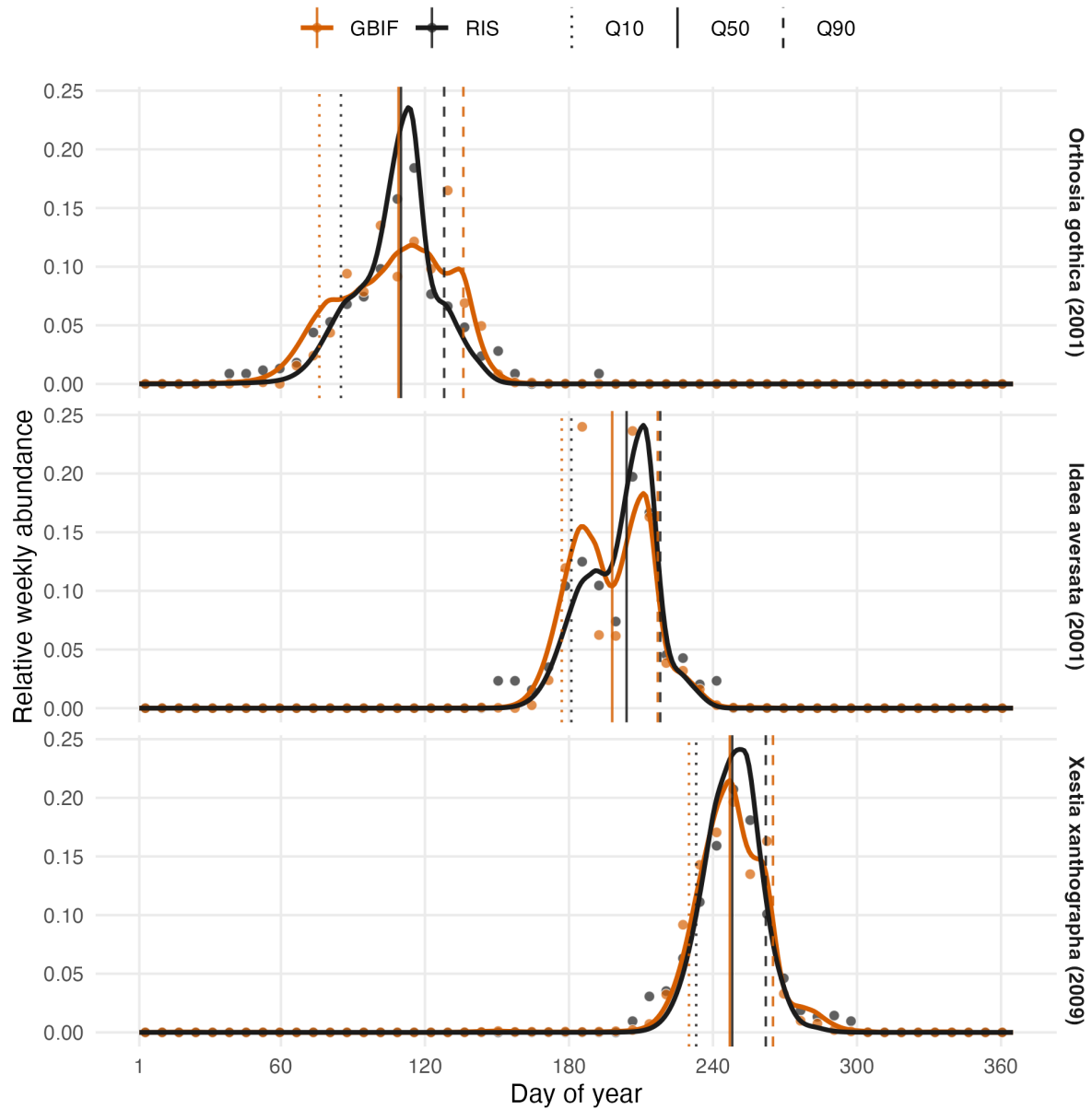
I chose the settings that control model flexibility (kernel spacing, left and right widths, and penalty strength) in two steps. I first tuned a single-scale asymmetric kernel. After inspecting species-years with sharp peaks and broad shoulders, I extended it to the two-scale form by adding separate broad and narrow banks, both with a 7-day spacing to match the weekly data. I chose among candidates using the simulations, where the truth is known. In both regimes, the two-scale model matched or beat the single-scale asymmetric kernel and the GAMs, and it also recovered seasonal shape (width, tails and peak), which a single-scale model cannot. I used it as the single estimator for both RIS and GBIF. I report its performance on the empirical RIS check, which has no known truth and ranks the models differently, as a validation (Results). Full tuning values and the effect of broader single-scale kernels on curve shape are in the Supplementary Methods (Fig. S6, Table S11).

### *GBIF phenology summaries*

The main GBIF estimates, in both the yearly and spatial comparisons, came from the same two-scale asymmetric-kernel model, with inferred weekly effort as exposure. I refer to this pairing of identical RIS and GBIF models as the matched workflow. I also kept two supplementary comparators: raw weekly GBIF summaries, and a simple per-year cyclic GAM with an effort

300 offset. These test whether raw GBIF data already contain phenological signal, and whether a  
301 simple effort-aware smoother improves on them. Example single-year fits show broad  
302 agreement between RIS and GBIF, with local differences in timing and shape (Figure 2).

303 The GBIF fits needed one further adjustment. I chose the model and its settings from the  
304 simulations and RIS alone, with no tuning to GBIF. However, very small inferred weekly effort  
305 values could produce implausibly high counts per unit effort, so I set a floor of 50 trap-nights on  
306 inferred weekly exposure. I chose 50 as a moderate value, not the value that gave the best  
307 agreement: agreement kept improving at higher floors, but only by overriding more and more of  
308 the effort estimate (Table S13).



**Figure 2.** Example single-year fits: GBIF generally reproduces the seasonal pattern seen in RIS, with local differences in timing and shape. Rows show *Orthosia gothica* in 2001, *Idaea aversata* in 2001 and *Xestia xanthographa* in 2009. Points show weekly relative abundance, coloured lines show the fitted curves for RIS and GBIF, and vertical lines mark the fitted Q10, Q50 and Q90.

#### *Metric extraction and abundance indices*

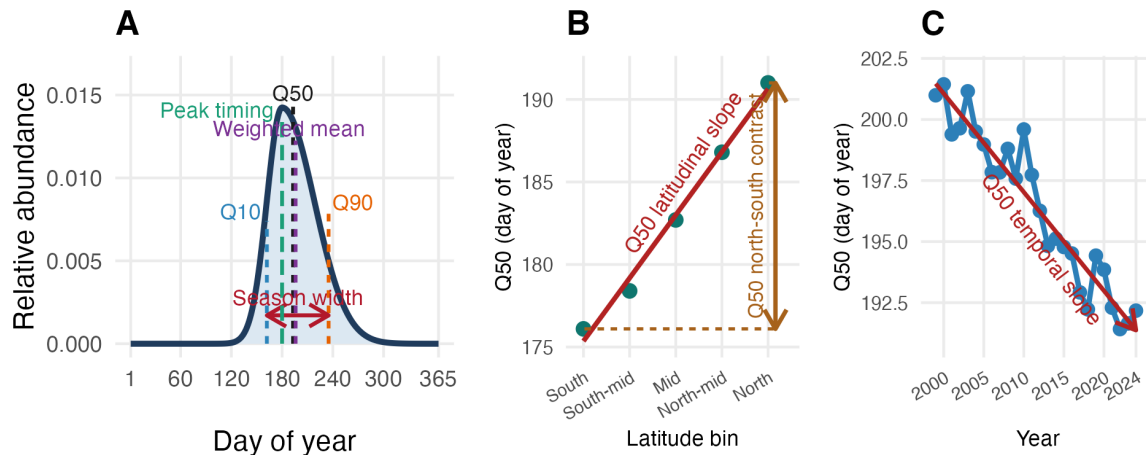
I estimated phenology from relative, not absolute, seasonal abundance. Within each species and year (and each latitude band, for the spatial analyses), I divided weekly counts or fitted values by

the seasonal total, giving a distribution that sums to one. I applied the same step to fitted daily curves before extracting metrics. All timing, width and peak metrics describe the timing and shape of the season, not how many moths were caught.

As a supplementary check, I tested whether GBIF recovered year-to-year variation in total seasonal abundance. For RIS, annual abundance was the seasonal catch divided by the number of trap-days that year. For GBIF, effort is not observed, so I compared two measures: the raw yearly number of individuals, and that number divided by the summed DBSCAN effort estimate. To put all sources on the same scale, I divided each species' yearly values by their mean across years, so that values above one indicate above-average years. The comparison asked whether good years in RIS were also good years in GBIF, not whether the counts matched. This abundance index measures change in total numbers between years, whereas the seasonal curves above describe how activity is spread within a year. I report the abundance results in the Supplementary Information, because this paper focuses on timing.

#### *Phenological metrics and derived summaries*

I assessed five main metrics. Q10, Q50 and Q90 are the 10th, 50th and 90th percentiles of the seasonal distribution, which mark the onset, median and end of the flight season. The weighted mean day of year is the average date, with each week weighted by its share of the season's activity. Season width is the span from Q10 to Q90 (Figure 3A). Together they describe the whole season: Q50 and the weighted mean give its centre, Q10 and Q90 its start and end, and season width its length. I also recorded peak timing, the day on which the fitted daily curve reached its maximum. From these yearly metrics I derived spatial contrasts between latitude bands and temporal trends (Figure 3B, C).



**Figure 3.** The phenological metrics and the spatial and temporal summaries derived from them.

**A.** A stylised flight season, showing Q10, Q50 and Q90 (early, median and late timing), the weighted mean day of year, peak timing, and season width (Q10 to Q90). **B.** Spatial summaries across the five latitude bands: the latitudinal slope and the north–south contrast in Q50. **C.** Temporal trend in yearly Q50.

I calculated spatial and temporal summaries from the band-specific metrics with simple regressions. For each species and year, the north–south contrast was the value in the northernmost band minus the value in the southernmost band, and the latitudinal slope was the slope of the metric against band midpoint latitude. For temporal trends, I regressed each yearly metric on year for each species, and took the slope as the rate of change. In the spatial workflow I did the same within each species and band, to give within-band trends.

### *Comparison criteria*

I assessed agreement separately for each species, because these metrics are used to make inferences about individual species, and treated pooled summaries as secondary. For each metric I used three measures of agreement: mean absolute error (MAE, the average difference between GBIF and RIS, ignoring its sign), Pearson correlation, and the slope of GBIF on RIS within each species. The slope tests whether GBIF matches the size of variation in RIS, not just its direction.

MAE is in each metric's own units (days for dates and season width, days per year for trends, days per latitude band for spatial gradients), so raw MAE cannot be compared across metrics.

I therefore also report a standardised error: the mean MAE divided by the standard deviation of the RIS values over the same units on which agreement was assessed. These units were years within species for the yearly, spatial-band and contrast metrics, and species for the trend slopes. The standardised error has no units, so it can be compared across metrics; a value near 1 means the error is as large as the signal being estimated.

Rather than combine these measures into one rating, I described each metric on two separate axes. *Pattern recovery* is whether GBIF tracks the RIS pattern, measured by the correlation. *Magnitude recovery* is whether GBIF reproduces the size of variation, measured by the slope. Two axes are needed because a metric can track the RIS pattern well but compress its size, or do neither. I also use the standardised error as a measure of *accuracy* when ranking metrics within species (see Results). These axes are practical guidance, not a pre-registered scheme: their purpose is to tell researchers which metrics they can use.

The regression slope needs one more qualification, because a low slope can mean either that GBIF misses the pattern or that it understates its size, and because the slope assumes the benchmark has no error. The slope combines the two because the least-squares slope equals the correlation multiplied by the ratio of standard deviations (slope =  $r \times SD_{GBIF}/SD_{RIS}$ ). The first factor measures how well GBIF *tracks* RIS; the second measures whether GBIF reproduces the *amplitude* of variation. A low slope can come from either, so I report both factors as well as their product. The assumption matters because random error in the RIS values pulls the slope towards zero (regression dilution), so the raw slope understates the true one. To correct for this, I estimated the sampling reliability of each RIS metric. Each RIS metric is read from a fitted curve, and its sampling error follows from the standard error of a quantile, which depends on the height of the fitted curve and the number of moths caught. I report the corrected slope (the raw

slope divided by the RIS reliability), with a sensitivity analysis for different levels of overdispersion in the catch (Tables S14 and S15).

I classed a metric's *pattern* recovery as good when the mean species-level correlation was at least 0.5, moderate when it was 0.4–0.5, and poor below 0.4. I classed its *magnitude* recovery as near 1:1 when the mean slope was at least 0.8, compressed when it was 0.4–0.8, and poor below 0.4. These thresholds are conveniences for describing a continuum, not formal cut-offs. A metric can be used to estimate the absolute size of change only if it does well on both axes.

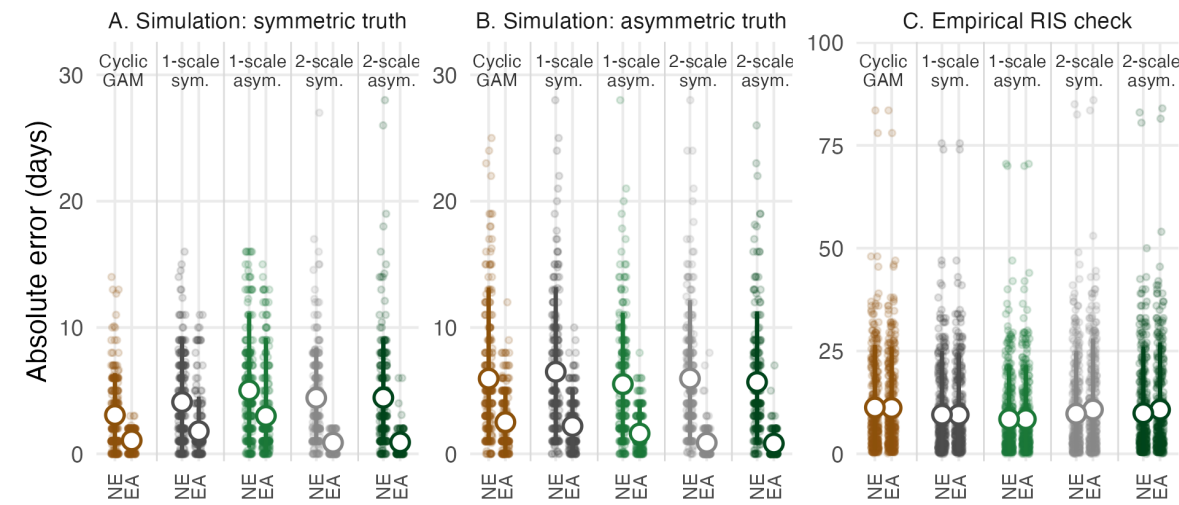
### *Software*

I ran all analyses in R version 4.6.1 (R Core Team 2026), using **dplyr** (Wickham et al. 2026) and **tidyr** (Wickham et al. 2025) for data handling, **mgcv** (Wood 2011, 2017) for GAMs, **sf** (Pebesma 2018) for spatial operations and **ggplot2** (Wickham 2016) for figures.

# Results

After filtering to the focal species, 6,609,289 GBIF records remained. Of these, 97.1% came from formal recording schemes or monitoring programmes, 1.9% from open citizen-science platforms, 1.0% from research projects or surveys, and a negligible fraction from museums or other sources (Table S12). The GBIF data tested here are overwhelmingly scheme data.

## Method validation



**Figure 4.** Model families rank differently in simulation and in the empirical RIS check. Panels A–C show the symmetric simulation, the asymmetric simulation and the empirical RIS check. Within each model family, versions without (NE) and with (EA) effort correction are shown side by side. Large points show the mean absolute error (MAE) across Q10, Q50, Q90, weighted mean and season width; bars show the 10th and 90th percentiles; pale points show individual errors (year  $\times$  area in A and B; species  $\times$  year in C). Lower MAE means better recovery.

I first tested whether the estimators could recover seasonal patterns, independently of GBIF, in two ways: on simulated data, where the true phenology is known, and on RIS data alone. Both tests assessed central timing and the fit of the whole seasonal curve, including broad shoulders and sharp peaks (Figure 4, Table S1). In simulation, effort-aware models always beat their no-

effort versions. With symmetric curves, the effort-aware two-scale symmetric kernel had the lowest average error across the five metrics (0.88 days). The asymmetric version was effectively tied (0.91 days), with a similar spread of errors (Figure 4). In the symmetric regime, then, the two two-scale kernels were equivalent.

The asymmetric regime separated the models more clearly. There, the effort-aware two-scale asymmetric kernel performed best (0.82 days on average), recovering Q50 with an MAE of 0.60 days, against 5.16 days without effort correction. Its errors were no more variable than those of the symmetric kernel. I therefore used the effort-aware two-scale asymmetric kernel for all further analyses (see Methods).

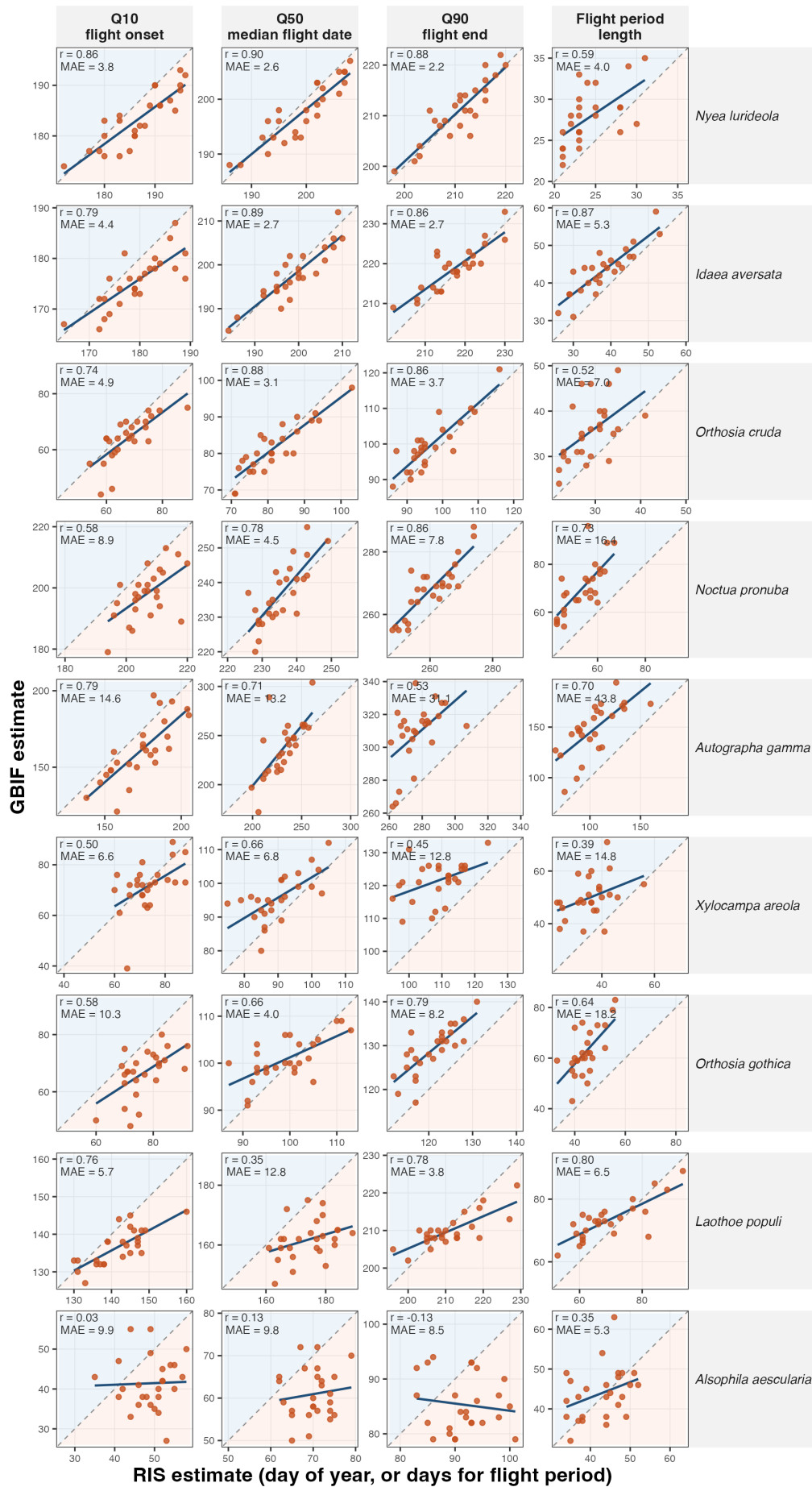
The empirical RIS check ranked the models differently: the single-scale asymmetric kernel without effort had the lowest average error and the narrowest spread. This check has no known truth. It scores each model against raw weekly quantiles, so a less-smoothed model can score well simply by following sampling noise and weekly binning. I therefore treat it as confirmation that all models give similar errors on real data, not as a basis for choosing between them. That observed RIS effort did not improve every model matches the later finding that effort correction does not always improve timing estimates from real data. Compared with the single-scale asymmetric kernel, the two-scale model fitted the whole seasonal curve much better. It reduced mean curve MAE by about 50%, season-width error from 2.42 to 1.35 days, and peak-height error from 0.0429 to 0.0169, and it won all 78 species-year comparisons on curve MAE and 74 on peak-height error.

