Developing statistical methodologies for analysing change and trends in environmental indicators

Ref: 2025-03-13-436-1

Report Authors:

John Addy, Gail Robertson & Nick Schurch

Date: 2025-04-10

BioSS at Edinburgh

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

Developing statistical methodologies for analysing change and trends in environmental indicators

March 2025

Biomathematics and Statistics Scotland (BioSS) were commissioned by the Office for Environmental Protection (OEP) to undertake a review of methodologies used to analyse trends in environmental indicators and define thresholds to summarise and communicate those trends. The views provided in this report are those of BioSS and do not represent the opinion or position of the OEP.

Table of Contents

1. Executive summary

- In this report we evaluate whether the statistical methods currently used by the OEP to inform their assessment of change over time in environmental indicators are fit-for-purpose and propose alternatives where they would improve the statistical rigour and robustness.
- Currently, the OEP assess the change over time for 46 (of 51) varied environmental indicators, primarily using a 3% 'rule-of-thumb' method that calculates the percentage change between the most recent years data and the data from a baseline year (usually five years previously). If the percentage change is greater than 3% the change is said to be 'significant' and a red or green rating is applied via a Red, Amber, Green (RAG) system depending on the direction of change. This method while straightforward to interpret, is overly reductionist, obscures critical nuances, and is not best practice. It's limitations including the arbitrary choice of the baseline year and the percentage used to determine whether significant change has occurred (Section 2).
- The current 3% methodology used by the OEP is based on a doubling or halving of values over a 25-year period and is widely used by UK agencies and organizations but has been altered sufficiently to have lost this original contextual meaning. This approach is not broadly used in an international context, however, there is a lack of consensus on best-practice methodology for assessing change over time in environmental indicators (Sections 2 & 3).
- Categorizing the change over time based on an arbitrary 3% threshold for the percentage change (whether calculated vs a single baseline year or vs an average baseline period), while straightforward to interpret, is overly reductionist, obscures critical nuances, and is not best practice. We recommend discontinuing this.
- We propose two new robust statistical methods to analyse short- and long-term changes in environmental indicator datasets (Section 5). These are:
    - I. a t-test based method, for time series where only short-term (>6 years) data are available.
    - II. a penalised spline smoother method, where long-term (>20 years) data are available.
- The new proposed methods represent a significant improvement over the current OEP practice by accounting for variation and trends in the data, supporting a more nuanced and evidence-based assessment of the change over time of environmental indicators.
- Applying these methods to a subset of ten environmental indicators, we find that the proposed new methods change the RAG rating classification for seven indicators, including for indicators with large percentage changes between their baseline year and the most recent year. The t-test method identifies these changes as not statistically significant because the inter-annual fluctuations over the baseline period are (relatively) large and are the driving a component of the measured percentage change and hence their current RAG rating classification[1]. Using the t-test method, these indicators are now classified as amber "little or no change". The 'Area of woodland' it the only indicator where the new proposed methods support an improved RAG assessment, moving this indicator from 'little or no change' to 'improving'. Here, despite a marginal yearly percentage change (2.9%), both tests identify the changes as statistically significant. This result is attributable to (relatively) small inter-annual fluctuations for this indicator that increase the confidence that even relatively small changes reflect real improvement (Section 6).
- We conclude that the new methods proposed are more robust alternatives to the 3% method, able to account for inter-annual variation and able to handle both short and long-term time series. They are broadly applicable across environmental indicators with suitable datasets, allowing the OEP and other organisations to detect changes in indicator values over time with more confidence.
- While presented as distinct approaches, these methods are, in fact, complementary and we recommend presenting both the effect size (*e.g.*, percentage change, not classified by the 3% method)

[1] Extent of UK area protected for nature on land and water, $PM_{2.5}$ & Sulphur Dioxide emissions, consumption-base greenhouse gas emissions, number of wildfire incidents and population-weighted annual mean concentrations of $PM_{2.5}$ in the air.

and statistical significance (*e.g.*, t-test and/or penalized spline smoother method), of the observed change over time together, allowing readers to assess whether annual changes are meaningful (sizable, statistically distinguishable, and indicative of consistent improvement or decline).

- We also discuss changes to the RAG rating system used if these methods are adopted, to bring it in line with best-practice scientific data visualization practices.

2. Introduction

Under the Environment Act 2021, the OEP must annually report on the UK governments progress towards achieving the targets of the Environmental Improvement Plan (EIP). The EIP assessment framework uses fifty-one environmental indicators drawn from a broad range of publications, largely based on national and official statistics, which are grouped into ten goals (GOV.UK, 2023).

The OEP's assessment of progress towards the EIP goals examines past trends and yearly progress for each of these fifty-one indicators and assesses the prospects of meeting the EIP targets and commitments based on these trends (GOV.UK, 2025; OEP, 2025). The current method used by the OEP for assessing the change over time for most indicators is based on a 3% "rule-of-thumb" approach, that calculates the percentage difference between the most recent year of interest and a specified baseline year and compares this against a +/-3% threshold to identify "significant" change (Table 1). This approach is applied and used widely across UK government departments including by the Department of the Environment, Food and Rural Affairs (DEFRA, see Section 3.1.1) which describes its assessment of change as (DEFRA, 2025a):

"For most indicators, a threshold of at least 3% change (positive or negative) is used. This is consistent with the approach adopted for assessment of some of the UK Biodiversity Indicators and some other government assessments, such as Forestry Commission Key Performance Indicators. Where existing official DEFRA assessments were available for the same time periods using a more tailored methodology, these were replicated in the Outcome Indicator Framework (OIF) assessment instead of applying the 3% threshold. The 'little or no change' category is intended to show indicators where any recorded change may be a result of random error in the dataset or due to chance, rather than a meaningful trend."

Our literature search suggests that the use of this threshold for categorizing change over time has its origins in the UK "Birds of Conservation Concern" (BOOC) reports. In the 2002 assessment of UK bird populations, bird species are placed onto either Red, Amber or Green lists for their population status depending on either the International Union for Conservation of Nature criteria given to the species, or whether the species were identified as being in either "Recent Decline", or "Historical Decline" (Galbraith, 2002). Galbraith defined "Recent Decline" as decline in either breeding or non-breeding population, or range contraction, by more than 50% over the last 25 years, corresponding to an ~2.8% year-on-year decline. This approach continued to be used for BOOC reports three through five and was adopted by DEFRA for its 2013 Annual Statistical Release "Wild Bird Populations in the UK, 1970 To 2013" (DEFRA, 2014), where it was applied to six bird population indices. This threshold was rounded up to 3% by the Joint Nature Conservation Committee for its 2013 report on 24 UK biodiversity Indicators for the Fifth National Report to the United Nations Convention on Biological Diversity (JNCC, 2013a). For further information on other current approaches organizations take to measuring change over time for environmental indicators, see Section 3.

Table 1: Threshold categories for indicator changes (OEP 2025a).

Comparison to baseline year	Description
> +3%	significant increase
-3% to +3%	no significant change
< -3%	significant decrease

Table 2: The Red/Amber/Green (RAG) rating method used to interpret changes in indicators variables (OEP 2025a).

RAG rating	Description	
		Goal-level rating

	Indicator-level rating	Past trend	Annual progress	Prospects
	Improvement	Improving trend	Good progress	On track
	Little or no change	Mixed picture	Mixed Progress	Partially on track
	Deterioration	Deteriorating trend	Limited Progress	Off track

The OEP evaluates short term trends are over a five-year period (for example, comparing 2024 data with a 2019 baseline), although the same method is also used over longer or shorter periods when necessary and when additional context is required in the assessment of the indicator (*i.e.,* when a new indicator has been introduced, or where there has been a significant methodological change in the way the indicator data are gathered or reported)[2]. To communicate the changes over time identified to a broad audience, the OEP use a Red-Amber-Green (RAG) rating and directional arrows based on the 3% threshold (Table 2)[3]. The 3% threshold, and it's associated RAG categorization, is also used by other public bodies to communicate change over time[4]. For example, JNCC uses it for "net removals of greenhouse gases by UK forests" (JNCC, 2025a), "public expenditure on biodiversity" (JNCC, 2025c), and "volunteer time spent on conservation activities" (JNCC, 2025b).

The 3% method has the advantage of being easy to interpret and communicate to a wide audience and requiring only basic statistical knowledge to understand, however it has four significant limitations:

I. The 3% value (or its associated threshold of a 25-year 50% decrease or 100% increase) is arbitrary (JNCC, 2013b).
II. Applying the 3% threshold to a baseline more than a year prior to the year of interest results in it losing its motivating association with a 25-year 50% decrease or 100% increase.
III. Whether an indicator falls above or below the 3% value depends as much on the choice of baseline year as on the current data of interest, which makes identifying the baseline year something that requires careful thought. Choosing a fixed period for the baseline (for example five years ago) makes the percentage change sensitive to year-on-year (inter-annual) variation in the baseline value (see Figure 1 and Figure 2 for year-on-year variations in $PM_{2.5}$ emissions data), as much as it is to the current year's value (an extreme example of the 'shifting baseline' problem). Conversely, fixing a baseline year to avoid a shifting baseline problem requires a choice out of (potentially) many options, and is likely to result in increasingly large changes being measured as the interval between the baseline and the year of interest grows (for example, for the air pollution by fine particulate matter, $PM_{2.5}$, indicator, comparing 2024 data with a baseline year of 1970 yields a decline of 88%, while comparing 2024 data with a 1990 baseline results in a decline of 72%).
IV. The 3% change threshold is only appropriate for indicators on population-type data (such as abundances, emissions concentrations, areas, etc), and even for those cases it's unlikely to be detailed enough to consistently estimate whether the scale of change is meaningful across a wide variety of indicators. For example, some indicators (like "percentage of woodland that is sustainably managed") are less likely to change significantly between years compared to others (such as those that may exhibit cyclical patterns, such as species abundances).

For some indicators (particularly those with long time series; >20 years), other methods are used to determine if there has been a change in the indicator. For example, the species abundance indicator uses a more complex methodology which examines the indicators five-year trend and assesses significant change based on calculating 95% credible intervals from a smoothing-based model, assigning a significant change if the current years sits outside the credible interval band for the baseline year of interest. These credible intervals are calculated from a state-space models within a Bayesian and Kalman filter framework, developed by (Freeman *et al.*, 2021) to assess multispecies abundance. The Freeman *et. al.* framework models smooth growth rates in individual species abundances via penalized splines, allowing the model to impute missing data, reducing the impact of

[2] We note that some other organizations calculate their indicator metrics using a three-year average of the indicator value centred around a baseline year, rather than using a single baseline year (for example, comparing 2024 data with the average value from 2018 to 2020; (*e.g.*, England Biodiversity Indicators dataset & (DEFRA, 2025a), however the OEP do not apply this methodology.
[3] For further information on the OEP methodology, see (OEP, 2025).
[4] We note that occasionally 5% is chosen by some organisations, particularly for indicators which are subject to large interannual variation (such as the "wintering waterbirds" indicator). The Office of National Statistics also use similar methods to assess whether a change constitutes an improvement/deterioration/no change, for example, although the specific thresholds used maybe be different.

short-term noise and integrating observation error, population stochasticity and species-specific variability into calculations of confidence intervals for the resulting smooth trends. The second derivatives of these trends are then used to identify years with significant directional changes.

