OEP027-02 Statistical approaches to analysing variation in species abundance data and testing whether biodiversity targets have been met

Ref: 2025-03-20-436-1

Report Authors:

Gail Robertson, John Addy, Anastasia Frantsuzova & Nick Schurch

Date: 2026-07-07

Biomathematics and Statistics Scotland

JCMB, The King's Buildings,

Peter Guthrie Tait Road,

EDINBURGH, EH9 3FD,

Scotland, UK.

Tel: +44 (0)131 650 4900

Email: enquiries@bioss.ac.uk

Table of Contents

1. Executive summary

This work was commissioned by the Office for Environmental Protection (OEP) to evaluate the robustness of proposed statistical methodologies for determining whether legislative targets designed to halt and then reverse the long-term decline in biodiversity have been met. The findings will support the OEP's assessment of whether legislative species abundance targets are appropriate and government progress towards meeting them.

The Environmental Targets (Biodiversity) (England) Regulations 2023 set two statutory targets for species abundance, assessed using the England Species Abundance Indicator (ESAI): first that the decline in species abundance has been halted by 2030, assessed by comparing ESAI values in 2029 with 2030; and second, that species abundance has increased by at least 10% between 2030 and 2042 and exceeds the 2022 ESAI value. The 2029 vs 2030 legislative test therefore relies on a comparison between adjacent annual values, whereas the 2042 legislative test assesses change over longer periods. Neither assessment takes account of the statistical uncertainty inherent in the indicator.

The ESAI is a complex proxy for biodiversity, derived from over 1,100 species with smoothing applied at multiple stages and final values produced under two general Smoothing Options SO1 and SO2. The indicator is strongly influenced by moths, which make up ~40% of the species included in the current indicator and historically comprised as much as 86% of the ESAI species. Other represented species groups do not reflect the same pattern of decline observed in moths. Historical analysis shows that the ESAI and the species sub-group indicators exhibit structured multi-year dynamics, persistent temporal autocorrelation, and characteristic timescales of change substantially longer than a single year. Consequently, consecutive year changes reflect a combination of ecological processes, data and modelling choices, and statistical properties of the methods used to assess change.

This study evaluates eight alternative candidate statistical methods described in a prior scoping study (Burns et al. 2025) to identify which can robustly assess whether the statutory targets have been achieved. We characterised the historical behaviour of the ESAI (1970–2023), developed a simulation framework to generate indicator trajectories under controlled scenarios, and used these to compare the performance of candidate approaches.

The multi-year dynamics of the ESAI indicator show that a comparison between two adjacent annual values (i.e. the 2029 v 2030 legislative test) is unlikely to robustly distinguish a genuine halt in decline from normal inter-annual variability. Our simulations confirm this. When the true underlying trend for the ESAI represents no change or an increase (the conditions set by the 2030 target), reliably detecting this from a single adjacent-year comparison requires a rise of approximately 2.9 percentage points (~4.3% relative change), exceeding recent historical precedent. In practice, this means that, short of an unprecedented year-to-year increase in species abundance, it will not be possible to reliably distinguish ongoing decline from the desired halting or reversal of that decline by comparing the 2029 and 2030 ESAI values alone. The secondary legislation governing the assessment of the species abundance target in 2030 is therefore not fit for purpose for determining whether the decline in species abundance has been halted.

Changes in the ESAI value are more robustly identifiable when simulation outputs are considered over longer timescales. The 2030–2042 assessment, which is based on a longer window, thus provides a stronger statistical basis for assessing whether the second statutory target has been met.

Based on our simulations, no single candidate statistical test performed well across both statutory targets. Within the bounds of the current smoothing approaches proposed by DEFRA, the analysis identified three tests which could provide greater understanding on whether the long-term decline in species abundance has been halted and then reversed.

The performance of the statistical methods examined here was shown to be sensitive to smoothing choices used in calculating the ESAI. We consider SO1 to be over-smoothing the ESAI, and we generally prefer SO2 for assessing progress on species abundance targets. However, both the 2030 and 2042 targets are sensitive to the shape of the smoothed ESAI at the end of the indicator time-series, precisely where the Generalised Additive Model approach that underlies the ESAI has weaker statistical power to identify trends reliably.

Utilising additional years of ESAI values beyond the point of the assessment are effective at mitigating this limitation, substantially improving the ability of all the statistical options to correctly identify a halt in species abundance decline. Where ESAI values are available beyond the date of assessment, SO2 is preferable because the model's greater flexibility reduces the impact of earlier ESAI trends and allows the model to reflect more recent changes in the indicator's behaviour. However, if an assessment must be made when ESAI values beyond the date of assessment are not available, SO1 is then preferable because the heavier smoothing helps limit the impact of shape variability at the end of the indicator time-series. These sensitivities indicate that the effect of smoothing parameters should be explored more fully before the ESAI is released as an official statistic.

Ideally, the legislative framework would incorporate uncertainty explicitly and defer definitive conclusions until post-assessment ESAI values are available. We recommend that each statutory target is reported with uncertainty in test performance, using specified confidence intervals, and continuous assessment of the targets that is updated as additional ESAI values become available. We recommend that conclusions about whether these targets have been met or not should only be drawn once at least two years of additional ESAI values beyond the date of assessment are available. For the 2030–2042 target, the ambiguity over whether "10% increase" refers to a relative percentage or an absolute percentage-point change should be resolved, as it materially affects detection reliability, with all methods performing better under the more stringent percentage-point interpretation.

2. Introduction

The UK is one of the most nature depleted countries in the world having seen significant loss of its plants, animals and fungi (Burns et al. 2023). Biodiversity has been in decline since the 1940's (Robinson and Sutherland 2002) resulting in reduced ecosystem functions and services as well as reduced resilience in the face of climate change (Oliver et al. 2015).

Recent legislation has been passed with the aim of halting this decline and improving the state of nature in the UK. Monitoring this aim requires a measurable indicator for biodiversity against which change can be assessed. While biodiversity is a multidimensional and contested concept (Toepfer 2019), trends in species abundance provide a practical and policy-relevant proxy, and the English statutory framework adopts this approach (devolved nations are developing similar measures).

The Environmental Targets (Biodiversity) (England) Regulations 2023 (Part 4) specifies statutory targets for species abundance in England under the powers conferred by the Environment Act 2021. Regulations 11 to 13 specify that, by 31 December 2030, the overall relative species abundance index must indicate that the long-term decline in species abundance has been halted. This is assessed by establishing whether the overall relative species abundance index for the year 2030 is the same as, or higher than, the overall relative species abundance index for the year 2029. Regulations 14 to 16 specify that, by 31 December 2042, the overall relative species abundance index must indicate that species abundance has increased. This is assessed by establishing whether the overall relative species abundance index for the year 2042 is at least 10% higher than the overall relative species abundance index for the year 2030, and greater than the overall relative species abundance index for the year 2022[1]. Hereafter in this report we refer to Regulations 11 to 13 and Regulations 14 to 16 as "the species abundance targets".

[1] https://www.legislation.gov.uk/uksi/2023/91/part/4

2.1. The ESAI and species level data

DEFRA has developed an indicator of multiple species abundance in England (DEFRA 2025a) which underpins the species abundance targets that aim to halt the decline in biodiversity and thereafter increase it. This metric is the average proportional change in relative abundance across a range of species over time, weighting all species equally, and can be created for all available species (global indicator) or individual taxonomic groups. DEFRA's multi-species abundance indicator is created using raw species abundance data collated from national monitoring schemes with standardised data collection protocols in England. The current version of this indicator includes time series abundance measurements of 1,176 species spanning birds, mammals, moths, butterflies, bumblebees, vascular plants, freshwater invertebrates, and fish (all of which are measured as number of individuals, apart from vascular plants which are measured on a categorical ground cover scale). The length of the time series varies between species (DEFRA 2025a). The creation of the indicator involves multiple steps which we outline briefly here.

1) Monitoring schemes collect data at the population level and these are processed and combined to a single time series for each species using dataset-specific modelling.
2) Some species-level time series undergo pre-smoothing using thin plate splines to remove short-term fluctuations. For example, raw moth data show a high amount of variation from year to year and exhibit significant seasonal variation. Across most species, the same smoothing parameters are used. Notably, applying the same smoothing to time series from species with different characteristics may result in over- or under-smoothing the data. Any resulting inaccuracies will be propagated into the time series for both the taxonomic groups and the global biodiversity indicator.
3) Individual species abundance time series are combined into a single global indicator of smoothed geometric means using the method described in Freeman et al. (2020), to create DEFRA's multi-species indicator and 95% credible intervals, called the England Species Abundance Indicator, ESAI.

The ESAI is computed using the Bayesian multi-species indicator method of Freeman et al. (2020), implemented via the *bma()* function in the R package *BRCindicators* (August et al. 2020). DEFRA's production code is publicly available[2] and consists of two R Markdown scripts; one to run the Freeman model and one to extract, summarise, and categorise the results.

The Freeman method uses a hierarchical state-space model in which each species has an unobserved true log-abundance trajectory, and observed species-level indices are treated as noisy realisations of these latent states. Species-specific annual growth rates are modelled as normally distributed around a shared multispecies average growth rate for each year. The multispecies indicator is then derived by accumulating these average growth rates over time, producing a composite measure of average proportional change in abundance across all contributing species. Penalised spline smoothing is applied to the multispecies growth rates within the model. The maximum number of knots (or, equivalently, the maximum number of basis splines minus one) sets an upper limit on the model's flexibility, while the degree of smoothing is determined by the penalty parameter, which is estimated within the model from a data-informed prior (see Freeman et al. 2020). The penalised spline is fitted using Markov chain Monte Carlo (MCMC) via the software program Just another Gibbs sampler (JAGS), and credible intervals on the smoothed indicator arise directly from the posterior samples.

In DEFRA's implementation, the input to the Freeman model is a set of species-level time series that have already been processed and, in most cases, smoothed upstream of the model (see Section 2.2). These pre-smoothed values are provided on the log-scale. The Freeman model is then run separately for each taxonomic group and for all species combined, with smoothing controlled by fixed priors on the smoothing penalty and

[2] https://github.com/BiologicalRecordsCentre/Abundance_Indicator_2025

a "years per knot" parameter: smoothing option 1 (SO1) uses one knot per ten years of data and smoothing option 2 (SO2) uses one knot per three years. For the approximately 54-year time series, this yields roughly 5–6 knots for SO1 and approximately 18-19 knots for SO2. A third configuration with one knot per year is also produced but is not used in the published indicator. The model is run with 5,000 MCMC iterations, and convergence is assessed using the Gelman-Rubin diagnostic. The smoothing parameterisation is a key component of the ESAI construction with implications across the indicator time series, particularly at the endpoint (e.g. 2022–2023). ESAI values derived under different smoothing options yield materially different estimated rates and even directions of change. These differences arise because alternative smoothing parameterisations allocate short-term variability differently between the underlying trend and residual fluctuations. Since the current statutory assessment of the species abundance targets depends on the estimated change between specific years at the series endpoint, conclusions regarding whether these have been met are inherently sensitive to the chosen smoothing approach (for more details on smoothing, see Sections 2.2 & 5.2.3).

The Freeman method provides a principled framework for combining heterogeneous species-level data into a single indicator; it propagates uncertainty from species-level indices through to the composite measure via the Bayesian posterior; and its state-space formulation naturally handles missing data, accommodating species that enter or leave the time series in different years without requiring imputation. The method does not require the contributing species-level indices to have been produced using identical statistical methods, only that each is proportional to abundance on the log-scale.

However, the method also has properties that are directly relevant to the statistical assessment undertaken here. First, the ESAI construction pipeline involves multiple stages of smoothing. Some species-level time series are pre-smoothed by their source monitoring scheme before entering the DEFRA pipeline, and additional smoothing is applied within the pipeline to species whose data are classified as unsmoothed (see Section 2.2). The Freeman model then applies its own penalised spline smoothing to the multispecies growth rates. For species whose input data have already been smoothed, the final indicator values therefore reflect at least two, and in some cases three, layers of smoothing. This compounding of smoothing stages can attenuate genuine short-term dynamics, inflate temporal autocorrelation in the final indicator beyond that present in the underlying ecological processes, and reduce the effective independence of adjacent annual values. These effects are directly relevant to the statistical tests evaluated in this report, particularly those that assess change between adjacent years. Second, the model treats all species as exchangeable contributors to the indicator, weighting them equally in the multispecies growth rate regardless of taxonomic relatedness, ecological similarity, or the precision of their underlying data (see Section 3.3.1). A critical evaluation or modification of the Freeman method is beyond the scope of this project, but these properties should be kept in mind when interpreting the results that follow.

We downloaded the current version of the global ESAI smoothed values[3] and DEFRA provided the individual species abundance data from 1970 to 2022 used to create the global indicator (details of these datasets are given in DEFRA (2025a)). For some species these values were already smoothed to remove short-term fluctuations and noise (see Section 3.2). The individual species dataset also includes aggregate species indicator values for the following taxonomic groups: bumblebees, fish, freshwater invertebrates, mammals, moths, birds, butterflies, and vascular plants. R code used to calculate the ESAI from individual species abundance data was made available by DEFRA.

2.2. Smoothing

Smoothing methods are used to reduce short-term variability in a time series to better reveal underlying long-term trends. A wide range of smoothing approaches exists, each with distinct assumptions and trade-offs (see Shumway & Stoffer (2017) for a comprehensive overview). However, inappropriate choice of smoothing

[3] https://oifdata.defra.gov.uk/themes/wildlife/D4/

parameters can distort inference. Over-smoothing may attenuate genuine short-term dynamics, whereas insufficient smoothing may allow short-term fluctuations or dynamics to dominate the estimated trend.

The individual species abundance dataset provided by DEFRA includes "pre-smoothed", "already smoothed", and "unsmoothed" values for species used to create both the ESAI as well as the taxonomic group indicators that each species belongs to, and the survey method used to estimate species abundance (which may be counts, presence/absence data, estimates based on counts). The pre-smoothed species time series are provided by DEFRA on a natural-log scale. For many species, abundance fluctuates from year to year, making it difficult to examine the underlying long-term trends. For this reason, some monitoring schemes apply smoothing methods to the species-level time series data. A cross-validation exercise as described in DEFRA (2025a) showed that the creation of the ESAI using the Freeman et al. (2020) method is more robust if the individual species data are pre-smoothed and part of the ESAI pipeline involves converting unsmoothed species-level time series data to a natural-log scale prior and applying smoothing prior to computing the ESAI values. Abundance values for some species are classed as "already smoothed", where smoothed time series are provided by the monitoring scheme (though smoothing method differs between some species included in this group). Pre-smoothed and already smoothed values are created using similar methods, though smoothing methods differ slightly for some species depending on species characteristics. Where there are too few data points for effective smoothing, values are classed as "unsmoothed" (this is the case mainly for vascular plants). All data from individual species datasets are included in the pipeline described in Freeman et al. (2020), regardless of whether these had been pre-smoothed, and are used to compute the ESAI values to which SO1 and SO2 are then applied.

During the process of creating the ESAI, two sets of values are created using a penalised spline Generalised Additive Model (GAM) for the smoothing, with different smoothing parameters. A key aspect of this kind of smoothing method is the number of "knots" used, which places an upper limit on some of the flexibility of the model. A higher limit on the number of knots allows the model to be more flexible where necessary, generally making the model less smooth and allowing it to capture shorter time scale variations; conversely a smaller limit on the number of knots more tightly constrains the model, generally making the model smoother. Smoothing option 1 (SO1) is a more constrained smoother, with the number of knots equal to one tenth of the total number of data points in the indicator time series. Smoothing option 2 (SO2) is a less constrained smoother, with one knot for every three years of data (DEFRA 2025a). The selection of knot density for each smoothing option represents a modelling choice that has a substantial influence on the behaviour of the ESAI, especially near the series endpoint. For example, the ESAI with SO1 rises smoothly from 2021 to 2023, however the ESAI with SO2 falls smoothly over this period. Where relevant, we examine the impact of the smoothing options on the statistical tests and compare the results for both.

3. Project aims

The current legislation mandates a simple comparison of the values of the ESAI in two consecutive years, 2029 and 2030. This comparison takes no account of the intrinsic variation or uncertainty in the ESAI and assumes that change can be detected over short timescales. Assessing long-term trends using this simple comparison carries a serious risk of misidentifying long-term trends in the ESAI and drawing incorrect conclusions regarding whether the species abundance targets have been met, or not.

In 2024, the Office for Environmental Protection (hereafter the OEP) commissioned preliminary research to examine alternative statistical approaches for assessing whether the species abundance targets are met. The resulting report (Burns et. al. 2025) assessed the strengths and weaknesses of 16 statistical options evaluating their conceptual suitability, statistical properties, and interpretability in the context of the ESAI. These options were structured around different types of inferential question, including tests of whether a recent rate of change is greater than or equal to zero, comparisons of change over defined time periods, equivalence-based assessments of negligible trend, and breakpoint or change-point approaches. The report considered the relative risk of false positive and false negative conclusions under each option, particularly considering uncertainty, smoothing, and short-term variability in the indicator. While Burns et al. 2025 provides a structured appraisal of strengths and weaknesses and highlighted limitations of the current year-on-year comparison specified in legislation, it does not undertake a formal simulation-based performance evaluation or recommend a single preferred method. Instead, it concludes that further analytical work is required to test how these candidate approaches perform under realistic scenarios of ecological change.

This project builds on the Burns et al. 2025 report by undertaking a simulation-based evaluation of the proposed statistical methodologies, informed by the current ESAI and available species abundance data. The aim is to identify approaches that can robustly assess whether the species abundance targets have been achieved, with an emphasis on methodological rigour, transparency, policy relevance, and explicit treatment of uncertainty. To do this, we carry out a historical assessment of indicator data created using the Freeman method from 1970 to 2023 to characterise timescales of change in the global indicator and across taxonomic groups under different smoothing parameters (Work Package 1). We then develop an R package to simulate indicator time series based on these characteristics under defined scenarios of change and use the simulated data to compare the performance of selected statistical options outlined in Burns et al. 2025 (Work Package 2). Finally, we assess which options were most appropriate for evaluating the species abundance targets under different smoothing approaches (Work Package 3).

4. WP1: Scoping & historical assessment:

4.1. Historical assessment versus expert elicitation

An initial meeting with OEP staff finalised the project aims and priorities, including data requirements and access to existing DEFRA code. Two possible approaches were presented for how we might best anticipate the scales of change for the ESAI indicator on the time scales required by the legislation: expert elicitation and historical assessment of available data.

Expert elicitation is a formal process that uses expert judgement to estimate distributions for a quantity or of interest. Protocols may be individual focused or consensus seeking, where the latter allows judgements of several experts to be collated to address challenges in which uncertainty plays an important role, either due to insufficiency of appropriate data or when data cannot be collected (O'Hagan et al. 2006). Established frameworks, such as the Sheffield Elicitation Framework (Gosling 2018), provide structured discussion and regular feedback from experts in a domain field. Expert elicitation can be carried out via in-person or online workshops, or via hybrid events. A key strength of this approach in this context is the expert's ability to anticipate scenarios that could drive ESAI changes that may not reflected in the historical record. By contrast, historical assessment uses the existing ESAI time series to quantify empirical characteristics of change, including trend magnitude, variability, and temporal dependence, and to use these characteristics to inform plausible future scenarios. This approach is quantitative, directly grounded in observed data, and well suited to parameterising simulation models within the scope of this project. However, it rests on the assumption that the statistical properties of the historical time series provide a reasonable basis for anticipating future behaviour.

Given the objectives, time frame, and resources of the current project, OEP staff agreed that a historical assessment provided a methodologically sufficient and proportionate basis for the simulation-based evaluation undertaken here, with the possibility of complementing this work through expert elicitation in future phases. The aims of the historical assessment are to:

1) Explore trends in the ESAI and indicator values for individual taxonomic groups
2) Examine historical ESAI to estimate three key parameters to inform WP2 simulations: the flexibility of the underlying trend; correlation among residuals; and the variance
3) Characterise the timescale over which change occurs in the ESAI and individual taxonomic indicators

4.2. Historical assessment

4.2.1. ESAI & taxonomic sub-group trends

Figure 1 shows the ESAI values over time from 1970 to 2023 for both smoothing options with their 95% confidence intervals shown as a shaded ribbon. Figures 2 shows the equivalent indicator values for taxonomic groups (birds, bumblebees, butterflies, fish, freshwater invertebrates, mammals, moths, vascular plants; species level data were only available through to 2022), calculated using the same method as the ESAI, with smoothing options 1 & 2.