Sensitivity checks showed why. Broadening the single-scale kernel filled out the shoulders of some curves but flattened the peak, and did not reduce error. The two-scale model instead reduced curve error, peak-height error and season-width error, while keeping fitted total abundance stable. Median timing was already stable under the single-scale model, so the main

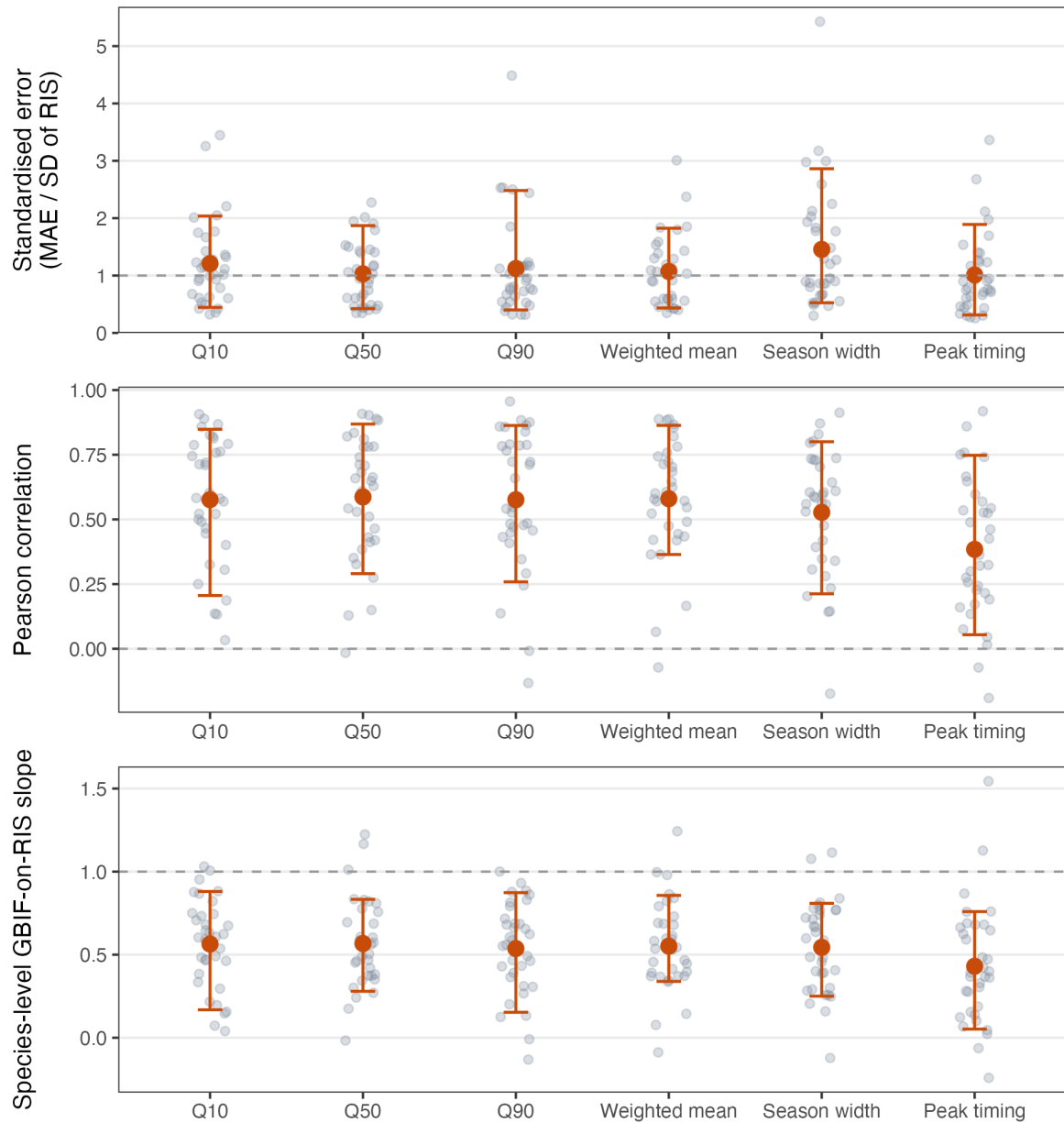
435 gain from the two-scale model was a better fit to the shape of the season, not a large change in  
436 Q50.

437 *GBIF vs RIS: yearly species-level phenology agreement*

438 The first question is whether, for a given species, RIS and GBIF give similar yearly metrics in the  
439 same years. I compared yearly estimates from the matched workflow, with the GBIF exposure  
440 floor of 50 trap-nights. I focused on agreement within species, because pooling across species  
441 inflates agreement: species differ widely in their mean timing. Figure 5 shows yearly comparisons  
442 of Q10, Q50, Q90 and season width for nine representative species, and Figure 6 summarises all  
443 yearly metrics across species (values in Table S2).



**Figure 5.** Yearly agreement between RIS and GBIF in Q10, Q50, Q90 and season width for nine of the 34 focal species, chosen to span strong, mixed and weak agreement (all 34 species are shown in Figures S7–S10). Columns show flight onset (Q10), median flight date (Q50), flight end (Q90) and flight period length (season width). Each point is one year, with RIS on the x-axis and GBIF on the y-axis. Dashed lines show 1:1 agreement, solid lines the fitted relationship, and panel text the correlation and mean absolute error. Shading marks years in which GBIF is later than RIS (blue, upper left) or earlier (orange, lower right); for season width, blue and orange mark a longer and a shorter GBIF flight period. Both datasets were fitted with the matched workflow, with a GBIF exposure floor of 50 trap-nights.



**Figure 6.** Yearly agreement between GBIF and RIS for each metric. Grey points show individual species; orange points and bars show the mean and the 10th–90th percentile range across species. Top: standardised error (mean absolute error divided by the standard deviation of the RIS values; the dashed line at 1 marks an error as large as the metric’s natural year-to-year variation). Middle: Pearson correlation. Bottom: slope of GBIF on RIS within species (the dashed line at 1 marks equal variation). Matched workflow, with a GBIF exposure floor of 50 trap-nights.

GBIF recovered the year-to-year pattern in median timing (Q50) reasonably well: a good pattern class by the thresholds in Table 2, though far from perfect. The mean correlation across species was 0.59 (median 0.64), all 34 species showed a positive relationship between RIS and GBIF, and the mean absolute error was 6.8 days. The mean slope was 0.57: GBIF compressed variation within species, even though it tracked its direction.

These averages hide large differences between species. Across the 34 species, the Q50 correlation ranged from -0.02 to 0.91 (interquartile range 0.42–0.78), the slope from -0.02 to 1.22, and the mean absolute error from 2.3 to 14.9 days. GBIF tracked RIS closely for 13 species ( $r > 0.7$ ), recovering median timing to within a few days, but recovered little yearly signal for 7 ( $r < 0.4$ ). Good but compressed agreement is an average, not a description of any one species: for the best species GBIF is a faithful record, and for the worst it captures little of the year-to-year signal.

The other quantile-based and central metrics performed similarly. Mean correlations were 0.58 for Q10, 0.58 for the weighted mean (mean MAE 6.3 days) and 0.58 for Q90, so Q10 was, if anything, slightly better recovered than Q50. Season width tracked the RIS pattern about as well (mean correlation 0.53) but had by far the largest errors. Its mean MAE of 11.7 days was larger than the natural year-to-year variation in the RIS values (standardised error 1.5; Table 2). Peak timing was the weakest yearly metric (mean correlation 0.38). Mean slopes were below one for every metric, so GBIF compressed year-to-year variation even where it ranked years correctly. GBIF captures clear yearly signal for all quantile-based and central metrics, but how well it does so varies among species.

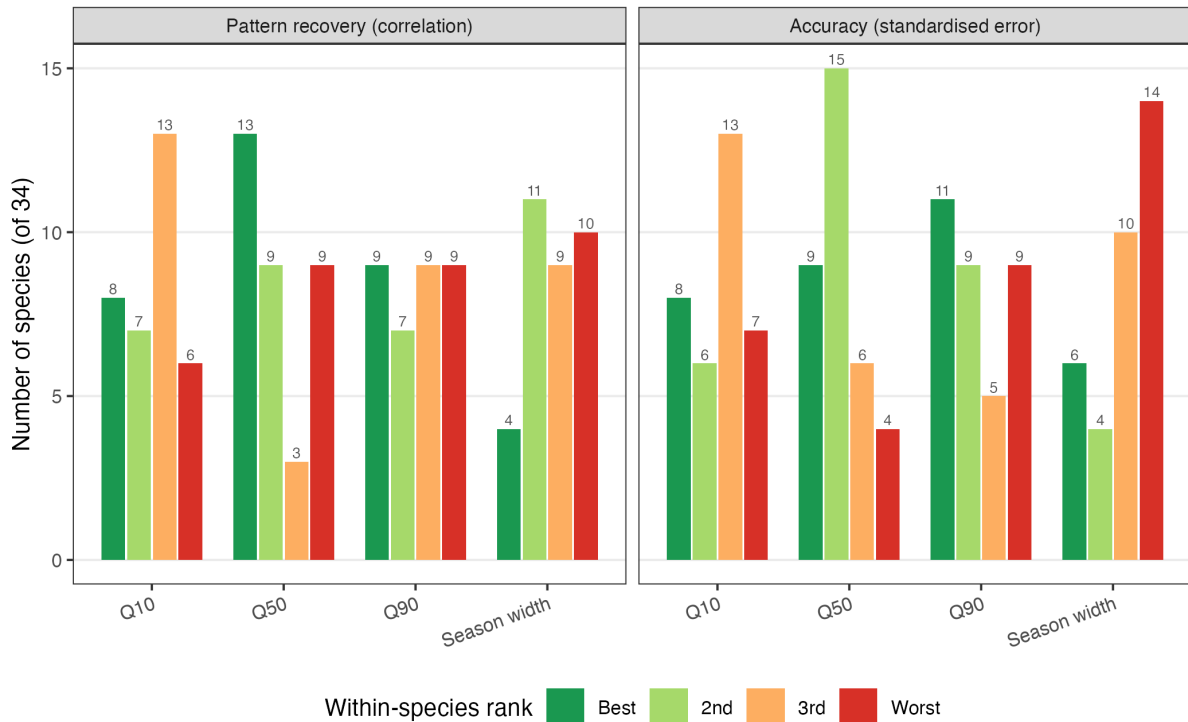
The low slope needs a closer look, for two reasons. On its own, it cannot distinguish true compression of the signal from imperfect tracking. It also assumes that the RIS values have no error. Splitting the Q50 slope into its two factors ( $\text{slope} = r \times \text{SD}_{\text{GBIF}}/\text{SD}_{\text{RIS}}$ ) shows that

486 GBIF's total year-to-year spread is 1.01 times that of RIS. At face value, then, the low slope  
487 (0.57) comes from the moderate correlation (0.59), not from a smaller spread in GBIF.

488 This near-equal spread has two possible explanations. GBIF could preserve the true signal but  
489 track it noisily, or it could compress the signal while observation noise makes up the lost  
490 variance. The two can be told apart only by knowing how much of the year-to-year variation in  
491 RIS is real signal rather than measurement error. I therefore estimated the sampling reliability of  
492 the RIS Q50 directly. Because each RIS metric is read from a fitted curve, its sampling error  
493 follows from the standard error of a quantile. The RIS Q50 has a sampling standard deviation of  
494 only about 1.6 days, against a year-to-year signal of about 7.0 days, giving a reliability of 0.95.

495 Almost all of the year-to-year variation in RIS is therefore real signal, which points to genuine  
496 compression. The corrected slope, an estimate of the true ratio of signal amplitudes, is 0.60,  
497 barely above the raw value. It stays below about 0.8 across plausible levels of overdispersion in  
498 the catch. It would exceed one only if the sampling error of the RIS Q50 were about five days,  
499 more than three times the estimate of 1.6 days, and the numbers of moths caught rule that out  
500 (Table S15). GBIF thus recovers only about 60% of the true year-to-year variation in median  
501 timing. Its near-equal total spread is compressed signal plus observation noise, not a faithful  
502 copy of the real variation. The same held for Q10 and Q90 (corrected slopes 0.63 and 0.60).

503 To test whether any one metric is the best summary within each species, not just on average, I  
504 ranked Q10, Q50, Q90 and season width from best to worst within each species (Figure 7),  
505 separately for pattern recovery and accuracy. No metric was best for every species. Q50 was  
506 most often ranked first on both measures, narrowly ahead of Q10, and season width was most  
507 often ranked last, especially for accuracy.

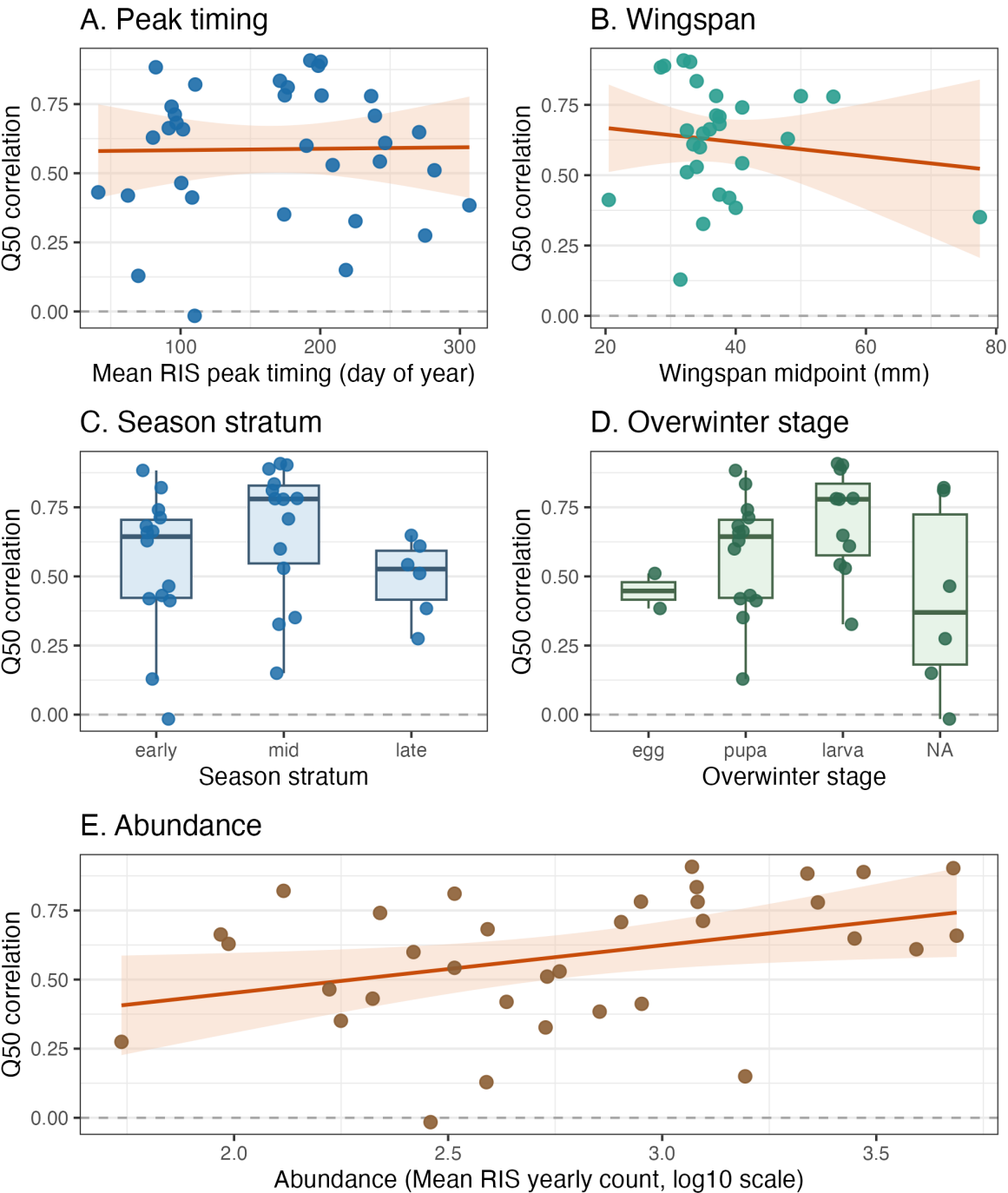


**Figure 7.** Ranking of Q10, Q50, Q90 and season width within each of the 34 species, from best (green) to worst (red). Left: pattern recovery (Pearson correlation; higher is better). Right: accuracy (standardised error, the mean absolute error divided by the standard deviation of the RIS values; lower is better). Bars show how many species give each metric each rank. Q50 is most often ranked first on both measures, but no metric is best for every species and its lead over Q10 and Q90 is small. Season width is most often ranked last, especially for accuracy. Many rankings within species are decided by small differences.

#### *Trait correlates of yearly Q50 agreement*

Given these differences between species, I asked whether simple traits predicted how well each species' Q50 was recovered (Figure 8). Across the 34 species with both agreement and trait data, the relationships with mean RIS peak timing ( $r = 0.02$ ) and wingspan ( $r = -0.13$ ) were weak. Agreement barely differed among early-, mid- and late-season species (mean correlations 0.55, 0.67 and 0.49). The clearest pattern was with abundance: commoner moths tended to have better

Q50 recovery ( $r = 0.37$  with mean yearly RIS catch). Abundance helps, but does not on its own determine how well a species is recovered.



**Figure 8.** Traits as predictors of how well GBIF recovers each species’ Q50. Each point is one focal species. Panels show the RIS–GBIF correlation in Q50 against mean RIS peak timing,

wingspan, season, overwintering stage and mean yearly RIS catch. Matched workflow, with a GBIF exposure floor of 50 trap-nights.