This is a best-practice robust and rigorous statistical modelling approach to this noisy data; however, limitations remain:

I. This statistical model is complex and difficult for a non-specialist to implement and draw conclusions from.
II. Where the model confidence interval range of the baseline year is wide even sizable effect sizes will not be reported as "significant" changes.
III. The change over time assessment still relies on the identification of, and difference to, a baseline year, effectively ignoring the longer-term trend information from the indicator.

These problems are amplified when assessing whether the 2030 species abundance target set under the Environment Act 2021 has been achieved, because this uses a one-year window (*e.g.,* comparing 2024 to 2023) to determine success, an approach that makes it difficult to assess the effectiveness of long-term policies designed to impact this indicator, and which requires large inter-annual changes in order for this modelling approach to identify changes as "significant".

Identifying methods that provide a more robust definition of statistical significance and which better account for both uncertainty in the annual indicator data and inter-annual variation, would improve confidence in the assessment of the environmental indicator changes over time. However, any proposed new methodology needs to remain easily communicable to non-specialists from various backgrounds, retaining the current RAG rating system (or with only slight modification) to ensure consistent communication between assessments, and needs to be relatively easy to implement for OEP staff. With these aims for improving its assessment of the government's progress toward achieving the EIP goals, the OEP commissioned BioSS to:

I. Review current methods for assessing trends and applying RAG system thresholds.
II. Propose alternative methodologies that would improve confidence in the assessment by improving statistical rigour and robustness.
III. Implement the alternative methodologies proposed for a subset of indicators, to demonstrate the potential improvements.

Here we report on the findings of this commission. In Section 3 we present a review of the current methods used by a range of organization for assessing change over time in environmental indicators. In Section 4 we group the 51 environmental indicators from the OEP's 2022/2023 progress report into a typology based on data type and method of data collection and (informed by discussions with OEP staff) identify a priority group of 14 indicators from this typology on which to focus this analysis (Appendix, Table A1), In Section 5 we propose alternative statistical approaches for measuring changes over time for these indicators, highlighting the improvements the methods offer over the current approach. In Section 6 we apply the proposed methods to the subset of indicators identified in Section 4, highlighting where the methods change the assessment of whether there has been significant change over time. In Section 7 we discuss the wider applicability of the methods and potential extensions to the project, and, in Section 9, we draw final conclusions.

3. Other national & international indicator assessment methods

Putting the OEPs current assessment practice into a national an internation context, we review the methods for evaluating changes over time in environmental indicators used by DEFRA, the Joint Nature Conservation Committee (JNCC), the Forestry Commission (FC), the European Environmental Agency (EEA), the Climate Change Committee (CCC), and the US Geological Survey, US Environmental Protection Agency, and the US National Interagency Fire Center.

3.1. Department of the Environment, Food and Rural Affairs

DEFRA's Outcome Indicator Framework (DEFRA, 2025b) summarises and hosts data for 66 indicators arranged into 10 themes and communicated via 16 'headlines', including Relative Abundance and/or distribution of species, Area of Woodland in England, Distribution of invasive non-native species, plant pests and diseases, and Waste Crime. The OIF assesses recent (<5 years) and long term (>5 years) trends and uses data for 2018 (or as close to this as possible) as its baseline. DEFRA uses a wide variety of statistical approaches for assessing indicator change over time including mixed effects models, multi-regional input-output modelling, hierarchical Bayesian modelling. Where a robust statistical assessment is not available, DEFRA use a similar 3% rule of thumb as the OEP but based on a three-year average around a year as the baseline value for the comparison, rather than a single year (DEFRA, 2025a). DEFRA uses the same classification as OEP, labelling indicator change over time as “Improving”, “Little or no change” or “Deteriorating”, However they do not present the classification using a RAG colour coding. DEFRA are continuously improving the approaches for assessing change over time for their indicators, aiming for best-practice and statistical robustness. For example, DEFRA (with UKCEH) are currently inviting broad community feedback on the use of smoothing options for the Relative Abundance and/or Distribution of Species (DEFRA code D4) [5], which currently identifies change over time by smoothing the time series data (1970-2022) and testing to see if the smoothed mean of reference year is within the 95% credible interval of the most recent year.

3.2. Joint Nature Conservation Committee

The JNCC produces a yearly assessment of 24 UK Biodiversity Indicators based on a wide variety of data provided by the UK government, research bodies, and the voluntary sector (*e.g.,* UK Biodiversity Indicators 2023 report, (JNCC, 2023a). Similar to DEFRA, JNCC assess the change over time on two timescales, long-term and short-term. The long-term assessment is only made for indicators with more than ten years of data and compares the current years data with data from the earliest year available. The short-term assessment compares the current years data with the data five years previous. Where possible, JNCC indicators are assessed using measures of statistical significance (*e.g.,* for Habitat Connectivity, Statis of UK Priority Species – Relative Abundance, Birds of the Wider Countryside and Sea, Insects of the Wider Countryside (Butterflies), Plants of the Wider Countryside, Mammals of the Wider Countryside (Bats) and Fish Size Classes in the North Sea), however where that is not possible JNCC use the same 3% rule of thumb as the OEP for assessing change over time. JNCC use the same categorisation system as the OEP RAG system, classifying significant changes as either “Improving” or “Deteriorating”, and other changes as “Little or no change”, provided data is available (see (JNCC, 2023b).

3.3. Forestry Commission

The Forestry Commission publishes an annual report which reviews the performance of 38 key indicators, comparing the difference between each indicator value in the most recent single year for which data are available with the data from five years earlier. The assessment method is similar to that adopted by DEFRA, JNCC and the OEP, including the RAG rating system.

3.4. Climate Change Committee

The Climate Change Committee produced a report in 2023, in collaboration with the Agricultural Development Advisory Service (a private independent consultancy), which reviews eight environmental indicators[6], updating the analysis of five indicators reported on in 2021/2022. For each indicator a high-level description was provided and a reference to the type of indicator it is categorised as under the Adaption Committee's assessment (Ffoulkes *et al.*, 2023). The report does not make a formal assessment of the recent change over time of the indicators, instead (where appropriate data are available for an indicator), the report presents indicator trend data in figures, maps and tables, typically showing the raw values of indicators[7].

---

[5] See https://defraenvironment.blog.gov.uk/2024/10/04/a-call-for-feedback-on-the-indicators-of-species-abundance-in-england/. The invitation for feedback on this indicator is currently open (2025-03-13).

[6] Rate of development of properties in areas at risk of flooding, Area of impermeable surfacing in urban areas, Area of urban greenspace, Wildfire incidents and area burnt, change in total hedgerow length, current crop production by area in UK & exposure to vulnerable groups to flooding.

[7] The exception is the “risk of flooding” indicator, where data for 2022 are presented along with percentage difference to 2020 as a baseline year with negative changes coded as red, positive changes coded as green, and no change left in black.

3.5. US Geological Survey

The US Geological Survey and US Federal Emergency Management Agency have online dashboards and resources to monitor protected areas and flood maps. The US Geological Society do not report on indicators directly; however, the data sources and statistics they generate inform indicators compiled and monitored by the US Environmental protection Agency.

3.6. US Environmental Protection Agency

The US Environmental Protection Agency monitors environmental indicators across five theme areas (Air, Water, Land, Human Exposure and Health, and Ecological Condition). The indicators are measured at specific locations across the US; however, the number and location of the measurements differs for each indicator and not all sites measure every indicator. The US EPA takes different analysis approaches to identifying changes in the aggregate statistics over time for each indicator, ranging from no statistical assessment, t-tests, and fitting trends to data including using linear regression and more complex statistical models. Basic percentage changes comparisons are often made for indicators, but the baseline year or period chosen for these comparisons is inconsistent across indicators. (*e.g.*, $PM_{2.5}$ measures are compared with values from 1980, 1990, 2000, and 2010 (US EPA, 2016), while acid deposition (acid rain, sulphur and nitrogen) compares the most recent data with values from 1989 (US EPA, 2015)). Where possible, the EPA makes a large effort to understand and communicate context and uncertainty in indicator variables, including explaining confidence intervals and how they refer to the possibility of errors in sampling, measurement, and/or reporting processes. They also explain how this uncertainty is associated with random variation of the indicators measured value. Explanations are given to describe visualising confidence intervals, such as errors bars in mean estimates and confidence bands for trendlines.

The EPA does not present a consistently assessed cross-indicator comparison or RAG classification, instead it lists 23 questions the indicators combine to address, and it offers both qualitative and quantitative assessments of the picture painted by the combined indicators that underlie each question. For example, on the question “What are the trends in greenhouse gas emissions and concentrations and their impacts on human health and the environment?” the EPA indicate that “For several greenhouse gases, the nation's estimated combined emissions that are directly attributable to human activity have decreased 7 percent between 1990 and 2020.” (US EPA, 2017).

3.7. US National Interagency Fire Center

Like the number of wildfire incidents indicator analysed by the OEP, the US National Interagency Fire Center provides yearly reports on the national wildfire activity in the US. The yearly reports consider both national- and regional-level analysis (for the latter, the US is separated into eleven regions) with summarised information for a range of data such as total number of wildfires, wildfires acres burned nationally, large fires, source of fire (*e.g.*, lightning, human), etc. National and regional summaries of number of wildfire incidents and acres burned nationally are presented as raw data with bar plots. Regional level analysis shows the percentage difference in wildfire incidents compared to national five-year, and ten-year, averages. The US National Interagency Fire Center does not assign statistical significance to changes over time.

3.8. European Environment Agency

The European Environment Agency produces a year report titled “The European Environment - state and outlook” (*e.g.*, EEA, 2019), which presents the state of 35 indicators assessed over three groups, “Protecting, conserving and enhancing natural capital”, “Resource-efficient, circular and low-carbon economy”, and “Safeguarding from environmental risks to health and well-being”. The analyses are available for each individual country within the bloc. Individual statistical and modelling approaches are used for quantifying trends over the past 10-15 years for each indicator. Building on these, a combination of modelled estimates of future developments (based on expected changes in environmental drivers of change) and expert opinion are used to assess the future development of each indicator. Several indicators are assessed using a percentage change comparison of current data against a baseline (for example air pollutant and Green House Gas emission are assessed vs 2000 and 1990 data), however none use a 3% threshold for classifying significance. Instead, indicators are classified as “improving trends/developments dominate”, “trends/developments show a mixed picture", and "deteriorating trends/developments dominate" (presented using a similar RAG rating representing), based on a contextual combination of baseline comparisons, trend analysis expert judgement and integration across themes.

Table 3: Modelling / count-based indicators selected from the 46 OEP indicator variables with available data.

