A counterfactual approach to measure the impact of wet grassland conservation on UK breeding bird populations

Sean Jellesmark[1,2], Malcolm Ausden[4], Tim M. Blackburn[1], Richard D. Gregory[4,1], Mike Hoffmann[2], Dario Massimino[5], Louise McRae[2] & Piero Visconti[3].

[1]Centre for Biodiversity and Environment Research, University College London, London, United Kingdom

[2]Institute of Zoology, Zoological Society of London, London, United Kingdom

[3]International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A-2361 Laxenburg, Austria

[4]RSPB Centre for Conservation Science, The Lodge, Sandy, Bedfordshire SG19 2DL

[5]British Trust for Ornithology, The Nunnery, Thetford, IP24 2PU, United Kingdom

Correspondence: Sean Jellesmark, Centre for Biodiversity and Environment Research, Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, United Kingdom; Email: sean.jellesmark.18@ucl.ac.uk

Running head: Counterfactual evaluation of wetland reserves

Article impact statement: Post-hoc quasi-experimental analysis shows that wet grassland conservation has positively affected targeted breeding bird populations.

Abstract

Wet grassland wader populations in the United Kingdom have experienced severe declines over the last three decades. To help mitigate these declines, the Royal Society for the Protection of Birds (RSPB) has restored and managed lowland wet grassland nature reserves to benefit these and other species. However, the impact that these reserves have on bird population trends has not been experimentally evaluated, as appropriate control populations do not readily exist. In this study, we compare population trends from 1994 - 2018 for five bird species of conservation concern that breed on these nature reserves with counterfactual trends using matched breeding bird survey observations. Our results showed positive effects of conservation interventions for all four wader species that these reserves aim to benefit: Lapwing (*Vanellus vanellus*), Redshank (*Tringa totanus*), Curlew (*Numenius arquata*) and Snipe (*Gallinago gallinago*). There was no positive effect of conservation interventions on reserves for the passerine, Yellow Wagtail (*Motacilla flava*). We compared reserve trends with three different counterfactuals, based on different scenarios of how reserve populations could have developed in the absence of conservation, and found that reserve trends performed better regardless of the counterfactual used. Our approach using monitoring data to produce valid counterfactual controls is a broadly applicable method allowing large-scale evaluation of conservation impact.

Introduction

Halting the decline of global biodiversity is currently one of humanity's greatest environmental challenges. Within animal populations, declines have predominantly been attributed to changes in land-use, invasive species, exploitation of species and habitats, pollution and climate change

2

(IPBES 2019). Policy makers, NGOs and conservation practitioners are addressing this global decline through a range of conservation actions, chiefly habitat and species management interventions within and outside protected areas. Between 2010 – 2018, £817 million was spent on average each year on promoting and protecting biodiversity within the United Kingdom (UK) alone (Department for Environment, Food and Rural Affairs 2019). Despite these efforts, biodiversity indicators in the UK continue to show declines (Hayhow et al. 2019), a directional trend borne out across much of the globe (IPBES 2019). An important question in understanding the impact of conservation interventions on target populations is the extent to which those interventions mitigate or reverse population declines (Hoffmann et al. 2010, 2015). However, limited resources often mean that evaluation efforts do not extend beyond simple measures of association. Population trends are often monitored within protected areas, but appropriate control trends are not. Thus, whether population changes in target species are caused by the conservation management methods employed within protected areas or represent changes that would have occurred in the absence of that management, remains untested.

To assess the impact of conservation, it is necessary to understand what would have happened in the absence of conservation, that is, the counterfactual conservation outcome (Ferraro & Pattanayak 2006; Ferraro 2009; Baylis et al. 2016; Bull et al. 2020). The exact form of the counterfactual can never be known for certain. Ideally, a robust study design such as a randomized controlled trial (RCT; random assignment of treatment and control groups) could be used to infer the causal effect of a treatment by approximating the counterfactual outcome. However, RCT designs are rarely used in conservation. This is because randomisation is often infeasible: for example, there can be legislative obligations to manage protected sites in ways

3

considered beneficial to conservation, which makes it difficult to include unmanaged controls. In addition, the scale of conservation interventions and sampling units may be too large to allow for sufficient replication (Margoluis et al. 2009; Baylis et al. 2016; Wiik et al. 2019).