The ESAI is constructed as a geometric mean across contributing species in each year. Taxonomic sub-groups with more species, or longer time series, exert greater influence on the indicator, and this influence has, and will continue to, vary over time as the composition of the contributing data evolves. The trends in the moth taxonomic sub-group (and birds to a lesser extent) correlate closely with the trend in the global indicator for both smoothing options (Figures 1 & 2, Pearson's correlation coefficient, with 95% confidence intervals: ($r_{moths,SO1} = 0.987^{+0.005}_{-0.010}$; $r_{moths,SO2} = 0.988^{+0.005}_{-0.009}$; $r_{birds,SO1} = -0.62^{+0.20}_{-0.14}$; $r_{birds,SO2} = -0.64^{+0.19}_{-0.14}$). Together, moths and birds account for ~50% (~40% & ~10%, respectively) of the species that comprise the current ESAI value and, unlike other taxonomic groups (e.g. freshwater invertebrates, ~21%, vascular plants, ~19%), these datasets span the full range of the indicator from 1970 to 2026 and have dominated historically, defining the shape of the indicator to-date. The percentage of the total ESAI species that are moths has declined over time (Figure 3), but today this sub-group still comprises a large percentage of total species. The time series length differs between taxonomic subgroups, and these differences in temporal coverage impact the statistical properties such as autocorrelation and variation that can be characterised for these sub-groups. It appears that change in ESAI is primarily responding to the decline in moth species abundance, and also bird species, though the correlation between the ESAI and bird species is negative, suggesting that bird species increased over time.

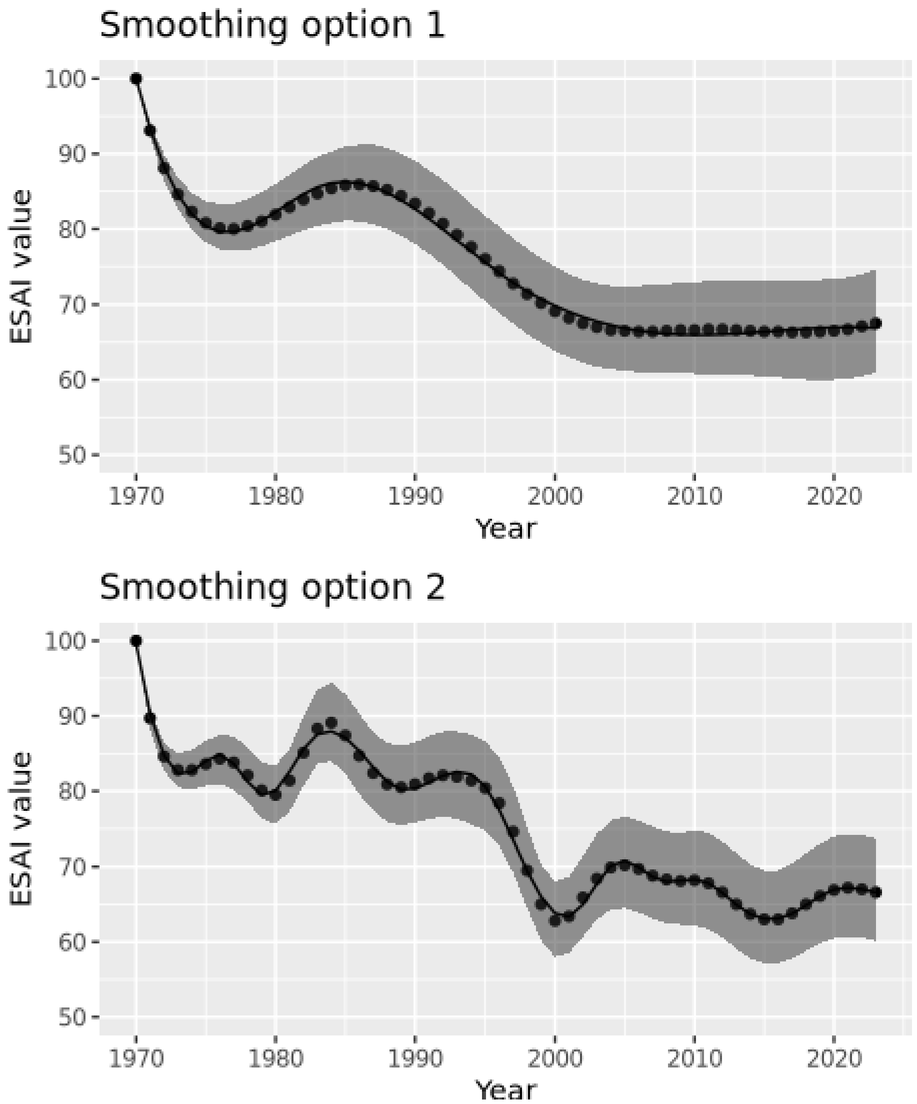


Figure 1: *The ESAI (black dots) and the associated 95% credible intervals (grey ribbon). The EASI is constructed using a penalised spline GAM smoother with two smoothing options, SO1 (top, knots = 5-6) and SO2 (bottom, knots = 18-19; see Section 2.1). We fit an equivalent penalised spline GAM model (black lines) to the ESAI values using the respective smoothing options.*

Figure 2: *The ESAI-equivalent SO1 (yellow) and SO2 (blue) indexes for individual taxonomic groups, with 95% fit confidence intervals (coloured ribbons). k defines the basis dimension used in each smoothing option.*

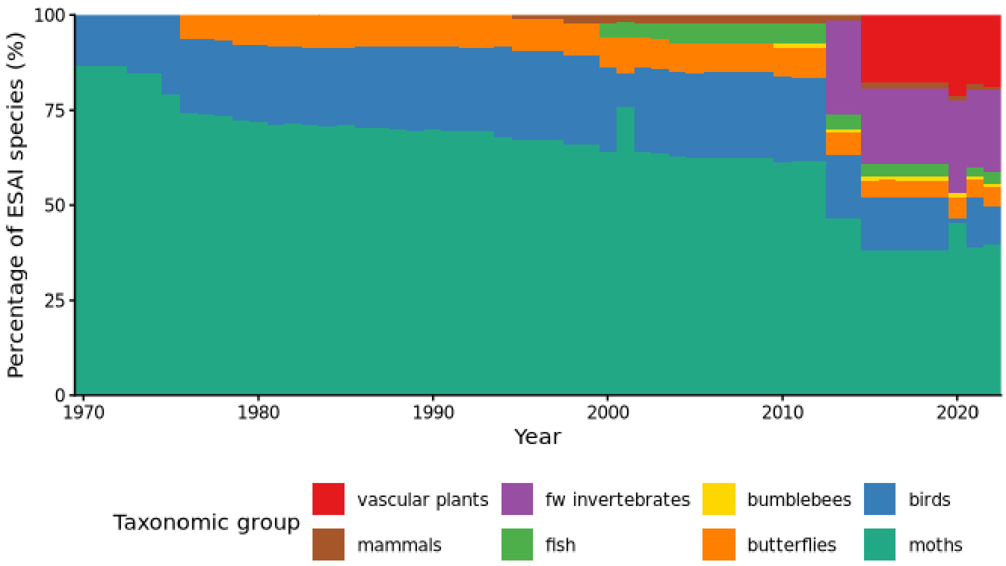


Figure 3: *Percentage of total ESAI species for each taxonomic group (colours) per year from 1970 to 2022*

4.2.2. Characterising smoothness

For credible simulation-based inference, the statistical tests evaluated in WP3 are applied to simulated future indicator trajectories whose behaviour depends on accurate characterisation of the key statistical properties of the historical ESAI values. We fitted a GAM to the historical ESAI to estimate three key simulation parameters: the basis dimension (*k*), which controls the flexibility of the underlying trend; the autocorrelation coefficient (AR(1), see Section 3.2.3), which captures correlation among residuals; and the variance, which quantifies unexplained variability (see Section 3.2.4).

Here, we do not fix the basis dimension, k, in the GAMs to match the knot specification used in the Freeman method, instead we treat k as an empirical tuning parameter. The ESAI curves were originally obtained from a penalised spline model fitted in JAGS, whereas here the GAMs are fitted using the *mgcv* package in R (Wood, 2017). These two implementations are not directly comparable, as they differ in their basis construction, penalty structure, and estimation framework (Bayesian MCMC versus frequentist generalised cross-validation or restricted maximum likelihood). Consequently, the same nominal number of knots or basis functions would not produce the same fitted curve under both approaches.

In particular, the basis dimension k in *mgcv* sets an upper bound on the number of basis functions available to the model, while the final smoothness of the fitted curve is determined through penalisation. The effective degrees of freedom are therefore typically smaller than k, and depend on the interaction between the chosen basis dimension and the smoothing penalty. For this reason, setting k in *mgcv* to one more than the number of knots used in the JAGS fits does not necessarily imply a comparable degree of smoothness and may not yield an adequate representation of the SO1 and SO2 ESAI curves.

To select an appropriate value of *k* for both SO1 and SO2, we fitted the historical ESAI values with both smoothing options across a range of *k* values and examined how the residual standard deviation (a measure of the goodness-of-fit) and the AR(1) autocorrelation coefficient (how much correlation remains in the residuals) varies with *k*. Figure 4 shows that the fit quality stabilises above *k* = 10 for SO1 and above *k* = 23 for SO2. Except where indicated, these values are used for all subsequent analyses.

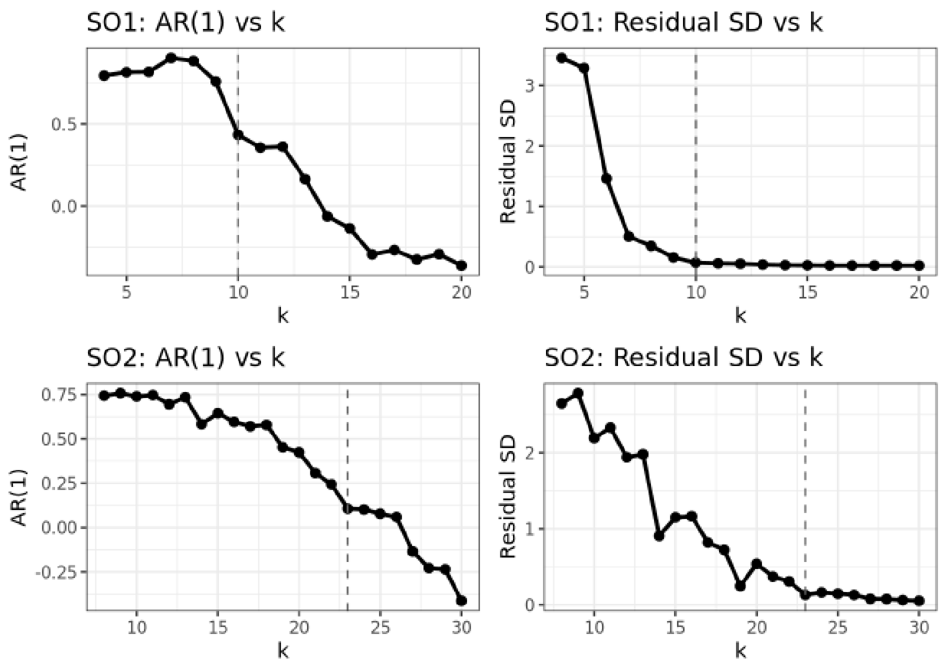


Figure 4. Sensitivity of GAM-estimated AR(1) autocorrelation coefficient (left panels) and residual standard deviation (right panels) to the value of the GAM basis dimensions k, for smoothing options 1 (top panels) and 2 (bottom panels). Dashed vertical lines indicate the adopted values of k (SO1: k = 10, SO2: k = 23), where the residual standard deviation stabilises.

4.2.3. Characterising autocorrelation

The GAM models fitted to the historical ESAI values with the selected basis dimensions estimate the lag-1 partial autocorrelation coefficient parameter for both smoothing options (SO1: AR(1) = 0.43, SO2: AR(1) = 0.11). We examined the residual structure and higher-order lag components fits with plots of the model residuals and their partial autocorrelation function (Figure 5). Some structure persists in the residuals after the fit. The SO1 residuals are small but show weak potential negative autocorrelation signals across higher lags. The SO2 time series residuals, by contrast, are larger and exhibit a stronger negative autocorrelation across higher lags, and particularly at a lag of 2 years. For SO2 in particular, this shows that some short and medium temporal structure remains after the GAM fit to the ESAI values, suggesting that the AR(1) model is a slight misspecification for SO2, which may impact the performance of downstream statistical tests. The differences between the smoothing options suggests that the temporal patterns in the model residuals are strongly influenced by the choice of smoothing parameters. Importantly, however, the scale of the residuals is small (≤ 0.3, equivalent to ~0.5%) suggesting that this downstream impact is likely to be small, also, we only simulate a maximum of 10 years of data to address questions in WP3 so autocorrelation will not have a large influence.

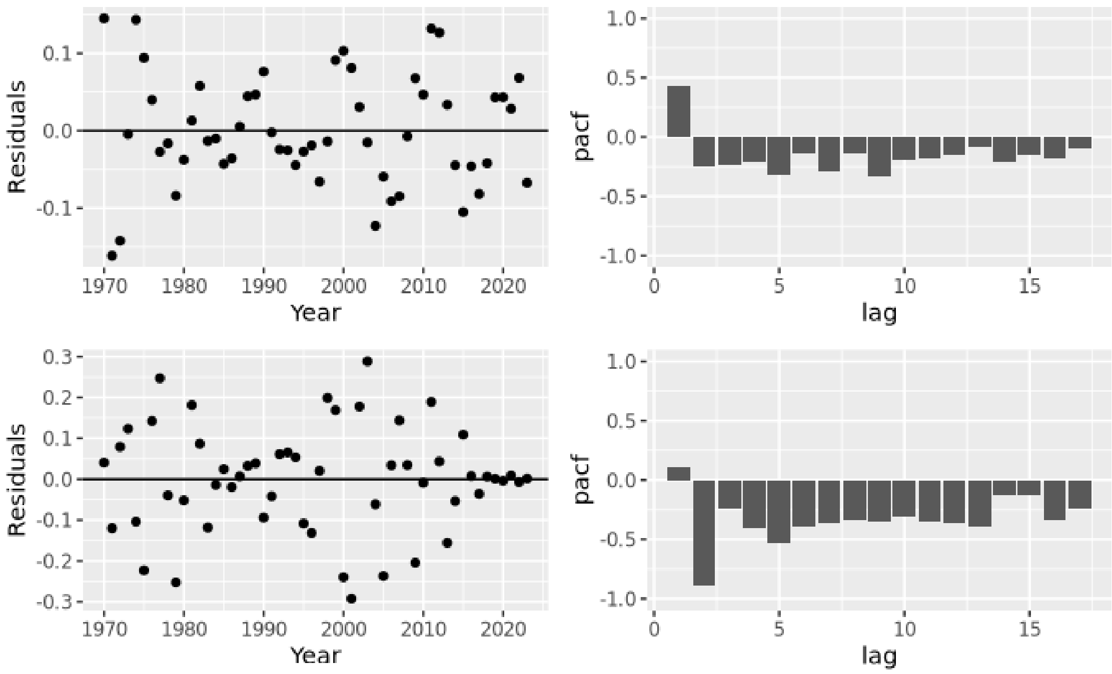


Figure 5: *Model fit residuals (left) and partial autocorrelation function (right) plots showing amount of residual autocorrelation in the GAM fit residuals at different time lags for each smoothing option (top: SO1, k = 10; bottom: SO2, k=23).*

Only the moth, bird, butterfly, and mammal taxonomic sub-groups had sufficiently long time series to allow reliable estimation of autocorrelation structure. All these groups show consistent patterns of autocorrelation to the ESAI; positive autocorrelation in the model fit residuals at a lag of 1 year (weaker for SO2 than SO1, except for mammals, which do not show a positive AR(1)), negative autocorrelations at higher lags and a stronger negative correlation for lags of 2 years (Figure 6). The similarity of the sub-group autocorrelation structures to those seen in the global indicator demonstrates that these patterns in the ESAI are not being driven by a single taxonomic sub-group. Both the correlation values, and the scale of the residuals, are larger than those for the global indicator (in particular for butterflies, whose model residuals are large for SO1, suggesting that this model over-smooths for these values). The larger correlations and residual values suggest that while short- and medium-timescale structure remains in all groups, these are weakened by the averaging process used in constructing the ESAI. This analysis reiterates that temporal dependence persists after the smoothing is applied across all the sub-groups, and that the choice of smoothing option has an important effect on the temporal characteristics of the ESAI.

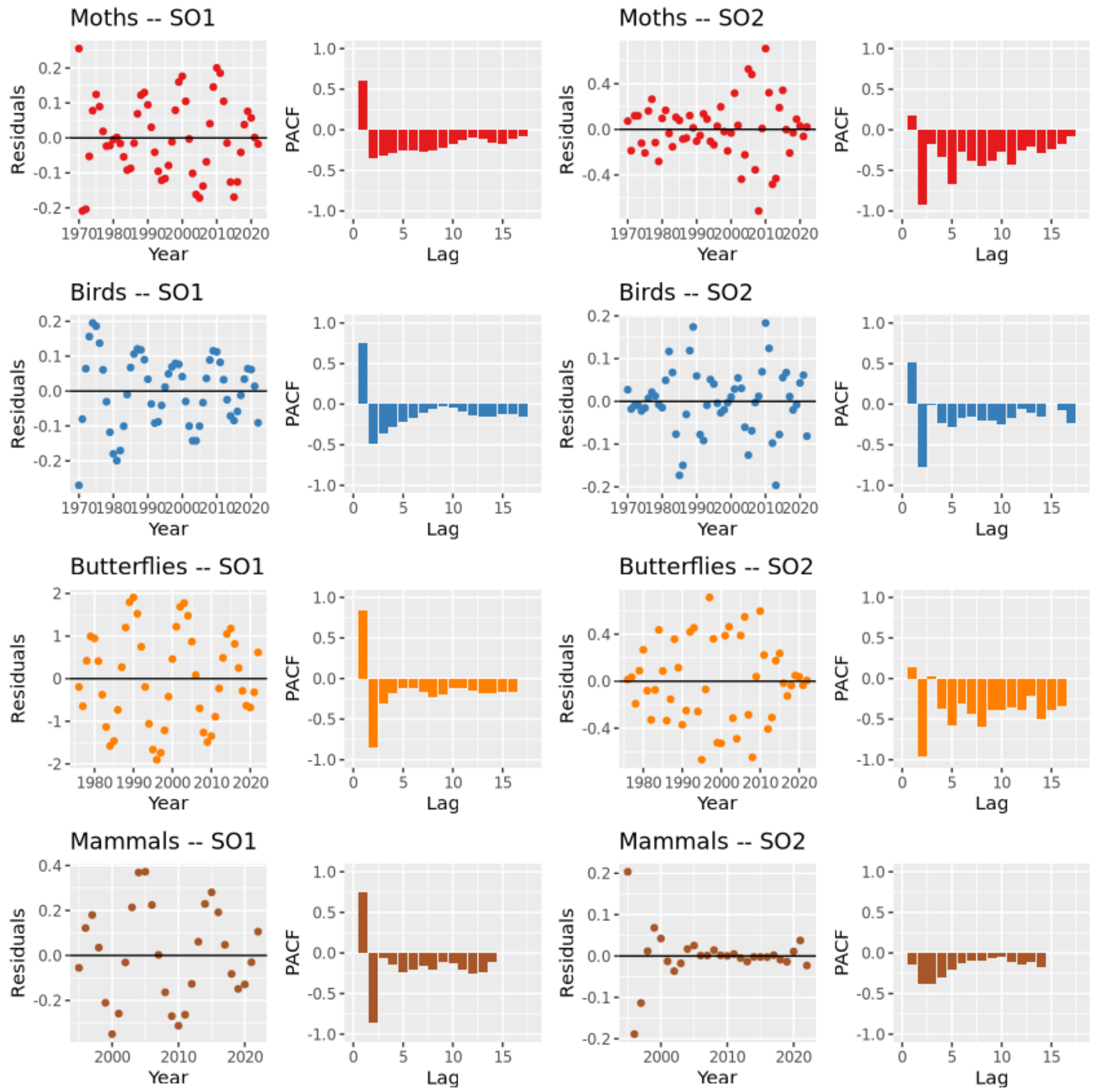


Figure 6: *Model fit residuals (columns 1 & 3) and partial autocorrelation function (columns 2 & 4) plots showing the degree of autocorrelation in the GAM residuals at different time lags for moths (top), birds (upper-middle), butterflies (lower-middle) and mammals (bottom) sub-group indicators under SO1 (columns 1 & 2) and SO2 (columns 3 &4).*

4.2.4. Characterising variance

We reflect the uncertainty in the ESAI by incorporating the scale of the 95% credible intervals into our simulations as year-to-year random noise added to the underlying trend in the simulated time series, after accounting for the autocorrelation structure of the AR(1) process as identified from the historical ESAI GAM fits. We convert each year's credible interval width to an implied standard error (assuming an approximately Gaussian posterior, $SE = CI_{width} / (2 \times 1.96)$). Figure 7 shows that the scale of the standard errors for both smoothing options is similar and grows over time. We model this relationship with a logarithmic model ($SE = b \log(\Delta y)$, where Δy is the year difference since 1969, the year before the start of the time series) and extrapolate into the simulation windows of interest for each of the questions in WP3 (e.g., Q1: 2024 – 2030, Q3: 2030-2042). We use the mean of the extrapolated standard errors over that window to derive the standard deviation of the AR(1) process, σ, for use in our simulations ($\sigma = \overline{SE}\sqrt{(1 - AR(1)^2)}$). We use this characterisation of the variance observed in the historical data in WP2 to generate simulations of ESAI values with realistic properties, and use these to test species abundance targets in WP3.