Indicator	Outcome Indicator Framework Goal	Method
Area of woodland	Thriving plants and wildlife	Roll-forward from 2011
UK emissions of 5 key air pollutants	Clean air	Emissions factors calculated from activity levels and emissions
Percentage of woodland that is sustainably managed	Using resources from nature sustainably	Count vs roll forward woodland accounting
Consumption-based greenhouse gas emissions	Mitigating and adapting to climate change	Modelling
Emissions of fluorinated gases	Mitigating and adapting to climate change	Modelling
Properties at high risk of flooding	Reduced risk of harm from environmental hazards	Count surveys and Modelling
Number of wildfire incidents	Reduced risk of harm from environmental hazards	Count surveys
Number of additional tree pests and diseases becoming established	Enhancing biosecurity	Survey and running average vs roll forward woodland accounting
Percentage of the total population in England living in close proximity of greenspace, as of October 2021	Enhancing beauty, heritage and engagement with the natural environment	Multiple indicators
Population-weighted annual mean concentrations of PM2.5 in the air	Clean air	Calculated from the average
Exceedance of damaging levels of nutrient nitrogen deposition in England	Clean air	Calculated using moving average

4. Indicator selection

The OEP's 2022/2023 progress report includes 51 environmental indicators, 46 of which have data available (see OEP, 2025). We grouped these indicators into a typology based on the underlying data type and the method of data collection (*i.e.,* survey, modelling/counts, combined emissions measures, abundance modelling, threshold/monitoring, economics, spatial mapping, other). A full breakdown for the 46 indicators is shown in Appendix (Table A1). With OEP staff we identified the modelling/count-based group of indicators (13 in total) and the Abundance of Priority Species as priorities for the analysis. The modelling/count-based indicators were prioritized despite the wide range of data types within this group because the raw data was accessible and unlikely to be processed, making it more straightforward to propose methodological improvements that could be implemented within the scope of the commissioned project.

5. Proposing new methodologies

We propose two statistical methodologies appropriate for analysing the change over time for the indicators selected in Section 4, that represent a significant improvement over the current OEP practice. For indicators where long-term (>20 years) data is available, we recommend a penalised spline smoother modelling approach. Where only short-term (6-20 years) data is available, we recommend a t-test based method. A guidance document accompanies this report showing how these methods can be applied to the Section 4 indicators, including a justification of method selection for each indicator.

In developing these recommendations, we aimed to select statistically robust methods that could:

I. Account for variability between annual indicator values in the data
II. Be used to analyse short-term and long-term time series.

III. Be straightforward to communicate to a non-statistical audience.
IV. Be straightforward to implement and use by OEP staff.
V. Be used with the OEPs existing RAG rating approach (or a lightly updated version thereof).

5.1. Short-term changes: the t-test method

The t-test method we propose accounts for inter-annual variations when assessing changes in environmental indicator data by using both the mean and standard deviation of the indicator over baseline period, comparing the current years data to these values using a one-sample t-test. The one-sample t-test compares a single value with a population sample[8], and assumes that the data used to calculate the mean and the standard deviation are independent (no autocorrelation), that the mean and variance are identically distributed (stationary, no trends) and that the deviation of the datapoints from the mean are normally distributed[9].

These assumptions define the hypothesis and null hypothesis being tested:

- ***Null Hypothesis:*** *the current year's value is equal to the mean of the previous 5 years.*
- ***Hypothesis:*** *the current year's value is significantly different from the mean of the baseline period.*

The method then calculates the p-value corresponding to the t-test statistic – that is, the probability of observing this test statistic, or a more extreme value, by chance, based on the t-distribution - and assigning statistical significance based on a predefined p-value threshold. This method is applicable to a broad range of environmental indicators with continuous data (as opposed to count, ordinal or categorical data), requiring as little as six datapoints to use.

The most significant weaknesses of this approach are the sample size constraints for the baseline period. A small sample size limits the statistical power of the t-test and may limit the ability of the method to identify meaningful changes. Conversely, using larger sample sizes increases the likelihood of the assumptions of the t-test being violated since the indicators time series are likely to be both autocorrelated, and non-stationary (*i.e,* to have trends over time) on longer time scales. We recommend that this method is used with the previous five years of data as baseline[10]. For indicators with less data that this we recommend that no assessment is made, including not using the 3% method. We anticipate that the characteristic timescales for long-term trends in environmental indicator data will be decades and that the impact of these trends on five-year timescales is likely to be small. For indicators have a significant trend over the chosen baseline period, the standard deviation measured will be inflated, reducing the statistical power of the method to detect meaningful changes. A final important limitation of the proposed method is that it can only be used with datasets where there is variation within the baseline period, because the t-test relies on being able to define a non-zero standard deviation[11], however the 3% method will produce a categorization in these instances. The most robust way to test for low variation is to calculate the coefficient of variation (CV, the standard deviation divided by the mean), and flagging those datasets with CV < 0.05, as having exceptional low variability. However, for small datasets such as these, the coefficient of variation is highly variable and only weakly informative and there is no clear quantitative threshold. We recommend calculating the coefficient of variation, counting the number of values that are different from the dataset mode (to identify datasets with mostly constant values), and visually inspecting dataset plots to assess whether there is sufficient variability for each indicator individually, as part of the exploratory data analysis undertaken prior to assessment. We note that for precisely constant data, attempting to use the t-test method will return an undefined value

[8] As opposed to a two-sample t-test, which compares two population samples.

[9] We note that, strictly speaking, t-tests (and ANOVA) do not make any direct assumption about the data distribution, they assume that the residuals around the mean value are normally distributed for each sample being used in the test. However, since the mean of each sample is a single value for all the data in that sample, this assumption is equivalent to assuming that the sample data are normally distributed within each sample.

[10] Although, in principle, any five-year period can be selected, the choice of which five-year period to use is driven by what question is being asked and using the previous five-years sets the current years value within this recent context. We also note that statisticians differ in their opinions regarding the minimum number of data points necessary to carry out a t-test, *e.g.*, (Winter, 2013) (Keselman *et al.*, 2004).

[11] This is a problem for two of the indicators select in Section 4 ("Extent of UK area protected for nature on land and water" & "Number of additional tree pests and diseases becoming established") which both have constant values for four out of the last five years, resulting in insufficient variation to apply the t-test method.

Alternative statistical methods to the t-test for short-term comparisons include z-scores (using rolling averages and predefined thresholds), Shewhart charts (control limits based on standard deviations), and CUSUM (cumulative deviations from a target mean with sequential thresholds). Like the t-test, these methods account for variance, but they differ in interpretation and application. While these approaches share the t-test's goal of identifying meaningful deviations from baseline variability, the t-test is preferred here because it provides a probabilistic framework familiar to diverse audiences (avoiding the arbitrariness of fixed z-score thresholds or CUSUM control limits) and it is generally applicable and applies flexibly snapshot comparisons without assuming temporal dependency (unlike CUSUM which is designed for sequential monitoring or Shewhart charts which is targeted at stable processes). An alternative non-parametric method is the Wilcoxon signed-rank test, which compares the current year's value to the median of a baseline period rather than to the mean. However, non-parametric tests typically have lower statistical power than parametric tests because they make fewer assumptions about the data's distribution. Given the small sample sizes in our analyses, this reduction in power is a significant drawback, which is why we prefer the t-test approach. Another alternative approach is to fit a regression model (either linear or non-linear) to the historical data, excluding the current data point, and then use the model to predict the current value along with its confidence interval. Significance is determined by comparing the actual current value to this prediction. This method is intuitive and does not require the five-year baseline to be stationary; however, it tests a subtly different hypothesis than the t-test. Specifically, it evaluates whether the current year's value significantly deviates from the expected value based on the trend over the previous five years. When there is no trend, the result is similar to a t-test, but if a trend is present, the interpretation becomes more complex.

To summarise, we prefer the t-test method to the existing 3% method because it incorporates variation between annual indicator values in a robust way and allows the statistical significance of observed changes to be quantified directly. We prefer the t-test method to the alternative methods discussed here because it balances statistical power, robustness to assumptions being violated, simplicity of implementation, is broadly applicable to indicators with continuous data, and because it's widely familiar, and straightforward to interpret and communicate to a broad audience.

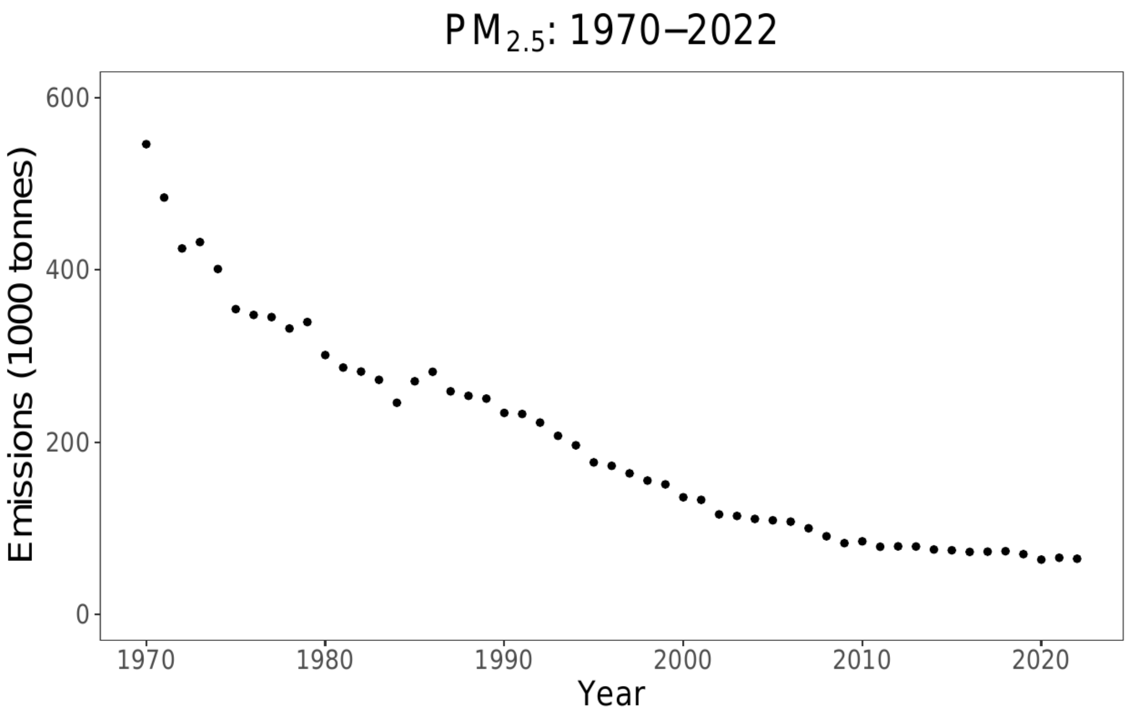

Figure 1: Full time series of annualized PM2.5 UK emissions from 1970-2022.

5.1.1. Example of applying the t-test method

Here we apply the t-test method to the fine-particulate matter ($PM_{2.5}$) emissions indicator as an example. Data from the UK is available from 1970 to 2022 (Figure 1). Using the current OEP 3% method, we compare 2022 with 2017 (Figure 2) and find a 8.2 kt decrease between 2002 (64.9 kt) and 2017 value (73.1 kt), corresponding to an 11.3% decrease (Table A9, OEP, 2025). This was identified as a significant decrease (>3% change). Highlighting the vulnerability of this approach to variability fluctuations in both the baseline and the current year, the percentage decrease measured against each of the individual years since 2005 (the reference year for emissions reduction commitments for this indicator) spans a range of 6.9% (2018 c.f. 2013) to 27% (2011-2006)[12].

Applying the t-test method to this indicator, we calculate a one sample t-test comparing the 2022 value (64.9 kt) with a mean calculated from 2017-2021 (69.4 kt). This test yields a T-statistic of 2.3, corresponding to a p-value (the probability of observing this test statistic, or a more extreme test statistic, by chance) of 0.08 (8%). Using a typical p-value threshold for significance of 0.05 (5%)[13] the proposed t-test method would classify this as "no significant change" between the 2017-2021 mean and the 2022 value. This is consistent with the 2022 value not falling outside the 95% confidence intervals for the 2017-2021 mean (Figure 2). Using the t-test result, and accounting for the inter-annual variation in the indicator, the RAG status of the $PM_{2.5}$ indicator would be amber ("Little or no change" and "mixed progress"), rather than green ("Improving").