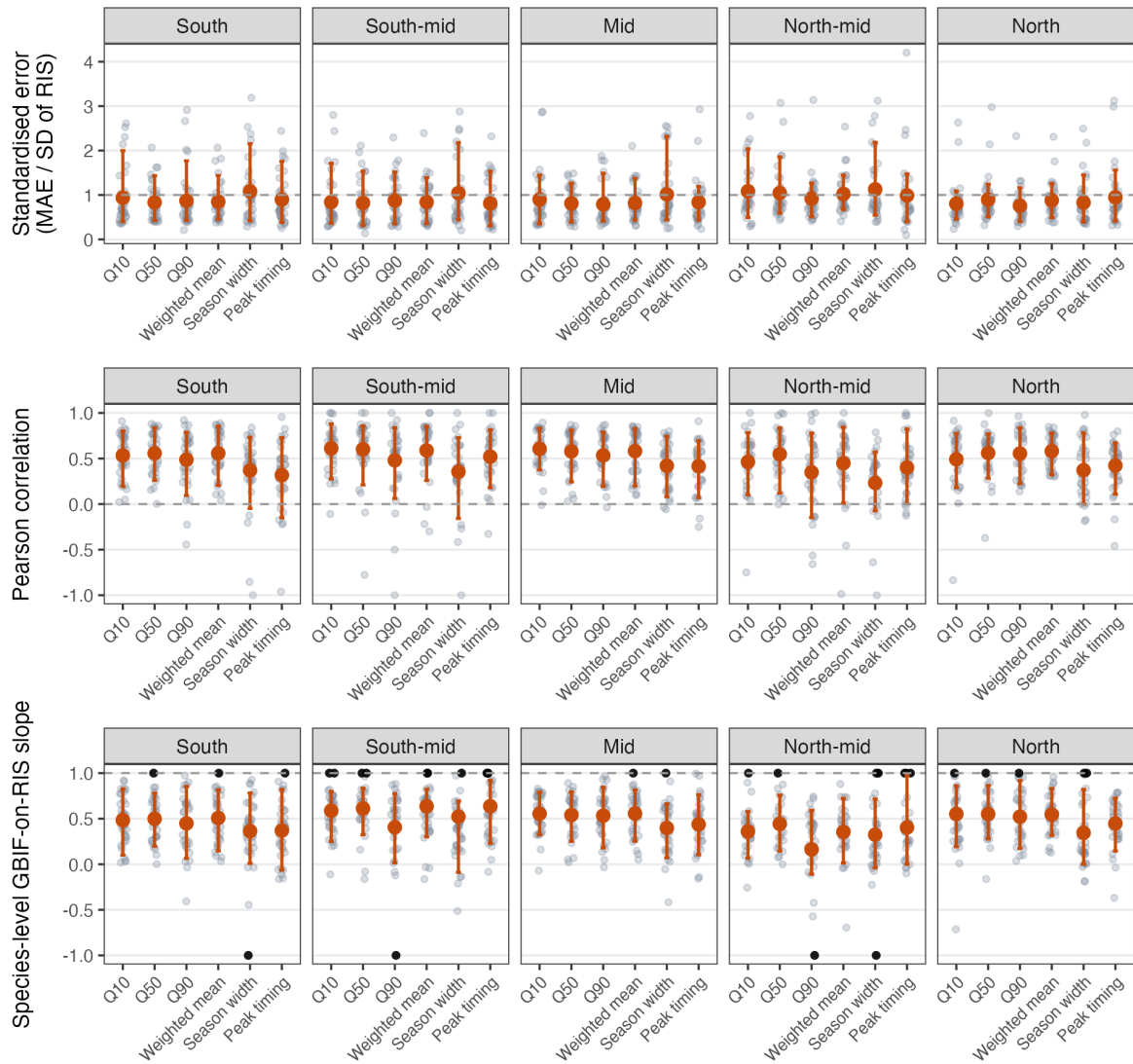
#### *Staged yearly GBIF model comparison*

I next asked whether a different GBIF model improved yearly metric recovery, comparing raw GBIF summaries with a simple per-year cyclic GAM (Fig. S1, Table S3). Effort correction did not improve yearly timing. Raw GBIF data already recovered Q50 well (MAE 6.2 days), and the cyclic GAM did much worse (13.4 days), with the same pattern for Q10, Q90, the weighted mean and season width. The simple GAM, the only effort-corrected alternative I tested for timing, was clearly worse than the raw summaries it was meant to improve. Effort correction therefore cannot be assumed to improve estimates of timing.

Effort correction did matter for abundance. The effort-standardised index clearly outperformed raw yearly totals (abundance-index MAE 0.25,  $r = 0.66$ , against 0.58 and  $r = 0.40$ ). The benefit of effort correction depends on the metric: real for abundance, marginal for timing (Fig. S2, Table S4).

#### *Derived yearly temporal trends*

Recovering trends through time is a harder test, because errors in single years add up over decades. For most metrics, trends were recovered far less well than the yearly values themselves. Correlations between GBIF and RIS trends were close to zero or negative for Q10, Q90, the weighted mean and season width, and weak for peak timing (0.19). The nearest to an exception was Q50, whose trend correlation (0.42; slope 0.40; 34 species) was moderate, but clearly lower than that of yearly Q50 itself (Fig. S3, Table S5). Yearly Q50 trends can therefore support only cautious, qualitative comparisons, and trends in other metrics from this GBIF workflow are not reliable.



551

552 **Figure 9.** Agreement between GBIF and RIS within latitude bands. Grey points show species

553 within each band; orange points and bars show the mean and the 10th–90th percentile range.

554 Top: standardised error (MAE divided by the standard deviation of the RIS values; the dashed

555 line at 1 marks an error as large as the metric’s natural variation). Middle: Pearson correlation.

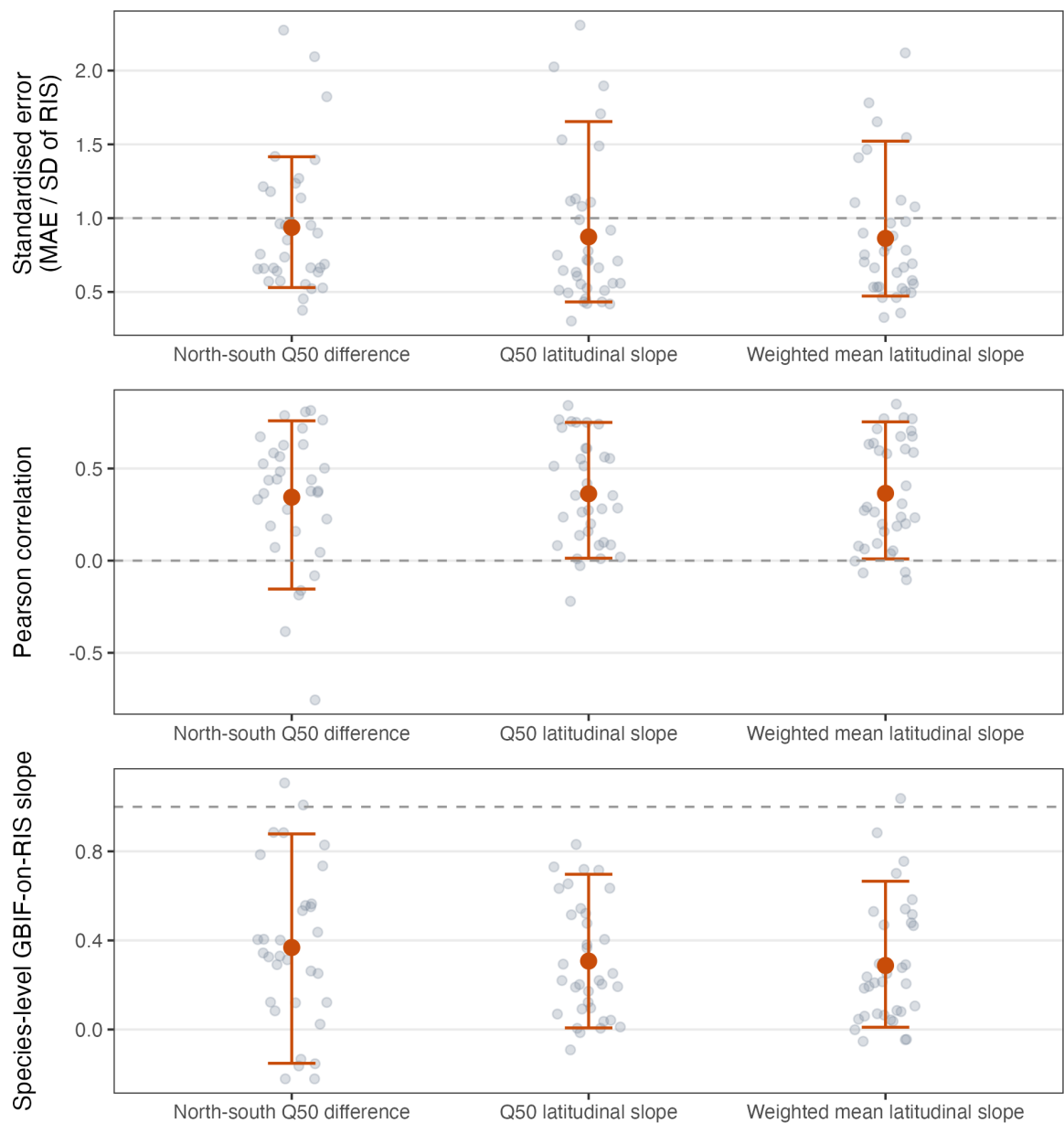
556 Bottom: slope of GBIF on RIS; a slope below 1 means GBIF compresses the variation seen in

557 RIS. The slope axis is limited to –1 to 1; the few values outside this range (4.3%) are shown as

558 black points at the axis limit. Matched workflow, with a GBIF exposure floor of 50 trap-nights.

I then asked whether RIS and GBIF give similar metrics within UK latitude bands, the scale at which the two data sources can be compared directly, again focusing on agreement within species (Figure 9, Table S6). Q50 showed the strongest spatial agreement, narrowly ahead of Q10, with a mean correlation of 0.57, mean MAE of 6.4 days and mean slope of 0.53 across all species  $\times$  band combinations with enough data (34 species overall, 32 to 34 per band). Q10 and the weighted mean were almost as strong (mean correlations 0.54 and 0.55), and Q90 and especially season width were weaker (0.48 and 0.35). So, unlike the yearly analysis, in which most timing metrics performed similarly, spatial agreement was highest for central timing and fell for metrics that depend on the late tail or on the length of the season.

RIS-only validation confirmed that the two-scale asymmetric kernel closely reproduced the empirical RIS values. Q50 (MAE 4.0 days,  $r = 1.00$ ) and the weighted mean (MAE 3.5 days,  $r = 1.00$ ) were recovered very closely and peak timing well. Q10 and Q90 kept high correlations despite larger absolute errors, and season width was again the weakest. The RIS estimator is unlikely to be the main source of disagreement between RIS and GBIF for central timing (Fig. S4, Table S7).



575

576 **Figure 10.** Recovery of spatial contrasts is weak and similar across contrasts. Grey points show  
577 species; orange points and bars show the mean and the 10th–90th percentile range. Top:  
578 standardised error (MAE divided by the standard deviation of the RIS values; the dashed line at  
579 1 marks an error as large as the contrast’s natural variation). Middle: Pearson correlation.  
580 Bottom: slope of GBIF on RIS, for the north–south difference in Q50, the latitudinal slope of

Q50 and the latitudinal slope of the weighted mean. As in Figure 9, slopes below 1 mean GBIF compresses the variation seen in RIS.

I next asked whether GBIF reproduces the north–south contrasts and latitudinal slopes derived from the band-specific metrics, again at the species level (Figure 10, Table S8). Spatial contrasts were recovered weakly; Table 2 classes all three as poor. The north–south Q50 difference had a mean correlation of 0.34 (mean MAE 7.4, mean slope 0.37). The Q50 latitudinal slope was similar (mean correlation 0.36, mean MAE 2.2), and the weighted-mean latitudinal slope was no better (0.37). These contrasts were recovered better than temporal trends, but clearly worse than the band-specific Q50 values they were built from.

#### *Derived temporal trends within the spatial workflow*

Trends within each latitude band depend on getting both the spatial and the temporal structure right, so they are harder to recover than either alone. They were much weaker than the yearly or band-specific metrics: even for Q50, agreement was only modest, and weighted-mean trends were effectively unrecoverable (Fig. S5, Table S9).

#### *Reliability synthesis*

Table 2 brings these results together (the same values are in Table S10). Each metric is described by its *pattern recovery* (from the correlation) and its *magnitude recovery* (from the slope), because the two can differ: a metric can track the RIS pattern while compressing its size. The classes use the descriptive thresholds set out in the Methods, and are practical guidance rather than formal statistical tests.

*Table 2: Reliability of GBIF phenology metrics. For each metric the table gives the number of species or species × band units (n), the mean absolute error (MAE; days for timing metrics, days per year for trends), the standardised error (MAE divided by the standard deviation of the RIS values; a value near 1 means the error is as large as the metric's natural variation), the mean Pearson correlation (r) and the mean slope of GBIF on RIS.*

Pattern recovery (does GBIF track the RIS pattern?) is classed from  $r$ : good  $\geq 0.5$ , moderate 0.4–0.5, poor  $< 0.4$ . Magnitude recovery (does GBIF reproduce the size of variation?) is classed from the slope: near 1:1  $\geq 0.8$ , compressed 0.4–0.8, poor  $< 0.4$ . Green = good or near 1:1, amber = moderate or compressed, red = poor. Within each class, metrics are ordered by correlation. The same values are given in Table S10.

| Metric class             | Metric                          | n   | MAE  | Std. error | r     | Slope | Pattern  | Magnitude  |
|--------------------------|---------------------------------|-----|------|------------|-------|-------|----------|------------|
| Yearly metric            | Q50                             | 34  | 6.8  | 1.0        | 0.59  | 0.57  | Good     | Compressed |
|                          | Weighted mean                   | 34  | 6.3  | 1.1        | 0.58  | 0.55  | Good     | Compressed |
|                          | Q10                             | 34  | 8.8  | 1.2        | 0.58  | 0.56  | Good     | Compressed |
|                          | Q90                             | 34  | 7.8  | 1.1        | 0.58  | 0.54  | Good     | Compressed |
|                          | Season width                    | 34  | 11.8 | 1.5        | 0.53  | 0.54  | Good     | Compressed |
|                          | Peak timing                     | 34  | 10.0 | 1.0        | 0.38  | 0.43  | Poor     | Compressed |
| Spatial band metric      | Q50                             | 166 | 6.4  | 0.9        | 0.57  | 0.53  | Good     | Compressed |
|                          | Weighted mean                   | 166 | 6.1  | 0.9        | 0.55  | 0.52  | Good     | Compressed |
|                          | Q10                             | 166 | 7.8  | 0.9        | 0.54  | 0.51  | Good     | Compressed |
|                          | Q90                             | 166 | 7.7  | 0.8        | 0.48  | 0.42  | Moderate | Compressed |
|                          | Peak timing                     | 166 | 9.1  | 0.9        | 0.42  | 0.46  | Moderate | Compressed |
|                          | Season width                    | 166 | 10.6 | 1.0        | 0.35  | 0.39  | Poor     | Poor       |
| Derived spatial contrast | Weighted mean latitudinal slope | 34  | 2.0  | 0.9        | 0.36  | 0.29  | Poor     | Poor       |
|                          | Q50 latitudinal slope           | 34  | 2.2  | 0.9        | 0.36  | 0.31  | Poor     | Poor       |
|                          | North-south Q50 difference      | 32  | 7.4  | 0.9        | 0.34  | 0.37  | Poor     | Poor       |
| Yearly temporal trend    | Trend Q50                       | 34  | 0.3  | 0.8        | 0.42  | 0.40  | Moderate | Poor       |
|                          | Trend peak timing               | 34  | 0.7  | 0.9        | 0.19  | 0.15  | Poor     | Poor       |
|                          | Trend weighted mean             | 34  | 0.3  | 0.8        | 0.06  | 0.05  | Poor     | Poor       |
|                          | Trend Q10                       | 34  | 0.4  | 0.7        | -0.01 | -0.01 | Poor     | Poor       |
|                          | Trend season width              | 34  | 0.8  | 0.8        | -0.06 | -0.02 | Poor     | Poor       |
|                          | Trend Q90                       | 34  | 0.6  | 0.7        | -0.16 | -0.05 | Poor     | Poor       |
| Spatial temporal trend   | Spatial trend weighted mean     | 170 | 0.3  | 0.7        | 0.33  | 0.30  | Poor     | Poor       |
|                          | Spatial trend Q50               | 170 | 0.3  | 0.7        | 0.29  | 0.20  | Poor     | Poor       |
|                          | Spatial trend Q10               | 170 | 0.3  | 0.7        | 0.20  | 0.12  | Poor     | Poor       |
|                          | Spatial trend Q90               | 170 | 0.5  | 0.4        | -0.01 | -0.00 | Poor     | Poor       |
|                          | Spatial trend season width      | 170 | 0.6  | 0.5        | -0.02 | -0.01 | Poor     | Poor       |

Overall, reliability has two dimensions. GBIF recovers the *pattern* of Q50, the weighted mean, Q10 and Q90 reasonably well, across both years and latitude bands. I recommend Q50 as the main summary, because it is the easiest to interpret and among the most consistently recovered. Ranking metrics within species confirms that none is best everywhere. Q50 is most often ranked first, but only narrowly, Q90 is the weakest of the quantiles, and season width is most often ranked last, especially for accuracy (Figure 7). Q50 is therefore a sensible default, not a uniformly superior choice. However, GBIF *compresses the magnitude* of variation, recovering only about 60% of the true year-to-year variation for Q50 and similar fractions for Q10 and Q90. Even the best-

617 recovered metrics are therefore better suited to relative comparisons than to estimating the  
618 absolute size of phenological change. Peak timing recovers neither pattern nor magnitude well.  
619 Season width tracks the yearly pattern, but its errors are larger than its natural variation, and it is  
620 poorly recovered across latitude bands. Temporal trends are poorly recovered. Yearly Q50 trends  
621 show moderate agreement but are strongly compressed, and trends should not otherwise be used  
622 for the main inference.

623

## Discussion

Occurrence databases such as GBIF are increasingly used to study phenology across taxa and regions. Recent studies have extracted onset, peak, end and season width from decades of citizen-science records of Lepidoptera, across continents, and used them to describe climate-driven shifts and latitudinal gradients in flight timing (Faltýnek Fric et al. 2020; Forsman et al. 2026). These studies assume that occurrence data can recover the same metrics as structured monitoring. That assumption has rarely been tested against an independent benchmark. I tested it by asking not whether GBIF data contain phenological signal, but which metrics from scheme-dominated GBIF data are accurate enough to support ecological inference.

Against the RIS light-trap network, agreement differed among species and depended strongly on the metric. From year to year, the timing quantiles (Q10, Q50 and Q90) and the weighted mean were recovered similarly and reasonably well (good by the thresholds in Table 2, though far from perfect). Season width was recovered with large errors, and peak timing poorly. Across latitude bands, agreement fell from central timing towards the late tail and season width. This matches simulation results showing that phenology estimators work best for central metrics and worse towards the tails when presence-only data are sparse (Belitz et al. 2020). Temporal trends were poorly recovered; yearly trends in Q50 were moderately recovered but strongly compressed.

### *What GBIF can deliver*

Median flight timing (Q50) was among the best-recovered metrics from year to year, the best across latitude bands, and the basis of the most consistent latitudinal contrasts. It is the most dependable single summary. Q50 marks the middle of the flight season, when most of a species' adults are on the wing, and is likely to reflect the temperature and day-length cues that control emergence. But responses to warming go beyond central timing. Season length, the number of generations per year, and the timing of insects relative to their food plants and enemies can all

respond in complex ways that a single central estimate cannot capture (Forrest 2016). That GBIF recovers central timing best and season width least accurately is an important limitation. The same pattern appears in simulations of presence-only data (Belitz et al. 2020) and in opportunistic plant records, where median observation date was the most robust metric across species (Rzanny et al. 2024). Even so, agreement for Q50 was good by the Table 2 thresholds but not strong in absolute terms: species-level correlations averaged 0.59, and, as discussed below, GBIF consistently compressed year-to-year variation.