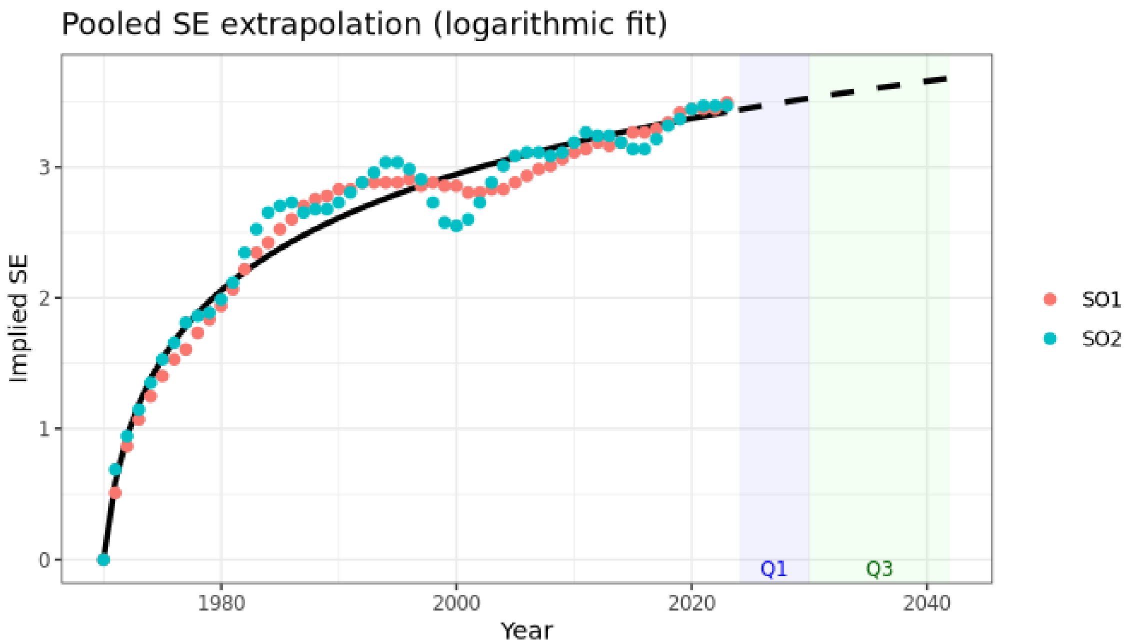


Figure 7: *Standard errors derived from the 95% credible interval widths of the historical ESAI (SO1, blue, SO2, red), with a logarithmic model fit for the observed period (solid black line), and extrapolated into the Q1 (2024–2030, blue) and Q3 (2030–2042, green) simulation windows respectively (dashed black line).*

4.2.5. Characterising timescales

We examined the variation on different timescales for the ESAI and, where data were available, for indicator values created from different taxonomic sub-groups, to identify the timescales where greater-than-expected change occurs. The time series were aggregated into non-overlapping bins spanning 2 to 15 years, with 15 years representing the longest scale supported by the 54-year series. For each binning scale, we calculated the standard deviation of the binned values and compared the resulting distribution with the expected distribution under a Cauchy model (Gelman et al., 2008) fitted to the 2-year binning. The 2-year scale was treated as representing predominantly random inter-annual fluctuations, on the assumption that substantive structural changes in the indicator are unlikely to occur over such short intervals. This comparison allowed us to assess whether variability at longer timescales exceeded that expected from short-term fluctuations alone. Figure 8 shows the distributions of the binned ESAI standard deviations in comparison with the expected 2-year binning Cauchy distribution for bins spanning up to 8 years, for SO1 and SO2 respectively.

Across increasing bin widths, the distribution of standard deviations shifts progressively towards larger values, with greater mass in the upper tail for longer timescales. The transition is gradual rather than abrupt, with no sharp discontinuity between consecutive bin sizes. For both smoothing options, the observed variability increasingly exceeds that expected under the 2-year baseline distribution beyond ~7 years, suggesting the presence of structured multi-year variation beyond short-term inter-annual fluctuations on long timescales. For SO2, divergence from the 2-year baseline distribution becomes apparent at shorter bin widths, with excess variability emerging from ~4 years, suggesting the presence of similar structured multi-year variation on intermediate timescales.

The smooth, progressive increase in variability across timescales provides no clear indication of abrupt, high-magnitude structural breaks dominating the ESAI time series. Instead, the observed patterns are more consistent with gradual, multi-year change. In this context, statistical methods designed primarily to detect discrete change points are less aligned with the dominant temporal structure of the indicator, which is characterised mainly by gradual change rather than abrupt shifts. Accordingly, such methods were excluded from formal evaluation in WP3. The subsequent analysis instead focuses on statistical approaches suited to assessing gradual change over time.

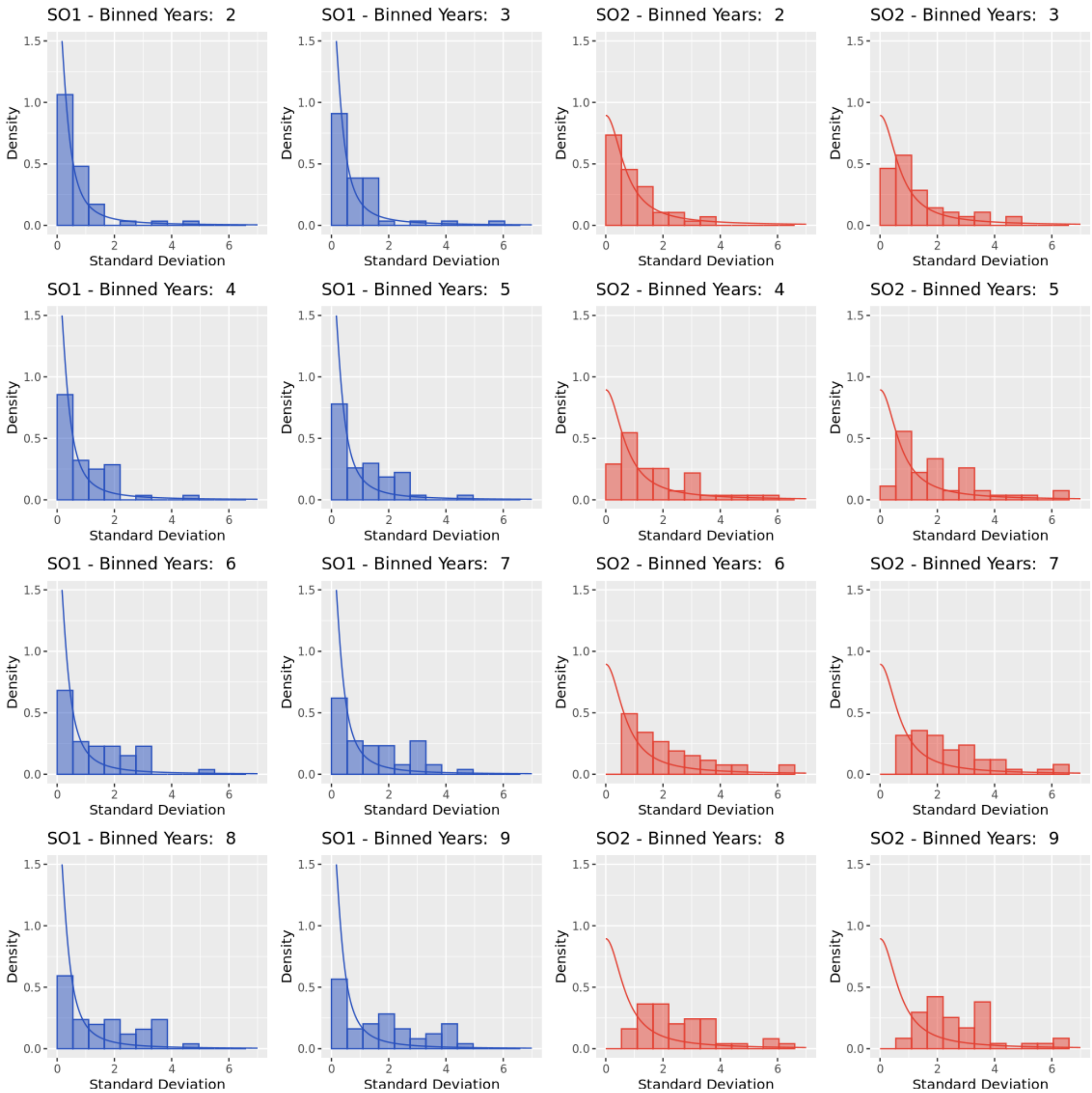


Figure 8: *Distribution of standard deviations of binned ESAI values for bin widths of 2 (top left) – 9 (bottom right) years under SO1 (left columns) and SO2 (right columns). Histograms show the empirical distribution of standard deviations calculated from non-overlapping bins at each timescale, overlayed with the expected Cauchy distribution fitted to the 2-year binning (solid line).*

Subgroup-specific analyses[4] show broadly similar scale-dependent patterns to those observed for the ESAI, with variability increasing progressively with bin width. However, the timescale at which deviation from the 2-year baseline becomes apparent differs between taxa. For moths and butterflies, excess variability emerges at shorter timescales (~4 – 6 years), whereas for mammals this occurs at longer timescales (~6 – 8 years), with birds showing intermediate behaviour. These results suggest that multi-year structure is present across all

[4] https://gitlab.bioss.ac.uk/grobert/oep-species-abundance-indicator/-/tree/main/Code/WP1

groups, however the characteristic timescales of change vary modestly between taxa and the characteristic timescales important for the ESAI are not being driven by a single sub-group.

4.3. Discussion

The historical assessment indicates that the ESAI exhibits structured multi-year dynamics rather than abrupt step changes, with temporal dependence persisting after smoothing and characteristic timescales of variation emerging beyond short-term inter-annual fluctuations. The allocation of temporal structure between trend and residual components depends materially on the smoothing parameterisation, and the characteristic timescales identified (~4 years under SO2 and ~7 years under SO1) therefore reflecting both properties of the underlying data and modelling choices. These findings motivate our focus on statistical approaches designed to assess gradual change over multi-year periods, rather than methods primarily intended to detect discrete change points.

The ESAI is constructed as a geometric mean across contributing species, and its composition changes over time as monitoring programmes begin and end. Moths and birds together comprise a substantial proportion of species in the indicator and span much of the temporal range of the series. The similarity between sub-group and global autocorrelation structures indicates that the characteristic temporal dynamics of the ESAI are not only an artefact of taxonomic dominance, but also because similar short- and medium-timescale structure is evident across multiple sub-groups. Nevertheless, dominance in representation means that certain taxa exert greater influence on the magnitude and trajectory of the global indicator. The implications of species non-independence and equal weighting for the interpretation of the ESAI are discussed in Sections 2.1 and 6.3.

Finally, the characteristic multi-year timescales identified here should not be interpreted as intrinsic ecological constants; they arise from the interaction between ecological dynamics and the chosen smoothing framework. Nonetheless, they provide an empirical basis for considering appropriate temporal windows over which change can be robustly assessed. Both identified scales are substantially longer than the single-year comparison currently specified in legislation (2029 vs 2030), reinforcing the need to evaluate statistical methods that appropriately account for temporal dependence and gradual change. The parameter estimates derived in WP1 therefore form the basis for constructing realistic simulation scenarios in WP2.

4.4. Conclusions

1) Informing simulation parameters from historical assessment: Historical assessment provided a proportionate and methodologically appropriate basis in WP1 for estimating the statistical characteristics of the ESAI time series to inform realistic future scenarios of change. This approach assumes that the principal statistical properties of the historical series provide a reasonable basis for anticipating future behaviour. This assessment indicates a long-term decline in species abundance over the period examined, persistent temporal autocorrelation in the detrended data, and increasing variation at longer timescales, consistent with structured multi-year dynamics characterised mainly by gradual change rather than abrupt shifts. The estimated trends, autocorrelation structures, and timescale-dependent variability identified in WP1 are important for the construction of appropriate simulations of the ESAI in WP2, and for evaluating the statistical methods outlined in Burns et al. 2025.

2) Smoothing option choice affects predicted trend meaning species abundance targets are sensitive to smoothing option: The statistical characteristics of the ESAI differ substantially depending on the smoothing parameterisation used. The two smoothing options used redistribute temporal structure differently between trend and residual components, influencing endpoint behaviour and the apparent timescale of change. These results have important implications for identifying appropriate statistical methods for assessing the species abundance targets. In particular, the presence of multi-year temporal structure and the sensitivity of results to smoothing parameterisation suggest that approaches designed to assess gradual change over appropriate multi-year windows are more appropriate than the current 2029 vs 2030 single-year comparison.

3) Change in ESAI is primarily responding to the decline in moth species abundance, and (to a lesser degree) the increase in bird species. The ESAI is less representative of other taxonomic groups, especially at the beginning of the time series: The temporal dynamics observed in the ESAI combine those seen across multiple taxonomic sub-groups of the indicator. Autocorrelation and multi-year variation are evident across the taxonomic groups indicating that the characteristic temporal structure of the ESAI is not solely attributable to the dominant taxa in isolation. However, moths and birds comprise a substantial proportion of the contributing species and moths in particular exert considerable influence on the global indicator. The dominance of the number of moth species ensures that the observed increases in other taxonomic groups are not strongly reflected in the behaviour of the ESAI.

5. WP2: Simulations

Here we develop an open source, publicly available, R package[5] for simulating indicator time series with controlled statistical characteristics identified from the ESAI in WP1, including long-term trend, variability, autocorrelation. This package, called *oepsims,* will be used in WP3 to evaluate the statistical approaches outlined in Burns et al. 2025 against the target criteria specified in the secondary legislation. This package was used in WP3 to create simulations of realistic time series data to address specific questions and scenarios of change in the ESAI relevant to testing whether species abundance targets can be met.

We developed various functions to simulate time series data with user-defined inputs describing:

1. Long-term linear trend (declining, flat, increasing)
2. Scale of the change over time
3. Variation
4. Autocorrelation
5. Taxonomic composition

Additionally, functions were developed to recover parameters from time series and provide tools for visualising the simulations.

The package includes extensive documentation describing the simulation framework and how the functions are implemented and used. As this documentation forms part of the package itself, these details are not repeated here. Instead, to illustrate the simulation process, we provide an example below demonstrating how the package is used to generate change scenarios for WP3.

5.1. Example

This example builds upon the brief examples included within the help files in the *oepsims* package and demonstrates how our functions can be used to simulate indicator time series data under specified scenarios. Here we simulate 1,000 ESAI time series using the parameters estimated from the historical assessment of the ESAI with SO1, carried out in WP1, which showed that change over time occurs over a characteristic timescale of approximately 7 years. This will be developed further in the section on WP3 to address specific questions related to targets in the secondary legislation.

We simulated 1,000 time series, each seven years in length, using the *n_samples_ts()* function from the package. This function generates a specified number of time series of a given length with a linear underlying trend, combined with stochastic variation and autocorrelation controlled by user-defined parameters. The simulated time series use parameter values estimated from the historical assessment of the real global indicator in WP1 for SO1 (Intercept = 67.5, AR(1) = 0.43, sigma = 3.14) and ran for a period of seven years after the historical ESAI values to 2023. To represent a scenario of significant positive change over the seven-year period, the slope parameter was set to 1.02, corresponding to a 2% year-on-year increase in the indicator[6]. This value was chosen because it represents an average change greater than the upper 95% confidence interval around the slope estimated for the ESAI in 2023, the current end of the historical indicator. The *plot_simulations()* function can be used to visualize the simulated time series (Figure 9), including the median simulated indicator value for each year, extracted from the simulations using the *summarize_simulations()* function.

[5] oepsims, https://gitlab.bioss.ac.uk/grobert/oep-species-abundance-indicator

[6] For clarity, this is percentage change from the current baseline value, not percentage-point change.

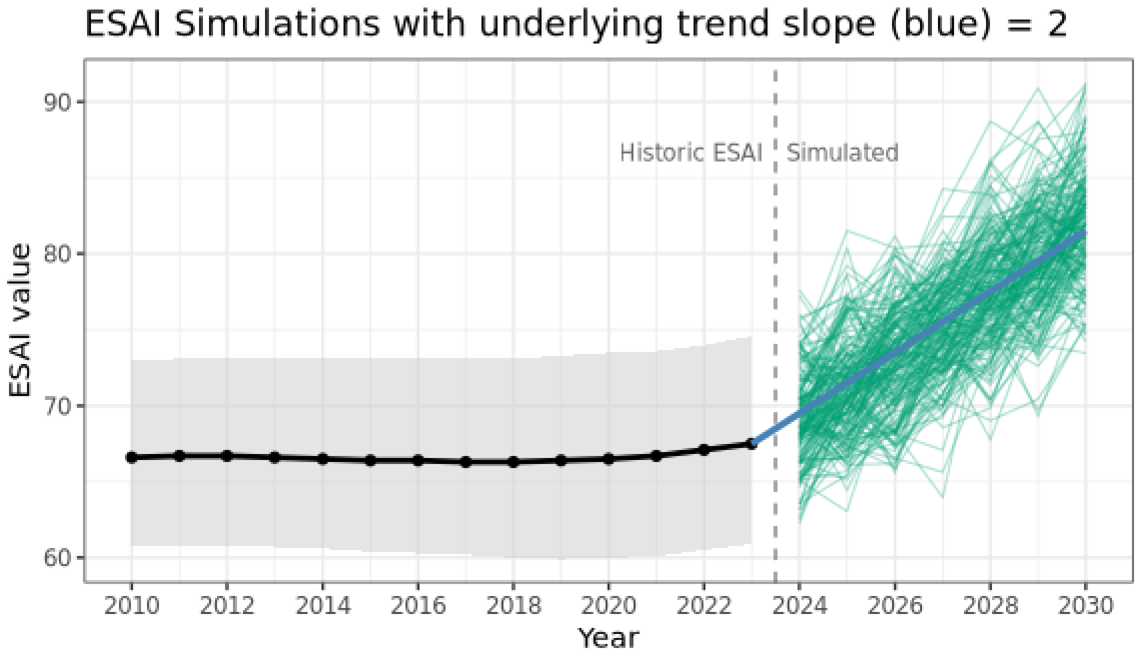


Figure 9: *Examples of ESAI simulations from the oepsims package. 200 simulated seven-year indicator time series (green) were generated with an underlying trend slope = 2%/yr (blue), variation = 3.14, AR(1) = 0.43, alongside the historical ESAI indicator values (black dots, intercept = 67.5) and their associated 95% credible intervals (grey ribbon), fit with a penalised spline GAM (SO1, black line).*

The function *recover_pars()* calculates key statistical properties across the simulation cohort, which can then be compared to the input values to verify that the simulations accurately correspond to the simulation intent. Figure 10 shows the distributions of recovered parameters from the simulated data, compared with the specified input parameter values. The intercept and slope parameters are recovered with good accuracy. In contrast, the AR(1) autocorrelation parameter is substantially underestimated, reflecting the expected difficulty of reliably estimating autocorrelation from short time series. Increasing the length of the time series improves recovery, particularly for the AR(1) parameter.