5.2. Long-term changes: the penalized spline smoother method

The penalized spline smoother method we propose here accounts for both inter-annual variation and long-term trends (autocorrelation) when assessing changes in environmental indicator data. This method fits a smooth curve to the long-term data, calculates the 95% credible interval for the rate of change of this smooth curve throughout the time series, then uses the credible interval to calculate a one-tailed p-value to quantify the chance of observing a rate of change of zero, or one with an opposite sign, in the current year. This test identifies whether the observed rate of change (decreasing or increasing) can be robustly distinguished from the opposite trend or from zero. Finally, the method assigns the significance of the changes observed based on comparing this probability with a predefined threshold.

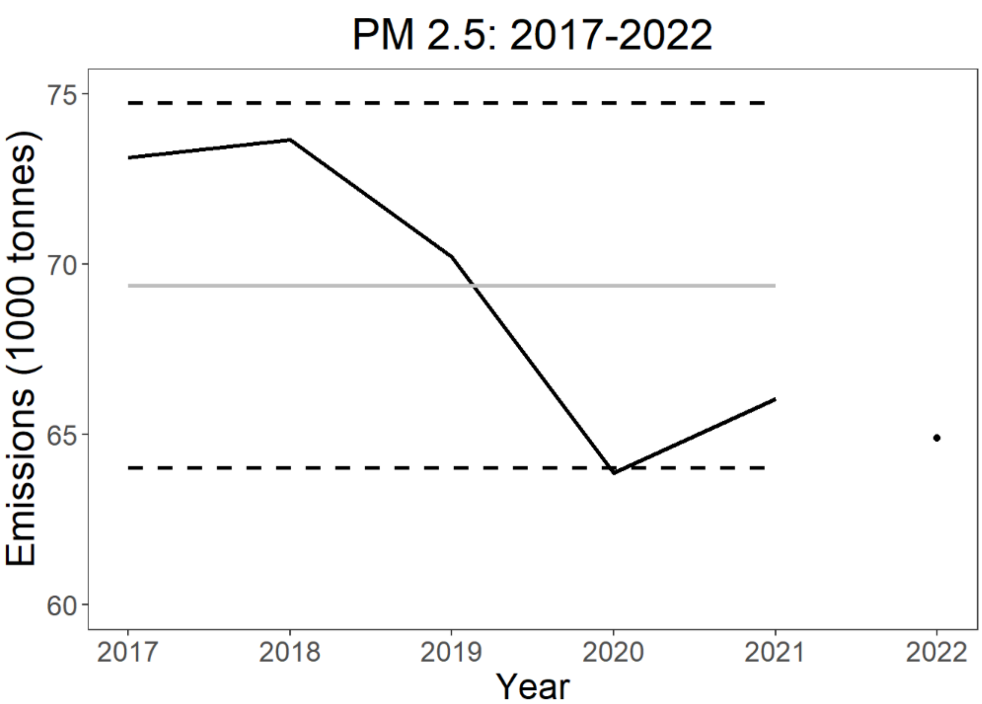

Figure 2: $PM_{2.5}$ data from 2017-2022 with 5-year mean (2017-2021, grey line) and the 95% confidence intervals on the mean (dashed lines),

[12] It is … their methodolo… …s adjustment is **not** simp… For example, the compa… …ease (11.3%). This sugge… …able.

[13] The choice of 5% as a p-value significance threshold is arbitrary but is widely used across many scientific fields as a threshold for reporting (McShane and Gal, 2017).

Penalized spline smoothers are a flexible non-parametric approach to modelling trends in time series data, that doesn't assume a specific shape (like a straight line or parabola). Splines are flexible curves constructed from polynomial segments connected smoothly at control points called knots. For non-penalized splines, the number of knots used for a spline fit is (to some extent) an arbitrary choice. Using fewer knots prioritizes capturing long-term trends in the time series, sacrificing the ability to fit short-term fluctuations, however too few knots can oversimplify the model, leading to underfitting that compromised the model's ability to capture meaningful trends and reduces predictive accuracy and generalizability to new data. Increasing the number of knots gives the spline the flexibility to model increasingly higher-frequence variation, improving the fit to the data. However, too many knots can result in the curve adapting to random noise fluctuations (overfitting) rather than modelling the underlying trend, again compromising predictive accuracy and generalizability to new data. Penalized spline models address the need to specify the number of knots in the fit by adding a penalization parameter (a "wiggliness" penalty) to the model, which discourages overly complex curves and balances both model fit and smoothness. The result is a smooth, data-driven trend that generalizes well to new observations and has good predictive power. For this method, we take a Bayesian approach to calculating confidence intervals for the penalized spline fits, using Metropolis-Hastings Markov Chain Monte-Carlo sampling to generate 10,000 draws from the posterior distribution for the spline coefficients[14] (using the best fit penalization parameter, which controls the number of knots available to the model). These draws can be thought of as 10,000 alternative model fits with the same overall smoothness (penalization parameter) but slightly different initial conditions that represent the uncertainty in the shape of the best-fit spline.

To test whether there is evidence for a an increasing or decreasing trend from the penalized spline fit, we calculate the first derivative of the model fit to get the rate of change curve for the indicator, which defines whether the indicator is increasing or decreasing for year, Repeating this process for the ensemble of posterior draws then defines the credible interval for the rate of change curve, characterizing the uncertainty in the rate of change each point as a histogram of values. Using the empirical cumulative distribution of the histogram, we calculate a one-tailed p-value to quantify the chance of observing a rate of change of zero, or one with an opposite sign, in the current year. We use a threshold value for this probability of <0.05 to assign whether the observed rate of change is significantly different from zero (*i.e.,* not consistent with little or no change, instead improving or deteriorating) or has a different sign that the mean - *i.e.*, the change of the rate of change being positive (increasing), where the current years value is negative (decreasing).

This method is a robust and efficient way of quantifying the evidence for a significant improvement or deterioration in the current years data, for environmental indicators with long-term data. It includes a robust estimation of uncertainty that incorporates inter-annual fluctuations in the data and, unlike the 3% method or the t-test method, works well with both autocorrelated data, data with long-term trends, and for indicators with a wide range of data types, including count-based indicator data[15]. The use of penalized splines makes the model flexible for modelling a wide variety of curve shapes, allowing the data to define the model complexity, while limiting pre-specification of specific tuning parameters to a single value. Most importantly, this approach avoids the need to choose a baseline year (as in the 3% method) or a baseline period (as in the t-test method) to assess whether there is clear evidence that the indicator is improving or deteriorating.

The performance of the penalized spline method (or, indeed, any other trend modelling method) is sensitive to both the strength of the short-term fluctuations in the time series, and their time scale relative to the length of the data series. This means that the model's ability to correctly capture long-term behaviour in the time series decreases as the length of the time series decreases and short-term (*i.e.,* inter-annual) fluctuations become dominant. This directly impacts the robustness of the modelling for shorter time series with the model becoming overparameterized (*i.e.,* a comparable number of model parameters than data). We recommend that the penalized spline smoother approach is only used with indicators that have >20 years of data (noting that this does significantly limit the number of OEP indicators this method can be used effectively for – only 4 of the 14 indicators in Table 3 have sufficient data, for example). The penalized smoother method is also a more complex method than either the current 3% method or the t-test method, requiring more specialist software to

[14] Assuming a normal distribution as the Bayesian prior for each spline coefficient. The number of posterior draws can be altered as necessary to meet competing timeliness/computational power constraints.

[15] Although we note that this does require adjustment of the R code to ensure proper distribution choice (*e.g.*, Poisson for counts, Negative Binomial for over-dispersed counts, etc).

implement (*R* (R Foundation for Statistical Computing, no date) , *mcgv* (Wood, 2011), etc[16]) and more expertise to run and interpret.

An inherent characteristic of penalized spline models (and many smoothing methods) is that credible intervals widen at the extremes of the dataset, where data are sparse and extrapolation uncertainty increases[17]. While this can be partially mitigated by penalizing the second derivative of the spline (increasing the smoothing of the curve and discouraging rapid trend changes), this approach imposes a global assumption that trends evolve gradually over time. Such an assumption may be reasonable for some environmental indicators but risks over-smoothing more rapid or nonlinear changes in others. An alternative approach to mitigate widening uncertainty at data extremes is to constrain the spline to linear behaviour beyond the observed range. However, this modifies the model specification asymmetrically, assuming the trend's gradient at the edge is both stable and known (a strong assumption lacking empirical justification). While linear extrapolation reduces interval width artificially, it risks significant bias if the underlying process deviates from this imposed structure (*e.g.*, accelerating declines or threshold effects). Another alternative way of working around this problem would be to reduce the threshold for significance when using credible intervals from the penalized spline method (*e.g.,* using 75% instead of 95%). While we acknowledge that the 95% threshold is inherently an arbitrary choice, this value is widely used across many scientific fields, and we propose using this with the broadly applicable t-test method in section 5.1. Altering the significance threshold would require a second arbitrary threshold choice, reducing the consistency of interpretation between indicators, and would require clear communication in the indicator results. To maintain consistency and generalizability across indicators with diverse dynamics, and avoid unjustified modelling assumptions, we prefer a consistent spline specification without additional penalization, constraints or altered significance thresholds, accepting wider intervals at extremes as a conservative reflection of uncertainty.

For completeness, we note that another alternative would be to ignore the final point of data when assessing the indicator status, assessing the previous year's value rather than the current year, where the confidence interval is better constrained. This is not a viable option for the OEP reporting, which needs to be as current as possible.

Alternative methods for assessing changes in environmental indicators with long-term datasets include parametric modelling methods (*e.g.,* linear and non-linear regression models), other non-parametric smoothing methods (*e.g.,* locally estimated/weighted scatterplot smoothing; LOESS & LOWESS), and time series decomposition approaches.

A straightforward alternative to penalized splines for modelling environmental indicator trends is parametric regression, such as linear regression. For example, the 2021 *State of the Thames* report modelled river temperature trends using a linear model (Institute of Fisheries Management, 2021). However, parametric approaches suffer from a critical limitation: they require *a priori* specification of the model's functional form (*e.g.*, linear, quadratic). This poses two challenges for environmental indicators. First, most indicator time series lack a simple parametric structure, often exhibiting non-linear locally varying trends, and fitting a parameterized regression will result in a poor model fit for many indicators. Second, the diversity of the environmental indicators ensures that a single parametric model is not applicable across indicators, forcing the modelling to be tailored on an indicator-by-indicator basis. We prefer the more generalizable approach of non-parametric models, which circumvents these issues by adapting flexibly to complex trends without rigid assumptions, offering a unified framework suitable for heterogeneous datasets.

Locally estimated scatterplot smoothing (LOESS) and locally weighted scatterplot smoothing (LOWESS) methods are widely used (including by government organisations (*e.g.*, DEFRA, 2004 & JNCC, 2023c)) non-parametric modelling methods that builds a smooth curve by performing weighted polynomial regressions within sliding data windows. Their primary tuning parameter, the *span* (the fraction of data included in each local window), determines smoothness: larger spans produce smoother curves by averaging over broader regions, while smaller spans prioritize local detail. LOESS/LOWESS smoothers excel at capturing spatially varying trends without assuming a global structure, making them highly adaptable to local nonlinearity. However, their reliance on a fixed span imposes uniform smoothness across the entire dataset, which can over-smooth abrupt changes or under-smooth stable regions. In contrast, penalized splines use globally distributed knots and a roughness

[16] We note that these software tools are all open source and freely available.