Compression was present throughout the GBIF data, and it is real, not a statistical artefact. Averaged across species, the within-species slope of GBIF Q50 on RIS Q50 was about 0.57: a ten-day shift in RIS timing appeared as only about a six-day shift in GBIF. Error in the RIS values could, in principle, bias this slope towards zero. However, the RIS Q50 is read from a curve fitted to many moths, so its reliability is high (about 0.95), and correcting for this error raises the slope only to about 0.60 (Table S14). The compression is therefore a property of the occurrence data, not of noise in the benchmark. One plausible mechanism is that in years when a species flies early, not all recorders are out early enough to record it, which pulls the estimated median back towards the long-term average. Simulations predicted this kind of attenuation under sparse or uneven presence-only sampling (Belitz et al. 2020); this study measures it against a structured benchmark.

There are two practical consequences. First, any study that regresses occurrence-derived timing on temperature will underestimate how much earlier a species flies per degree of warming. A true sensitivity of about five days per °C, the order estimated for UK flowering plants from 250 years of structured records (Amano et al. 2010), would appear as about 3.0 days per °C. Second, because GBIF tracks RIS only moderately from year to year, sensitivities estimated from occurrence data are also imprecise. Individual estimates, and especially rankings of species by sensitivity, should be treated with caution. Studies that use GBIF to compare phenological

sensitivity across taxa or to parameterise climate projections (e.g. Forsman et al. 2026) need to allow for both effects. Compression is not a statistical curiosity: it limits what occurrence data can tell us about how species respond to warming.

How well GBIF performed also varied widely among species, so the averages above describe no single species well. For median timing, the best species came close to the benchmark, with correlations near 0.9, errors of two to three days and slopes around 0.8. The weakest species retained little year-to-year signal. The key question for a user is not whether GBIF can recover a metric in general, but for which species it can. Simple traits gave only a partial answer. Wingspan and mean peak timing predicted little, while commoner moths tended to have somewhat better Q50 recovery, in line with the general finding that occurrence data are more reliable for common, well-recorded species (Rzanny et al. 2024).

Two caveats apply. First, each species' agreement rests on only about 26 years of paired values, and fewer for some species. The individual correlations and slopes are uncertain, and the spread among species somewhat overstates the true differences, so no single low-scoring species should be over-interpreted. Second, the well-recovered species are robust: high record density, high correlation, low error and a slope near one do not occur together by chance. Some species are clearly recorded faithfully by GBIF, even if their exact ranking is uncertain. In practice, restricting an analysis to certain kinds of species does not avoid the compression problem, and how well GBIF performs for a given species is hard to predict without testing it against a structured benchmark.

The tail quantiles, Q10 (flight onset) and Q90 (flight end), were recovered about as well as Q50 from year to year. They differed across space: Q10 stayed close to Q50 across latitude bands, but Q90 was clearly weaker. Agreement for the tails was good enough to detect broad differences in onset and end dates among years, species and latitude bands. As for Q50, it was not good enough to measure phenological change precisely, and trends in both tails were not recovered at

all (see below). Studies that use tail metrics from occurrence data to track the advance or delay of the flight season should allow for this uncertainty.

#### *What GBIF cannot reliably deliver*

Season width was the least accurately recovered timing metric. From year to year it tracked the RIS pattern about as well as the quantiles, but its errors were larger than its natural year-to-year variation, and it was poorly recovered across latitude bands. Season width is the difference between two quantiles, so the errors in each tail add together, while the real year-to-year change in season length is small. A metric can rank years correctly and still misjudge the length of the season by more than that length varies. Biologically, season width reflects processes that are hard to capture from sparse records at the edges of the season: how synchronised emergence is, how long adults live, and how quickly the population declines. It matters ecologically because it sets how long interacting species overlap in time. Changes in that overlap, through unequal shifts in predator and prey timing or a shorter period of resource availability, are an important route by which climate change can affect food webs (Burgess et al. 2018). Whether such mismatches are damaging or buffered depends partly on the length of the overlap itself (Weir and Phillimore 2024). Whether species' flight seasons are lengthening or shortening is among the most important questions to ask of phenological data, and it is exactly the question that the GBIF data used here cannot answer reliably.

The most important negative result concerns temporal trends. Estimating trends is one of the most common aims of large-scale phenology studies, and occurrence databases are widely used to track rates of change across species and regions (e.g. Forsman et al. 2026). A clear signal in each year does not guarantee an accurate estimate of change over time. For most metrics, correlations between GBIF and RIS trends were close to zero or negative, even where mean absolute errors were moderate. Low correlation means that rankings of species by their rate of

change are likely to be unreliable, even when average errors look acceptable. Such rankings are central to asking which species respond fastest to warming or which traits predict sensitivity. Trends in median timing were the one partial exception. Across the 34 species, GBIF Q50 trends correlated with RIS trends less well than yearly Q50 values did ( $r = 0.42$ ; slope 0.40). With only 34 species this estimate is imprecise (approximate 95% confidence interval 0.1–0.7), and Q50 trends within latitude bands were weaker ( $r = 0.29$ ). Yearly Q10 values were recovered as well as Q50, but Q10 trends were not, so the problem is not simply that yearly errors add up over time. One possible explanation is that the edges of the recorded flight season are sensitive to long-term changes in recording itself, such as more recorders or changes in when during the season they are active. These add slow, non-biological drift to trend estimates, whereas the median, anchored by the bulk of records, is less affected. For comparing rates of change across species, or for identifying traits linked to faster responses to warming, Q50 is the only metric I can tentatively support. Occurrence data like these may not be enough on their own.

### *The spatial dimension*

GBIF retained clear geographic structure in flight timing across UK latitude bands, with central timing recovered better than the late tail or season width. Within bands, mean Q50 correlations averaged 0.57 across species. This was clearly better than the north–south contrasts and latitudinal slopes derived from the same band-level values, whose mean correlations were 0.34–0.37. This is expected: a contrast between two imperfect estimates combines their errors. Comparing central flight timing directly between northern and southern bands is more trustworthy than estimating the steepness of the latitudinal gradient from GBIF. Spatial contrasts were still recovered better than most trends. This is probably because the difference in timing between Scotland and southern England is large enough for even noisy data to detect, whereas year-to-year shifts of a few days are hard to detect when recording effort varies between years.

Even so, the spatial signal is compressed: where RIS showed a given north–south difference in flight timing, GBIF showed a difference only about 0.37 times as large. GBIF captures the direction of geographic gradients more reliably than their size. This matters for studies that ask whether latitudinal gradients in phenology are becoming shallower or steeper as the climate warms, a question with direct implications for how species interactions will change. Continental-scale studies using citizen-science Lepidoptera data (e.g. Forsman et al. 2026) may recover the broad geographic pattern, but they are likely to underestimate the steepness of the gradient, and therefore any change in it over time.

#### *Effort correction: useful but not transformative*

Accounting for observation effort improved recovery in simulations, but in practice its value depended on how it was done. The cyclic GAM did worse than raw GBIF summaries. The likely reason is that GBIF effort must be inferred, and the inferred effort is sometimes close to zero in some weeks. When a model corrects for effort using these values, a week with few records and a tiny estimated effort is treated as a week of intense activity, which inflates the corrected count and distorts the seasonal curve. The kernel model avoided this by setting a floor on the smallest effort values, which stopped the most extreme distortions from reaching the phenology estimates. Effort correction helps only when it is designed for the quirks of the data, and a more elaborate correction is not automatically better. This fits the wider observation that occurrence records carry several biases that are hard to correct at once (Isaac and Pocock 2015). Recorders cluster in accessible places, some years are covered more intensively than others, and some species are reported more consistently than others. A well-fitting effort model does not guarantee unbiased phenology.

The timing results depend little on the inferred effort. Raw GBIF records, with no effort adjustment, recovered the timing of the flight season about as well as the effort-adjusted model, and agreement held steady, or improved slightly, the less weight the inferred effort carried (Table

S13). The exact rules used to group records into traps therefore have little effect on the conclusions about timing. They matter more for the year-to-year abundance estimates, which depend directly on effort and which I report only as a supplementary check.

#### *When a well-fitting model is not enough*

When a phenology model fits well and produces plausible curves, it is tempting to treat every metric derived from it as equally valid. These results show that this is unsafe. Even the RIS benchmark needed a more flexible curve model, and even then the reliability of the derived metrics differed widely. A study that estimates season width or trends from GBIF without independent validation is making stronger assumptions than the data support, however well the model appears to fit. The same applies to individual species: a method that works well on average can still fail for particular species.

Researchers should therefore validate the specific metrics they intend to use, against the best structured benchmark available for their taxon and region, before treating them as evidence of a biological pattern. This applies well beyond moths and GBIF. Herbarium collections are increasingly used to measure flowering phenology at global scales (Williamson et al. 2025), acoustic monitoring networks are expanding rapidly (Cretois et al. 2026), and automated camera traps are starting to record insect phenology at fine time scales (Høye et al. 2021). All of these data sources share the same problem: some summaries are more robust to their biases than others, and a well-fitting model does not show that a given metric is reliable. The three-stage framework used here (simulation, then structured-data benchmarking, then direct comparison of metrics) can be applied directly to these data sources. As opportunistic and automated data sources grow, the need for this kind of metric-by-metric validation will only increase.

### *Limitations*

The main limit on these findings is the source of the GBIF data: 97.1% of records came from formal recording schemes. The results apply mainly to scheme data, not to the unstructured platform observations that dominate GBIF downloads for many other taxa and regions. RIS is a strong benchmark, but it is itself a monitoring system, not a perfect record of the true seasonal distribution. The validation measures agreement with the best available benchmark, not absolute accuracy. The reliability classes apply to this combination of taxa, region, method and data sources, and should be applied to other systems with care. In particular, researchers using GBIF data dominated by open citizen-science platforms such as iNaturalist, where biases in recording effort may differ substantially (Isaac and Pocock 2015), should treat these classes as a framework to adapt, not as thresholds to import. The underlying argument, that each metric must be validated against a structured benchmark before large-scale phenological inference, applies regardless of taxon, region or data source.

## **Funding**

This work was supported by a RECRUIT Sabbatical grant from Novo Nordisk Foundation (NNF24OC0090844), *Harnessing big data to unravel the impact of climate change on global phenological shifts*.

## **Author contributions**

ORJ conceived and designed the study, conducted the analyses, interpreted the results, and wrote the manuscript.

## **Declaration of AI use**

Claude (Anthropic) was used to assist with developing, debugging, and refactoring analysis code, including consolidating multiple scripts into a more streamlined and reproducible workflow. It was also used to help develop code for generating tables directly from analytical outputs and to improve the clarity and readability of author-written text. Outputs from the revised scripts were systematically compared with those from the originals to verify consistency. All AI-assisted work was reviewed and validated by the author, who takes full responsibility for the accuracy, integrity, and scientific content of the manuscript.

## **Conflict of interest**

The author declares no conflict of interest.

## **Data and code availability**

All code required to reproduce the analyses, tables, and figures is archived at Zenodo (<https://doi.org/10.5281/zenodo.23233550>). GBIF occurrence data were obtained from GBIF and can be accessed via the download DOI <https://doi.org/10.15468/dl.bh2286> (GBIF.org)

2026). Raw trap-level RIS data are held by the Rothamsted Insect Survey and are available to researchers on request. Contains data supplied by the Rothamsted Insect Survey, Rothamsted Research (RIS 2021; extract supplied March 2026). The archive's code is released under the MIT licence. The data and results, which include the RIS-derived tables, are shared under CC BY-NC 4.0 in line with the RIS data sharing statement, that is, for non-commercial use with attribution; RIS trap locations in the archive are rounded to 0.1° and carry no trap names. To support reproducibility, the archived repository includes the derived RIS summaries, processed comparison datasets, and figure inputs used in this study, together with the scripts required to regenerate the manuscript outputs.

## Acknowledgements

I am grateful to Ally Phillimore, his research group, and the wider Institute of Ecology and Evolution at the University of Edinburgh for hosting me during the sabbatical on which this work was carried out, and for valuable discussions that shaped the project. I thank the Rothamsted Insect Survey and the recorders, curators, and data publishers whose observations are mobilised through GBIF. I also thank the GBIF team for curating and managing the database and for providing access through the GBIF website. I specifically thank Tracey Kruger at Rothamsted Research for preparing the RIS data for this study. Finally, I thank my wife, Julia Jones, and our children for their patience, love, and support during the extended periods I spent away from home during my sabbatical.

## Supplementary information

The supplementary materials include the following items:

- **Figure S1:** Raw versus effort-aware GBIF yearly timing-metric recovery against RIS.
- **Figure S2:** Yearly abundance-index recovery against RIS for raw GBIF totals and simple effort-standardised GBIF totals.
- **Figure S3:** Recovery of species-specific yearly temporal slopes from GBIF (poor; moderate for Q50).
- **Figure S4:** RIS-only spatial validation of the two-scale asymmetric-kernel RIS estimator against empirical RIS.
- **Figure S5:** Recovery of derived temporal slopes within the UK spatial workflow.
- **Figure S6:** Sensitivity of single-scale asymmetric-kernel curve shape to broader kernel settings in focal species-years.
- **Figure S7:** Full species-level yearly Q10 (onset) agreement between modelled RIS and GBIF, for all 34 species.
- **Figure S8:** Full species-level yearly Q50 (median) agreement between modelled RIS and GBIF, for all 34 species.
- **Figure S9:** Full species-level yearly Q90 (end) agreement between modelled RIS and GBIF, for all 34 species.
- **Figure S10:** Full species-level yearly season width (flight period length) agreement between modelled RIS and GBIF, for all 34 species.
- **Table S1:** Summary of method-validation performance across simulation and empirical RIS benchmarks.
- **Table S2:** Species-level agreement between modelled yearly RIS and GBIF phenology metrics.

- **Table S3:** Comparison of raw and effort-aware GBIF yearly phenology metrics against RIS.
- **Table S4:** Comparison of yearly abundance-index recovery for raw GBIF totals and simple effort-standardised GBIF totals against RIS.
- **Table S5:** Agreement between RIS and GBIF for species-specific yearly temporal slopes.
- **Table S6:** Species-level agreement between modelled spatial RIS and GBIF phenology metrics within UK latitude bands.
- **Table S7:** RIS-only spatial validation of the two-scale asymmetric-kernel RIS model against the empirical RIS benchmark.
- **Table S8:** Species-level agreement between RIS and GBIF for derived modelled spatial contrasts.
- **Table S9:** Species-level agreement between RIS and GBIF for derived temporal slopes within UK latitude bands.
- **Table S10:** Underlying values for the two-axis reliability summary (Table 2): MAE, standardised error, correlation, slope, and the pattern- and magnitude-recovery classes.
- **Table S11:** Focused sensitivity summary for alternative single-scale asymmetric-kernel hyperparameter settings in focal species-years.
- **Table S12:** Source composition of the focal UK GBIF benchmark by dataset category and record type.
- **Table S13:** Sensitivity of yearly Q50 agreement to alternative GBIF exposure-floor values.
- **Table S14:** Decomposition of the yearly GBIF-on-RIS slope (tracking versus amplitude) and its correction for measurement error in the RIS benchmark.

- 895       • **Table S15:** Sensitivity of the RIS reliability and reliability-corrected slope to the assumed  
896       overdispersion of the catch.
- 897       • **Table S16:** Effect of the RIS-derived GBIF records on spatial agreement with RIS  
898       (records kept versus removed).
- 899

## References

- Amano, T., Smithers, R. J., Sparks, T. H. and Sutherland, W. J. 2010. A 250-year index of first flowering dates and its response to temperature changes. – *Proc. R. Soc. B* 277: 2451–2457.
- Aono, Y. and Kazui, K. 2008. Phenological data series of cherry tree flowering in Kyoto, Japan, and its application to reconstruction of springtime temperatures since the 9th century. – *Int. J. Climatol.* 28: 905–914.
- Beck, J., Böller, M., Erhardt, A. and Schwanghart, W. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. – *Ecol. Inform.* 19: 10–15.
- Belitz, M. W., Larsen, E. A., Ries, L. and Guralnick, R. P. 2020. The accuracy of phenology estimators for use with sparsely sampled presence-only observations. – *Methods Ecol. Evol.* 11: 1273–1285.
- Both, C., Artemyev, A. V., Blaauw, B., Cowie, R. J., Dekhuijzen, A. J., Eeva, T., Enemar, A., Gustafsson, L., Ivankina, E. V., Järvinen, A., Metcalfe, N. B., Nyholm, N. E. I., Potti, J., Ravussin, P.-A., Sanz, J. J., Silverin, B., Slater, F. M., Sokolov, L. V., Török, J., Winkel, W., Wright, J. and Visser, M. E. 2004. Large-scale geographical variation confirms that climate change causes birds to lay earlier. – *Proc. R. Soc. B* 271: 1657–1662.
- Both, C., van Asch, M., Bijlsma, R. G., van den Burg, A. B. and Visser, M. E. 2009. Climate change and unequal phenological changes across four trophic levels: constraints or adaptations? – *J. Anim. Ecol.* 78: 73–83.
- Burgess, M. D., Smith, K. W., Evans, K. L., Leech, D., Pearce-Higgins, J. W., Branston, C. J., Briggs, K., Clark, J. R., du Feu, C. R., Lewthwaite, K., Nager, R. G., Sheldon, B. C., Smith, J. A., Whytock, R. C., Willis, S. G. and Phillimore, A. B. 2018. Tritrophic phenological match–mismatch in space and time. – *Nat. Ecol. Evol.* 2: 970–975.

923 Cereghetti, E., Ajvazi, X., Bossart, R., Keck, F., Patrignani, A., Tartini, N., Altermatt, F. and  
 924 Carraro, L. 2026. Tracking the phenology of riverine insect communities using environmental  
 925 DNA. – *Environ. DNA* 8: e70249.

926 Charmantier, A. and Gienapp, P. 2013. Climate change and timing of avian breeding and  
 927 migration: evolutionary versus plastic changes. – *Evol. Appl.* 7: 15–28.

928 Cretois, B., Rosten, C. M., Wiel, J., Barile, C., McEwen, B., Bernard, C., [AUTHOR LIST TO  
 929 COMPLETE] and Sethi, S. 2026. TABMON: design and deployment of a transnational passive  
 930 acoustic monitoring network for European birds. – *Methods Ecol. Evol.* 17: 1867–1879.

931 Crick, H. Q. P. 2004. The impact of climate change on birds. – *Ibis* 146: 48–56.

932 Crick, H. Q. P. and Sparks, T. H. 1999. Climate change related to egg-laying trends. – *Nature*  
 933 399: 423–424.

934 Di Cecco, G. J., Barve, V., Belitz, M. W., Stucky, B. J., Guralnick, R. P. and Hurlbert, A. H. 2021.  
 935 Observing the observers: how participants contribute data to iNaturalist and implications for  
 936 biodiversity science. – *BioScience* 71: 1179–1188.

937 Donnelly, A., Caffarra, A. and O'Neill, B. F. 2011. A review of climate-driven mismatches  
 938 between interdependent phenophases in terrestrial and aquatic ecosystems. – *Int. J. Biometeorol.*  
 939 55: 805–817.

940 Edwards, M. and Richardson, A. J. 2004. Impact of climate change on marine pelagic phenology  
 941 and trophic mismatch. – *Nature* 430: 881–884.

942 Faltýnek Fric, Z., Rindoš, M. and Konvička, M. 2020. Phenology responses of temperate  
 943 butterflies to latitude depend on ecological traits. – *Ecol. Lett.* 23: 172–180.