The simulation framework provides a practical way to generate indicator time series with statistical properties similar to those observed in historical data and is intended to explore plausible scenarios rather than provide precise forecasts. It assumes that the residual dynamics of the indicator are approximately stationary (i.e., that variance and autocorrelation structure remain broadly constant through time), a common assumption for many time series simulation studies but one will be impacted by ecological dynamics, monitoring protocols, or indicator construction change. The current implementation uses an AR(1) autocorrelation process only. The WP1 residual analysis (Section 3.3.2) identified additional structure at longer lags, but the strength and pattern of these higher-order correlations varied substantially between smoothing options, and restricting the simulations to AR(1) provides a parsimonious representation of the dominant temporal dependence without over-complicating the framework. The AR(1) coefficient is difficult to recover accurately for short time series. This is not critical for assessing relative statistical method performance but may impact absolute performance for these series. For the analysis in WP3 the simulation parameters are derived from the full historical ESAI, which spans a longer period and allows the AR(1) process to be estimated more reliably.

The simulations generate indicator values directly rather than simulating species-level time series and aggregating them via the Freeman method. This simplifies the framework and avoids additional assumptions about species-level dynamics and covariance structures. The variance and autocorrelation parameters estimated from the historical ESAI implicitly capture the aggregate variability arising from the underlying species dynamics. A species-level simulation approach could provide a more mechanistic representation but would introduce substantial additional complexity without a clear improvement for the purposes of WP3.

The simulation functions developed in WP2 are provided as a freely available, open-source R package, *oepsims*[7], and are used in WP3 to assess the performance of candidate statistical methods against simulated scenarios of ESAI change.

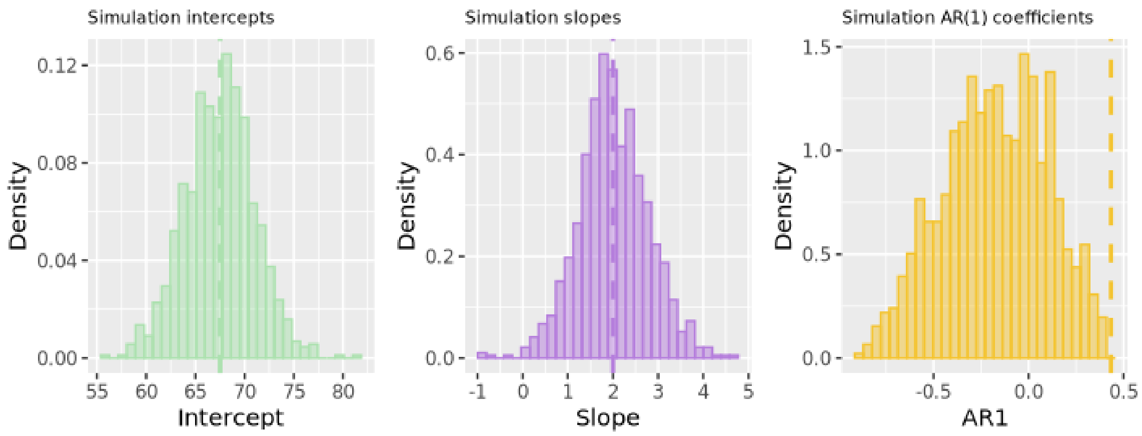


Figure 10: *Distributions of intercept (green), slope (purple) and AR(1) (yellow) parameters recovered from the 1,000 simulated seven-year time series. Dashed lines show the true input parameter values. Intercept and slope are recovered well, while the AR(1) parameter is strongly underestimated due to the short series length.*

[7] https://gitlab.bioss.ac.uk/grobert/oep-species-abundance-indicator

6. WP3: Comparison of Statistical methods

In Work Packages 1 and 2 we examined historical multi-species indicator data to determine the characteristics of the time series, and developed functions in R to simulate new ESAI time series using parameters estimated from the historical assessment in WP1. In Work Package 3 (WP3) we evaluate the effectiveness of candidate statistical methods proposed in Burns et al. 2025 for detecting change in biodiversity indicator datasets simulated according to specified scenarios of change.

Evaluation focuses on a subset of the sixteen statistical approaches outlined in Burns et al. 2025 (hereafter referred to as *options*), excluding those deemed unsuitable based on the characteristics of the ESAI identified in WP1. Here we directly address the following questions:

1. a) Which statistical approaches can robustly determine whether the simulated ESAI shows a decline between 2029 and 2030 (the comparison specified in current legislation), and b) whether including additional ESAI values after 2030 increases confidence in assessing whether this species abundance target has been met.
2. Which statistical approaches can identify significant change in the global indicator over the characteristic timescales identified in WP1 for each ESAI smoothing option, a) SO1: ~7 years, and b) SO2: ~4 years.
3. Which statistical approaches can reliably detect whether the 2042 ESAI value satisfies the species abundance target that the value exceeds the value in 2022 and has increased at least 10% between 2030 and 2042.

The statistical options will be assessed using performance metrics evaluating accuracy (i.e. probability of correct identification of halt/increase), together with false positive and false negative rates. Where possible, emphasis will be placed on minimising both false positives (which risk undermining public and stakeholder trust, premature policy relaxation, and misallocation of conservation efforts) and false negatives (which risk delaying recognition of success, failing to reward effective policy, or misrepresenting ecological trends), reflecting the importance of detecting genuine progress towards biodiversity targets. These metrics will identify the options that are sensitive to noise, smoothing and short-term fluctuations. All analyses were conducted in R (version 4.5.0, R Core Team 2025), and model fitting was carried out using the *mgcv* package (Wood 2017) unless otherwise stated. Code and result summaries are fully documented to ensure transparency and reproducibility.

6.1. Selecting statistical options

Burns et al. 2025 describes a suite of statistical approaches that could be applied to a species abundance indicator to assess whether biodiversity targets under the secondary legislation have been met. There, sixteen candidate options were collated from the literature and characterised according to several properties relevant to policy assessment, including the propagation of uncertainty, risk of false positive and false negative outcomes, ease of communication, and compatibility with the legislative framework. These options represent combinations of different model simplification approaches and statistical questions that can be asked of indicator time series. Detailed descriptions of all options are provided in Burns et al. 2025.

The options span a range of analytical approaches. Six test whether a single rate of change is greater than or equal to zero, including the approach specified in the secondary legislation and related variants that incorporate uncertainty around the estimated rate of change. Other options assess change over longer time periods, for example by fitting linear trends or comparing rates of change between different periods. A further group of methods categorises the rate of change relative to predefined thresholds, while several approaches assess whether recent change differs from historical trends using breakpoint or piecewise regression techniques.

From this set of sixteen candidate methods, we select a subset for evaluation using simulated indicator time series. Selection was guided by the characteristics of the ESAI identified in WP1 and by the relevance of each option to the policy questions addressed in WP3. In particular, the historical analysis of the ESAI (1970–2023) found no evidence of abrupt structural changes in the indicator and instead indicated gradual change occurring over multi-year timescales. Methods designed primarily to detect abrupt change points or piecewise shifts in trend were therefore not considered appropriate for evaluating the scenarios simulated here. Similarly, some options in Burns et al. 2025 were intended primarily as interpretative or descriptive frameworks rather than formal statistical tests (e.g. options based on categorical interpretation of trend magnitude). These were not included as core candidate tests, although one representative categorisation approach was retained to assess whether such methods might nonetheless provide useful diagnostic information.

Applying these criteria resulted in a set of eight candidate methods for evaluation (Options 1–8 in Burns et al. 2025 terminology). Different subsets of these options were applied to address the specific policy questions outlined below. Descriptions of each method and their original references are provided in Burns et al. 2025; we therefore provide only brief summaries of the selected approaches here. The objective of this selection was not to identify a preferred method *a priori*, but to evaluate a representative set of candidate approaches spanning the principal classes of statistical tests proposed in Burns et al. 2025.

6.1.1. Options 1 – 3: Rate of change assessment

Options 1–3 share the same underlying model structure; estimating the rate of change between the penultimate and final years of interest using a penalised spline GAM fitted to the ESAI time series (e.g. Wood 2017, GAM parameters match SO1 for Q1, Q2a & Q3, SO2 for Q2b). These methods differ in the decision rule used to interpret uncertainty in that estimate. Although SO1 and SO2 are only directly compared in Q2, the general effect of smoothing parameterisation was explored in Q1 as part of the analysis of the sensitivity of the species abundance target to small-scale trend variation (see Section 6.2.3).

Option 1: estimates the rate of change as the difference between the GAM-fitted values in 2029 and 2030. The decline is considered to have halted if this estimated change is non-negative. This corresponds directly to the legislative test, which evaluates whether the ESAI value in the final year is equal to or greater than that in the preceding year. Notably, Option 1 uses only the point estimate of the change — it does not account for the uncertainty around that estimate. For each simulated scenario, we recorded whether the estimated change between 2029 and 2030 was non-negative (decline halted) or negative (decline continuing).

Option 2: assesses whether the decline has halted using a Bayesian framework. The posterior distribution of the estimated rate of change in 2030 is used to quantify uncertainty. The probability that the decline has halted is defined as the proportion of the posterior distribution that lies at or above zero. To make large-scale simulation computationally tractable, the Bayesian posterior is approximated using the frequentist optimisation built into the *mgcv* GAM framework, which yields estimates equivalent to the maximum *a posteriori* solution conditional on priors (Miller 2025). For each simulation, we recorded the posterior probability that the rate of change in 2030 was positive (one-tailed) (greater or less than zero) and the posterior probability that it differed from zero in either direction (two-tailed).

Option 3: assesses whether the decline has halted using a frequentist hypothesis test (Gregory et al. 2004). The rate of change in 2030 is estimated from the model derivative. Unlike Option 2, which interprets uncertainty probabilistically via the posterior distribution, Option 3 applies a conventional significance test: the decline is considered to have halted if the null hypothesis of no change is rejected in favour of a non-negative rate of change at the 95% confidence level. For each simulation, we recorded whether the null hypothesis was rejected under three alternative hypotheses: (i) rate of change greater than zero, (ii) less than zero, and (iii) different from zero (two-tailed).

6.1.2. Option 4: Multi-year comparison

Option 4: compares the ESAI value at two specified time points using GAM-fitted values (DEFRA 2025c). This option defines the indicator as increasing or decreasing where the 95% uncertainty interval for the later year lies entirely above or below that of the earlier year. We calculate the difference between fitted values at the two years of interest with its standard error derived from the GAM's variance–covariance matrix, which correctly accounts for the correlation between predictions at nearby years. This is used to compare the two values using one- and two-sided t-tests. For Q1 we compare 2029 with 2030 (the legislative halt-in-decline test); for Q2 we compare 2025 with 2030 (a five-year interval). For each simulation, we recorded whether the null hypothesis of no change was rejected under three alternative hypotheses: (i) the later value is greater than the earlier value, (ii) less than, and (iii) different from (two-tailed).

6.1.3. Options 5 & 6: Linear trend estimation

Options 5 and 6 assess whether the ESAI exhibits a directional trend over a specified period using linear models fitted to the indicator time series. Unlike Options 1–3, which compare specific time points, these methods evaluate the overall trajectory of the indicator over a defined window.

Option 5: tests whether the ESAI shows a significant linear trend over a specified period. A linear model is fitted to the indicator time series with year as the explanatory variable, and the slope coefficient is tested against a null hypothesis of zero change. The original formulation in Burns et al. 2025 assessed whether the distribution of average annual growth rates across taxonomic groups differed from a zero-centred distribution; here we apply the equivalent test to the ESAI using the slope of the fitted linear model. For each simulation, we recorded the p-value of the slope coefficient and whether it was significant at the 5% level.

Option 6: fits a linear model including time as a covariate and compares it to an intercept-only model to assess whether a temporal trend is present in the ESAI time series. Following Schmidt and Meyer (2008), a secondary test evaluates whether population growth rate depends on population size (density dependence). Akaike weights from the two model comparisons are combined to produce a single probability that the ESAI exhibits no directional trend. For each simulation, we recorded the combined probability of no trend derived from the Akaike weights.

6.1.4. Options 7 & 8: Categorization and equivalence approaches

Options 7 and 8 assess change in the ESAI relative to predefined thresholds rather than testing directly whether change differs from zero. These approaches therefore provide an interpretation of indicator change relative to specified bounds representing meaningful increases, decreases, or negligible trends.

Option 7: classifies the GAM-estimated rate of change at the final year as a proportion of the indicator level, applying predefined thresholds used in several UK biodiversity indicators (Eaton et al. 2015; DEFRA 2025a,c). Categories represent strong increase (≥ +100%), weak increase (+33% to +100%), little change (−25% to +33%), weak decline (−25% to −50%), or strong decline (≤ −50%). This approach provides an interpretable classification of ESAI change but does not constitute a formal statistical hypothesis test. For each simulation, we recorded the assigned category. The decline was considered to have halted if the simulation was classified in any category other than weak or strong decline.

Option 8: uses an equivalence test to determine whether a linear trend differs meaningfully from zero (Dixon and Pechmann 2005). A linear model is fitted to the log-transformed indicator over a 20-year window. The null hypothesis is that the estimated trend lies outside a predefined equivalence region, and this is tested using two one-sided t-tests (TOST) applied to the slope. Following Burns et al. 2025, the equivalence bounds correspond to a halving or doubling of the ESAI over the 20-year window (2010–2030 for Q1a, 2012–2032 for Q1b; see Section 5, Q1). For each simulation, we recorded the decline as halted if the estimated trend fell inside the equivalence region (negligible change) or if the estimated slope was non-negative (indicating stability or improvement). Only trends that were both outside the equivalence region and negative were classified as continuing decline.

6.2. Which statistical approaches can robustly determine whether the simulated ESAI shows a decline between 2029 and 2030 (Q1a), and does

including additional years of ESAI values beyond 2030 increase confidence in identifying this (Q1b)?

We simulated the ESAI from 2024 to 2030 to compare the performance of the selected options assuming scenarios of no change, positive change, or negative change between 2029 and 2030. We chose slope parameters ranging from a 4% annual decrease to a 4% annual increase, in steps of 0.25%, (33 slope parameters) to explore the performance of the options across a range of plausible change scenarios. 1000 simulations were performed for each slope value, with other simulation parameters set from the parameter estimations from the historical assessment (Specifically SO1: intercept = 67.5; sigma = 3.14; AR(1) = 0.43). We also assessed the effect of including additional years of ESAI values beyond the 2030 target on the options capacity to robustly identify changes performing the same simulation but extending them from 2024 to 2032.

Options 1, 2, 3 and 4, are appropriate for this assessment, and we include Options 7 and 8 as examples of categorical tools. Each evaluates change at, or between, specific time points. To assess their performance, we combine the historical ESAI values and the simulated values together as one series and fit them using a penalised GAM with the same parameters as the fits used for the historical assessment in WP1 (for SO1: k=10, AR(1) = 0.43, sigma = 3.14). Options 1–3 estimate the rate of change (derivative) of fit at 2030, while Option 4 compares the GAM fit values in 2029 and 2030 explicitly. Options 7 and 8 are included as categorical and equivalence-based alternatives that apply different decision rules to the same underlying quantities (the estimated rate of change or trend magnitude near the assessment date, respectively). Options 5 and 6 are not included in the Q1 assessment because they estimate a linear trend over a defined multi-year window, which addresses a different question; whether there is a directional trend over a period, rather than whether the indicator value has changed between two specific adjacent years. These options are better suited to the multi-year assessments considered in Q2 and Q3.

Figure 11 shows the performance power curves for Options 1, 2, 3, 4, calculated from 1000 simulated ESAI SO1 realizations with underlying trend slopes ranging from -4 to +4 (33 steps, slope increments = 0.25%), out to either 2030 or 2032. We fit the results with a binomial generalised linear model allowing interactions between the Options and the slope. These curves can be used to quantify how large the underlying driving trend must be before each option detects the trend at a chosen confidence level. To use the curves, identify the confidence level required, then read off the corresponding underlying trend slope that the indicator would have to achieve for each test to meet this confidence level. The dashed lines show an example of this for a 90% confidence level (see Table 1 for values).

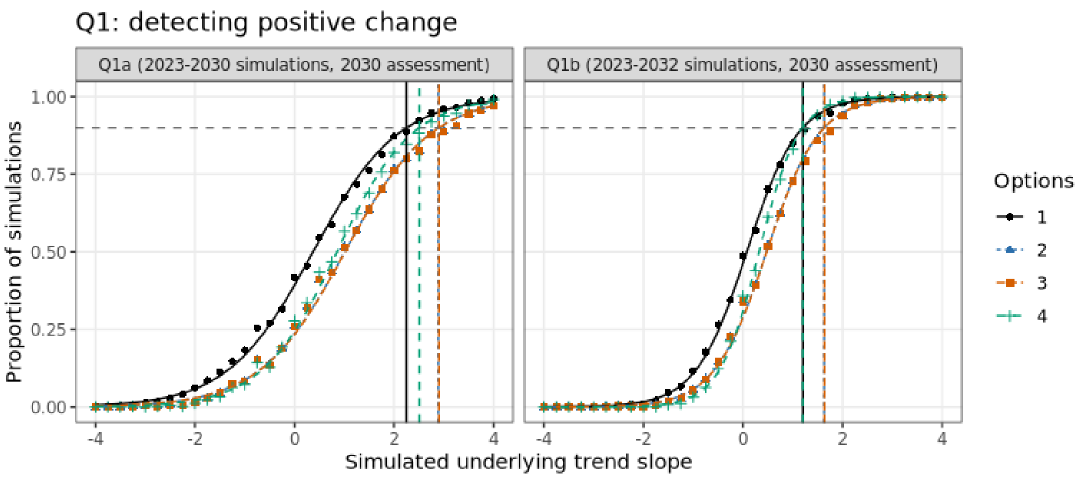


Figure 11: *Q1 power curves showing the proportion of 1,000 simulations, at each underlying trend slope, in which the statistical tests conclude that decline has halted between 2029 and 2030. Results are shown for Options 1–4 (black, dark blue, orange and cyan, respectively). The left panel (Q1a) uses simulations extending to 2030, while the right panel (Q1b) uses simulations extending to 2032. Points represent simulated values and solid lines show binomial generalised linear model fits to these points for each option.*

The horizontal dashed line (grey) marks the 90% threshold, while vertical lines indicate the minimum simulated slope required for each option to achieve this level of power.

Figure 11 shows the performance power curves for Options 1, 2, 3, 4, calculated from 1000 simulated ESAI SO1 realizations with underlying trend slopes ranging from -4 to +4 (33 steps), out to either 2030 or 2032. We fit the results with a binomial generalised linear model allowing interactions between the Options and the slope. These curves can be used to quantify how large the underlying driving trend must be before each option detects the trend at a chosen confidence level. To use the curves, identify the confidence level required, then read off the corresponding underlying trend slope that the indicator would have to achieve for each test to meet this confidence level. The dashed lines show an example of this for a 90% confidence level (see Table 1 for values).