[17] We note that for short time series, this characteristic is strongly impacted by the scale of, and presence or absence of, short-timescale fluctuations in the individual dataset (see Section 5.2.2).

penalty to balance fit and smoothness. Rather than relying on local windows, splines adapt the knot density along the data to the data's local complexity, while the global roughness parameter penalizes rapid curvature changes, allowing flexible non-uniform, yet controlled, smoothness across the entire data range. While LOESS/LOWESS are simpler to implement due to a single intuitive parameter (span), penalized splines offer finer control over smoothness trade-offs and better accommodate heterogeneous trends. For environmental indicators, where gradual trends and uncertainty quantification are critical, we prefer to prioritize penalized splines' global smoothness control and flexibility over LOESS/LOWESS's localized adaptability.

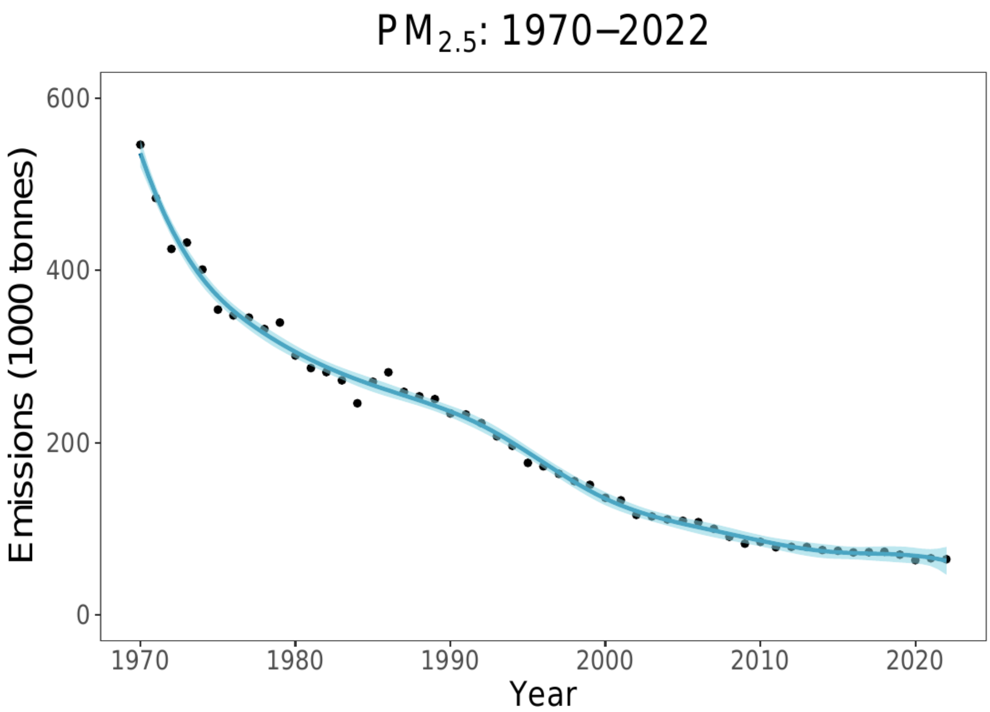

Figure 3: Penalised spline fit to $PM_{2.5}$ indicator data from 1970-2022 (dark blue), 95% credible interval (light blue)

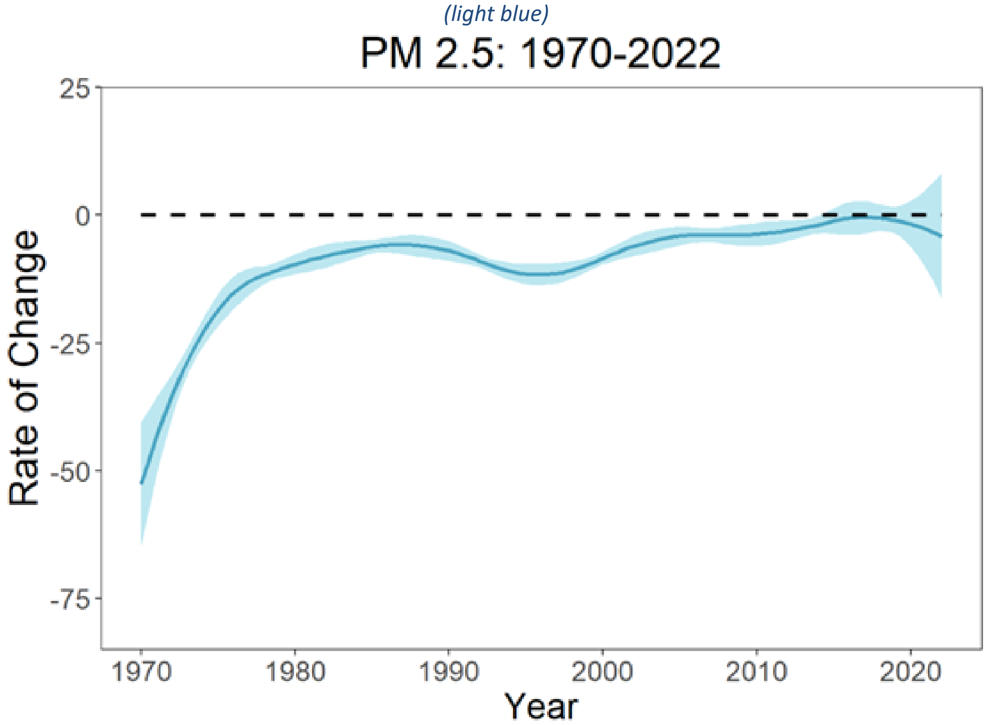

Figure 4: Rate of change (first derivative) of the penalized spline model fit of rate of change $PM_{2.5}$ data from 1970-2022 (dark blue) with 95% credible intervals (light blue).

Time series decomposition methods offer an alternative framework for modelling long-term indicator data by separating a series into three (or more) components: a trend, periodic (seasonal) fluctuations, and residual noise (*e.g.*, Cowpertwait and Metcalfe, 2009). Decomposition improves interpretability by explicitly modelling periodic contributions to the data, a potential advantage for data with strong periodic signals, however this added complexity is unnecessary for annualized environmental indicator data. By design, these indicators aggregate data into yearly summaries, averaging over daily/seasonal cycles and minimizing residual periodicity. Consequently, the trend component becomes the primary focus for assessing long-term change. Notably, decomposition methods often rely on non-parametric smoothers (*e.g.*, LOESS in the *stl* R package; Cleveland, Cleveland and Terpenning, 1990) to estimate trends, mirroring the penalized spline approach. The penalized splines method proposed here provided comparable trend estimation without the overhead of disentangling extraneous components, streamlining analysis while retaining flexibility to adapt to irregular, non-linear trends.

5.2.1. Example of applying the Penalized Spline Smoother Method

Using the fine-particulate matter ($PM_{2.5}$) emissions indicator as an example, we applied the penalised spline smoother with the full 1970-2022 dataset to assess whether this indicator is improving significantly in 2022. Figure 3 shows the best-fit penalized spline smoother to the dataset and the 95% credible interval for this fit based on 10,000 MCM iterations. From these, we compute the rate-of-change for the best-fit curve, and it's associated 95% credible intervals (Figure 4). The negative estimates for the rates of change suggesting continually declining emissions of $PM_{2.5}$[18], however the plot highlights that the 95% credible interval of the rate of change of this indicator has included zero (*i.e.*, no significant improvement or deterioration) since 2013, suggesting that there has been little meaningful improvement in this indicator in recent years. Figure 5 shows the histogram of rate of changes values from the 10,000-model ensemble for the $PM_{2.5}$ rate of change for 2022, highlighting that a value of zero falls comfortably within the body of the distribution of values for this year (the current year, for this example dataset). We use the probability distribution to define a credible interval, and to robustly quantify the probability of observing a rate of change of zero, in this example, for 2022, yielding a probability of 0.25 (25%). Using a probability threshold of 0.05 (5%) for assigning statistical significance, we

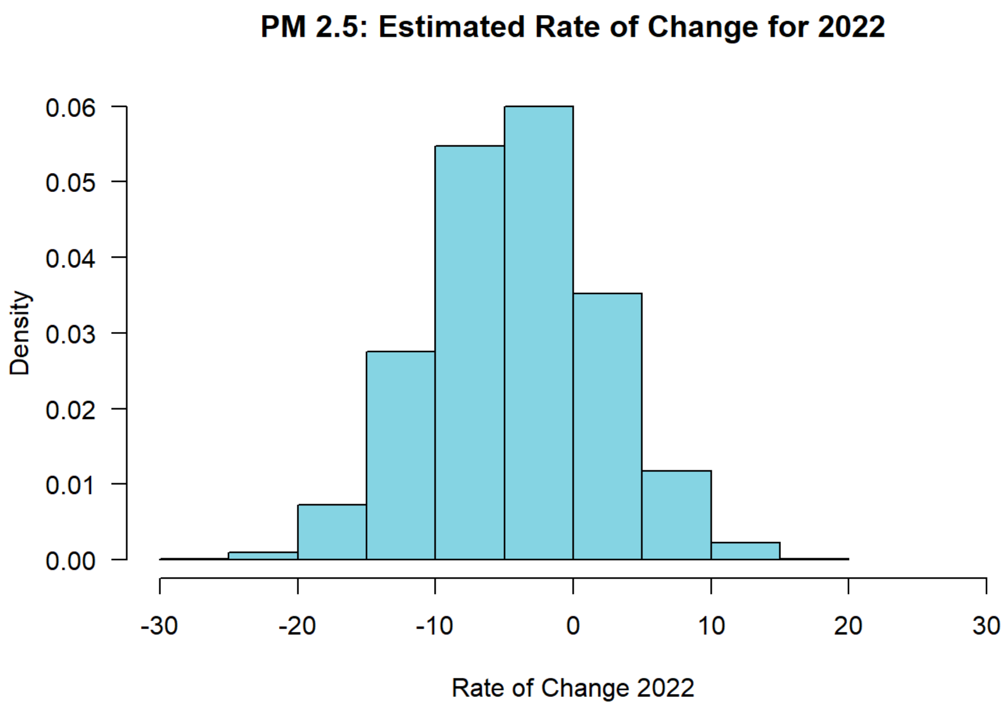

Figure 5: Histogram of the 2002 $PM_{2.5}$ rate of changes values from the 10,000-model ensemble from the penalized spline modelling method.

[18] We note that negative rates of change do not mean emissions were zero, or below zero, only that the emissions are reducing.

conclude that there is no evidence for any significant change in this indicator instead classifying this indicator as "Little or no change" and giving it an amber RAG rating. We note that the expanding confidence intervals at the extremes of the data represent our uncertainty in the rate of change, but non-zero rates of change are still detectable (for example, applying this method to the "Area of woodland" indicator yields a probability of significant change of <0.01, resulting in a green RAG rating). Table 3 shows the results of the current and proposed methods applied to the $PM_{2.5}$ emissions data as well as their RAG rating.

Table 3: Summary of the proposed methods to identify change in indicators applied to the PM2.5 emissions dataset along with their resulting RAG ratings.