944 Fitter, A. H. and Fitter, R. S. R. 2002. Rapid changes in flowering time in British plants. –  
 945 *Science* 296: 1689–1691.

946 Forrest, J. R. K. 2016. Complex responses of insect phenology to climate change. – *Curr. Opin.*  
 947 *Insect Sci.* 17: 49–54.

948 Forsman, A., Karimi, B. and Franzén, M. 2026. Contrasting phenological shifts in diurnal and  
 949 nocturnal Lepidoptera under climate change. – *Commun. Biol.* 9: 538.

950 GBIF.org 2026. GBIF occurrence download, 4 April 2026. –  
 951 <https://doi.org/10.15468/dl.bh2286>.

952 Gienapp, P., Reed, T. E. and Visser, M. E. 2014. Why climate change will invariably alter  
 953 selection pressures on phenology. – *Proc. R. Soc. B* 281: 20141611.

954 Goodenough, A. E., Hart, A. G. and Stafford, R. 2010. Is adjustment of breeding phenology  
 955 keeping pace with the need for change? Linking observed response in woodland birds to changes  
 956 in temperature and selection pressure. – *Clim. Change* 102: 687–697.

957 Gordo, O. 2007. Why are bird migration dates shifting? A review of weather and climate effects  
 958 on avian migratory phenology. – *Clim. Res.* 35: 37–58.

959 Graves, S. and Manzitto-Tripp, E. A. 2026. Observing shifts in phenology of tropical flowering  
 960 plants. – *PLoS One* 21: e0342105.

961 Hahsler, M., Piekenbrock, M. and Doran, D. 2019. dbscan: fast density-based clustering with R.  
 962 – *J. Stat. Softw.* 91: 1–30.

963 Hällfors, M. H., Pöyry, J., Heliölä, J., Kohonen, I., Kuussaari, M., Leinonen, R., Schmucki, R.,  
 964 Sihvonen, P. and Saastamoinen, M. 2021. Combining range and phenology shifts offers a  
 965 winning strategy for boreal Lepidoptera. – *Ecol. Lett.* 24: 1619–1632.

966 Harrington, R. and Woiwod, I. 2007. Foresight from hindsight: the Rothamsted Insect Survey. –  
 967 *Outlooks Pest Manage.* 18: 9–14.

968   Hickinbotham, E. J., Pattison, Z., Fox, R. and Rushton, S. P. 2024. Drivers of moth phenology  
 969   in England and Wales. – *J. Insect Conserv.* 28: 969–979.

970   Høye, T. T., Ärje, J., Bjerger, K., Hansen, O. L., Iosifidis, A., Leese, F., Mann, H. M., Meissner,  
 971   K., Melvad, C. and Raitoharju, J. 2021. Deep learning and computer vision will transform  
 972   entomology. – *Proc. Natl Acad. Sci. USA* 118: e2002545117.

973   Isaac, N. J. B. and Pocock, M. J. O. 2015. Bias and information in biological records. – *Biol. J.*  
 974   *Linn. Soc.* 115: 522–531.

975   Jarić, I., Correia, R. A., Brook, B. W., Buettel, J. C., Courchamp, F., Di Minin, E., Firth, J. A.,  
 976   Gaston, K. J., Jepson, P., Kalinkat, G., Ladle, R., Soriano-Redondo, A., Souza, A. T. and Roll, U.  
 977   2020. iEcology: harnessing large online resources to generate ecological insights. – *Trends Ecol.*  
 978   *Evol.* 35: 630–639.

979   Jones, T. and Cresswell, W. 2010. The phenology mismatch hypothesis: are declines of migrant  
 980   birds linked to uneven global climate change? – *J. Anim. Ecol.* 79: 98–108.

981   Klosterman, S. and Richardson, A. D. 2017. Observing spring and fall phenology in a deciduous  
 982   forest with aerial drone imagery. – *Sensors* 17: 2852.

983   Larsen, E. A. and Shirey, V. 2021. Method matters: pitfalls in analysing phenology from  
 984   occurrence records. – *Ecol. Lett.* 24: 1287–1289.

985   Mason, B. M., Mesaglio, T., Barratt Heitmann, J., Chandler, M., Chowdhury, S., Gorta, S. B. Z.,  
 986   Grattarola, F., Groom, Q., Hitchcock, C., Hoskins, L., Lowe, S. K., Marquis, M., Pernat, N.,  
 987   Shirey, V., Baasanmunkh, S. and Callaghan, C. T. 2025. iNaturalist accelerates biodiversity  
 988   research. – *BioScience* 75: 953–965.

989   Mommott, J., Craze, P. G., Waser, N. M. and Price, M. V. 2007. Global warming and the  
 990   disruption of plant–pollinator interactions. – *Ecol. Lett.* 10: 710–717.

991 Menzel, A., Sparks, T. H., Estrella, N., Koch, E., Aasa, A., Ahas, R., Alm-Kübler, K., Bissolli, P.,  
 992 Braslavská, O., Briede, A., Chmielewski, F. M., Crepinsek, Z., Curnel, Y., Dahl, Å., Defila, C.,  
 993 Donnelly, A., Filella, Y., Jatzczak, K., Måge, F., Mestre, A., Nordli, Ø., Peñuelas, J., Pirinen, P.,  
 994 Remišová, V., Scheifinger, H., Striz, M., Susnik, A., Van Vliet, A. J. H., Wielgolaski, F.-E., Zach,  
 995 S. and Züst, A. 2006. European phenological response to climate change matches the warming  
 996 pattern. – *Glob. Change Biol.* 12: 1969–1976.

997 Parmesan, C. and Yohe, G. 2003. A globally coherent fingerprint of climate change impacts  
 998 across natural systems. – *Nature* 421: 37–42.

999 Pebesma, E. 2018. Simple features for R: standardized support for spatial vector data. – *R J.* 10:  
 1000 439–446.

1001 Poloczanska, E. S., Brown, C. J., Sydeman, W. J., Kiessling, W., Schoeman, D. S., Moore, P. J.,  
 1002 Brander, K., Bruno, J. F., Buckley, L. B., Burrows, M. T., Duarte, C. M., Halpern, B. S., Holding,  
 1003 J., Kappel, C. V., O'Connor, M. I., Pandolfi, J. M., Parmesan, C., Schwing, F., Thompson, S. A.  
 1004 and Richardson, A. J. 2013. Global imprint of climate change on marine life. – *Nat. Clim.*  
 1005 *Change* 3: 919–925.

1006 R Core Team 2026. R: a language and environment for statistical computing. – R Foundation for  
 1007 Statistical Computing, <https://www.R-project.org/>.

1008 Reséndiz-Infante, C. and Gauthier, G. 2020. Temporal changes in reproductive success and  
 1009 optimal breeding decisions in a long-distance migratory bird. – *Sci. Rep.* 10: 22067.

1010 Richardson, A. D., Hufkens, K., Milliman, T. and Froliking, S. 2018. Intercomparison of  
 1011 phenological transition dates derived from the PhenoCam dataset V1.0 and MODIS satellite  
 1012 remote sensing. – *Sci. Rep.* 8: 5679.

1013 RIS 2021. Rothamsted Insect Survey Online Database, extract supplied March 2026. –  
 1014 Rothamsted Research, <https://doi.org/10.23637/rothamsted.987q7>.

- 1015 Root, T. L., Price, J. T., Hall, K. R., Schneider, S. H., Rosenzweig, C. and Pounds, J. A. 2003.  
1016 Fingerprints of global warming on wild animals and plants. – *Nature* 421: 57–60.
- 1017 Roy, D. B., Ploquin, E. F., Randle, Z., Risely, K., Botham, M. S., Middlebrook, I., Noble, D.,  
1018 Cruickshanks, K., Freeman, S. N. and Brereton, T. M. 2014. Comparison of trends in butterfly  
1019 populations between monitoring schemes. – *J. Insect Conserv.* 19: 313–324.
- 1020 Roy, D. B. and Sparks, T. H. 2000. Phenology of British butterflies and climate change. – *Glob.*  
1021 *Change Biol.* 6: 407–416.
- 1022 Rzanny, M., Mäder, P., Wittich, H. C., Boho, D. and Wäldchen, J. 2024. Opportunistic plant  
1023 observations reveal spatial and temporal gradients in phenology. – *npj Biodivers.* 3: 5.
- 1024 Sanz, J. J. 2002. Climate change and breeding parameters of great and blue tits throughout the  
1025 western Palearctic. – *Glob. Change Biol.* 8: 409–422.
- 1026 Templ, B., Koch, E., Bolmgren, K., Ungersböck, M., Paul, A., Scheifinger, H., Rutishauser, T.,  
1027 Busto, M., Chmielewski, F.-M., Hájková, L., Hodžić, S., Kaspar, F., Pietragalla, B., Romero-  
1028 Fresneda, R., Tolvanen, A., Vučetić, V., Zimmermann, K. and Zust, A. 2018. Pan European  
1029 Phenological database (PEP725): a single point of access for European data. – *Int. J.*  
1030 *Biometeorol.* 62: 1109–1113.
- 1031 Thackeray, S. J., Sparks, T. H., Frederiksen, M., Burthe, S., Bacon, P. J., Bell, J. R., Botham, M.  
1032 S., Brereton, T. M., Bright, P. W., Carvalho, L., Clutton-Brock, T., Dawson, A., Edwards, M.,  
1033 Elliott, J. M., Harrington, R., Johns, D., Jones, I. D., Jones, J. T., Leech, D. I., Roy, D. B., Scott,  
1034 W. A., Smith, M., Smithers, R. J., Winfield, I. J. and Wanless, S. 2010. Trophic level asynchrony  
1035 in rates of phenological change for marine, freshwater and terrestrial environments. – *Glob.*  
1036 *Change Biol.* 16: 3304–3313.
- 1037 Thorup, K., Tøttrup, A. P. and Rahbek, C. 2007. Patterns of phenological changes in migratory  
1038 birds. – *Oecologia* 151: 697–703.

- 1039 Tulloch, A. I. T., Possingham, H. P., Joseph, L. N., Szabo, J. and Martin, T. G. 2013. Realising  
1040 the full potential of citizen science monitoring programs. – *Biol. Conserv.* 165: 128–138.
- 1041 Van Buskirk, J., Mulvihill, R. S. and Leberman, R. C. 2009. Variable shifts in spring and autumn  
1042 migration phenology in North American songbirds associated with climate change. – *Glob.*  
1043 *Change Biol.* 15: 760–771.
- 1044 Visser, M. E. and Both, C. 2005. Shifts in phenology due to global climate change: the need for a  
1045 yardstick. – *Proc. R. Soc. B* 272: 2561–2569.
- 1046 Weir, J. C. and Phillimore, A. B. 2024. Buffering and phenological mismatch: a change of  
1047 perspective. – *Glob. Change Biol.* 30: e17294.
- 1048 Wickham, H. 2016. *ggplot2: elegant graphics for data analysis*. – Springer.
- 1049 Wickham, H., François, R., Henry, L., Müller, K. and Vaughan, D. 2026. *dplyr: a grammar of*  
1050 *data manipulation*. – R package ver. 1.2.0, <https://doi.org/10.32614/CRAN.package.dplyr>.
- 1051 Wickham, H., Vaughan, D. and Girlich, M. 2025. *tidyr: tidy messy data*. – R package ver. 1.3.2,  
1052 <https://doi.org/10.32614/CRAN.package.tidyr>.
- 1053 Williamson, D. R., Prestø, T., Westergaard, K. B., Trascau, B. M., Vange, V., Hassel, K., Koch,  
1054 W. and Speed, J. D. 2025. Long-term trends in global flowering phenology. – *New Phytol.*,  
1055 <https://doi.org/10.1111/nph.70139>.
- 1056 Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation  
1057 of semiparametric generalized linear models. – *J. R. Stat. Soc. B* 73: 3–36.
- 1058 Wood, S. N. 2017. *Generalized additive models: an introduction with R*, 2nd edn. – Chapman  
1059 and Hall/CRC.

# Supplementary Information

## Supplementary Methods

This section provides the detailed parameter values and model specifications referred to in the main-text Methods. Sections follow the order in which the corresponding material appears in the main text. Supplementary figures and tables referenced here are provided in the sections that follow.

**Data sources.** Contains data supplied by the Rothamsted Insect Survey, Rothamsted Research (RIS 2021, <https://doi.org/10.23637/rothamsted.987q7>; extract supplied March 2026). GBIF occurrence data are from GBIF.org (2026), GBIF occurrence download, 4 April 2026, <https://doi.org/10.15468/dl.bh2286>.

### *S1. Simulation parameters*

**Seasonal abundance curves.** Synthetic weekly phenology data were generated under two regimes. In the symmetric regime, the true daily abundance curve was Gaussian, with annual peak date drawn uniformly on day 60 to 250, peak width (standard deviation) on 10 to 20 days, and total annual abundance on 10,000 to 800,000 individuals. In the asymmetric regime, the curve was split-normal, with peak date drawn on day 70 to 240, left-side width on 18 to 32 days, right-side width on 7 to 14 days, and the same total abundance range. In both regimes I simulated ten years across five latitudinal areas, imposing fixed area-level shifts in timing and abundance so that the benchmark contained both year-to-year and spatial structure.

**Observer effort curves.** On top of each true abundance curve I imposed a separate Gaussian effort curve representing observer effort through the year. In the symmetric regime the effort peak date was drawn uniformly on day 150 to 200 and its spread on 20 to 60 days; in the asymmetric regime the effort peak date was drawn on day 150 to 210 and its spread on 20 to 60

days. Maximum daily sampling effort was fixed at 20 traps in both regimes. I added stochastic noise on the log scale so that effort varied randomly around this seasonal pattern. Observed counts were then drawn from a Poisson distribution whose expected value was proportional to the product of true abundance, sampling effort, and a fixed detection probability of 0.06.

## *S2. Two-scale asymmetric-kernel model specification*

For a given species, year, and (where relevant) latitude band, weekly expected abundance was represented as the sum of two non-negative kernel banks, one broad and one narrow:

$$\mu_t = E_t \left[ \sum_{j=1}^J w_j^{(B)} K_j^{(B)}(t) + \sum_{j=1}^J w_j^{(N)} K_j^{(N)}(t) \right],$$

where  $t$  is week,  $E_t$  is the exposure (the trapping effort behind each weekly count),  $w_j^{(B)} \geq 0$  and  $w_j^{(N)} \geq 0$  are the broad-bank and narrow-bank kernel weights, and  $K_j^{(B)}(t)$  and  $K_j^{(N)}(t)$  are kernels centred at regularly spaced seasonal positions. Each kernel was asymmetric, with separate left and right standard deviations:

$$K_j(t) \propto \begin{cases} \exp \left[ -\frac{(t - c_j)^2}{2\sigma_L^2} \right], & t < c_j \\ \exp \left[ -\frac{(t - c_j)^2}{2\sigma_R^2} \right], & t \geq c_j, \end{cases}$$

where  $c_j$  is the kernel centre,  $\sigma_L$  is the left-side width, and  $\sigma_R$  is the right-side width. I estimated kernel weights under non-negativity constraints with separate penalties,  $\lambda_B \sum_j w_j^{(B)} + \lambda_N \sum_j w_j^{(N)}$ , ensuring the model cannot predict negative abundance. I fitted the model on the weekly scale and then converted the fitted seasonal distribution to day-of-year summaries for comparison.

**Hyperparameter tuning.** The settings that control how flexible the fitted curves can be (the hyperparameters) were chosen in two steps. I first tuned a single-scale asymmetric kernel by varying kernel spacing, left and right widths, and penalty strength, yielding a 7-day spacing, left standard deviation of 21 days, right standard deviation of 10 days, and sparsity penalty  $\lambda = 0.25$ . After inspecting focal species-years with sharp peaks and broad shoulders, I then extended this to a two-scale form combining a broad bank (left SD 28 days, right SD 14 days,  $\lambda_B = 0.10$ ) and a narrow bank (left SD 10 days, right SD 5 days,  $\lambda_N = 0.35$ ), both sharing the same 7-day spacing. The 7-day spacing matched the weekly support of the RIS data and avoided introducing artificial sub-week structure into the fitted curves. The broad bank captured the shoulders of the seasonal distribution and the narrow bank the sharp central peaks. Across symmetric simulation, asymmetric simulation, and empirical RIS validation, this two-scale variant outperformed both the single-scale asymmetric kernel and the GAM-based alternatives, so I retained it as the preferred estimator.

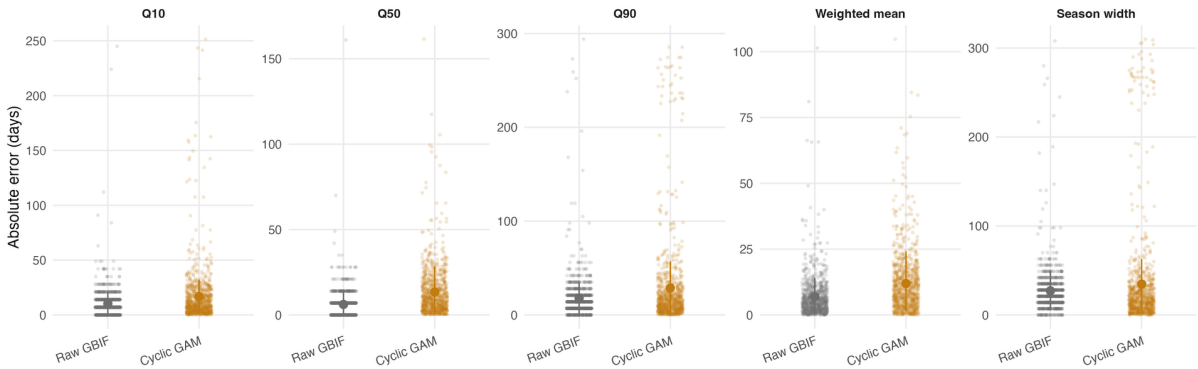
### *S3. GBIF exposure floor sensitivity*

For the matched kernel-based GBIF comparisons, inferred weekly exposure  $E_t$  was floored at 50 inferred trap-nights, that is, replaced with  $\max(E_t, 50)$ . This stabilised the effort-aware implementation against the implausibly large count-per-effort rates that arose when inferred weekly effort was very small. The floor of 50 was adopted as a moderate regularisation choice rather than the value that optimised yearly Q50. A sensitivity analysis across alternative floor values (Table S13) showed that yearly Q50 agreement continued to improve up to floors of around 200 inferred trap-nights, but at the cost of progressively more aggressive modification of the inferred effort proxy. The floor of 50 therefore represents a conservative compromise between stabilising the fit and preserving fidelity to the inferred effort signal.