6.2.1. Q1a: 2024-2030 simulations; 2030 assessment

Table 1 shows the key performance characteristics for Options 1-4 for detecting "decline halted". We count true positives (TP; simulations with underlying trend slopes > 0, which the methods identify as "decline halted") and false positives (FP; simulations with underlying trend slopes < 0, which the methods identify as "decline halted") and compare the Options' ability to identify TPs (Figure 11, underlying trend slopes > 0; Table 1, slope at 90% values) and their capacity to balance this against rejecting FPs (Figure 11, underlying trend slopes < 0; Table 1, FP/TP ratio), where lower values indicate fewer false alarms relative to real detections

Option 1 is the most powerful of the Options (lowest required slope to achieve a given proportion of simulations identifying "decline halted"), however, this is misleading. Option 1 does not account for, or quantify, uncertainty, resulting in a substantially higher FP/TP ratio than Options 2, 3 & 4 (0.14 c.f. ~0.08, Table 1). This marks Options 2, 3 & 4 as substantially more discriminating, particularly for small absolute values of the underlying trend. Options 2, 3 and 4 improved discrimination performance is driven by their explicit treatment of uncertainty, which makes them more statistically robust, but also more conservative (requiring larger signals for them to detect change, 2.5 - 2.9 c.f. 2.3 at 90% Table 1). Options 2 and 3 produce near-identical power curves because both test the same GAM derivative at 2030, but with different statistical approaches. Option 2 uses the proportion of posterior samples and Option 3 using a frequentist t-test, both applying a $p < 0.05$ significance threshold. These converge to the same result when Options 2's prior is uninformative and with a large number of posterior samples. Option 4 constructs a linear contrast between the GAM fit's predicted values in 2029 and 2030, computing the standard error of that difference from the model's variance-covariance matrix, and testing significance via a t-statistic, again using a $p < 0.05$ significance threshold. Its performance is marginally better than Options 2 & 3, with a similar FP/TP ratio but better detection characteristics (Table 1, slope at 90% values).

Option	Slope at 90%		FP/TP ratio
	Q1a: 2023-2030	Q1b: 2023-2032	
1	2.3	1.2	0.14
2	2.9	1.6	0.08
3	2.9	1.6	0.08
4	2.5	1.2	0.07

Table 1: *Q1 performance characteristics (i.e. slope at 90% confidence threshold and false positive/true positive ratio) for Options 1-4 under SO1*

For the SO1 GAM, the power curves are not centrally aligned around a slope of zero. The SO1 GAM is a relatively strong smoother, with limited flexibility over the ~65-year combined series (historic + simulated). The long historic decline of the ESAI results in the GAM fits across the time series carrying some downward momentum into the 2024 to 2030 simulation period. When the underlying simulation trend slope is zero or weakly positive, this momentum biases the estimated derivative negative, shifting the ability of the methods to identify "decline halted" resulting in lower performance than expected for a given underlying trend slope. This is reflected in the crossing point of these curves for an underlying trend slope of zero (~44% for Option

1, compared to an expectation of ~50%, and ~26-28% for the more conservative Options 2-4). This highlights the sensitivity of these tools to the GAM smoothing parameterisation and the historic ESAI conditions (see section 5.2.3).

6.2.2. Q1b: 2024-2032 simulations; 2030 assessment

All the Options show improved performance when the simulations extend out to 2032 rather than 2030 (Q1b vs Q1a, Figure 11, Table 1). This improvement is driven by the additional two years of post-2030 data reducing the GAM's 'end-effect' at the 2030 test point. Smoother fits such as GAM fits are inherently less well constrained at the boundaries of the data, where the curve of the model is informed by observations on only one side, resulting in significant flexibility. Here, the SO1 GAM fit is parameterised to have ten basis functions (approx. 1 for every 6 years of data) and the simulated data is influenced by the downward momentum from the fit to the historic portion of the ESAI (see Section 5.2.1), with no data beyond 2030 to help anchor the fit. In Q1b, 2030 becomes an interior point and the two subsequent years of data help constrain the smooth more tightly, counteracting the bias (Option 1: ~49% for an underlying trend slope of zero, compared to 50% expectation).

To explore this further, we ran the simulations from 2024 to 2037 and examined the performance of Options 1-4 as the length of the simulated indictor extension beyond 2030 increased (Figure 12). Extending the indicator series beyond 2030 improves the uncertainty of the estimated derivative of the GAM model, resulting in narrower confidence intervals and improving detection power across all Options for detecting positive change (Figure 12, lower panel). The change in detection power decreases smoothly, but not linearly, levelling off after 2034 (+4 years of data). In subsequent years, the performance does not improve significantly suggesting that the GAM fit is well constrained by this point. The proportion of simulations returning decline halted for an underlying trend slope of zero rises slightly as the length of the additional data increased, showing that the additional data also helps mitigate some of the impact of the historic decline on performance. Together, these suggests that statements made about whether the decline has been halted, made in each subsequent year after 2030, will increase in confidence and that assessment of the legislative test can be continuously improved.

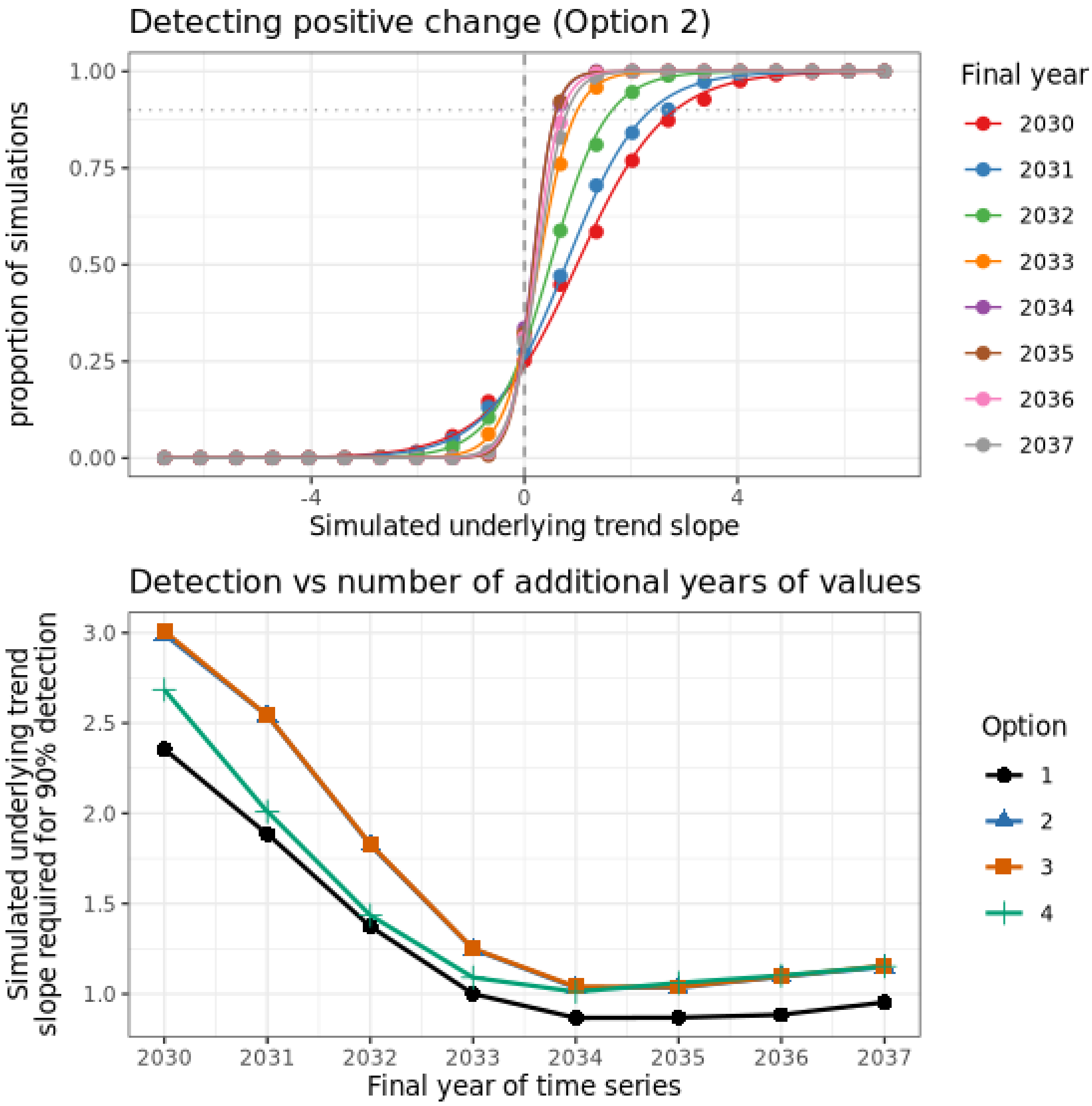


Figure 12: *Sensitivity of decline detection to time series length. Top panel: Option 2 (Bayesian credible interval, shown as a representative example) power curves showing the proportion of simulations that detect positive change across the simulated underlying trend slopes for series ending in 2030-2037 (red, through grey, respectively). The dotted grey line marks 90% detection power. Bottom panel: The simulated underlying trend slope required for 90% of simulations to identify positive change as a function of the final year of the simulated indicator, for Options 1-4 (black through green).*

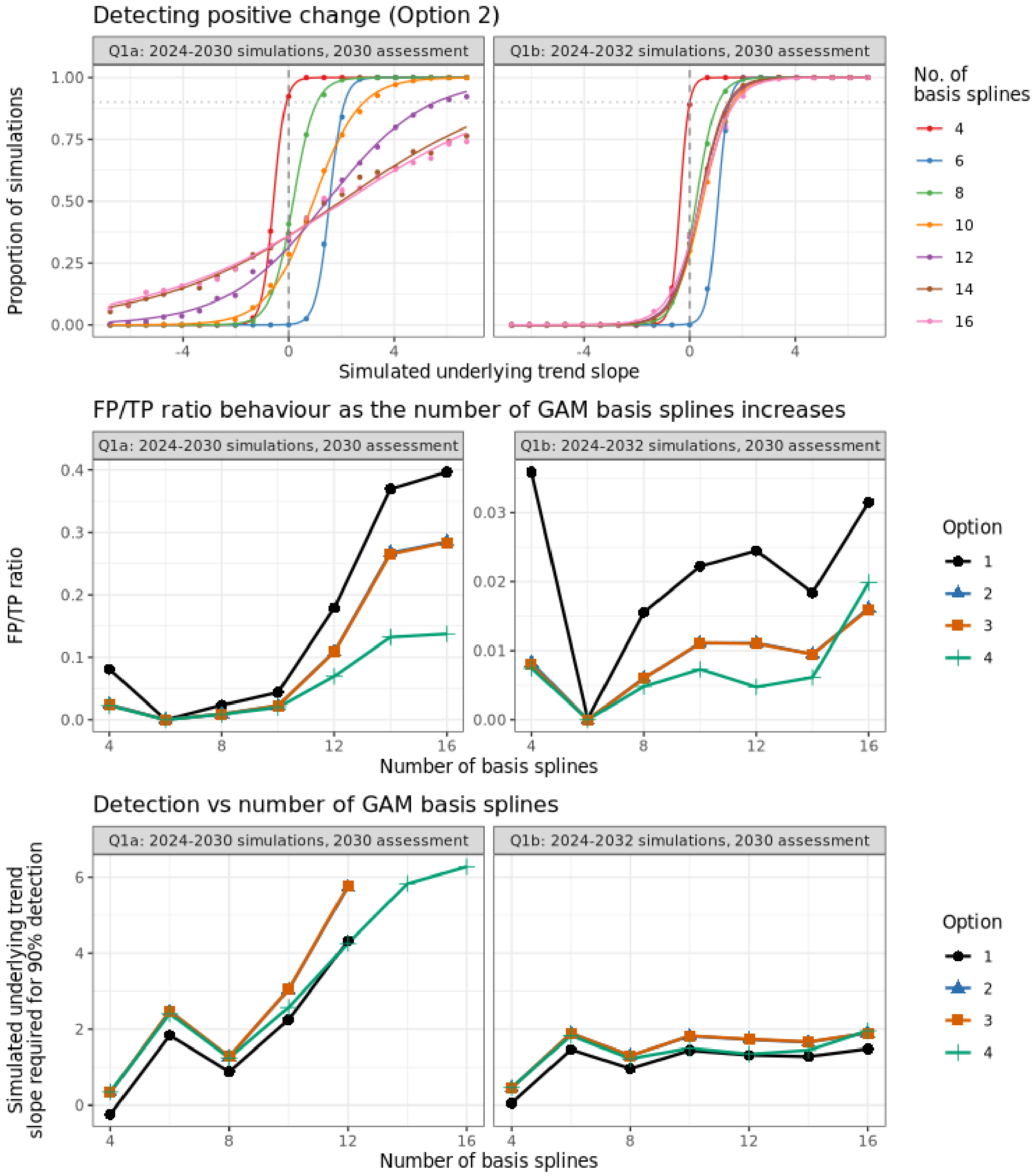


Figure 13: *Sensitivity of Q1 Option performance to the GAM basis dimension (the maximum number of basis splines, k), for simulations from both 2023 - 2030 (left panels, Q1a) and 2023 – 2032 (right panels, Q1b). Top row: Option 2 (Bayesian credible interval, shown as a representative example) power curves showing the proportion of simulations that detect positive change at each simulated slope, for k = 4 – 16 (SO1 has k = 10). Middle row: FP/TP ratio integrated across all underlying trend slopes. Note the different y-axis scales between Q1a and Q1b. Bottom row: The simulated underlying trend slope required for 90% of simulations to identify positive change as a function of k for Options 1-4 (black through green). Missing points (left panel, k=14 & k=16) indicate that the power curve does not reach 90% within the simulated underlying trend slope range.*

6.2.3. Q1: The impact of smoothing parameterisation

WP1 (Section 3) and Sections 5.2 and 5.2.1 all highlight that, for Q1, Options 1-4 are all sensitive to the smoothing parameterisation of the penalised-spline General Additive Model used to smooth the data during the construction of the ESAI.

The GAM basis dimension, k, controls the maximum flexibility available to the fitted spline while the realised smoothness is determined jointly by this basis dimension and the smoothing penalty. When k is too small, the fitted curve is constrained and may be unable to represent genuine short-term variation in the time series. With small k, local changes are absorbed into the broader long-term trend, and historical patterns carry momentum into the endpoint of the fitted series. This can bias point estimates of recent change and affect uncertainty quantification. Increasing k gives the model greater capacity to represent local features, where these are supported by the data. Although this reduces underfitting, larger k increases the model's sensitivity to noise if the fitted smooth is insufficiently penalised and greater flexibility near the boundary of the fit increases uncertainty in both fitted values and derivatives, amplifying end-effects and inflating standard errors. The choice of basis dimension therefore involves a trade-off between bias (too few basis functions) and variance (too many), with the optimal value depending on the question being asked, the length of the time series, and the timescale of the changes of interest.

Penalised splines can resolve the bias-variance trade-off using a smoothing parameter. The smoothing parameter shrinks the smoother's basis parameters by penalising how the model fits the data (provided the maximum number of basis splines, k, is overspecified, serving now as an upper bound on complexity). The smoothing penalty thus balances between fit and smoothness (via generalised cross-validation or Restricted Maximum Likelihood estimation).

Here we explore the sensitivity of Options 1-4 by running simulations from 2023 out to 2030 and 2032, using the SO1 historical ESAI values both for the historical values and for the underlying variance and autocorrelation characteristics of the simulation. We fit the historical plus simulated indicator values with penalised-spline GAM models with k = 4-16 basis functions (for reference, SO1, k = 10; SO2 k = 23) and examine the performance of Options 1-4 as the maximum number of basis splines increases (Figure 13).

Since the model fit underlying SO1 is based on a maximum of 10 basis functions, we expect that model parameterisations with $k > 10$ should constrain their fits through their penalisation, resulting in similar curves and performance. For low values (k = 4 & 6) we see no evidence of penalisation since k is below the true model complexity (e.g. the number of basis coefficients used in the DEFRA code). The fits for both Q1a and Q1b simulations struggle to model the indicator, and significant bias is introduced to the test Options as historical ESAI patterns carry momentum into the endpoint of the fit. Further, small changes in k result in large shifts in the x-axis location of knots and position of the power curves, as the stiff GAM fits shift this bias. At values (k = 8 & 10) the impact of the position of the knots stabilises for both Q1a and Q1b simulations. At higher values (k = 12 – 16), the curves for Q1b remain stable as penalisation reduces the complexity of the more flexible models.

For Q1a simulations, although the detection power for a slope of zero remains approximately constant as k increases, the false positive ratio rises sharply (approaching ~40% for Option 1, ~30% for Options 2 & 4, and ~15% for Option 4; Figure 13, middle panels), as the Options increasingly lose power to detect a halt in the decline of the indicator. The Option performance is driven by the flexibility of the GAM fit inflating the variance of the end point rate-of-change estimate. The behaviour does not stabilise as k increases, suggesting the penalization applied by the GAM fits does not help control this end-effect. By contrast, for the Q1b simulations, both the position and the detection power of the options stabilise above $k = 10$, with the FP ratio remaining low and stable and a constant stable value for the underlying trend slope required for 90% detection of positive change (Figure 13, lower panels). This suggests that the fits are not over-fitting the indicator values and that the stabilisation is doing a good job of converging the model complexity to a single equivalent model. This highlights the importance of having additional years of values beyond 2030 to help constrain the GAM fits and improve the performance of the statistical tools (see Section 5.2.2) and suggests a potential diagnostic for empirically deriving an appropriate smoothing for the raw data underlying the ESAI.

We caution that substantial temporal autocorrelation presents a challenge for the penalisation and could lead to GAM smoothing penalties to be poorly calibrated, with knock on consequences for the statistical assessment in 2030. Typically, penalised spline GAM fits assume independent residuals when estimating the smoothing parameter. This could cause the model to overestimate the amount of independent information in the data, under-penalising *k*. Without additional years of data post-2030, this will inflate the variance of the end point rate-of-change measurements, decreasing the statistical power of the Options.