Method	Baseline period	2022 value (kt)	Baseline Value	2022 Change Statistic	Probability	RAG rating
3%	2017	64.9	73.1	-11.26%[19]	NA	
t-test	2017 - 2021	64.9	69.4 ± 1.9	2.32[20]	0.08	
penalized spline smoother	NA	64.9	NA	-4.19[21]	0.25	

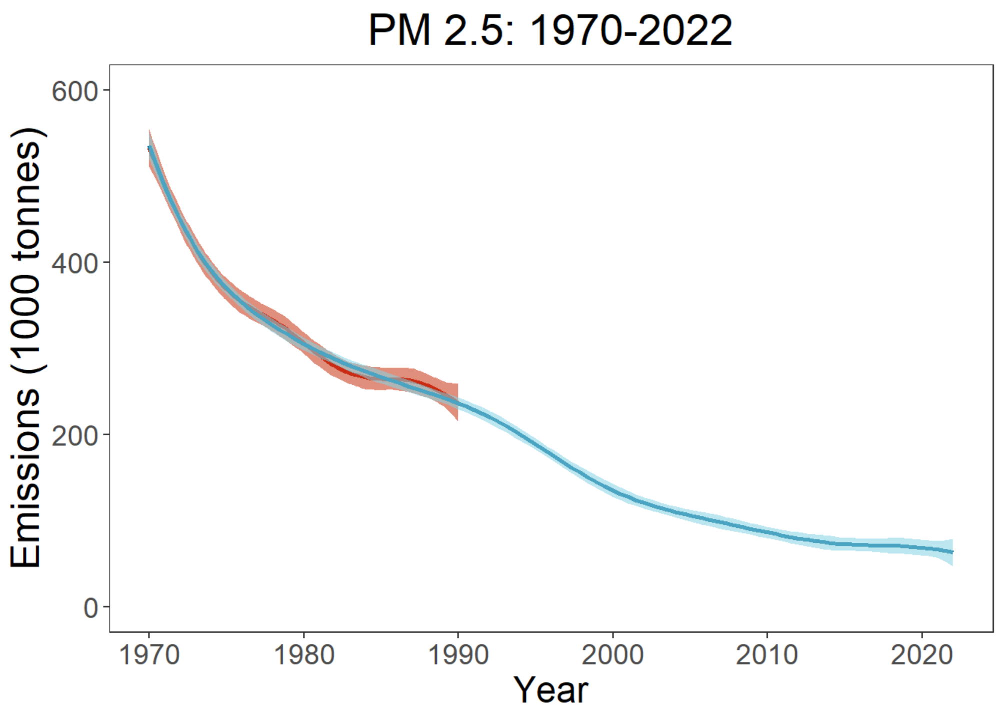

Figure 6: Penalized spline smoother fits to $PM_{2.5}$ data from 1970-2022 (dark blue) and 1970-1990 (red), with 95% credible intervals (light blue & brown, respectively).

[19] Percentage difference
[20] t-statistic
[21] Rate of change

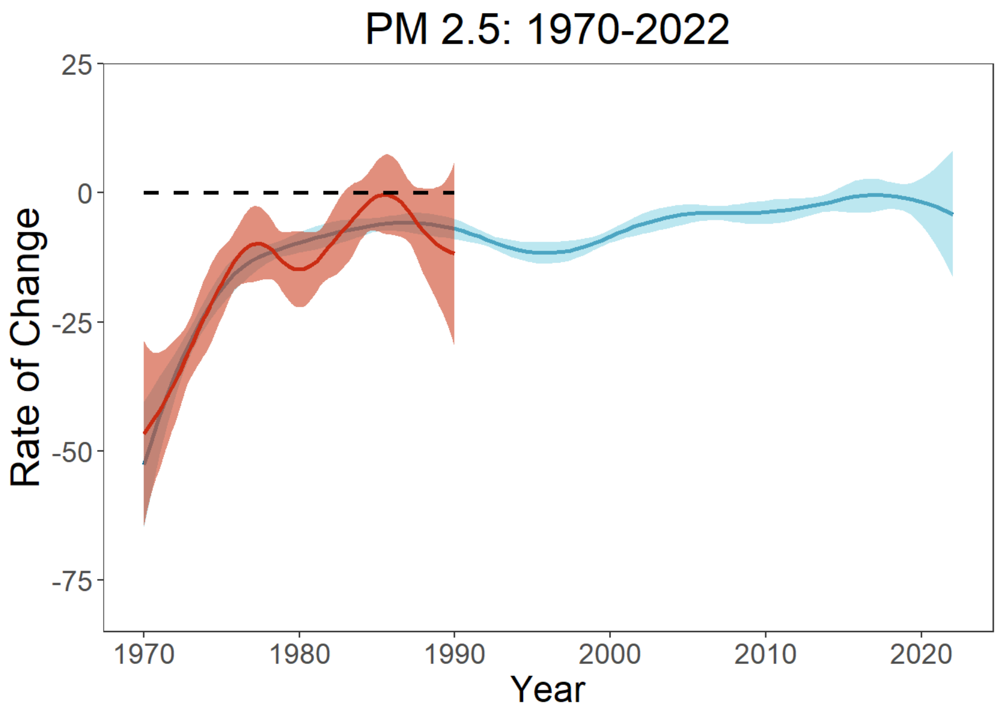

Figure 7: Rate of change curved for the penalized spline smoother fits to $PM_{2.5}$ data from 1970-2022 (blue) compared with 1970-1990 (red), with 95% credible intervals (light blue & brown, respectively).

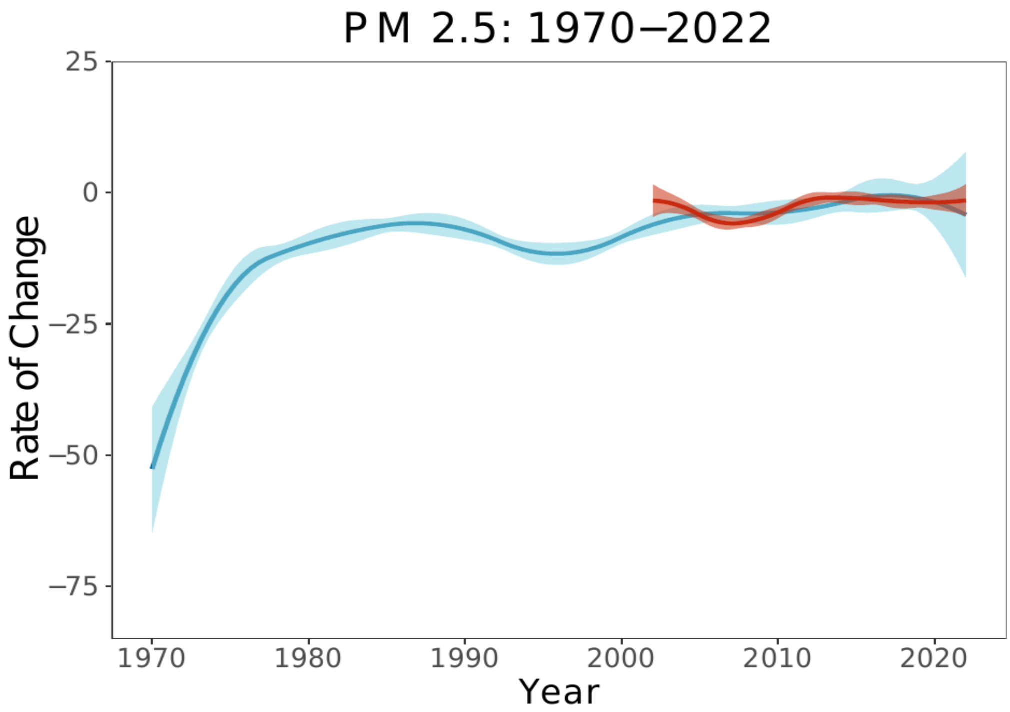

Figure 8: Rate of change curved for the penalized spline smoother fits to $PM_{2.5}$ data from 1970-2022 (blue) compared with 2002-2022 (red), with 95% credible intervals (light blue & brown, respectively).

5.2.2. Use of the smoother method over Shorter Time series

To highlight the importance of a longer time series for the penalized spline method, in figures 6 & 7 we contrast the $PM_{2.5}$ emissions data fit to both the full dataset (blue) and to the 1970-1990 sub-series (red). The model fit to just the 1970-1990 period is more variable than the model fit to the full dataset, because the spline knots are distributed over a shorter period allowing the model the flexibility to respond to higher frequency fluctuations in this period of the data. The estimated rate of change in 1990 for the shorter time series model is -11.8 (95%

CI: -29.2, 5.5) compared to -7.0 (95% CI: -17.0, -5.0) for the longer time series. Using the shorter time series, we would conclude that the estimated rate of change in 1990 is not different from zero and we would assign an amber RAG rating (little or no change), rather than a green RAG rating (improving significantly). Interestingly, choosing a different subset of the time series yields in a different result. Using the penalised spline method for the 2002-2022 sub-series, where the data are not changing rapidly and the short-timescale fluctuations are dominated by relatively small inter-annual variations, results in apparently tighter confidence intervals that those from the method applied to the whole 1970-2022 dataset (figure 8). This contrast in apparent performance for different short sub-series highlights how the penalised spline method becomes increasingly sensitive to both the presence or absence of short timescale fluctuations, their amplitude, and their timescale relative to the length of the dataset, motivating our recommendation to restrict this method to indicators with long time series.

6. Application to subset of indicators

We applied the standard 3% method, using both a one-year baseline and a three-year average baseline, and the proposed the t-test method to nine of the modelling/count-based indicators[22] (see Section 4), and applied the proposed long-term smoother method to four of these indicators. We excluded the indicators "Properties at high risk of flooding", "Percentage of the total population in England living in close proximity of greenspace, as of October 2021", "Exceedance of damaging levels of nutrient nitrogen deposition in England" and "Number of additional tree pests and diseases becoming established", because these have only been established as indicators recently and don't yet have sufficient data for us to apply either of the alternative methods proposed here. The Abundance of Priority species indicator was excluded here because this is already assessed using a sophisticated statistical modelling approach, similar to the penalized spline smoother approach, and unsmoothed data were not available. See Appendix (Table A2) for a final list of the indicators the methods have been applied to, and the resulting change over time assessments.

Using a three-year average as the baseline for the 3% method, rather than a one-year baseline, did significantly change the percentage differences measured for some indicators (for example, "Extent of UK area protected for nature on land and water" has 42% change with the one-year baseline, and a 26% change with the three-year average baseline), however the only RAG assessment that changed as a result was for "Percentage of woodland that is sustainably managed", driven by a marginal shift in the percentage change from -3.0% to -2.9%. This highlights the sensitivity if the percentage change measure to changes in the baseline, how the hard classification thresholds that do not account for uncertainty can strongly impact the headline behaviour reported for trends.

Comparing the RAG classifications from the 3% method (both baselines) with those from the t-test method, six of the nine indicators were classified differently. This includes indicators that exhibit large percentage changes between their baseline year and the most recent year, even though the t-test identifies these changes as not statistically significant, indicating that large inter-annual changes over the baseline period is a driving a component of the measured percentage changes and the subsequent RAG rating classification[23]. Using the t-test method, these indicators are classified as amber "little or no change". The 'Area of woodland' indicator is the only case where the t-test's robust handling of baseline uncertainty reversed the RAG assessment from 'little or no change' to 'improving'. Here, despite a marginal percentage change, the t-test identified statistical significance, a result attributable to smaller inter-annual fluctuations during the baseline period, which increases confidence that even minor changes reflect real improvement.