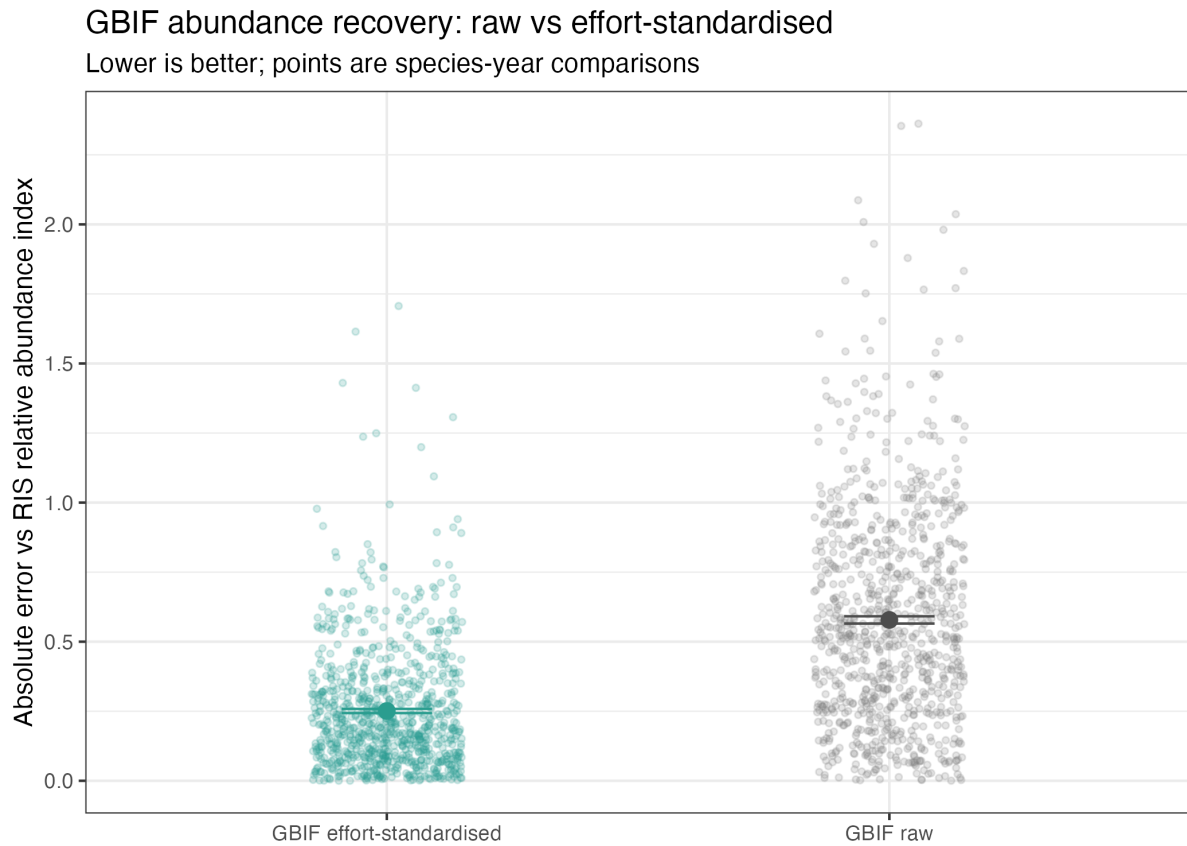
Supplementary Figures

This supplement provides additional figures and numerical tables supporting the main text.

Figure S1



**Figure S1.** Effort-aware smoothing does not uniformly improve yearly timing recovery against RIS. Panels show absolute-error distributions for the five focal yearly metrics under raw GBIF summaries and a simple per-year cyclic GAM. Points show species-year comparisons, and overlaid pointranges show the mean absolute error with the 10th and 90th percentiles.

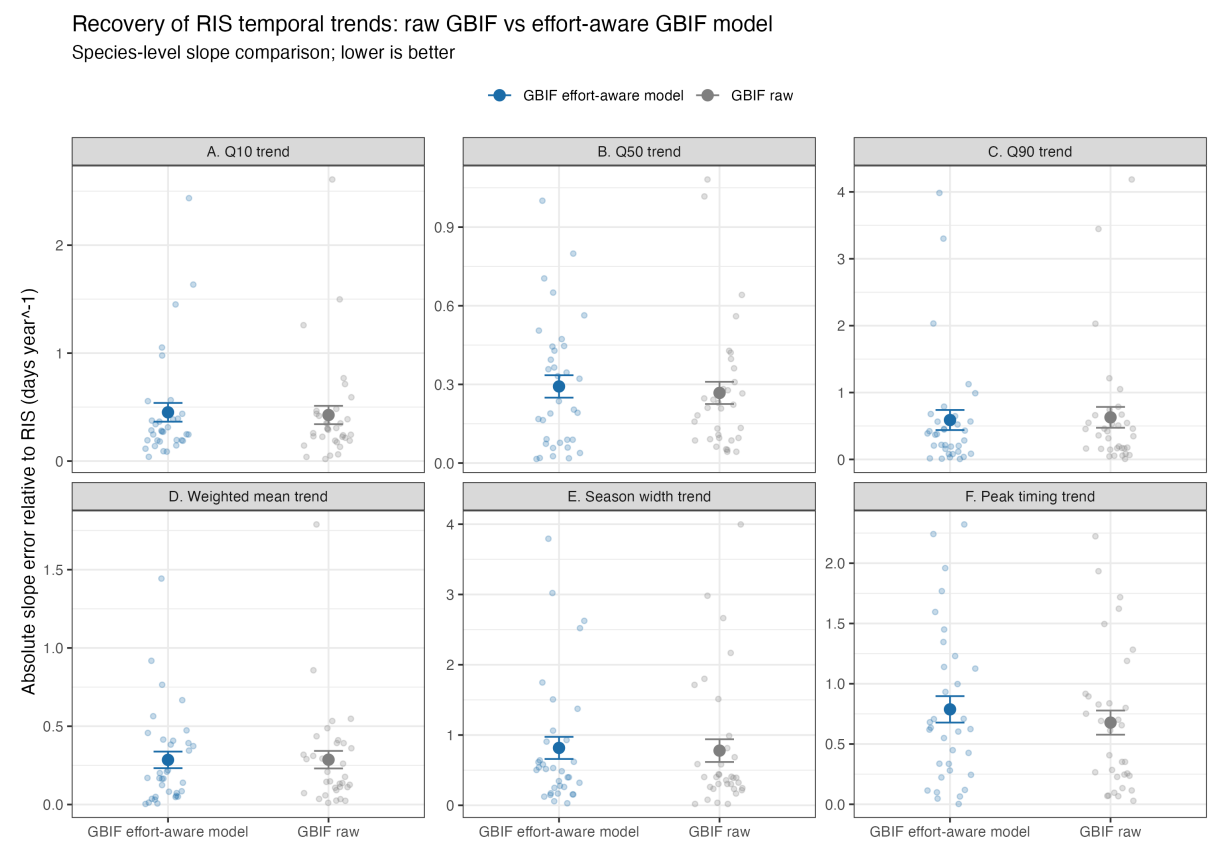


**Figure S2.** Simple effort standardisation improves empirical recovery of yearly abundance. The plotted distributions summarise absolute error in the yearly abundance index for raw GBIF totals and simple effort-standardised GBIF totals, both evaluated against RIS. Points show paired species-year values and overlaid summaries show mean error and uncertainty.

This abundance analysis was intended as a supporting validation rather than a main phenology result. In the empirical RIS–GBIF comparison, simple effort standardisation clearly improved recovery of the yearly abundance index: the effort-standardised index achieved Pearson  $r = 0.657$  against the RIS abundance index, compared with  $r = 0.404$  for raw GBIF individual totals, and mean absolute error fell from 0.578 for raw totals to 0.251 after standardisation. At the species level, agreement was mixed but again improved markedly with standardisation, with species-level correlations averaging  $0.56 \pm 0.25$  for the effort-standardised index against  $0.30 \pm$

0.37 for raw GBIF. Some species showed strong year-to-year agreement, for example *Orthosia*  
*cruda* ( $r = 0.801$  raw;  $r = 0.894$  effort-standardised) and *Agrotis exclamationis* ( $r = 0.691$  raw;  
 $r = 0.770$  effort-standardised), but this was not simply a matter of abundance: mean RIS  
abundance was only weakly associated with species-level agreement ( $r = -0.106$  for raw GBIF  
abundance and  $r = -0.131$  for effort-standardised abundance). Together these results indicate  
that abundance recovery is feasible in principle, but is materially less robust than recovery of  
central timing metrics in the empirical GBIF benchmark unless the counts are standardised by  
the inferred effort proxy.

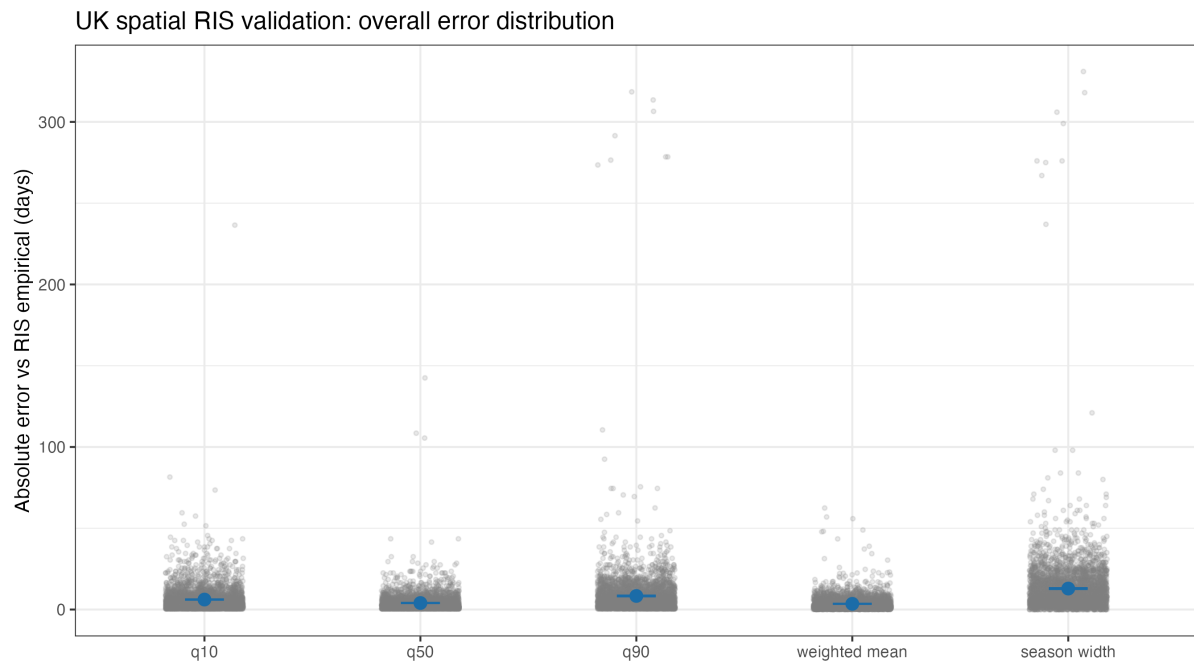
Figure S3



**Figure S3.** GBIF recovers through-time phenological slopes much less reliably than yearly  
timing itself, including for Q50, where agreement is moderate but strongly compressed. Panels  
compare raw and effort-aware GBIF estimates of species-specific temporal slopes for each

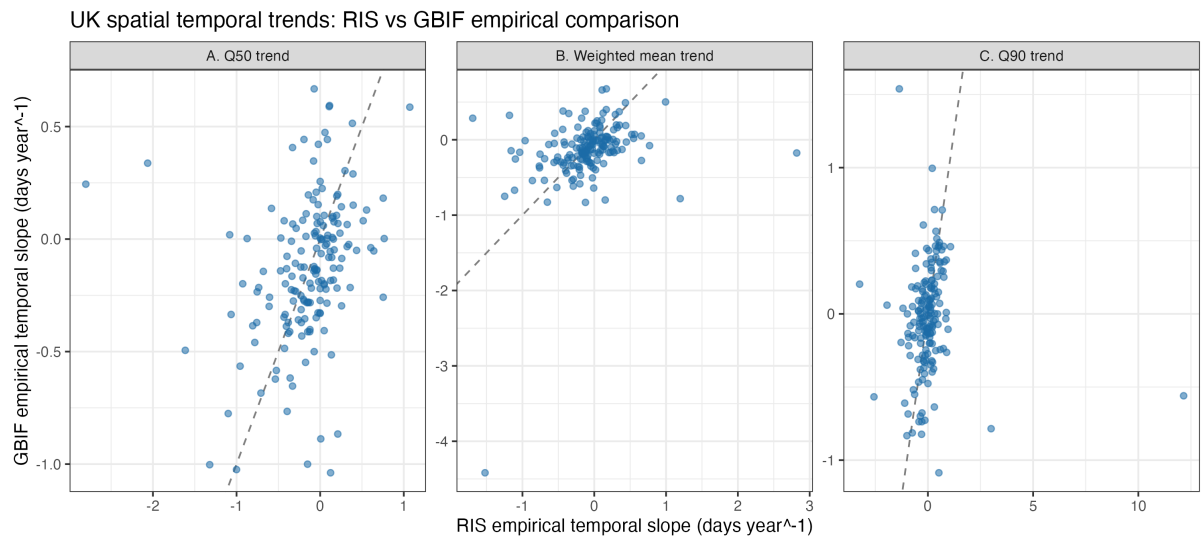
phenological metric. Panels A-F show Q10 trend, Q50 trend, Q90 trend, weighted mean trend, season width trend, and peak timing trend, respectively. Points show species-level slope comparisons and overlaid summaries show mean error and standard error.

*Figure S4*



**Figure S4.** The two-scale asymmetric-kernel RIS estimator closely reproduces the empirical RIS benchmark for the central timing metrics used in spatial validation. The figure summarises absolute error for the RIS phenology model relative to empirical RIS phenology within UK latitude bands. Points show species-year-band values and summary markers show average error across the benchmark set.

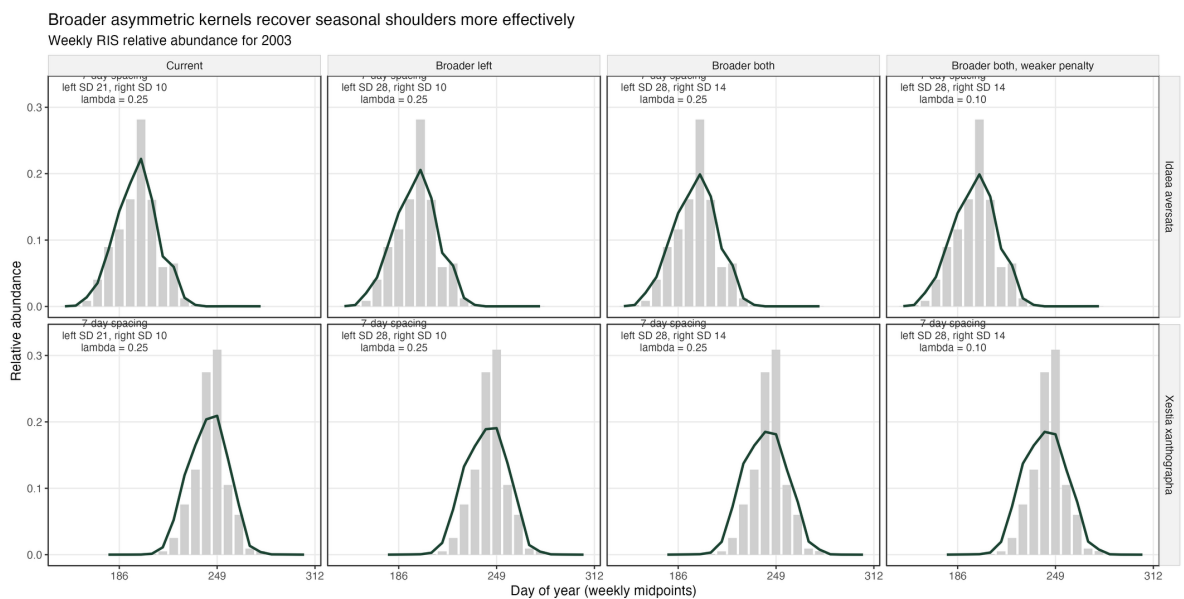
110 *Figure S5*



111

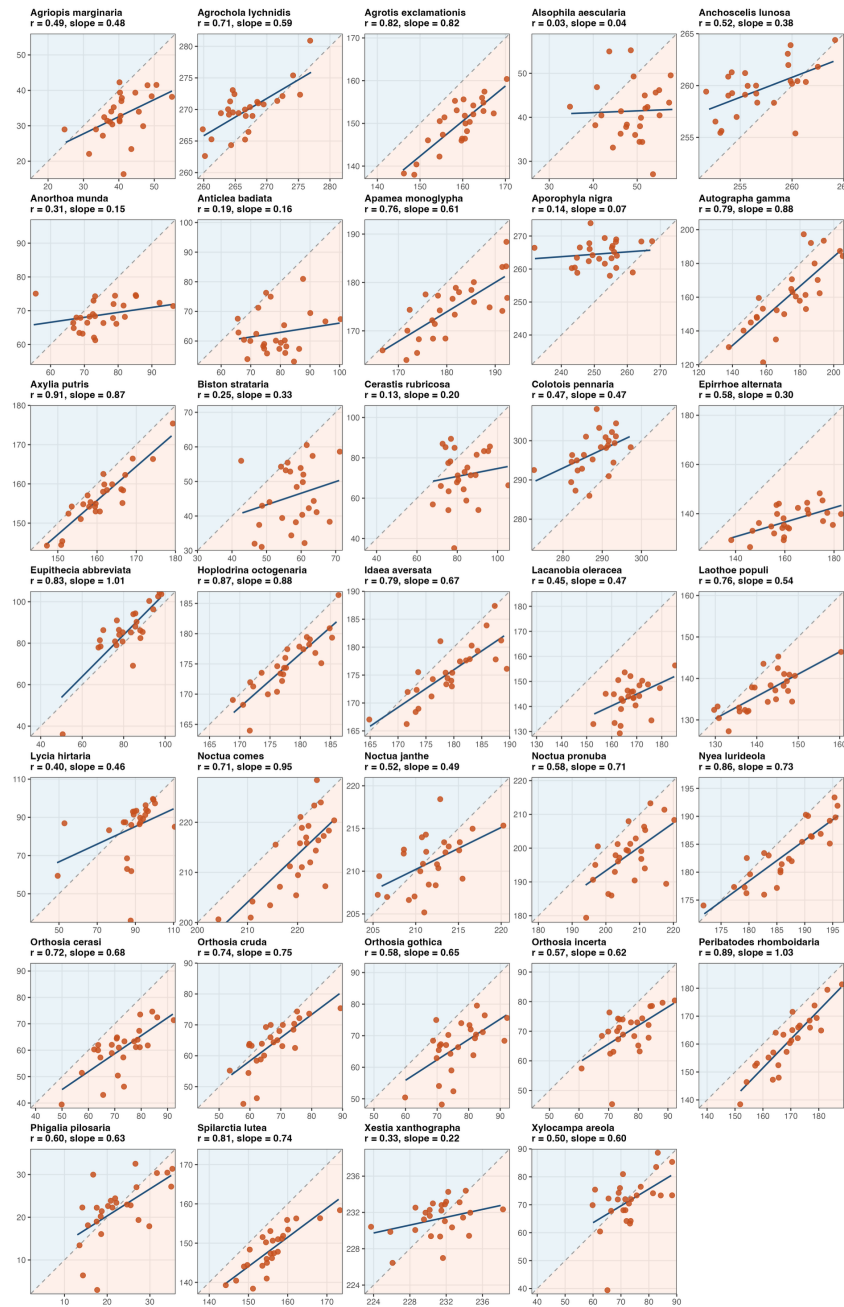
112 **Figure S5.** Derived temporal slopes remain weakly recovered even within the empirical spatial  
113 workflow. Each panel summarises species-by-band trend comparisons for one phenological  
114 metric, with the dashed line indicating 1:1 agreement between RIS and GBIF. Panels A-C show  
115 Q50 trend, Q90 trend, and weighted mean trend, respectively. Points show the underlying  
116 comparisons.

117 *Figure S6*



118

**Figure S6.** Broadening the single-scale asymmetric kernel fills some shoulders but also flattens the peak. Grey bars show weekly empirical RIS relative abundance and green lines show exposure-adjusted fitted weekly expected abundance under alternative single-scale asymmetric-kernel settings. Columns compare the tuned single-scale setting with broader-kernel alternatives, and rows show focal species-years for *Idaea versata* and *Xestia xanthographa* in 2003.