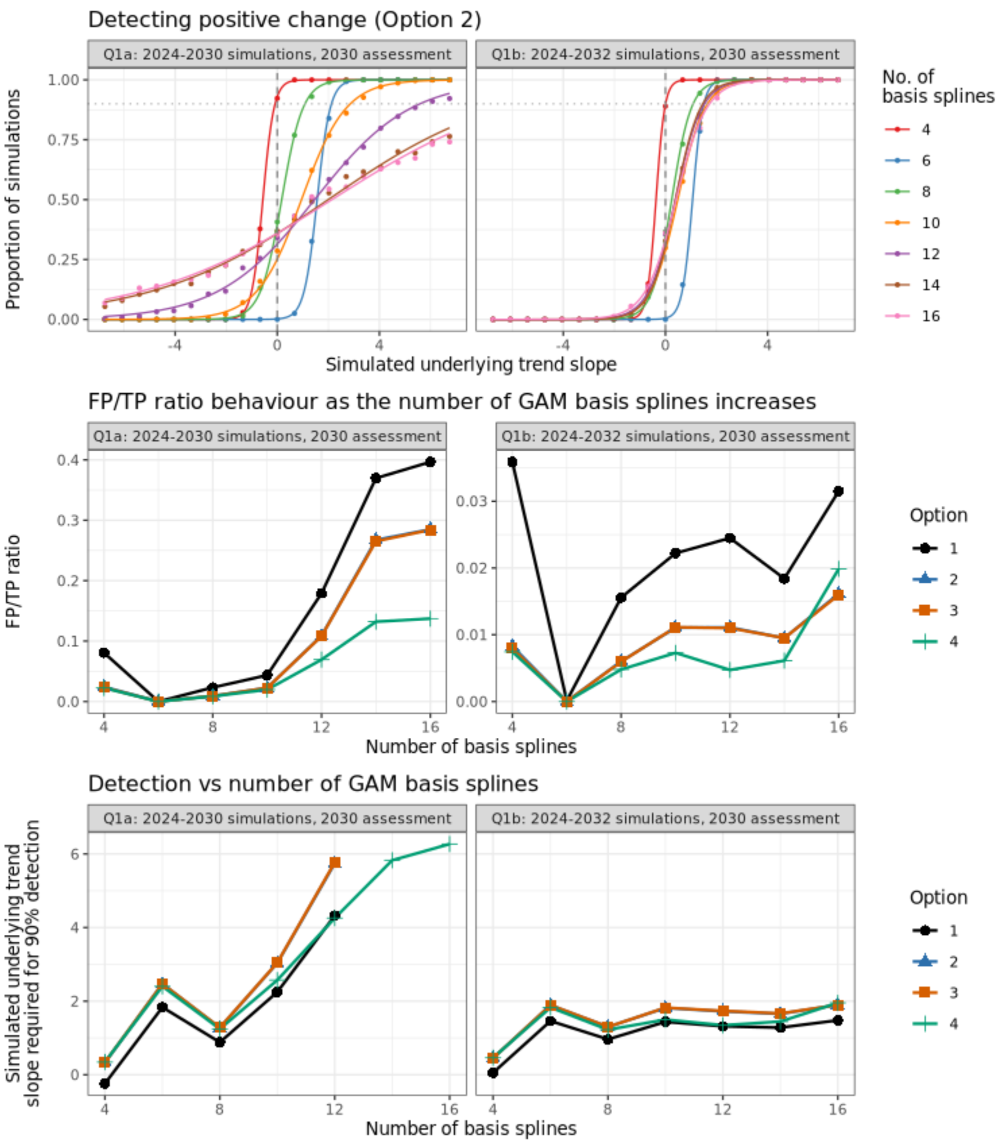
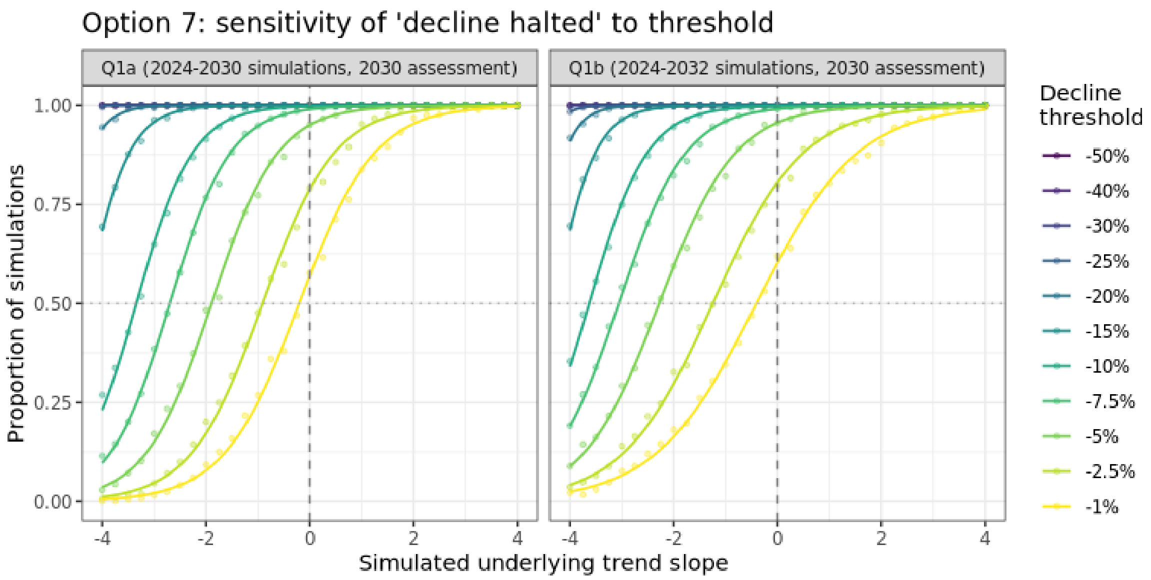


Figure 13: *Sensitivity of Q1 Option performance to the GAM basis dimension (the maximum number of basis splines, k), for simulations from both 2023 - 2030 (left panels, Q1a) and 2023 – 2032 (right panels, Q1b). Top row: Option 2 (Bayesian credible interval, shown as a representative example) power curves showing the proportion of simulations that detect positive change at each simulated slope, for k = 4 - 16. Middle row: FP/TP ratio integrated across all underlying trend slopes. Note the different y-axis scales between Q1a and Q1b. Bottom row: The simulated underlying trend slope required for 90% of simulations to identify positive change as a function of k for Options 1-4 (black through green). Missing points (left panel, k=14 & k=16) indicate that the power curve does not reach 90% within the simulated underlying trend slope range. Note that Options were tested using simulated ESAI values assuming SO1 (where k = 10).*

6.2.4. Q1: Categorical tests

The categorical tools, Options 7 and 8, in the original configurations specified from their references, are largely insensitive to detecting change in this range, classifying all the change values as either “little change” or “inside the equivalence region”. The Default categories for these tools were designed for very different purposes than quantifying change in the ESAI (wild bird population monitoring, Gregory et al. 2024; amphibian population monitoring Dixon & Pechmann 2005), and these results demonstrate that these defaults are hugely conservative and inappropriate for use in this context. The threshold values hold meaning, such as a doubling in 10 years (Option 8), however the choice of these thresholds is essentially arbitrary and can be adjusted to alter the behaviour of the test - in the same way that using a p <= 0.05 threshold for Options 2, 3 & 4 is an arbitrary choice and can be adjusted to alter the behaviour of the test. Figure 14 shows how the statistical power of these options to identify “decline halted” changes as the category thresholds they use are reduced, for the SO1 2023-2030 and 2023-2032 simulations (Q1a and Q1b).

Option 7 computes the ratio of the point estimate of the derivative of the GAM fit in 2030 to the value of the fit in 2029 for its classification, requiring this be above a threshold value. This approach allows the tool to be flexible to recent short timescale changes in the indicator and makes the test responsive to threshold tuning. However, the point estimate on the percentage scale makes the same absolute change more impactful for small indicator values than for larger indicator values, resulting in a power curve that flattens as the thresholds are tightened. The result is that although the model gains sensitivity, it also loses some discriminatory power as the thresholds are tightened. Despite improving performance as the threshold approaches ~1% (a slope of 0.7 for an indicator value of 70), Option 7 never performs as well as Options 2, 3 or 4. Uniquely, this effect is more prominent for the Q1b simulations, making the performance of Option 7 worse with additional data after the assessment point. Here, the more constrained GAM fit with the 2023-2032 simulations tends to attenuate the derivative of the GAM fit at the assessment point, shrinking the equivalent percentage change toward zero, making the test more conservative and reducing performance.



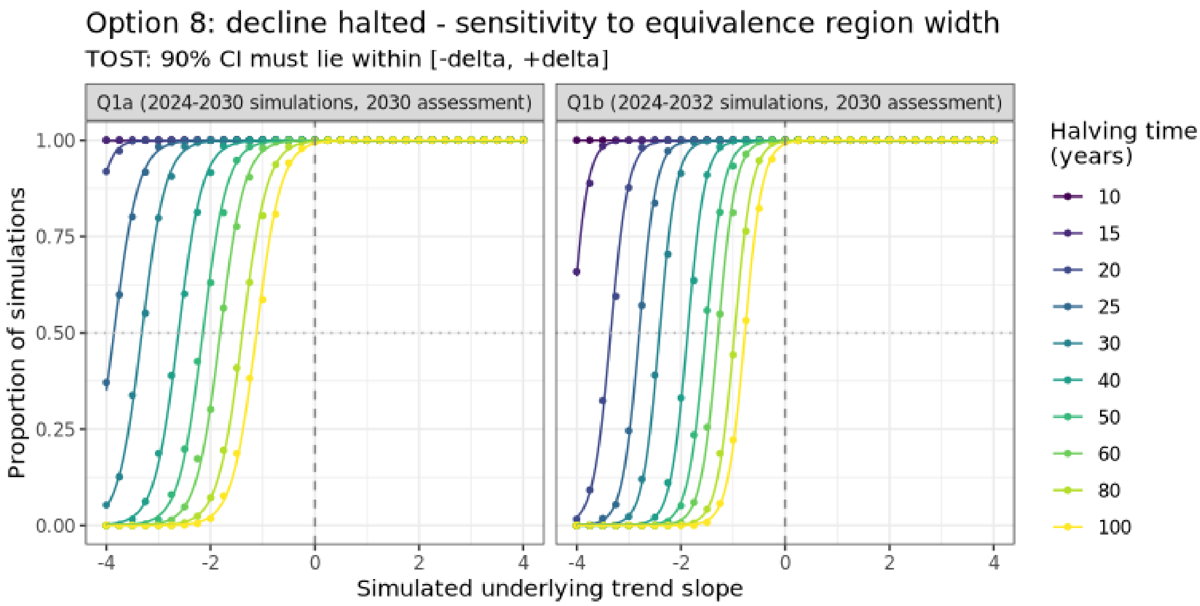


Figure 14: *Q1 power curves showing the proportion of 1000 simulations, for each underlying trend slope, in which the statistical tests conclude that decline has halted between 2029 and 2030. Results are shown for Option 7 (top panels) and Option 8 (bottom panels). For Option 7, curves correspond to different decline thresholds (the lower limit of the "little change" category, default -25%). For Option 8, curves correspond to different halving times defining the width of the equivalence region used in the Two One-Sided Test (TOST; default = 20 years). Colours indicate the threshold or equivalence-region width used by the test: darker colours represent broader, more conservative category widths, while lighter colours represent narrower, more discriminating thresholds. The left panels (Q1a) use simulations extending to 2030, while the right panels (Q1b) use simulations extending to 2032. Points show simulated values and solid lines show binomial generalised linear models.*

Option 8 fits a linear model to the log-transformed indicator over the full simulation window and tests whether the estimated slope falls within an equivalence region that corresponds to a halving or doubling time of *t* years. Decline is classified as halted if the slope is inside this region or falls outside it and is non-negative. Because this approach estimates a single slope from all data points in the window rather than a local derivative, its decision statistic is well-determined, and the test is discriminatory (resulting in power curves ~3x steeper than Option 7). Similarly to Option 7, as the equivalence region width is narrowed (the halving/doubling time increased) the sensitivity is increased, shifting the power curves to the right. However, unlike Option 7, Option 8 shows no loss of discriminatory performance flattening. Increasing the halving time to ~50–100 years brings the test into a useful operating range, though the choice of threshold remains reliant on a subjective decision about what constitutes a negligible rate of change. Unlike Option 7, Option 8 benefits from additional data after the assessment point, where the additional values improve the precision with which the linear slope that underlies the test can be measured. An interesting characteristic of this method is that the non-negative aspect of the decision rule means that narrowing the equivalence region never reduces the true positive rate, because even at a true change of zero roughly half of all estimated slopes are non-negative. This provides a natural limit on performance that Option 7 lacks, but it also means that Option 8 cannot fully eliminate false positives at mildly negative slopes.

Regardless of which Option is used, when the true underlying trend for the ESAI represents no change or an increase (the conditions set by the 2030 target), reliably detecting this from a single adjacent-year comparison would require an unprecedented year-to-year increase in species abundance. In the absence of such a significant change, it will not be possible to reliably distinguish ongoing decline from the desired halting or reversal of that decline by comparing the 2029 and 2030 ESAI values. It follows that the current secondary legislation governing the assessment of the species abundance target in 2030 is, therefore, not fit for purpose for determining whether the decline in species abundance has been halted.

6.3. Which statistical approaches can identify significant change in the global indicator over the characteristic timescales identified in WP1 for each ESAI smoothing option (Q2a: SO1, ~7 years, Q2b: SO2, ~4 years)?

Here we simulate the ESAI to compare the performance of the selected options assuming scenarios of no change, positive change, or negative change. We chose three underlying trend slopes representing significant change over seven years for simulations based on the characteristic timescale identified under SO1, and four years for simulations based on the characteristic timescale identified under SO2. Significant change in this context was defined as the final indicator value from the underlying trend being outside the current 95% credible intervals of the 2023 ESAI value (approx. ±10%), and the simulation underlying trend slopes were set accordingly (SO1: ±1.02 over 7 years, SO2: ±1.78 over 4 years). 1000 simulations were performed for each slope value, for each smoothing option, with the other simulation parameter set from the parameter estimations from the historical assessment (Specifically SO1: intercept = 67.5; sigma = 3.14; AR(1) = 0.43. SO2: intercept = 66.6; sigma = 3.46; AR(1) = 0.11).

Options 4, 5 and 6 are appropriate for this assessment, and we include Option 8 as an example of a categorical tool. Figures 15 and 16 show the performance characteristics of Options 4 and 5, and Options 6 and 8, respectively. For Options 4 and 5 (Figure 15), good performance requires both: detection - the statistical test correctly identifies a significant change (or no significant change in the case of a slope of zero) and coverage - that the 95% confidence intervals on the estimate of the magnitude of the underlying long-term trend include the true simulation trend value.

Option 4 achieves only 31–44% coverage across both Q2 time series, indicating that this method substantially underestimates the true uncertainty in the simulated indicator. Option 4 accounts for the covariance between GAM-fitted values at the two comparison years by using the full variance-covariance matrix of the smooth to construct confidence intervals on the difference, but it assumes that the model residuals are independent and does not account for their autocorrelation. Consequently, the method underestimates the size of the confidence intervals on the fit parameters, reducing coverage performance. For SO2, where the autocorrelation is weaker, the coverage performance is better (although still not good). For both SO1 and SO2, Option 4's detection performance is good for simulations of positive or negative change, however its false positive performance is poor for simulations with no change (high detection rate for underlying trend slope of zero; Figure 15). Option 4's detection rate is ~10% worse across all trends for the SO2 simulations. Despite the steeper slope in these simulations, the shorter test window in Q2b (4 years vs 7 years) significantly reduces the detection performance. Because the combined metric requires both correct change detection and accurate coverage of the magnitude of change, it is necessarily lower than either individual metric. Consequently, Option 4 rarely both identifies whether change has occurred and accurately estimates its magnitude. Options 4's performance demonstrates that fitting a GAM to the full ESAI time series and using it for formal inference about change over a sub-period can lead to unreliable uncertainty estimates and highlights the sensitivity of downstream statistical inference to the choice of smoothing parameters.

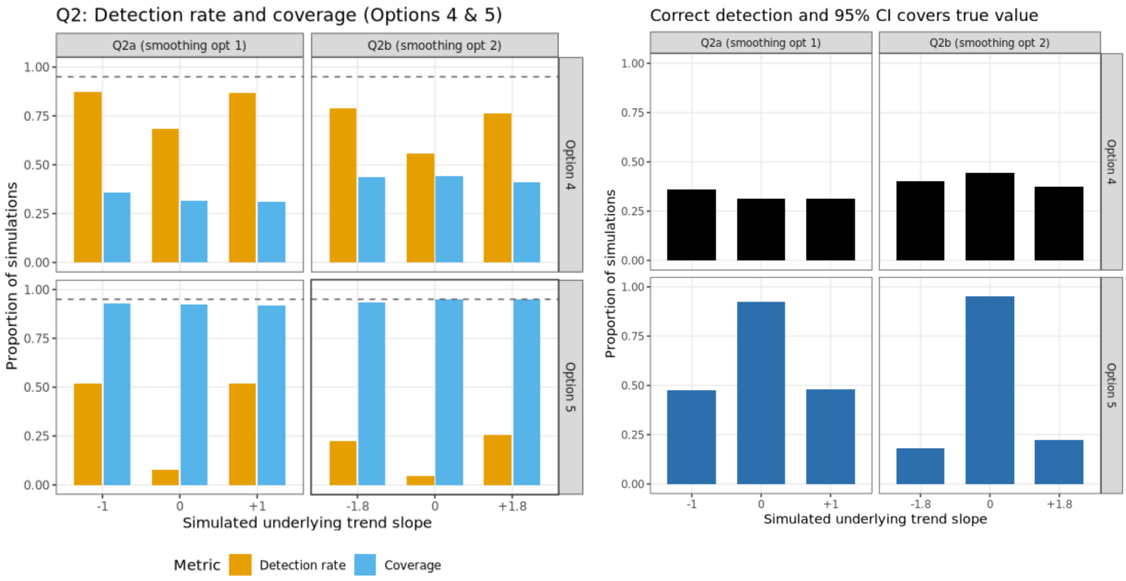


Figure 15: *Q2 performance histograms for Options 4 (top row) and 5 (bottom row) for three contrasting ESAI scenarios that capture 'significant change' (decrease, left bars; no change, central bars; increase, right bars) over timescales appropriate for the ESAI smoothing options (SO1, ~7 years, left columns; SO2, ~4 years, right columns). Significant change in this context was defined as +/-10%, the scale of the current 95% credible intervals of the ESAI indicator values. Left Panel: Detection rate (statistical test is significant) and Coverage (confidence intervals on trend estimation include the true value) metrics. Right Panel: combined performance requiring correct detection (significance for the positive or negative underlying trend slopes, not significant for slope = 0) and coverage*

In contrast, Option 5 assesses significance using a simpler linear model fitted only over the test window rather than a GAM fitted to the full time series. As a result, its performance is largely insensitive to the historical trend. Option 5 achieves 90-95% coverage across both Q2a and Q2b, substantially higher than Option 4 and approximately achieving the nominal 95% expected from a well-calibrated test. For the SO1 simulations, Option 5's detection rates are moderate and broadly symmetric between positive and negative slopes, suggesting that the test is approximately unbiased, and the low false positive rate when the slope equals zero indicates that the test controls Type I error reasonably well. For SO2 simulations the coverage remains good, but the shorter test window reduces the detection rate by ~50% for all change categories. Consequently, the combined metric for Option 5 is limited primarily by moderate detection power rather than by underestimation of uncertainty. An autocorrelation-corrected estimator (e.g. Generalized Least Squares with AR(1) errors rather than Ordinary Least Squares) may improve both coverage and power by producing more accurate standard errors.

Options 6 and 8 (Figure 16) are classification-based methods that classify change over time into coarse bins. Here we use the standard bins for these methods (i.e. those described in Burns et al. 2025), but we note that altering the bin boundaries will directly alter the perceived tool performance (see Section 5.2.4). Both Options show good specificity at zero slope (~85-95% correct classification) across both Q2a and Q2b series, but limited power to detect actual change. For non-zero slopes Option 6 achieves only ~40% correct classification for both Q2a and Q2b simulated indicator series, highlighting the conservative nature of this test. For non-zero slopes Option 6 achieves only ~5% correct classification for Q2a and ~35% for Q2b. The improved performance in Q2b highlighting that the ±0.035 equivalence bounds used by Option 8 are calibrated for much larger effects than the Q2a simulated slopes produce.

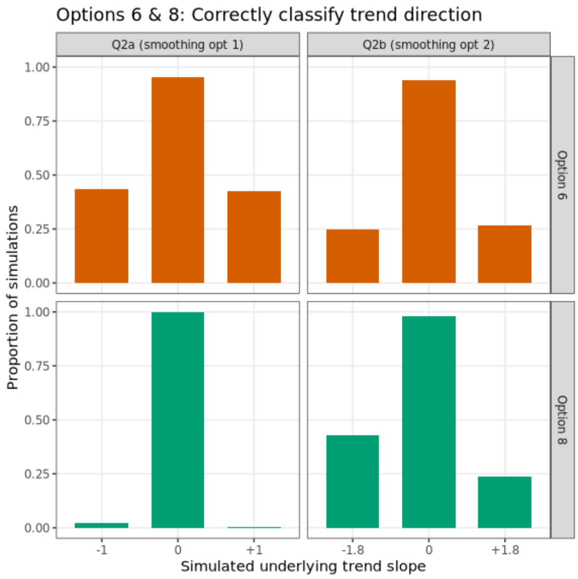


Figure 16: *Q2 Performance histograms for Options 6 (top row) and 8 (bottom row) for three contrasting ESAI scenarios that capture 'significant change' (decrease, left bars; no change, central bars; increase, right bars) over timescales appropriate for the ESAI smoothing options (SO1, ~7 years, left columns; SO2, ~4 years, right columns). Significant change in this context was defined as +/-10%, the scale of the current 95% credible intervals of the ESAI indicator values. Options 6 & 8 are classification-based methods and correctly classify the trend direction when the direction of the underlying trend slope matches the classification direction.*

6.4. Which statistical approaches can reliably detect an ESAI in 2042 that is excess of the 2022 level, and has increased by at least 10% between 2030 and 2042?