For two of the three indicators with long-term datasets there is good agreement between the t-test and penalized spline smoother methods, with the more sophisticated method confirming the updated RAG classifications for three of the four indicators. The remaining indicator ("Area under agri-environment schemes") exhibits a substantial (61.3%) increase between 2018 and 2023, classified as 'improving' by both the 3% threshold method and the t-test ($p < 0.01$). However, the penalized spline smoother, which evaluates trends within the full historical context, did not identify this change as statistically significant ($p = 0.25$) indicating that

[22] The methods were applied separately for the five individual components of the indicator "UK emissions of 5 key air pollutants".

[23] Extent of UK area protected for nature on land and water, $PM_{2.5}$ & Sulphur Dioxide emissions, consumption-base greenhouse gas emissions, number of wildfire incidents and population-weighted annual mean concentrations of $PM_{2.5}$ in the air. Also, see Section 7 for a discussion around the impact of change points on time series modelling.

while the recent short-term increase is notable, in the context of the long-term trend variations of indication there isn't clear evidence that it represents a consistent directional momentum.

7. Discussion

Here we proposed two new methods to enhance the interpretation of environmental indicator trends used by the OEP, moving beyond the current reliance on percentage change and the 3% rule-of-thumb threshold. The first method, a t-test based approach, contextualizes yearly changes against the preceding five years of data, testing whether recent fluctuations exceed natural variability observed in that baseline period. The second method, a penalized spline smoother based approach, evaluates long-term datasets to determine whether the current rate of change aligns with a sustained directional trend or represents transient noise. While presented as distinct tools, these methods are, in fact, complementary. The percentage change (ideally measures vs. a baseline average not a single year) quantifies the magnitude of change, the t-test establishes whether the change can be robustly distinguished from the indicators short-term variability, and the spline assesses whether the current rate of change represents a consistent directional trend (improvement or decline), filtering out transient anomalies[24]. Together, these approaches allow readers to assess whether annual changes are sizable, statistically distinguishable, and indicative of consistent improvement or decline. This contrasts with the 3% rule, which lacks strong statistical grounding and fails to account for variability or long-term patterns, offering simplicity at the cost of analytical depth. By integrating effect size, robustness, and coherence, our approach supports more nuanced and evidence-based assessments of environmental progress. For shorter datasets, we recommend presenting two sets of information for each indicator: the percentage change and the t-test p-value. For longer datasets, we recommend presenting all three sets of information, adding the spline results to represent whether the longer-term trend is reliable.

Although more statistically sophisticated methods exist for analysing time series data (such as Markov Switching Autoregressive Models or machine learning techniques like Long Short-Term Memory (LSTM) networks), their complexity poses significant barriers to implementation, maintenance, and interpretation for non-specialists. The OEP's mandate prioritizes accessible, high-impact reporting: indicators must be summarized in ways that are transparent, easily understood by the public and actionable for staff without extensive statistical expertise. The methods proposed here balance statistical rigor, generalizability, interpretability with ease of implementation by non-specialists.

Despite the robustness and generalizability of the methods proposed here, their performance will be significantly disrupted where indicators have abrupt step changes in their time series. These change points can be caused by events such as COVID-19 or shifts in measurement methodology or classification. For example, the simple percentage change year-to-base comparison might show an apparent large improvement immediately around and across the discontinuity, while indicating little change on either side, thereby distorting the true trend. The t-test method is similarly affected because the inflated variance across the step change does not represent natural variability, leading to large p-values even when the change is meaningful[25]. Although a penalized smoother can better capture the long-term trend by averaging over the step change, it may still distort the trend if the discontinuity occurs in a very recent year. We recommend that the OEP consider using specialized change point detection methods (for example, Pruned Exact Linear Time, Killick, Fearnhead and Eckley, 2012) to identify datasets with step changes and then apply alternative or adjusted analyses to ensure an accurate assessment of indicator trends.

The OEP's current Red-Amber-Green (RAG) rating system provides a simple but reductive summary of environmental trends. While intuitive, this approach obscures critical nuances. It fails to convey uncertainty in the magnitude or direction of change and collapses continuous trends into discrete categories. If adopting the proposed methods, we strongly recommend evolving the RAG system into a more nuanced framework that communicates both the effect size (*e.g.*, percentage change) and statistical significance (*e.g.,* p-value) of trends,

[24] Except where they occur near the end of the time series.

[25] The Extent of UK area protected for nature on land and water is an example of this. A large percentage change caused by discontinuity in the data in 2019 results in the 3% method identifying a large change while this year remains in the five-year window regardless of little recent improvement while the t-test and penalized spline smoother models both identify no significant change, despite a meaningful change having occurred.

ideally including uncertainty intervals. Where color-coding remains essential, we urge replacing the Red-Amber-Green palette with a perceptually uniform, colourblind-friendly scheme, such as a Blue-Gray-Yellow gradient. This scale (continuous or discrete) avoids the accessibility pitfalls of red-green contrasts while maintaining the intuitive associations of cooler tones (blue) for positive trends, neutral tones (grey) for stability, and warmer tones (yellow) for negative shifts. These updates would align the OEP's communication with best practices in data visualization, ensuring transparency, inclusivity, and scientific fidelity.

The two new methods proposed here are appropriate for assessing change over time for a wide range of environmental indicators with continuous numerical values, including those not included in the subset of indicators we focus on here. For the t-test method, a key assumption is that the residuals around the mean are normally distributed, implying data that are normally distributed, limiting the potential range of indicators it should be used with. In particular, when used with count-based indicators, the t-test method is only applicable where the data follow a Poisson distribution when the values are high (>30) where a Poisson distribution approximates a normal distribution; below this value the Poisson distribution violates this assumption and a square-root or logarithmic transformation should be applied to the data before the t-test is used. Similarly, for log-normally distributed data such as molecular concentrations, data values should have an appropriate transformation applied to move the values into a space where they are approximately normally distributed before analysis. The penalized spline smoother is a flexible and adaptable modelling method that can be applied to a wide variety of data types[26], but it requires tailoring specifically for indicators whose values are not normally distributed[27].

8. Conclusion

In this report, we review the current approach the OEP uses to assess change over time in environmental indicator variables, put this approach in a wider national and international context. We propose and implement two alternative methodologies that would improve the statistical rigour and robustness of the OEP assessment, by explicitly including the scale of recent variation and the long-term trend shape in the assessment. The new methods proposed are a i) t-test based approach, which contextualizes the most recent yearly change against the variation observed in the preceding five years of data, and ii) a penalized spline smoother based approach, which evaluates whether the current rate of change of an indicator aligns with a sustained directional trend or represents transient noise, long-term datasets to determine.

We conclude that:

I. The current 3% methodology used by the OEP is based on a doubling or halving of values over a 25-year period and is widely used by UK agencies and organizations but has been altered sufficiently to have lost this original contextual meaning. This approach is not broadly used in an international context, however, there is no international consensus agreement on best-practice methodology for assessing change over time in environmental indicators.
II. The assessment of percentage change from a recent baseline year is a useful, but noise-prone, measure of the effect size of the observed change. Measuring the effect size using a three- or five-year mean centred on the baseline year would be a significant improvement, diluting the impact of inter-annual fluctuations.
III. Categorizing the change over time based on an arbitrary 3% threshold for the percentage change (whether calculated vs a single baseline year or vs an average baseline period), while straightforward to interpret, is overly reductionist, obscures critical nuances, and is not best practice. We recommend discontinuing this.
IV. The new proposed methods represent a significant improvement over the current OEP practice by accounting for variation and trends in the data, supporting a more nuanced and evidence-based assessment of the change over time of environmental indicators.
V. Using the proposed new methods changes the RAG rating classification for seven of the sub-set of ten indicators focussed on for this report.

[26] Including continuous, log-normal, (over-dispersed) count, binary/proportional, positive continuous and ordered categorical data types.
[27] The modelling code used here can be tailored using the "family" argument for the *mgcv* modelling framework, see https://www.rdocumentation.org/packages/mgcv/versions/1.9-1/topics/gam

VI. While presented as distinct approaches, the methods are, in fact, complementary and we recommend presenting both the effect size (*e.g.*, percentage change, not classified by the 3% method) and statistical significance (*e.g.*, t-test and/or penalized spline smoother method), of the observed change over time together, allowing readers to assess whether annual changes are meaningful (sizable, statistically distinguishable, and indicative of consistent improvement or decline).

VII. The OEP should consider updating the RAG rating system (particularly if these methods are adopted), to bring it in line with best-practice for scientific visualizations.

9. References

Cleveland, R.B., Cleveland, W.S. and Terpenning, I. (1990) 'STL: A Seasonal-Trend Decomposition Procedure Based on Loess', *Journal of Official Statistics*, 6(1), p. 3.

Cowpertwait, P.S.P. and Metcalfe, A.V. (2009) *Introductory Time Series with R*. New York, NY: Springer. Available at: https://doi.org/10.1007/978-0-387-88698-5.

DEFRA (2004) *D4: Relative abundance and/or distribution of species*. Available at: https://oifdata.defra.gov.uk/themes/wildlife/D4/ (Accessed: 13 March 2025).

DEFRA (2014) *Wild bird populations in the UK 1970 to 2013*.

DEFRA (2025a) *England Biodiversity Indicators - Assessing indicators (Section 2)*, *GOV.UK*. Available at: https://www.gov.uk/government/statistics/england-biodiversity-indicators/assessing-indicators (Accessed: 13 March 2025).

DEFRA (2025b) *Outcome Indicator Framework 2025 - Indicator Assessment*. Available at: https://oifdata.defra.gov.uk/assessment/ (Accessed: 13 March 2025).

EEA (2019) *The European environment — state and outlook 2020*. Available at: https://www.eea.europa.eu/en/analysis/publications/soer-2020 (Accessed: 13 March 2025).

Ffoulkes, C. *et al.* (2023) 'Research to update indicators which monitor progress in adaptation in England'.

Freeman, S.N. *et al.* (2021) 'A Generic Method for Estimating and Smoothing Multispecies Biodiversity Indicators Using Intermittent Data', *Journal of Agricultural, Biological and Environmental Statistics*, 26(1), pp. 71–89. Available at: https://doi.org/10.1007/s13253-020-00410-6.

Galbraith, C. (2002) 'The Population Status of Birds in The UK: Birds of Conservation Concern: 2002–2007'.

GOV.UK (2023) *Environmental Improvement Plan 2023*, *GOV.UK*. Available at: https://www.gov.uk/government/publications/environmental-improvement-plan (Accessed: 13 March 2025).

GOV.UK (2025) *Progress in improving the natural environment in England 2023 to 2024*, *GOV.UK*. Available at: https://www.gov.uk/government/publications/progress-in-improving-the-natural-environment-in-england-2023-to-2024 (Accessed: 13 March 2025).

Institute of Fisheries Management (2021) *THE STATE OF THE THAMES 2021 – Institute of Fisheries Management*. Available at: https://ifm.org.uk/the-state-of-the-thames-2021/ (Accessed: 13 March 2025).

JNCC (2013a) *Fifth National Report to the United Nations Convention on Biological Diversity: United Kingdom (Appendix 5: UK Biodiversity Indicators)*.

JNCC (2013b) *UK Biodiversity Indicators 2013 - Assessing indicators*. Available at: https://webarchive.nationalarchives.gov.uk/ukgwa/20131221104026/http://jncc.defra.gov.uk/page-4230 (Accessed: 13 March 2025).