**Figure S7.** Full species-level yearly agreement in modelled Q10 (flight onset) between RIS and GBIF under the matched two-scale asymmetric-kernel workflow, for all species in the focal set. Figure 5 in the main text shows nine of these species; this figure shows the complete set. Each panel is one species, with one point per year, using the RIS two-scale asymmetric-kernel Q10 on the x-axis and the GBIF effort-aware two-scale asymmetric-kernel Q10 on the y-axis. Dashed lines indicate 1:1 agreement, fitted lines summarise the within-species relationship, panel titles

133 report the species-level correlation and GBIF-on-RIS slope, and the shaded triangles indicate  
134 years where GBIF is later than RIS (blue, upper left) or earlier than RIS (orange, lower right).

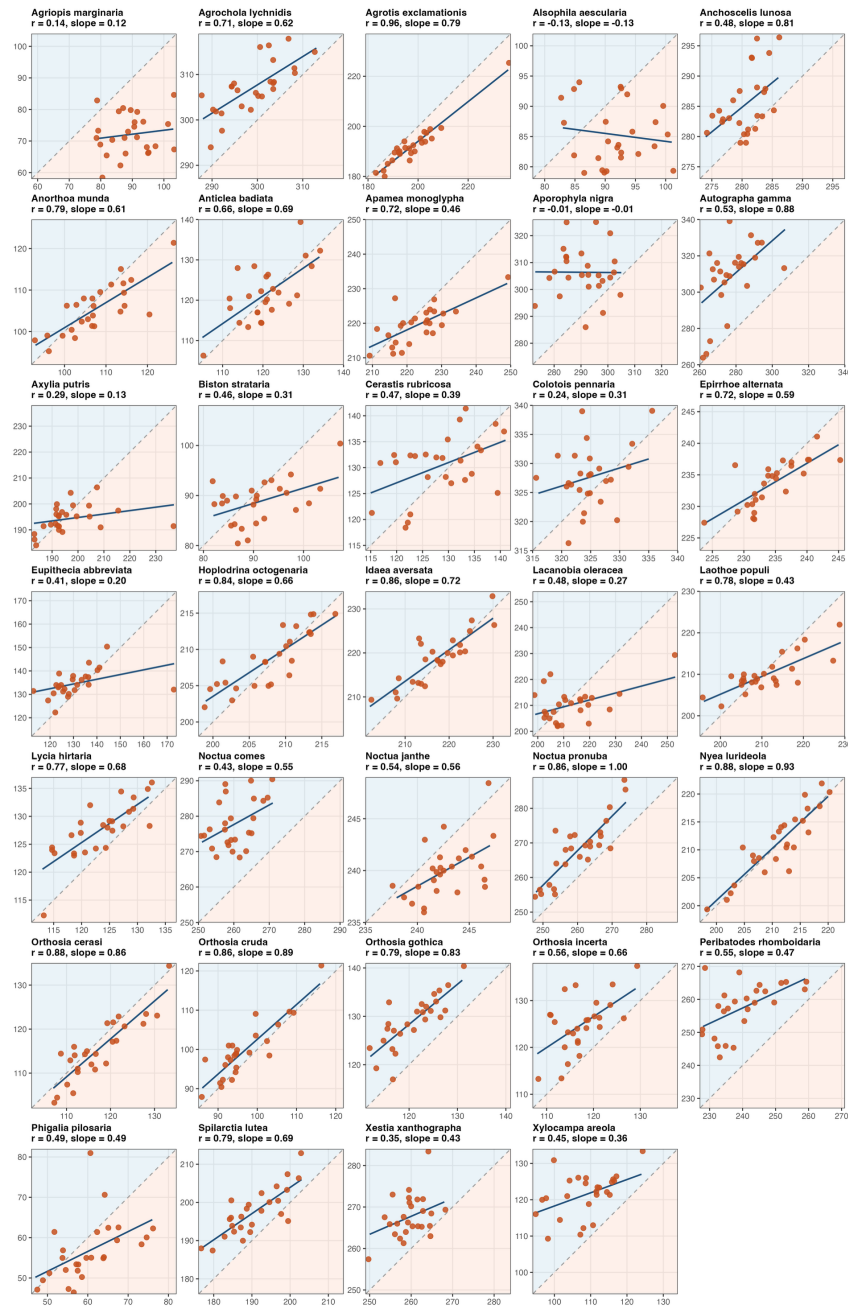
135



**Figure S8.** Full species-level yearly agreement in modelled Q50 (median flight date) between RIS and GBIF under the matched two-scale asymmetric-kernel workflow, for all species in the focal set. Figure 5 in the main text shows nine of these species; this figure shows the complete set. Each panel is one species, with one point per year, using the RIS two-scale asymmetric-kernel Q50 on the x-axis and the GBIF effort-aware two-scale asymmetric-kernel Q50 on the y-axis. Dashed lines indicate 1:1 agreement, fitted lines summarise the within-species relationship,

144 panel titles report the species-level correlation and GBIF-on-RIS slope, and the shaded triangles  
145 indicate years where GBIF is later than RIS (blue, upper left) or earlier than RIS (orange, lower  
146 right).

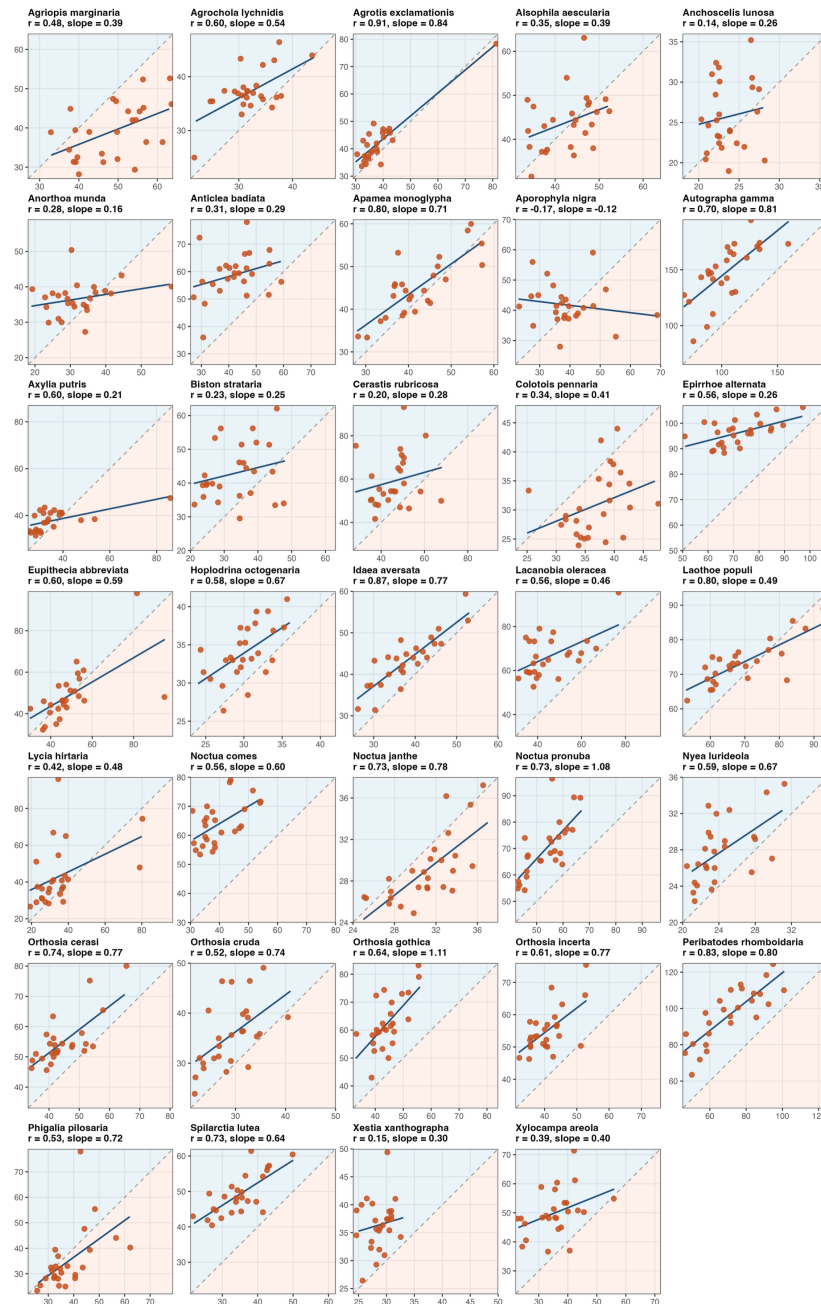
147



**Figure S9.** Full species-level yearly agreement in modelled Q90 (flight end) between RIS and GBIF under the matched two-scale asymmetric-kernel workflow, for all species in the focal set. Figure 5 in the main text shows nine of these species; this figure shows the complete set. Each panel is one species, with one point per year, using the RIS two-scale asymmetric-kernel Q90 on the x-axis and the GBIF effort-aware two-scale asymmetric-kernel Q90 on the y-axis. Dashed lines indicate 1:1 agreement, fitted lines summarise the within-species relationship, panel titles

156 report the species-level correlation and GBIF-on-RIS slope, and the shaded triangles indicate  
157 years where GBIF is later than RIS (blue, upper left) or earlier than RIS (orange, lower right).

158



160

161 **Figure S10.** Full species-level yearly agreement in modelled season width (flight period length)  
 162 between RIS and GBIF under the matched two-scale asymmetric-kernel workflow, for all species  
 163 in the focal set. Figure 5 in the main text shows nine of these species; this figure shows the  
 164 complete set. Each panel is one species, with one point per year, using the RIS two-scale  
 165 asymmetric-kernel season width on the x-axis and the GBIF effort-aware two-scale asymmetric-  
 166 kernel season width on the y-axis, both in days. Dashed lines indicate 1:1 agreement, fitted lines

167 summarise the within-species relationship, panel titles report the species-level correlation and  
168 GBIF-on-RIS slope, and the shaded triangles indicate years where the GBIF flight period is  
169 longer than RIS (blue, upper left) or shorter than RIS (orange, lower right).

170

## Supplementary Tables

This supplement also provides the numerical tables corresponding to the main-text and supplementary figures.

### Table S1

**Table S1.** Summary of method-validation performance across simulation and empirical RIS benchmarks. For each regime, model family, and effort treatment, columns report the mean absolute error (MAE) across the focal phenology metrics, together with the 10th and 90th percentiles of that absolute-error distribution (Error P10 and Error P90), all in days; lower values indicate better recovery. This table corresponds to Fig. 4 in the main text.

| Regime                          | Method                         | Effort       | MAE    | Error P10 | Error P90 |
|---------------------------------|--------------------------------|--------------|--------|-----------|-----------|
| A. Simulation: symmetric truth  | Cyclic GAM                     | Effort-aware | 1.064  | 0.000     | 2.000     |
| A. Simulation: symmetric truth  | Cyclic GAM                     | No effort    | 3.073  | 0.000     | 6.214     |
| A. Simulation: symmetric truth  | Single-scale asymmetric kernel | Effort-aware | 3.001  | 0.294     | 9.000     |
| A. Simulation: symmetric truth  | Single-scale asymmetric kernel | No effort    | 5.018  | 0.684     | 11.000    |
| A. Simulation: symmetric truth  | Single-scale symmetric kernel  | Effort-aware | 1.794  | 0.000     | 4.352     |
| A. Simulation: symmetric truth  | Single-scale symmetric kernel  | No effort    | 4.100  | 0.000     | 9.000     |
| A. Simulation: symmetric truth  | Two-scale asymmetric kernel    | Effort-aware | 0.911  | 0.000     | 1.317     |
| A. Simulation: symmetric truth  | Two-scale asymmetric kernel    | No effort    | 4.443  | 1.000     | 9.000     |
| A. Simulation: symmetric truth  | Two-scale symmetric kernel     | Effort-aware | 0.882  | 0.000     | 1.266     |
| A. Simulation: symmetric truth  | Two-scale symmetric kernel     | No effort    | 4.430  | 1.000     | 8.251     |
| B. Simulation: asymmetric truth | Cyclic GAM                     | Effort-aware | 2.508  | 0.000     | 5.077     |
| B. Simulation: asymmetric truth | Cyclic GAM                     | No effort    | 5.952  | 1.000     | 13.000    |
| B. Simulation: asymmetric truth | Single-scale asymmetric kernel | Effort-aware | 1.613  | 0.000     | 4.000     |
| B. Simulation: asymmetric truth | Single-scale asymmetric kernel | No effort    | 5.520  | 1.000     | 11.000    |
| B. Simulation: asymmetric truth | Single-scale symmetric kernel  | Effort-aware | 2.201  | 0.074     | 5.000     |
| B. Simulation: asymmetric truth | Single-scale symmetric kernel  | No effort    | 6.495  | 1.000     | 13.013    |
| B. Simulation: asymmetric truth | Two-scale asymmetric kernel    | Effort-aware | 0.817  | 0.000     | 2.000     |
| B. Simulation: asymmetric truth | Two-scale asymmetric kernel    | No effort    | 5.682  | 1.000     | 11.100    |
| B. Simulation: asymmetric truth | Two-scale symmetric kernel     | Effort-aware | 0.882  | 0.000     | 2.000     |
| B. Simulation: asymmetric truth | Two-scale symmetric kernel     | No effort    | 5.961  | 1.000     | 12.000    |
| C. Empirical RIS check          | Cyclic GAM                     | Effort-aware | 11.158 | 1.500     | 25.600    |
| C. Empirical RIS check          | Cyclic GAM                     | No effort    | 11.270 | 1.490     | 26.000    |
| C. Empirical RIS check          | Single-scale asymmetric kernel | Effort-aware | 8.439  | 0.500     | 21.000    |
| C. Empirical RIS check          | Single-scale asymmetric kernel | No effort    | 8.437  | 0.607     | 21.000    |
| C. Empirical RIS check          | Single-scale symmetric kernel  | Effort-aware | 9.496  | 0.940     | 24.000    |
| C. Empirical RIS check          | Single-scale symmetric kernel  | No effort    | 9.518  | 0.828     | 24.500    |
| C. Empirical RIS check          | Two-scale asymmetric kernel    | Effort-aware | 10.736 | 1.002     | 26.550    |
| C. Empirical RIS check          | Two-scale asymmetric kernel    | No effort    | 9.868  | 0.701     | 25.500    |
| C. Empirical RIS check          | Two-scale symmetric kernel     | Effort-aware | 10.771 | 1.197     | 27.500    |
| C. Empirical RIS check          | Two-scale symmetric kernel     | No effort    | 9.750  | 0.672     | 24.550    |

*Table S2*

**Table S2.** Species-level agreement between modelled yearly RIS and GBIF phenology metrics under the matched two-scale asymmetric-kernel workflow. Columns report the number of species, the mean and median species-level mean absolute error (MAE), the mean and median species-level Pearson correlation, and the mean and median species-level GBIF-on-RIS slope. Pooled across-species correlations are intentionally omitted because they can be inflated by between-species differences in mean timing. This table corresponds to Fig. 6 in the main text.

| Metric        | n Species | Mean MAE | Median MAE | Mean Corr | Median Corr | Mean Slope | Median Slope |
|---------------|-----------|----------|------------|-----------|-------------|------------|--------------|
| Q10           | 34        | 8.774    | 7.769      | 0.576     | 0.582       | 0.564      | 0.606        |
| Q50           | 34        | 6.762    | 6.481      | 0.587     | 0.639       | 0.567      | 0.564        |
| Q90           | 34        | 7.785    | 5.942      | 0.576     | 0.556       | 0.537      | 0.573        |
| Weighted mean | 34        | 6.349    | 6.207      | 0.580     | 0.590       | 0.551      | 0.540        |
| Season width  | 34        | 11.747   | 8.692      | 0.527     | 0.573       | 0.544      | 0.563        |
| Peak timing   | 34        | 10.001   | 7.635      | 0.384     | 0.343       | 0.430      | 0.380        |

*Table S3*

**Table S3.** Comparison of yearly GBIF phenology metrics against RIS for raw GBIF summaries and a simple per-year cyclic GAM. Columns report sample size (n), mean absolute error (MAE), median absolute error, and the 10th and 90th percentiles of the absolute-error distribution for each metric and method. This table corresponds to Fig. S1.

| Metric        | Method     | n   | MAE    | Median Abs Error | Q10   | Q90    |
|---------------|------------|-----|--------|------------------|-------|--------|
| Q10           | Raw GBIF   | 884 | 10.421 | 7.000            | 0.000 | 21.000 |
| Q10           | Cyclic GAM | 884 | 16.800 | 9.500            | 1.500 | 32.200 |
| Q50           | Raw GBIF   | 884 | 6.153  | 7.000            | 0.000 | 14.000 |
| Q50           | Cyclic GAM | 884 | 13.367 | 9.500            | 1.500 | 28.500 |
| Q90           | Raw GBIF   | 884 | 18.023 | 14.000           | 0.000 | 35.000 |
| Q90           | Cyclic GAM | 884 | 28.699 | 14.500           | 2.500 | 57.200 |
| Weighted mean | Raw GBIF   | 884 | 6.897  | 4.843            | 0.898 | 14.075 |
| Weighted mean | Cyclic GAM | 884 | 11.942 | 9.127            | 1.497 | 24.088 |
| Season width  | Raw GBIF   | 884 | 27.161 | 21.000           | 7.000 | 49.000 |
| Season width  | Cyclic GAM | 884 | 34.666 | 16.000           | 3.000 | 63.000 |

*Table S4*

**Table S4.** Comparison of yearly abundance-index recovery for raw GBIF totals and simple effort-standardised GBIF totals against RIS. Reported values summarise absolute error in the abundance index. This table corresponds to Fig. S2.

| n   | Raw<br>MAE | Effort Standardised<br>MAE | Mean<br>Improvement | Median<br>Improvement | Prop Effort Standardised<br>Better | Wilcox<br>p | Sign<br>Test p |
|-----|------------|----------------------------|---------------------|-----------------------|------------------------------------|-------------|----------------|
| 884 | 0.578      | 0.251                      | 0.327               | 0.331                 | 0.844                              | 0.000       | 0.000          |

*Table S5*

**Table S5.** Species-level comparison of yearly temporal slopes between RIS and GBIF in the supplementary staged workflow. The reported correlations are correlations of species-specific slopes across species, not pooled correlations across all yearly points. This table corresponds to Fig. S3.

| Metric           | n<br>Species | Raw<br>MAE | Reference<br>MAE | Mean<br>Improvement | Prop<br>Reference<br>Better | Raw<br>Sign<br>Match | Reference<br>Sign<br>Match | Raw<br>Slope<br>Corr | Reference<br>Slope Corr | Wilcox<br>p | Sign<br>Test<br>p |
|------------------|--------------|------------|------------------|---------------------|-----------------------------|----------------------|----------------------------|----------------------|-------------------------|-------------|-------------------|
| peak             | 34           | 0.678      | 0.788            | -0.110              | 0.412                       | 0.618                | 0.500                      | 0.188                | 0.114                   | 0.867       | 0.885             |
| Q10              | 34           | 0.427      | 0.452            | -0.025              | 0.412                       | 0.618                | 0.588                      | -0.012               | 0.178                   | 0.736       | 0.885             |
| Q50              | 34           | 0.268      | 0.292            | -0.025              | 0.412                       | 0.618                | 0.735                      | 0.420                | 0.486                   | 0.867       | 0.885             |
| Q90              | 34           | 0.629      | 0.589            | 0.040               | 0.647                       | 0.382                | 0.529                      | -0.160               | -0.009                  | 0.009       | 0.061             |
| Season<br>width  | 34           | 0.778      | 0.817            | -0.038              | 0.471                       | 0.618                | 0.618                      | -0.057               | 0.106                   | 0.595       | 0.696             |
| Weighted<br>mean | 34           | 0.286      | 0.285            | 0.001               | 0.412                       | 0.618                | 0.647                      | 0.063                | 0.232                   | 0.575       | 0.885             |