This legislative target has two components, an absolute threshold and a percentage increase. Because the current indicator level is approximately the same as the 2022 value, if the ESAI is flat or grows between now and 2030, all scenarios passing the 10% increase requirement will automatically pass the absolute threshold component of the legislative test. If the ESAI declines between now and 2030 then, if the decline is less than or equal to 10%, then any subsequent increase meeting the 10% requirement will also necessarily meet the 2022 threshold. If the decline is >10%, then a growth rate of >10% will be required to meet the 2022 absolute threshold. Since quantifying larger changes is generally easier for statistical tools than smaller changes, the change that is relevant for evaluating the performance of the statistical tool in the hardest case is their ability to quantify a 10% increase, and the absolute threshold component is not directly relevant. Nevertheless, because several of the statistical Options incorporate historical information in their test (as a GAM fit the entire indicator series, or as a fit to a long timescale window, for example), the history of the indicator is likely to be relevant for the performance assessment of the statistical Options.

To assess Q3, we generate 1000 simulated ESAI time series using SO1 parameters (intercept = 67.5; sigma = 3.14; AR(1) = 0.43) for the minimum increase requirement under three scenarios with more complex series histories:

1. A 10% increase between 2023 and 2030, followed by a 10% increase between 2030 and 2042.
2. No change between 2023 and 2030, followed by a 10% increase between 2030 and 2042.
3. A 10% decrease between 2023 and 2030, followed by a 10% increase between 2030 and 2042.

When this legislative target is assessed in 2042, the 2030 ESAI value will be a historical point, similar to the 2023 value for Q1 and Q2, and we treat it similarly by simulating change out to 2030, computing summary statistics (median value and confidence intervals), and using these summary values as the starting point for simulating change from 2030 out to 2042 (this allows us to use the same simulation functions from *oepsims*).

Unfortunately, the wording of the legislative target is ambiguous about what a 10% increase represents. The ESAI is a percentage of the value of the indicator in 1970, so a 10% increase could be interpreted as either a relative percentage (i.e. requiring ESAI = 50 to rise to 55) or as absolute percentage point (i.e. requiring ESAI = 50 to rise to 60). Here, we simulate the three scenarios under both definitions and contrast the results.

Options 4, 5 and 6 are appropriate for this assessment, and we include Option 8 as an example of a categorical tool. Figure 17 shows the performance of Options 4 and 5 for the Q3 assessment. As with Q2 (Section 5.3), we evaluate both detection rate (the proportion of simulations in which the statistical test correctly identifies a significant trend) and coverage (the proportion of simulations in which the 95% confidence interval on the estimated trend includes the true value). The combined metric requires both correct detection and coverage.

Option 4 achieves high detection rates across all pre-2030 scenarios and both interpretations of the 10% target (~78–98%), however the coverage is consistently poor (~37–42%), substantially below the nominal 95%. This mirrors the Q2 finding (Section 5.3) and reflects the same underlying statistical issues. The poor coverage is not scenario-dependent, persisting regardless of the pre-2030 history and regardless of the 10% interpretation. As a result, the combined performance metric remains low (~37–42%) and, despite high detection rates, Option 4 rarely both identifies the correct trend and produces confidence intervals that contain the true value. Detection rates increase from the negative to the positive pre-2030 scenario under Q3a, suggesting that when the pre-2030 trajectory is positive, the GAM more readily detects the continuation of this positive trend. Under the absolute interpretation of a 10% increase (Q3b), the effect is less pronounced, and detection is high across all scenarios, likely because the larger change produces a stronger signal.

Option 5 achieves more moderate detection rates (~54–85%) and strong coverage (~84–89%), exceeding the results for Option 5 under Q2 (Section 5.3) and demonstrating the benefit of statistical assessment over a longer test window (12 years, c.f. either 7 or 4 years). Performance for the combined metric for Option 5 exceeds Option 4 in all scenarios (~46–79%), and the improvement is particularly pronounced under the absolute interpretation (Q3b, ~74–79%). This is driven by the better coverage; Option 5 detects change less well than Option 4., but when it does detect change it is more likely to also produce well-calibrated confidence intervals that encompass the true underlying trend value.

The difference between Q3a and Q3b is notable across both Options. Under the relative interpretation (Q3a), a 10% increase represents a smaller change in absolute terms (e.g. requiring an ESAI of 50 to rise to 55), whereas the absolute interpretation (Q3b) requires a larger change (e.g. from 50 to 60). The steeper underlying trend implied by the absolute interpretation provides a stronger statistical signal, boosting detection rates for Option 5 from ~54–64% (Q3a) to ~82–85% (Q3b), with corresponding improvements in combined performance. This demonstrates that the legislative ambiguity about what “10% increase” means has direct consequences for the reliability of the statistical assessment. The effect of pre-2030 history on Option 5 is modest: under Q3a, detection increases slightly from the negative to the positive pre-2030 scenario, while under Q3b detection is more uniform across scenarios. Coverage is approximately stable regardless of the pre-2030 trajectory.

Figure 18 shows the results for Option 6, a classification-based method that tests for the presence of any trend (positive or negative) using Akaike weights. Correct classification here means rejecting the null hypothesis of no trend. Option 6 correctly classifies the simulated trend direction in approximately 44–55% of simulations under Q3a (relative 10%) and approximately 75–78% under Q3b (absolute 10 units). Performance is therefore better under the absolute interpretation, consistent with the pattern observed for Options 4 and 5. Under the relative 10% interpretation (Q3a), there is a modest gradient across scenarios (correct classification increases from ~44% for the negative pre-2030 scenario to ~55% for the positive pre-2030 scenario), but that gradient is not apparent under the alternative absolute 10% interpretation. Similarly to Options 4 & 5, this suggests that for weaker changes, Option 6 more readily detects a continuation of a positive trend rather than a reversal. Overall, Option 6 shows moderate detection power for Q3, better than its Q2 performance for non-zero slopes, but remains limited. Its value as a supplementary diagnostic, rather than a primary test, is consistent with its performance across all the questions examined in this study.

Option 8 is omitted from this assessment as it classifies 0% of simulations as outside the equivalence region across all scenarios and both interpretations. The log-scale slopes corresponding to 10% growth over 12 years (~0.008–0.014/yr) fall well within its default equivalence bounds of ±0.0346, corresponding to halving or doubling over 20 years. All simulations are therefore classified as showing negligible change, regardless of the pre-2030 scenario or the 10% interpretation. This confirms the finding from Q1 (section 5.2.4) that Option 8's default thresholds are not calibrated for changes of the magnitude relevant to the legislative targets.

Across all Options assessed for Q3, Option 5 provides the most reliable combined performance, particularly under the absolute interpretation of the 10% target. Option 4 achieves higher detection rates but its consistently poor coverage makes it unreliable for formal inference about the magnitude of change. Option 6 offers moderate detection power as a supplementary diagnostic. Option 8 is ineffective in its default configuration. More broadly, the longer 12-year assessment window for Q3 provides a fundamentally stronger statistical basis than the Q1 adjacent-year comparison: all trend-estimation methods show better absolute performance here than in Q1 or Q2, and the difference between Q3a and Q3b results demonstrates that resolving the legislative ambiguity about the meaning of "10% increase" is important for ensuring that the assessment methodology is appropriately calibrated.

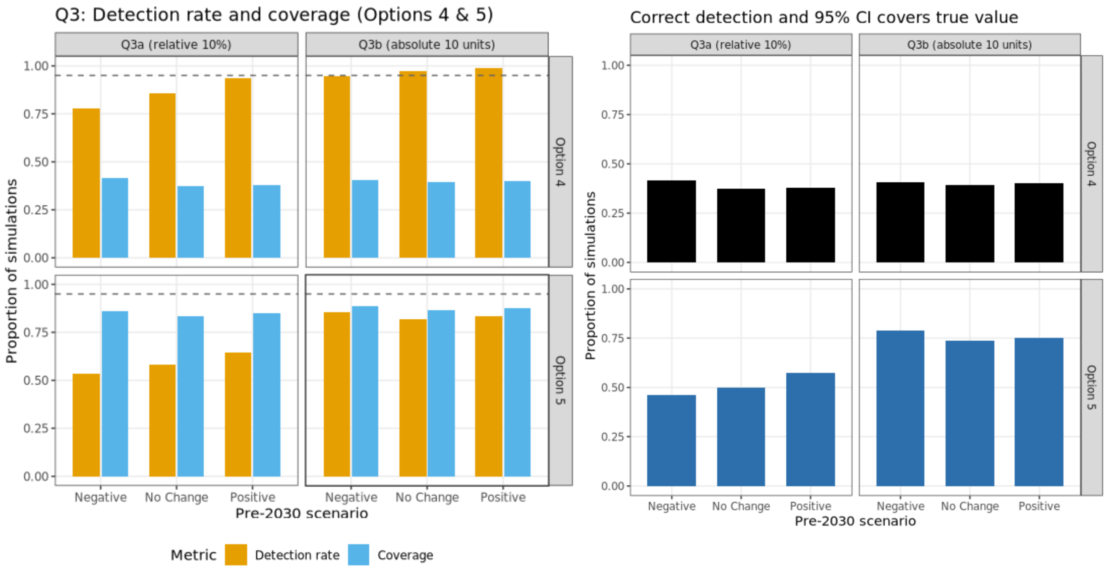


Figure 17: *Q3 Performance histograms for Options 4 (top row) and 5 (bottom row) for three contrasting ESAI scenarios that capture alternative ESAI pre-2030 histories (decrease, left bars; no change, central bars; increase, right bars), for two interpretations of 10% change (relative 10% change, left columns; absolute 10% change, right columns). Left Panel: Detection rate (statistical test is significant) and Coverage (confidence intervals on trend estimation include the true value) metrics. Right Panel: combined performance requiring correct detection (significance for the positive or negative underlying trend slopes, not significant for slope = 0) and coverage.*

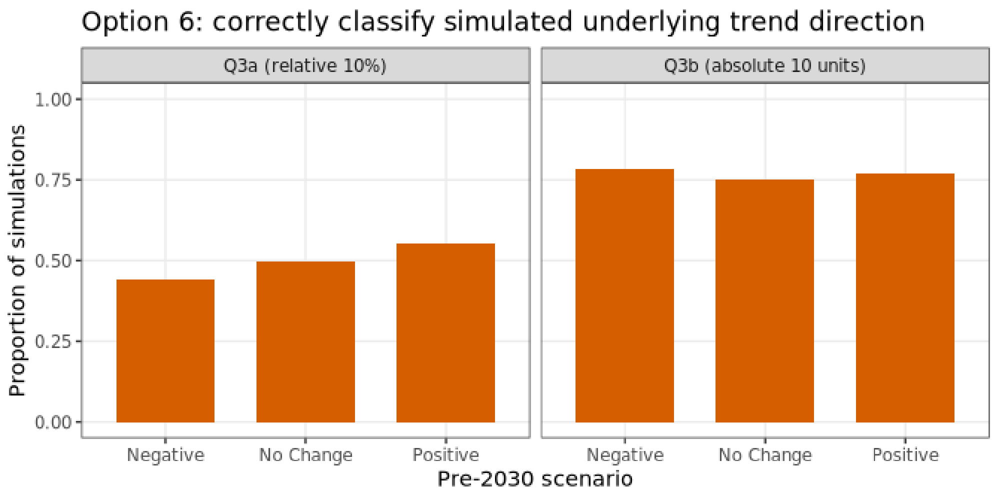


Figure 18: *Q3 performance histogram for Option 6 showing the proportion of simulations that correctly detect a trend across three pre-2030 scenarios (negative, stable, and positive prior trajectory) under relative (Q3a, left) and absolute (Q3b, right) interpretations of the 10% growth target.*

7. Discussion

This study evaluated a set of candidate statistical approaches for assessing whether the legislative biodiversity targets associated with the ESAI have been met. Using a simulation framework parameterised from the historical ESAI time series, the analysis generated synthetic indicator trajectories under controlled rates of change and applied the statistical methods proposed in Burns et al. 2025 to those trajectories. The simulations examined three policy-relevant questions: whether decline has halted between 2029 and 2030 (Question 1), whether the rate of change differs from zero over the characteristic multi-year windows identified in the historical analysis (Question 2), and whether the indicator shows a sustained 10% increase between 2030 and 2042 (Question 3). By evaluating how often each method correctly detects change under known underlying trends, the simulations allow the statistical properties of the candidate approaches, including sensitivity, false positive rates, and uncertainty calibration, to be compared directly.

7.1. Interpreting the simulations

The simulations reveal that the statistical challenge of the three legislative questions differs substantially. For Question 1 (adjacent-year comparison between 2029 and 2030) the inferential requirement is the most challenging. The ESAI exhibits structured variation over multiple years (Section 3), so the signal of a genuine change in trajectory is inherently weak over a single annual step. Under SO1 parameterisation, Options 2 and 3 required an underlying trend slope of approximately 2.9 (i.e. a 2.9 percentage point year on year change, or an ~4.3% relative change) to achieve 90% detection, a rate that substantially exceeds anything observed in the recent historical record. Option 1 has higher raw detection power but also a higher false positive rate because it does not account for uncertainty. Options 2, 3, and 4 converge on similar power curves for adjacent-year comparisons because all three estimate the same GAM derivative in 2030. Extending the simulated series to 2032 materially improved power across all Options (Section 5.2.2), confirming that additional post-target data reduces GAM end-effects and strengthens inference.

For Question 2, the assessment operates over longer characteristic timescales (~7 years for SO1, ~4 years for SO2), providing a stronger statistical basis. Option 4, which derives inference from a GAM fitted to the full historical series, achieved only 31–44% coverage (true slope lying within the 95% confidence interval of the model fit slope) across both Q2a and Q2b, substantially underestimating uncertainty due to historical bias. By contrast, Option 5, a linear model fitted only over the assessment window, achieved approximately 90-95% coverage, a substantial improvement though still below the nominal 95% on average, likely reflecting temporal autocorrelation structure that the method is not accounting for correctly. Option 5's detection rates were moderate and approximately symmetric between positive and negative slopes, indicating an approximately unbiased test. These results demonstrate that windowed linear models produce better-calibrated uncertainty estimates than full-series GAM approaches for multi-year trend assessment.

The results for Question 3 reinforce and extend these findings. Over the 12-year 2030–2042 assessment window, Option 5 again outperformed Option 4, achieving combined detection rates and coverage of approximately 74–79% under the absolute interpretation of the 10% target, compared with approximately 39-41% for Option 4. Option 4's poor coverage (~37–42%) persisted across all pre-2030 scenarios and both interpretations of "10% increase," confirming that the full-series GAM's coverage failure is systematic. An important Q3 finding is that the legislative ambiguity about whether "10% increase" refers to a relative percentage or an absolute percentage-point change impacts the results. Under the absolute interpretation, Option 5's detection rates improved ~54–64% (Q3a) to ~82–85% (Q3b). The effect of pre-2030 history on Q3 performance was modest.

A general finding is that indicator construction and statistical inference cannot be treated independently. The ESAI is produced by smoothing a composite index using a full-series GAM, and the simulations show that this directly influences downstream assessment. The k-sensitivity analysis (Section 5.2.3) for Q1a simulations (ending in 2030), showed that false positive rates rose sharply as the GAM basis dimension increased and did not stabilise, whereas for Q1b simulations (ending in 2032) both position and detection power stabilised above k = 10 (SO1). This demonstrates that the interplay between smoothing and inference depends on whether sufficient post-target data are available to anchor the GAM fit. The indicator production pipeline and the assessment methodology should therefore be considered as parts of the same, unified, analytical framework.

7.2. Methodological Considerations

The simulations were parameterised using characteristics estimated from the historical ESAI time series and the statistical methods were evaluated using the ESAI as we understood them to be applied for policy assessment This is appropriate for comparing candidate methods, but the conclusions are conditional on how well the historical ESAI model captures the statistical properties that will matter for future assessment. The ESAI is not a direct ecological measurement but a constructed composite indicator, and the performance of the methods depend on the specifics of the indicator construction framework.

This analysis highlights that the indicator values, the downstream statistical inference, and ultimately conclusions about the legislative targets, are impacted by smoothing choices for the ESAI. In particular, we show that statistical inference for the Options considered here is highly sensitive to the smoothing specification. Methods can become conservative under one parameterisation and anti-conservative under another, because smoothing changes the statistical properties of the quantity being tested. Because the ESAI is defined using a full-series GAM smooth, inference about recent change is derived from the same fitted object that defines the reported indicator values, creating a close link between construction and testing. The Q1 simulations used SO1 parameterisation throughout, including for the smoothing sensitivity analysis, and the recommendation to prefer SO2 (see Section 8) rests primarily on inference from this analysis (Section 5.2.3). Direct simulation under SO2 parameterisation would strengthen this recommendation.

The simulations also inherit limitations from the historical record. The composition of the ESAI has not been constant through time, monitoring effort differs across groups, and earlier portions of the series rely more heavily on hindcasting. The ESAI aggregates information from monitoring schemes that differ in both design and statistical treatment, and not all forms of uncertainty are carried transparently through to the final smoothed series. Species are treated equally in the geometric mean despite unequal informational contributions and ecological non-independence. In particular, the absence of species weightings and the numerical dominance of key phylogenetic groups is expected to reduce the effective information content of the indicator to something lower than the nominal species count suggests. The simulations should therefore be understood as historically informed evaluations rather than exact forecasts.

The linear trend form used in the simulations provides a transparent benchmark for method comparison, but the results should not be over-interpreted as covering every conceivable future trajectory. Similarly, the candidate methods were evaluated individually; some may work better in combination, which was the original intention for the suite of methods outlined in Burns et al. 2025. The analysis answers a foundational question: how do these procedures behave, on their own, when applied to ESAI-like time series under known patterns of change?

Finally, the low detection power and sensitivity to smoothing parameterisation identified in Section 5 for the 2029–2030 assessment are not artefacts of the simulation design. They are inherent consequences of assessing a smoothed, autocorrelated composite indicator at its endpoint using a single adjacent-year comparison. These are inherent properties of the indicator that the simulations make visible, rather than create. The value of the simulations is that they allow these issues to be quantified under controlled conditions, separating weaknesses of specific methods from constraints imposed by the ESAI and the policy questions being asked of it.

7.3. Policy implications and communication

The simulation results have implications for how the ESAI should be interpreted and communicated in a policy setting. A central tension is that the legislation defines success using decision rules at specific dates, whereas the indicator is a smoothed, composite summary with multi-year temporal structure. For the 2030 target, this means that a formally simple comparison (whether the ESAI in 2030 equals or exceeds the 2029 value) may be harder to assess robustly than it appears. The simulations show that short-horizon tests can be weak, sensitive to smoothing, and difficult to calibrate.

A related implication concerns the balance between timeliness and reliability, and the importance of communicating uncertainty. Including additional years of data beyond the point of assessment improves both detection power and statistical confidence in conclusions around legislative targets being met. Delaying assessment, however, reduces the time available to respond if progress is inadequate. This trade-off should be made explicit. Furthermore, the type of error a method is most likely to make (false positives, risking premature policy relaxation, or false negatives, delaying recognition of progress), becomes a policy characteristic when results support legal or strategic decisions. A pragmatic approach would be to report the indicator at the target date with its full uncertainty, making clear that the initial assessment carries wide confidence intervals, and then refine the conclusion in subsequent years as additional data narrow those intervals. This would preserve the timeliness of reporting while allowing the statistical basis of the assessment to strengthen progressively.