JNCC (2023a) *UK Biodiversity Indicators 2023*. Available at: https://webarchive.nationalarchives.gov.uk/ukgwa/20240906064122/https://jncc.gov.uk/our-work/uk-biodiversity-indicators-2023/ (Accessed: 13 March 2025).

JNCC (2023b) *UK Biodiversity Indicators 2023 - Assessing Indicators*. Available at: https://webarchive.nationalarchives.gov.uk/ukgwa/20240906062116/https:/jncc.gov.uk/our-work/assessing-indicators/ (Accessed: 13 March 2025).

JNCC (2023c) *UK Biodiversity Indicators 2023 - C5. Birds of the wider countryside and at sea*. Available at: https://webarchive.nationalarchives.gov.uk/ukgwa/20240906065550/https:/jncc.gov.uk/our-work/ukbi-c5-birds-of-the-wider-countryside-and-at-sea/ (Accessed: 13 March 2025).

JNCC (2025a) *UK Biodiversity Indicators 2025 - Biodiversity expenditure*. Available at: https://jncc.gov.uk/our-work/ukbi-biodiversity-expenditure/ (Accessed: 13 March 2025).

JNCC (2025b) *UK Biodiversity Indicators 2025 - Conservation Volunteering*. Available at: https://jncc.gov.uk/our-work/ukbi-conservation-volunteering/ (Accessed: 13 March 2025).

JNCC (2025c) *UK Biodiversity Indicators 2025 - Greenhouse gas removal*. Available at: https://jncc.gov.uk/our-work/ukbi-greenhouse-gas-removal/ (Accessed: 13 March 2025).

Keselman, H.J. *et al.* (2004) 'The New and Improved Two-Sample t Test', *Psychological Science*, 15(1), pp. 47–51. Available at: https://doi.org/10.1111/j.0963-7214.2004.01501008.x.

Killick, R., Fearnhead, P. and Eckley, I.A. (2012) 'Optimal Detection of Changepoints With a Linear Computational Cost', *Journal of the American Statistical Association*, 107(500), pp. 1590–1598. Available at: https://doi.org/10.1080/01621459.2012.737745.

McShane, B.B. and Gal, D. (2017) 'Statistical Significance and the Dichotomization of Evidence', *Journal of the American Statistical Association*, 112(519), pp. 885–895. Available at: https://doi.org/10.1080/01621459.2017.1289846.

OEP (2025) 'OEP Methodological Statement – Progress in improving the natural environment in England 2023/2024'.

R Foundation for Statistical Computing (no date) 'R: A Language and Environment for Statistical Computing.'

US EPA (2015) *US EPA Report on the Environment*, *US EPA*. Available at: https://www.epa.gov/report-environment (Accessed: 13 March 2025).

US EPA (2016) *US EPA Air Quality - National Summary*. Available at: https://www.epa.gov/air-trends/air-quality-national-summary (Accessed: 13 March 2025).

US EPA (2017) *US EPA Greenhouse Gases*. Available at: https://www.epa.gov/report-environment/greenhouse-gases (Accessed: 13 March 2025).

Winter, J.C.F. de (2013) 'Using the Student's t-test with extremely small sample sizes', *Practical Assessment, Research, and Evaluation*, 18(1). Available at: https://doi.org/10.7275/e4r6-dj05.

Wood, S.N. (2011) 'Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models', *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1), pp. 3–36. Available at: https://doi.org/10.1111/j.1467-9868.2010.00749.x.

10. Appendix

Indicator	Goal	Method	Method Group
UK emissions of 5 key air pollutants	Clean air	Emissions factors calculated from activity levels and emissions	Modelling/counts
Population-weighted annual mean concentrations of PM2.5 in the air		Calculated from the average	Modelling/counts
Exceedance of damaging levels of nutrient nitrogen deposition in England		Calculated using moving average	Modelling/counts
Percentage of Monitoring Stations above 10 μg/m³ annual mean PM2.5 concentrations		Percentage exceedances	Threshold/Monitoring
Incidents of exceedances against Air Quality Standards Regulations in England		Count of exceedances	Threshold/Monitoring
State of the water environment [Water Framework Directive Ecological Status]	Clean and plentiful water	Combined metrics	Combined emissions measures
Condition of bathing waters		Percentage meeting criteria	Threshold/Monitoring
Loads discharged to rivers from water company sewage treatment works [of three key pollutants]		Concentration + Flow	Threshold/Monitoring
Pollution incidents to water [Environment Agency as Category 1 to 3]		Count	Other
Per capita drinking water consumption in England		Moving average	Other
Water company security of supply performance		"stars"	Other
Water leakage in England [from water company drinking water networks]		Moving average	Other
Visits to the natural environment	Enhancing beauty, heritage and engagement with the natural environment	Survey	Survey
Condition of geological and geomorphological heritage features of Sites of Special Scientific Interest in England		Survey	Survey
Environmental attitudes and behaviours		Survey	Survey
Environmental Health and well-being benefits		Survey	Survey
Percentage of the total population in England living in close proximity of greenspace, as of October 2021		Multiple indicators	Modelling/counts
Number of additional tree pests and diseases becoming established	Enhancing biosecurity	Survey and running average vs roll forward woodland accounting	Modelling/counts
Number of invasive nonnative species becoming established		Survey + expert opinion	Abundance modelling
Emissions of persistent organic pollutants	Exposure to chemicals and pesticides	Combined metrics	Combined emissions measures
Emissions of mercury to air, land and water		Combined metrics	Combined emissions measures
Chemical Status of Surface and Groundwater [Water Framework Directive]		Classification	Threshold/Monitoring
Hazardous Waste Disposal		Count	Other
Percentage of sampled fulmars having more than 0.1g of plastic in their stomach, Greater North Sea, 2004-2008 to 2017-2021 [Marine Good Environmental Status Descriptor Marine Litter]		Count	Threshold/Monitoring

Resource Productivity	Maximise our resources, minimise our waste	Material Flow Accounting	Economics
Amount of raw material consumed		Material Flow Accounting	Economics
Residual Waste		Survey	Survey
Number of Fly-Tipping Incidents		Survey	Survey
Number of illegal waste sites		Survey	Survey
Consumption-based greenhouse gas emissions	Mitigating and adapting to climate change	Modelling	Modelling/counts
Emissions of fluorinated gases		Modelling	Modelling/counts
UK emissions of greenhouse gases		Combined (sectoral estimated) metrics	Combined emissions measures
Properties at high risk of flooding	Reduced risk of harm from environmental hazards	Count surveys and Modelling	Modelling/counts
Number of wildfire incidents		Count surveys	Modelling/counts
Achievement of marine 'good environmental status'	Thriving plants and wildlife	Survey (many indicators)	Survey
Condition of Marine Protected Areas		Survey	Survey
Extent of land cover more likely to support nature friendly habitat		Mapping	Mapping
Area of woodland		Roll-forward from 2011	Modelling/counts
Abundance of priority species		Average from metrics	Abundance modelling
Threat of extinction to UK species		Surveys + models	Abundance modelling
Condition of Sites of Special Scientific Interest [which are in favourable or unfavourable recovering condition]		Area + Survey	Survey
Extent of UK area protected for nature on land and water		Area counts and smoothed trends	Modelling/counts
Area under agri-environment schemes		Area counts	Modelling/counts
Percentage of woodland that is sustainably managed	Using resources from nature sustainably	Count vs roll forward woodland accounting	Modelling/counts
Fish stocks that are sustainably harvested [Good Environmental Status Descriptor Commercial Fish]		Sampling + models	Threshold/Monitoring
Soil Health		Sampling + models	Threshold/Monitoring

Table A1: Details of the 46 OEP environmental indicators that have data, including indicator name, associated goal (as stated in the Outcome Indicator Framework), and a brief description of the method and data type underlying each indicator and the group (8 in total) each indicator was assigned. Priority indicators selected for the analysis in this report are highlighted in green.

		Current year		One-year Baseline			Three-year mean baseline			t-test vs baseline		Long-term Penalized Spline Smoother		
Indicator name	**Units**	**Year**	**Value**	**Year**	**Value**	**% change vs baseline (RAG assessment)**	**Years**	**Value**	**% change vs baseline (RAG assessment)**	**Years**	**p-value**	**Years**	**Rate of Change**	**p-value**
Extent of UK area protected for nature on land and water	Mha	2024	40.6	2019	28.6	42.0	2018 - 2020	32.3	25.7			1950 - 2024	-0.27	0.27
Area under agri-environment schemes (England)	kha	2023	4487	2018	2781	61.3	2017 - 2019	2925	53.4	2018 - 2022	<0.01	1992 - 2023	541.5	0.25
Area of woodland (England)	kha	2024	3279	2019	3198	2.5	2018 - 2020	3196	2.6	2019 - 2023	0.01	1998 - 2024	17.75	<0.01
UK emissions of 5 key air pollutants														
- Ammonia	kt (indexed to 2012)	2022	99.5	2017	105.3	-5.5	2016 - 2018	104.4	-4.7	2017 - 2021	0.02			
- Non-Methane Volatile Organic Compounds	kt (indexed to 2013)	2022	54.4	2017	74.1	-26.6	2016 - 2018	74.0	-26.5	2017 - 2021	0.04			
-Nitrogen Oxides	kt (indexed to 2014)	2022	84.6	2017	93.6	-9.6	2016 - 2018	94.0	-10.0	2017 - 2021	<0.01			
- $PM_{2.5}$	kt (indexed to 2015)	2022	81.9	2017	92.2	-11.2	2016 - 2018	92.3	-11.3	2017 - 2021	0.08			
- Sulphur Dioxide	kt (indexed to 2016)	2022	25.2	2017	39.3	-35.9	2016 - 2018	38.8	-35.1	2017 - 2021	0.07			
Percentage of woodland that is sustainably managed	%	2024	57.0	2019	58.8	-3.0	2018 - 2020	58.7	-2.9	2019 - 2023	<0.01			
Consumption-based greenhouse gas emissions	$MtCO_2e$	2021	158	2016	198	-20.2	2015 - 2017	203	-22.2	2016 - 2020	0.08			
Emissions of fluorinated gases	$MtCO_2e$	2021	9.3	2016	11.5	-19.1	2015 - 2017	11.5	-19.1	2016 - 2020	<0.01			
Number of wildfire incidents	Counts	2020/21	26870	2015/16	23834	12.7	2014/15 - 2016/17	21934	22.5	2015/16 - 2019/20	0.90			
Population-weighted annual mean concentrations of $PM_{2.5}$ in the air	kt	2022	64.9	2017	73.1	-11.2	2016 - 2018	73.2	-11.3	2017 - 2021	0.81	1970-2022	-4.19	0.25

Table A2: Indicator assessments for priority indicators (see Section 4). Change over time for current year values were assessed using the 3% method applied vs a one-year baseline, the mean of a three-year baseline period, using the proposed t-test method and using the proposed penalized spline smoother method where sufficient data exist (cells shaded in blue indicate insufficient data). Cell shading indicated the RAG rating assigned to each indicator for each method (using a p-value threshold of 0.05 for the t-test and penalized spline smoother). Note that the large percentage change for the "Extent of UK area protected for nature on land and water" indicator is flagged as not-significant by the t-test methos and the penalized spline smoother method because it is due to a recent change point in the data (see Section 7).