*Table S6*

**Table S6.** Species-level agreement between modelled spatial RIS and GBIF phenology metrics within UK latitude bands under the matched two-scale asymmetric-kernel workflow. Columns report the number of species-by-band units, the mean and median species-by-band mean absolute error (MAE), the mean and median species-by-band Pearson correlation, and the mean and median species-by-band GBIF-on-RIS slope. Pooled across-species correlations are intentionally omitted because they can be inflated by between-species differences in mean timing. This table corresponds to Fig. 9 in the main text.

| Metric        | n Species Band Units | Mean MAE | Median MAE | Mean Corr | Median Corr | Mean Slope | Median Slope |
|---------------|----------------------|----------|------------|-----------|-------------|------------|--------------|
| Peak timing   | 166                  | 9.113    | 7.293      | 0.416     | 0.443       | 0.459      | 0.439        |
| Q10           | 166                  | 7.817    | 6.160      | 0.542     | 0.595       | 0.508      | 0.475        |
| Q50           | 166                  | 6.438    | 5.489      | 0.569     | 0.626       | 0.529      | 0.544        |
| Q90           | 166                  | 7.694    | 6.523      | 0.482     | 0.551       | 0.417      | 0.459        |
| Season width  | 166                  | 10.609   | 8.287      | 0.351     | 0.397       | 0.391      | 0.335        |
| Weighted mean | 166                  | 6.082    | 5.274      | 0.552     | 0.611       | 0.521      | 0.528        |

*Table S7*

**Table S7.** RIS-only spatial validation of the two-scale asymmetric-kernel RIS model against the empirical RIS benchmark. Columns report sample size (n), mean absolute error (MAE), root mean squared error (RMSE), and correlation for each metric. This table corresponds to Fig. S4.

| Metric        | n     | MAE    | RMSE   | Corr  |
|---------------|-------|--------|--------|-------|
| Q10           | 3,549 | 6.141  | 9.883  | 0.992 |
| Q50           | 3,549 | 4.021  | 6.641  | 0.996 |
| Q90           | 3,549 | 8.352  | 17.848 | 0.975 |
| Weighted mean | 3,549 | 3.508  | 5.391  | 0.997 |
| Season width  | 3,549 | 12.894 | 21.980 | 0.766 |
| peak          | 3,549 | 9.043  | 17.200 | 0.972 |

*Table S8*

**Table S8.** Species-level agreement between RIS and GBIF for derived modelled spatial contrasts. Columns report the number of species, the mean and median species-level mean absolute error (MAE), the mean and median species-level Pearson correlation, and the mean and median species-level GBIF-on-RIS slope for each contrast metric. Pooled across-species correlations are intentionally omitted because they can be inflated by between-species differences in mean timing. This table corresponds to Fig. 10 in the main text.

| Metric                          | n Species | Mean MAE | Median MAE | Mean Corr | Median Corr | Mean Slope | Median Slope |
|---------------------------------|-----------|----------|------------|-----------|-------------|------------|--------------|
| North-south Q50 difference      | 32        | 7.385    | 5.878      | 0.344     | 0.407       | 0.368      | 0.337        |
| Q50 latitudinal slope           | 34        | 2.224    | 1.749      | 0.363     | 0.319       | 0.307      | 0.220        |
| Weighted mean latitudinal slope | 34        | 2.036    | 1.717      | 0.365     | 0.281       | 0.287      | 0.212        |

*Table S9*

**Table S9.** Species-level agreement between RIS and GBIF for derived temporal slopes within UK latitude bands in the empirical spatial workflow. Columns report the number of species, the

mean and median species-level mean absolute error (MAE), the mean and median species-level Pearson correlation, and the mean and median species-level GBIF-on-RIS slope for each slope metric. Pooled across-species correlations are intentionally omitted because they can be inflated by between-species differences in mean timing. This table corresponds to Fig. S5.

| Metric              | n Species | Mean MAE | Median MAE | Mean Corr | Median Corr | Mean Slope | Median Slope |
|---------------------|-----------|----------|------------|-----------|-------------|------------|--------------|
| Q10 trend           | 34        | 0.333    | 0.306      | 0.178     | 0.185       | 0.198      | 0.108        |
| Q50 trend           | 34        | 0.312    | 0.263      | 0.198     | 0.288       | 0.089      | 0.135        |
| Q90 trend           | 34        | 0.491    | 0.333      | 0.029     | -0.015      | 0.028      | -0.003       |
| Season width trend  | 34        | 0.609    | 0.380      | -0.091    | -0.144      | 0.038      | -0.053       |
| Weighted mean trend | 34        | 0.313    | 0.223      | 0.215     | 0.335       | 0.128      | 0.138        |

*Table S10*

**Table S10.** Underlying values for the two-axis reliability summary in the main text (Table 2). Columns report the number of species or species-band units, mean absolute error (MAE), the standardised error (Std MAE; MAE divided by the standard deviation of the RIS values), that standard deviation (SD RIS), mean Pearson correlation, mean GBIF-on-RIS regression slope, and the assigned pattern- and magnitude-recovery classes.

| Metric Class             | Metric                          | n Units | Mean MAE | Std MAE | Sd RIS | Mean Corr | Mean Slope | Pattern Class | Magnitude Class |
|--------------------------|---------------------------------|---------|----------|---------|--------|-----------|------------|---------------|-----------------|
| Yearly metric            | Q10                             | 34      | 8.770    | 1.210   | 7.260  | 0.576     | 0.564      | Good          | Compressed      |
| Yearly metric            | Q50                             | 34      | 6.760    | 1.030   | 6.560  | 0.587     | 0.567      | Good          | Compressed      |
| Yearly metric            | Q90                             | 34      | 7.790    | 1.120   | 6.940  | 0.576     | 0.537      | Good          | Compressed      |
| Yearly metric            | Weighted mean                   | 34      | 6.350    | 1.070   | 5.930  | 0.580     | 0.551      | Good          | Compressed      |
| Yearly metric            | Season width                    | 34      | 11.750   | 1.460   | 8.070  | 0.527     | 0.544      | Good          | Compressed      |
| Yearly metric            | Peak timing                     | 34      | 10.000   | 1.010   | 9.890  | 0.384     | 0.430      | Poor          | Compressed      |
| Spatial band metric      | Peak timing                     | 166     | 9.110    | 0.900   | 10.180 | 0.416     | 0.459      | Moderate      | Compressed      |
| Spatial band metric      | Q10                             | 166     | 7.820    | 0.910   | 8.570  | 0.542     | 0.508      | Good          | Compressed      |
| Spatial band metric      | Q50                             | 166     | 6.440    | 0.880   | 7.320  | 0.569     | 0.529      | Good          | Compressed      |
| Spatial band metric      | Q90                             | 166     | 7.690    | 0.840   | 9.140  | 0.482     | 0.417      | Moderate      | Compressed      |
| Spatial band metric      | Season width                    | 166     | 10.610   | 1.020   | 10.410 | 0.351     | 0.391      | Poor          | Poor            |
| Spatial band metric      | Weighted mean                   | 166     | 6.080    | 0.880   | 6.910  | 0.552     | 0.521      | Good          | Compressed      |
| Derived spatial contrast | North-south Q50 difference      | 32      | 7.390    | 0.930   | 7.940  | 0.344     | 0.368      | Poor          | Poor            |
| Derived spatial contrast | Q50 latitudinal slope           | 34      | 2.220    | 0.880   | 2.540  | 0.363     | 0.307      | Poor          | Poor            |
| Derived spatial contrast | Weighted mean latitudinal slope | 34      | 2.040    | 0.860   | 2.350  | 0.365     | 0.287      | Poor          | Poor            |

| Metric Class           | Metric                      | n Units | Mean MAE | Std MAE | Sd RIS | Mean Corr | Mean Slope | Pattern Class | Magnitude Class |
|------------------------|-----------------------------|---------|----------|---------|--------|-----------|------------|---------------|-----------------|
| Yearly temporal trend  | Trend Q10                   | 34      | 0.430    | 0.740   | 0.570  | -0.012    | -0.006     | Poor          | Poor            |
| Yearly temporal trend  | Trend Q50                   | 34      | 0.270    | 0.800   | 0.330  | 0.420     | 0.399      | Moderate      | Poor            |
| Yearly temporal trend  | Trend Q90                   | 34      | 0.630    | 0.660   | 0.960  | -0.160    | -0.046     | Poor          | Poor            |
| Yearly temporal trend  | Trend peak timing           | 34      | 0.680    | 0.880   | 0.770  | 0.188     | 0.147      | Poor          | Poor            |
| Yearly temporal trend  | Trend Season width          | 34      | 0.780    | 0.760   | 1.020  | -0.057    | -0.020     | Poor          | Poor            |
| Yearly temporal trend  | Trend Weighted mean         | 34      | 0.290    | 0.820   | 0.350  | 0.063     | 0.048      | Poor          | Poor            |
| Spatial temporal trend | Spatial trend Q10           | 170     | 0.330    | 0.660   | 0.500  | 0.198     | 0.119      | Poor          | Poor            |
| Spatial temporal trend | Spatial trend Q50           | 170     | 0.310    | 0.670   | 0.470  | 0.294     | 0.198      | Poor          | Poor            |
| Spatial temporal trend | Spatial trend Q90           | 170     | 0.490    | 0.440   | 1.130  | -0.005    | -0.002     | Poor          | Poor            |
| Spatial temporal trend | Spatial trend Weighted mean | 170     | 0.310    | 0.660   | 0.470  | 0.327     | 0.297      | Poor          | Poor            |
| Spatial temporal trend | Spatial trend Season width  | 170     | 0.610    | 0.500   | 1.210  | -0.018    | -0.006     | Poor          | Poor            |

Table S11

**Table S11.** Focused sensitivity summary for alternative single-scale asymmetric-kernel settings in focal species-years. Columns report curve mean absolute error (Curve MAE), absolute error in Q50 and season width, absolute error in peak height, and the ratio of total fitted expected abundance to observed abundance. Lower values indicate better recovery for all columns except total fitted / observed, where values close to 1 indicate stable abundance scaling. This table corresponds to Fig. S6.

| Species                    | Setting                      | Curve MAE | Q50 Abs Error | Width Abs Error | Peak Height Abs Error | Total Ratio |
|----------------------------|------------------------------|-----------|---------------|-----------------|-----------------------|-------------|
| <i>Orthosia gothica</i>    | Current                      | 0.012     | 0             | 7               | 0.032                 | 0.999       |
| <i>Orthosia gothica</i>    | Broader left                 | 0.013     | 0             | 7               | 0.037                 | 0.999       |
| <i>Orthosia gothica</i>    | Broader both                 | 0.014     | 0             | 7               | 0.037                 | 0.999       |
| <i>Orthosia gothica</i>    | Broader both, weaker penalty | 0.014     | 0             | 7               | 0.037                 | 1.000       |
| <i>Idaea aversata</i>      | Current                      | 0.009     | 0             | 0               | 0.059                 | 0.999       |
| <i>Idaea aversata</i>      | Broader left                 | 0.009     | 0             | 0               | 0.076                 | 0.999       |
| <i>Idaea aversata</i>      | Broader both                 | 0.010     | 0             | 0               | 0.083                 | 0.999       |
| <i>Idaea aversata</i>      | Broader both, weaker penalty | 0.010     | 0             | 0               | 0.083                 | 1.000       |
| <i>Xestia xanthographa</i> | Current                      | 0.020     | 0             | 0               | 0.100                 | 0.999       |
| <i>Xestia xanthographa</i> | Broader left                 | 0.024     | 0             | 0               | 0.118                 | 0.999       |
| <i>Xestia xanthographa</i> | Broader both                 | 0.026     | 0             | 7               | 0.124                 | 0.999       |

| Species                    | Setting                      | Curve MAE | Q50 Abs Error | Width Abs Error | Peak Height Abs Error | Total Ratio |
|----------------------------|------------------------------|-----------|---------------|-----------------|-----------------------|-------------|
| <i>Xestia xanthographa</i> | Broader both, weaker penalty | 0.026     | 0             | 7               | 0.124                 | 1.000       |

*Table S12*

**Table S12.** Source composition of the focal UK GBIF benchmark after filtering to the focal species and applying the filters described in the main-text Methods. Categories are assigned at the dataset level using **datasetKey/datasetName** with supporting publisher metadata, then summarised here by source category and record type.

| Source Category                                | Record Type Category | n Datasets | n Records | Percent Records |
|------------------------------------------------|----------------------|------------|-----------|-----------------|
| Formal recording scheme / monitoring programme | Human observation    | 64         | 6,418,026 | 97.106          |
| Citizen science platform                       | Human observation    | 7          | 126,463   | 1.913           |
| Research project / survey                      | Human observation    | 33         | 63,244    | 0.957           |
| Museum / institutional collection              | Specimen / material  | 5          | 659       | 0.010           |
| Other / mixed / unknown                        | Human observation    | 2          | 468       | 0.007           |
| Other / mixed / unknown                        | Other                | 2          | 425       | 0.006           |
| Museum / institutional collection              | Human observation    | 1          | 4         | 0.000           |

*Table S13*

**Table S13.** Sensitivity of yearly Q50 agreement to the GBIF exposure floor used in the matched two-scale asymmetric-kernel workflow. Columns report the tested exposure floor in inferred trap-nights, the number of species contributing to the comparison, and the mean and median species-level Q50 mean absolute error (MAE) and Pearson correlation against RIS. Lower MAE and higher correlation indicate better agreement.

| Exposure Floor | n Species | Mean Q50 MAE | Median Q50 MAE | Mean Q50 Corr | Median Q50 Corr |
|----------------|-----------|--------------|----------------|---------------|-----------------|
| 0              | 34        | 11.941       | 8.135          | 0.449         | 0.465           |
| 10             | 34        | 8.238        | 7.327          | 0.537         | 0.595           |
| 25             | 34        | 7.406        | 6.981          | 0.566         | 0.617           |
| 50             | 34        | 6.762        | 6.481          | 0.587         | 0.639           |
| 100            | 34        | 6.077        | 5.692          | 0.607         | 0.641           |
| 200            | 34        | 5.440        | 5.154          | 0.620         | 0.687           |

*Table S14*

**Table S14.** Decomposition of the yearly GBIF-on-RIS regression slope and its correction for measurement error in the RIS benchmark. For each metric the table reports the number of

species ( $n$ ); the mean species-level Pearson correlation ( $r$ , the *tracking* factor); the ratio of standard deviations  $SD(GBIF)/SD(RIS)$  (the *amplitude* factor); their product, the ordinary least-squares slope (OLS); the reduced major axis slope ( $RMA = OLS / r$ ), which rescales amplitude under the assumption that RIS and GBIF are equally reliable; the estimated sampling reliability of the RIS metric (from the quantile standard error of its smooth-curve estimate); and the reliability-corrected slope (OLS divided by the RIS reliability). Reliability is reported for the three quantile metrics, where the quantile standard error applies exactly. Because the SD ratio is near one for the central quantiles, the sub-unity OLS slope is driven by the correlation rather than by a smaller GBIF amplitude; because the RIS reliability is high, the correction for regression dilution is small, so the compression is largely genuine. See Table S15 for its sensitivity to the assumed overdispersion of the catch.

| Metric        | n  | r (Tracking) | SD Ratio (Amplitude) | OLS Slope | RMA Slope | RIS Reliability | Corrected Slope |
|---------------|----|--------------|----------------------|-----------|-----------|-----------------|-----------------|
| Q10           | 34 | 0.576        | 0.980                | 0.564     | 0.980     | 0.899           | 0.628           |
| Q50           | 34 | 0.587        | 1.012                | 0.567     | 0.949     | 0.946           | 0.600           |
| Q90           | 34 | 0.576        | 0.948                | 0.537     | 0.824     | 0.900           | 0.597           |
| Weighted mean | 34 | 0.580        | 0.963                | 0.551     | 0.892     |                 |                 |
| Season width  | 34 | 0.527        | 1.068                | 0.544     | 1.027     |                 |                 |

*Table S15*

**Table S15.** Sensitivity of the RIS reliability and the reliability-corrected slope to the assumed overdispersion of the catch. The RIS quantile sampling standard deviation is inflated by a factor  $\sqrt{k}$  to mimic clustering beyond the weekly Poisson estimate ( $k = 1$  is the model's own estimated overdispersion). For each central quantile and inflation factor the table reports the implied RIS sampling standard deviation (days), the resulting reliability, and the reliability-corrected slope. The qualitative conclusion (real, moderate compression) holds while the RIS Q50 sampling standard deviation stays below about four days (corrected slope up to about 0.76); only an implausibly large RIS error, about five days or more, would erase the compression.

| Metric | Overdispersion Factor ( $k$ ) | RIS Sampling SD (Days) | RIS Reliability | Corrected Slope |
|--------|-------------------------------|------------------------|-----------------|-----------------|
| Q10    | 1                             | 2.369                  | 0.899           | 0.628           |

| Metric | Overdispersion Factor (k) | RIS Sampling SD (Days) | RIS Reliability | Corrected Slope |
|--------|---------------------------|------------------------|-----------------|-----------------|
| Q10    | 2                         | 3.350                  | 0.823           | 0.686           |
| Q10    | 3                         | 4.103                  | 0.763           | 0.739           |
| Q10    | 5                         | 5.297                  | 0.645           | 0.875           |
| Q10    | 10                        | 7.491                  | 0.425           | 1.327           |
| Q50    | 1                         | 1.592                  | 0.946           | 0.600           |
| Q50    | 2                         | 2.252                  | 0.891           | 0.636           |
| Q50    | 3                         | 2.758                  | 0.837           | 0.678           |
| Q50    | 5                         | 3.560                  | 0.746           | 0.760           |
| Q50    | 10                        | 5.035                  | 0.559           | 1.015           |
| Q90    | 1                         | 2.329                  | 0.900           | 0.597           |
| Q90    | 2                         | 3.294                  | 0.799           | 0.672           |
| Q90    | 3                         | 4.034                  | 0.699           | 0.768           |
| Q90    | 5                         | 5.207                  | 0.527           | 1.018           |
| Q90    | 10                        | 7.365                  | 0.322           | 1.669           |

*Table S16*

**Table S16.** Effect of the RIS-derived GBIF records on spatial agreement with RIS. The National Moth Recording Scheme includes records from the Rothamsted (RIS) light traps, identified by the sampling protocol “Rothamsted Trap”; these were removed from all analyses so that the benchmark and the data under test share no records. This table compares model-free (empirical) agreement between RIS and GBIF in each latitude band and year for the same GBIF download processed with these records kept and removed, and so isolates the records themselves (no model or effort estimate is involved). For each metric it reports the mean, over species and latitude bands, of the Pearson correlation between RIS and GBIF across years and of the mean absolute error (days), with the records kept and removed, and the change caused by removing them.

| Metric        | Correlation, Records Kept | Correlation, Records Removed | Change in Correlation | MAE, Records Kept (Days) | MAE, Records Removed (Days) | Change in MAE (Days) |
|---------------|---------------------------|------------------------------|-----------------------|--------------------------|-----------------------------|----------------------|
| Q10           | 0.434                     | 0.425                        | -0.010                | 8.544                    | 8.676                       | 0.132                |
| Q50           | 0.453                     | 0.442                        | -0.011                | 7.473                    | 7.691                       | 0.218                |
| Q90           | 0.338                     | 0.320                        | -0.017                | 11.457                   | 11.641                      | 0.184                |
| Weighted mean | 0.499                     | 0.484                        | -0.016                | 7.698                    | 7.948                       | 0.250                |
| Season width  | 0.252                     | 0.237                        | -0.014                | 15.337                   | 15.492                      | 0.155                |