The ESAI is an aggregate indicator, based on the geometric mean, that can conceal important underlying structure whose composition changes over time as monitoring programmes begin and end. Opposing changes across taxa or regions may offset one another, and equal mathematical weighting does not imply equal informational contribution. Species groups which contain many species, or with long running monitoring efforts, or which are charismatic and easy to count (such as moths, butterflies and birds), dominate the indicator and exert disproportionate influence on its characteristics. The similarity between sub-group and global autocorrelation structures indicates that the characteristic temporal dynamics of the ESAI are not only an artefact of taxonomic dominance, but also a consequence of similar short- and medium-timescale structure in the species and sub-group data the ESAI uses (including smoothing parameterisations). More fundamentally, the ESAI treats species as independent units, despite shared phylogenetic histories, ecological interactions, and common environmental drivers. The effective information content of the indicator is therefore likely to be lower than the raw number of contributing species would imply. While evaluation of alternative indicator constructions is beyond the scope of this project, there may be value in exploring approaches that account for ecological similarity or phylogenetic relatedness, or that weight species according to shared drivers, enhancing the indicators interpretation as a proxy for biodiversity.

The global indicator represents a high-level summary rather than a complete account of biodiversity change, and greater transparency would improve interpretability and support more nuanced assessment. For example, an interactive reporting format allowing users to move from the global ESAI to taxonomic sub-groups, species-level trajectories, or regional summaries, would allow stakeholders to interrogate the indicator and understand its strengths and weaknesses. There may also be value in exploring approaches that account for ecological similarity or shared drivers among species.

Challenges to interpreting the ESAI are likely to occur wherever biodiversity policy relies on smoothed composite abundance indicators, including in devolved administrations developing similar measures. Continued methodological coordination across the UK, including shared statistical principles and transparent documentation of differences, would improve comparability and ensure that differences in reported progress reflect genuine ecological change rather than methodological divergence.

8. Conclusions

Statutory assessment of the ESAI depends on choices about smoothing, timescale, uncertainty, and decision rules. This study used simulation analysis, informed by the historical behaviour of the ESAI, to evaluate candidate statistical approaches for assessing the legislative biodiversity targets. The simulations support a cautious and explicit approach to policy interpretation: one in which the statistical properties of the indicator are acknowledged openly, the trade-offs between timeliness and robustness are transparent, and the global trend is accompanied by access to the underlying structure it compresses.

The results presented here show that the suitability of the statistical methods considered depends on the temporal relevance of the question, the way the indicator is constructed, and the alignment between the two. Specifically:

1. The legislative test of whether decline has halted between 2029 and 2030 is poorly aligned with the statistical structure of the ESAI. Under SO1 parameterisation, Options 2 and 3 required an underlying trend slope of ~2.9, corresponding to a 2.9% percentage point annual increase (equivalent to a 4.3% relative annual increase) to achieve 90% detection confidence, a rate that substantially exceeds recent historical precedent.
2. The statistical behaviour of the candidate methods depends strongly on the smoothing used to construct the indicator, and the choice of smoothing should be regarded as a central component of the assessment framework. The sensitivity analysis presented in Section 5.2.3 shows that for a 2029-2030 comparison, test performance does not stabilise as the basis dimension of the penalised GAM smoother increases. However, adding additional years of values after the 2030 assessment point does result in the test performance stabilising as the basis dimension of the penalised GAM smoother increases.
3. The statistical properties of GAMs mean that inferential assessments based on estimates at the end of the fit are weakest, and confidence improves progressively as additional post-target data become available. For Q1, extending the simulated series from 2030 to 2032 materially improved detection power across all Options except Option 7, and the basis dimension sensitivity analysis (Section 5.2.3) showed that test performance stabilised only when post-2030 data were available to anchor the GAM fit (Section 5.2.3). This suggests that conclusions about whether decline has halted should be revisited in subsequent years following the initial assessment, and that the initial assessment at the target date should be reported with its full uncertainty rather than treated as definitive. Conclusions on whether legal targets have been achieved should ideally be drawn after the assessment date, once sufficient data are available for the statistical performance of the test to be stable.
4. Short-horizon tests that derive inference from the full-series GAM are vulnerable to smoothing-induced bias, particularly for low basis dimension splines, and inference on ESAI for legislative targets should be made on longer time horizons (10+ years) where possible.
5. The longer 2030–2042 assessment window provides a fundamentally stronger statistical basis than the 2029–2030 comparison. However, the legislative ambiguity about whether “10% increase” refers to a relative percentage or an absolute percentage-point change has direct consequences for assessment reliability, with all Options performing better under the absolute interpretation.
6. No single method performs well across all three legislative questions. Options 2 and 3 performed well for the adjacent-year Q1 comparison, but Option 5 was better suited to the multi-year assessments in Q2 and Q3.
7. Option 5, based on multi-year linear trend estimation over a defined assessment window, provided the most reliable combination of detection and coverage across Q2 (~92% coverage) and Q3 (~74–79% combined performance under the absolute 10% interpretation).

8. Threshold-based and equivalence-based approaches (Options 7 and 8) are poorly calibrated to the ESAI in their default configurations, but the Q1 results (Section 5.2.4) show that re-calibrated thresholds could provide useful supplementary diagnostics.

9. Recommendations

The recommendations below set out practical approaches for assessing each legislative test in a way that is consistent with the statistical properties of the indicator identified in this study.

1. Assessing whether decline has halted (2029 vs 2030)

The adjacent-year comparison required by this legislative test represents the most statistically challenging setting examined in the simulations. The ESAI exhibits multi-year dynamics and is produced using a smoothed composite index and averaged, meaning that a comparison between two adjacent annual values provides only a weak signal of a genuine change in trajectory. If the comparison between 2029 and 2030 must be made, the following approach is recommended:

- **Review the legislative test to consider assessing progress over a longer time window.**

  The current 2030 species abundance target (whether decline has halted between 2029 and 2030) is poorly aligned with both the statistical properties of the ESAI and the level of uncertainty in annual estimates and is therefore not fit for purpose for determining whether the decline in species abundance has been halted. The ESAI is smoothed over timescales longer than a year, reflecting multi-year dynamics, and therefore has limited sensitivity to change over short timescales. Uncertainty in the ESAI annual values means that large year-to-year changes are required for robust detection. Where possible, species abundance targets using the ESAI should assess change over longer periods where signal can be more reliably distinguished from noise.

- **Consider adjusting the legislative test to incorporate uncertainty explicitly.**

  The current statutory comparison between the 2029 and 2030 ESAI values makes no provision for the uncertainty inherent in a smoothed composite indicator. If this assessment must be made, we recommend a more robust formulation requiring an assessment framed in terms of a statistical test with a defined confidence level (we recommend 90%), rather than a simple point comparison.

- **Where possible, delay drawing conclusions regarding the legally binding targets until at least two years post-target data are available.**

  Ideally, the legislative framework would also allow the formal assessment to be made at a point after the target date when the underlying models are better constrained by additional data, rather than requiring a definitive conclusion at a time when the statistical evidence is at its weakest. Waiting for a short period after the target year substantially improves the reliability of the statistical inference b providing timely policy feedback. The current reporting date for assessing the species abundance indicator in 2030, is 2032. Delaying the assessment to accumulate two additional years of data after 2030 to improve the reliability of the statistical inference would push this reporting date back to 2034.

- **Use smoothing parameters conditional on data availability when applying the 2029–2030 legislative test.**

  If the assessment is based only on ESAI values up to 2030, use SO1. If ESAI values beyond 2030 (e.g. to 2032 or later) are available, use SO2.

  The simulations show that more constrained smoothing (SO1) can introduce bias by propagating the historical decline into the assessment period, whereas the more flexible SO2 configuration reduces this bias. However, SO2 is sensitive to end-effects when fitted near the boundary of the time series. When additional data beyond 2030 are available, the fit is better constrained and penalisation controls model complexity effectively, making SO2 the preferred option. In contrast, when only data up to 2030 are available, SO1 is preferred, as the more constrained model reduces end-effects at the boundary, albeit at the cost of some bias. This recommendation is based primarily on inference from the k-sensitivity analysis of SO1 simulations (Section 5.2.3); direct simulation under SO2 parameterisation would further strengthen the evidence base.

- **Use 'rate of change' based approaches as primary statistical test (Option 2 or Option 3).**

  If the 2029 to 2030 comparison must be made, we recommend using Options 2 or 3 to make the assessment. These approaches assess the rate of change between two time points, in this case subsequent years, 2029 and 2030, using either a Bayesian or frequentist statistical framework. They show broadly consistent behaviour in the simulations and provided a clear and interpretable test of whether the rate of change differs from zero. Option 4 is not recommended. While its performance for the question appears good, it achieved only 31–44% % coverage in Q2 and 37–42% in Q3, suggesting that it systematically underestimating uncertainty.

- **Reassess the target in subsequent years after 2030.**

  The current reporting date for the 2030 species abundance target is 2032, and includes data recorded up to 2030. The statistical signal of a genuine change in trajectory becomes clearer as additional post-2030 data accumulate. Assessments should therefore be updated and reported on each year as new data become available, refining and improving the robustness of the statistical inference around the assessment

Even under these recommendations, the power of the adjacent-year test remains limited. Conclusions drawn from the 2029–2030 comparison should therefore be interpreted cautiously and considered alongside subsequent assessments.

2. Assessing sustained improvement over the recovery period (2030–2042)

The legislative assessment of a 10% increase in the ESAI between 2030 and 2042 provides a stronger statistical basis for inference. The simulations indicate that methods based on explicit trend estimation perform best in this setting. The recommended approach is:

- **Clarify the legislative interpretation of "10% increase".**

  The simulations show that whether this is interpreted as a relative percentage or an absolute percentage-point change affects detection rates across all Options tested (Section 5.4). Option 5's detection rates improved from approximately 54-64% assuming a relative increase to approximately 82-85% under the absolute increase interpretation. This ambiguity should be resolved to ensure the assessment methodology is appropriately calibrated.

- **Consider assessing progress using data extending beyond 2042.**

  Including one or two additional years of data (e.g. through 2044) can further stabilise the statistical assessment, although this is less critical than for the 2029–2030 comparison.

- **Use smoothing option SO2 when constructing the indicator.**

  The increased flexibility of the SO2 GAM minimises the impact of bias from the historical decline propagating into the assessment period, while the extended assessment window minimises the impact of GAM end effects. Over the test window, the fit is better constrained and penalisation controls model complexity effectively, making SO2 the preferred option. This recommendation is based primarily on inference from the k-sensitivity analysis of SO1 simulations (Section 5.2.3); direct simulation under SO2 parameterisation would further strengthen the evidence base.

- **Use Option 5 as the primary statistical method for assessment.**

  This method tests whether the ESAI shows a significant linear trend over a specified period by fitting a linear model to the EASI time series with year as the explanatory variable and testing the slope coefficient against a null hypothesis of zero change. This Option provided the most reliable combination of detection performance and uncertainty calibration, achieving combined performance of 74–79% under the absolute interpretation of the 10% target in Q3.

- **Optionally use Option 6 as a supplementary test.**

  Like Option 5, this method tests whether the ESAI shows a significant linear trend and then performs a second test to determine whether the rate of change depends on the current value of the indicator. The results of both tests are combined into a single probability that no overall trend is present. Can provide supporting evidence on the presence of a trend but should not be used as the sole basis for inference because of its higher false positive classification rate.

3. General recommendations

The analysis also suggests several broader principles for assessing biodiversity targets using the ESAI:

- **Match the statistical method to the legislative question.**

  Different questions require different analytical approaches; no single statistical method performs well across all assessment scenarios. More than one statistical method may be appropriate for addressing a single legislative question.

- **Reassess the smoothing approach used in constructing the indicator.**

  The simulations indicate that the choice of smoothing can materially influence inference, with more constrained configurations propagating historical trends into the assessment period and introducing bias. This highlights that smoothing parameters should be selected with the specific statistical objective in mind, rather than fixed *a priori*. In particular, smoothing used for indicator construction should be appropriate for the timescale over which change is being assessed, balancing bias and variance to avoid both over-smoothing (which can mask genuine change) and under-smoothing (which can introduce spurious variation). Further evaluation of smoothing choices under different simulation scenarios would strengthen the evidence base for selecting an appropriate approach.

- **Recognise the role of indicator construction in statistical inference.**

  The smoothing used to construct the ESAI has a direct influence on the statistical properties of downstream tests and should be considered an integral part of the assessment framework.

- **Update assessments as new data become available.**

  Because the ESAI is a time series indicator with inherent variability, conclusions about progress toward targets should be revisited as additional years of data accumulate.

- **Consider the role of categorical approaches.**

  The default thresholds used in Options 7 and 8 are poorly calibrated for the ESAI: Option 8 classified 0% of Q3 simulations outside its equivalence region. However, re-calibrated categorical classifications could provide useful supplementary diagnostics alongside the primary statistical tests.

- **Invest in longer-term methodological development.**

  The analysis points to several directions for strengthening future assessments, including exploration of approaches that account for ecological similarity or shared drivers among species, interactive or disaggregated reporting of the ESAI, and further investigation of how the assumption of independence in the indicator construction pipeline affects downstream uncertainty quantification.

Taken together, these recommendations aim to ensure that the statistical assessment of biodiversity targets using the ESAI reflects both the structure of the indicator and the timescale of the policy questions being addressed.

10. References

Bence, J.R. (1995) Analysis of short time series: correcting for autocorrelation. Ecology 76: 628-639

Burns, F., Bradfer-Lawrence, T. & Gregory, R.D. (2025) Exploring options to test a 'halt the decline' nature target using a multi-species abundance indicator. RSPB Research Report to the Office for Environmental Protection RSPB Centre for Conservation Science, RSPB, The Lodge, Sandy, Bedfordshire, SG19 2DL.

Burns, F., Mordue, S., al Fulaij, N., Boersch-Supan, P.H., Boswell, J., Boyd, R.J., Bradfer-Lawrence, T., de Ornellas, P., de Palma, A., de Zylva, P., Dennis, E.B., Foster, S., Gilbert, G., Halliwell, L., Hawkins, K., Haysom, K.A., Holland, M.M., Hughes, J., Jackson, A.C., Mancini, F., Mathews, F., McQuatters-Gollop, A., Noble, D.G., O'Brien, D., Pescott, O.L., Purvis, A., Simkin, J., Smith, A., Stanbury, A.J., Villemot, J., Walker, K.J., Walton, P., Webb, T.J., Williams, J., Wilson, R. & Gregory, R.D. (2023) State of Nature report 2023.

Burns, F., Eaton, M.A., Hayhow, D.B., Outhwaite, C.L., Al Fulaij, N., August, T.A., Boughey, K.L., Brereton, T., Brown, A., Bullock, D.J., Gent, T., Haysom, K.A., Isaac, N.J.B., Johns, D.J., Macadam, C.R., Mathews, F., Noble, D.G., Powney, D.G., Sims, D.W., Smart, S.M., Stroh, P., Walker, K.J., Webb, J.R., Webb, T.J. & Gregory, R.D. (2018) An assessment of the state of nature in the United Kingdom: A review of findings, methods and impact. *Ecological Indicators* 94: 226-236. https://doi.org/10.1016/j.ecolind.2018.06.033

DEFRA. (2023) [Environmental Improvement Plan 2023.](https:// www.gov.uk/government/publications/environmental-improvement-plan)

DEFRA. (2025a) [Indicators of species abundance in England.](https://www.gov.uk/government/statistics/indicators-of-species-abundance-in-england/indicators-of-species-abundance-in-england)

DEFRA. (2025b) [Indicators of species abundance in England: Response to feedback.](https://www.gov.uk/government/statistics/indicators-of-species-abundance-in-england/indicators-of-species-abundance-in-england-response-to-feedback#trend-breakdowns)

DEFRA (2025c) Wild bird populations in the UK and England, 1970 to 2023: Accredited official statistics. Available at: https://www.gov.uk/government/statistics/wild-bird-populations-in-the-uk-and-england-1970-to-2024

Dixon, P.M. and Pechmann, J.H. (2005) A statistical test to show negligible trend. Ecology, 86(7): p. 1751-1756.

Eaton, M.A. et al. (2015) The priority species indicator: measuring the trends in threatened species in the UK. Biodiversity, 16: p. 108-119.

Freeman, S.N., Isaac, N.J.B., Besbeas, P. et al. (2020) A generic method for estimating and smoothing multispecies biodiversity indicators using intermittent data. JABES 26: 71–89. https://doi.org/10.1007/s13253-020-00410-6

Gelman, A., Jakulin, A., Pittau, M.G. and Su, Y.S. (2008). A weakly informative default prior distribution for logistic and other regression models. https://doi.org/10.1214/08-AOAS191

Gosling, J.P. (2018). SHELF: The Sheffield Elicitation Framework. In: Dias, L., Morton, A., Quigley, J. (eds) Elicitation. International Series in Operations Research & Management Science 261. Springer, Cham. https://doi.org/10.1007/978-3-319-65052-4_4

Gregory, R.D., Noble, D.G., and J. Custance (2004) The state of play of farmland birds: population trends and conservation status of lowland farmland birds in the United Kingdom. Ibis, 146: p. 1-13.

Johnson, T.F., Beckerman, A.P., Childs, D.Z. et al. (2024) Revealing uncertainty in the status of biodiversity change. Nature 628: 788–794. https://doi.org/10.1038/s41586-024-07236-z

McRae, L., Cornford, R., Marconi, V., Puleston, H., Ledger, S.E.H., Deinet, S., Oppenheimer, P., Hoffmann, M. & Freeman, R. 2025. The utility of the Living Planet Index as a policy tool and for measuring nature recovery. *Philos Trans R Soc Lond B Biol Sci* 380: 20230207. https://doi.org/10.1098/rstb.2023.0207

Miller, D.L. (2025) Bayesian views of generalized additive modelling. Methods in Ecology and Evolution, 16: p. 446-455.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T. (2006) Uncertain Judgements: Eliciting Experts' Probabilities. John Wiley & Sons, Ltd

Oliver, T.H., Heard, M.S., Isaac, N.J.B., Roy, D.B., Procter, D., Eigenbrod, F., Freckleton, R., Hector, A., Orme, C.D.L., Petchey, O.L., Proença, V., Raffaelli, D., Suttle, K.B., Mace, G.M., Martín-López, B., Woodcock, B.A. and Bullock, J.M. (2015) Biodiversity and Resilience of Ecosystem Functions. Trends in Ecology & Evolution, 30: p. 673-684.

R Core Team (2025). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Robinson, R.A. and Sutherland, W.J. (2002), Post-war changes in arable farming and biodiversity in Great Britain. Journal of Applied Ecology, 39: 157-176. https://doi.org/10.1046/j.1365-2664.2002.00695.x

Schmidt, B.R. and A.H. Meyer (2008) On the analysis of monitoring data: Testing for no trend in
Shumway, R. H. & Stoffer, D. S., (2017), Time Series Analysis and Its Applications: With R Examples, 4th edition., Journal for Nature Conservation, 16, DOI: 10.1007/978-3-031-70584-7

Shumway, R.H. and Stoffer, D.S. (2017) Time Series Analysis and Applications using the R Statistical Package. Free dog publishing.

Toepfer, G. (2019) On the Impossibility and Dispensability of Defining ‘‘Biodiversity’’ in E. Casetta et al. (eds.), From Assessing to Conserving Biodiversity, History, Philosophy and Theory of the Life Sciences 24, https://doi.org/10.1007/978-3-030-10991-2_16

UK Government, The Environmental Targets (Biodiversity) (England) Regulations 2023, Statutory Instruments 2023 No. 91. 2023. Available at: https://www.legislation.gov.uk/uksi/2023/91/made

Wood, S.J. (2017) Generalised Additive Models: An Introduction with R, Second Edition. Chapman and Hall/CRC, New York.

11. Acknowledgements

The work presented in this report was informed by helpful discussions throughout from Fiona Burns (RSPB), Andrew Gill (OEP) and Robbie McDonald (OEP). Additionally, we’d also like to thank Clare Betts, Claire Harris and Madeline May (DEFRA) for support in accessing and utilising the data underpinning the ESAI and together with Rob Boyd (UKCEH), thank them for their contribution to the technical workshop on the report’s findings. Also Maniv Pathak (DEFRA) for contributions to a high-level discussion on the implications of the report’s findings.