Conservation practitioners resort to other evaluation designs because of the financial, practical, and logistical challenges of the RCT design. These include "After" (A) methodology (e.g. increasing or decreasing post-treatment population size), "Before-After" (BA) (e.g. pre-treatment population changes are compared with post-treatment population changes), and "Control-impact" (CI) (e.g. comparing population densities within reserves with population densities outside reserves). Such approaches are important in determining the extent to which conservation objectives are being achieved and are a prerequisite for adaptive management. However, if potential biases are not properly addressed, these approaches cannot be used to determine cause and effect with any high level of confidence. The "After" study design describes the post-treatment rate of change and direction but does not provide insight into whether the change would have differed without the treatment. The "Before-After" study design assumes that temporal variability and confounding factors before and after the intervention are comparable, and "Control-impact" assumes time-for-space substitution and comparability between groups. The validity of such inferences is therefore compromised if a population would have developed similarly regardless of conservation (e.g. A), if the effect of confounding variables is not homogenous across time (e.g. BA), and/or local variation is systematically different between impact and control groups (e.g. CI) (Joppa & Pfaff 2010; Geldmann et al. 2013; Ferraro & Pressey 2015; De Palma et al. 2018; Adams et al. 2019). To improve the credibility of an inference, the "BA" and the "CI" study design can be combined, forming the "Before-After-Control-Impact" (BACI) study design (e.g. comparing pre-treatment and post-treatment

4

densities in a treated and a control group while accounting for the pre-treatment density difference between treated and control group). Using simulated ecological data, the BACI design has been shown to estimate the true effect size better than RCTs (1.3-1.8 times more likely to estimate ±30% of the true effect and direction), CIs (3.2-4.6 times more likely) and A study designs (7.1-10.1 times more likely) (Christie et al. 2019). However, this study design has many of the same limitations as the RCT and is further limited if appropriate controls cannot be identified ex-ante (e.g. appropriate controls cannot be selected if confounders are unknown or poorly understood).

To produce reliable conservation effect estimates, matching techniques are increasingly being used in conservation science (Sills et al. 2017; Schleicher et al. 2019; Sonter et al. 2019). The intent of matching is to create treatment and control groups with similar covariates by sub-setting treated and control samples so that comparisons are carried out using groups that have similar characteristics (e.g. comparing the outcome of a treated group to the outcome of a control group where both groups are from the same habitat type, altitude and country). The post-matching control group then represents the counterfactual outcome of the treated group, and the effect of a given treatment can be inferred as the difference between outcomes. For example, Ferraro et al. (2007) tested the effectiveness of US Endangered Species Act listing and funding on species recovery, using matching on a set of observable covariates to account for bias in the listing and funding process. They found listing to be effective only when accompanied by adequate funding. Geldmann et al. (2019) assessed whether protected areas (PAs) reduce anthropogenic pressure. They used ten variables linked to PA selection to match PAs to similar unprotected areas and found that, on average, PAs did not reduce human pressure. Nevertheless, while the theoretical potential of these methods has been highlighted, examples of

their application remain scarce (Ferraro & Pattanayak 2006; Margoluis et al., 2009; Joppa & Pfaff 2010).

Here, we adopt a matching approach to explore the impact of specific conservation interventions on a particular habitat of conservation concern in Europe: lowland wet grassland (Franks et al. 2018). Conversion to other habitat types, changes in grazing regimes, drainage and agricultural intensification have adversely affected these grasslands (Wilson et al. 2004). In particular, wetland bird species using this habitat to breed and over-winter, such as wading birds (Charadriiformes), have exhibited severe breeding population declines as a result of these habitat changes (Wilson et al. 2005; Boatman et al. 2007; Colhoun et al. 2017; Harris et al. 2019; Hayhow et al. 2019). For example, Lapwing (*Vanellus vanellus*) populations, once abundant in the countryside of the UK, have declined by 42% between 1995-2017 (Harris et al. 2019). To help mitigate these declines, the Royal Society for the Protection of Birds (hereafter RSPB) has allocated resources to purchasing, restoring and managing reserves in lowland wet grassland habitats to benefit breeding waders within the UK. Conservation interventions such as raising and manipulating water levels, beneficial stock grazing regimes, control and/or exclusion of generalist predators, and mechanical vegetation control are implemented on these reserves (Ausden et al. 2019). Conservation efforts of this type have been shown to be associated with increasing wader populations (Ausden & Hirons 2002; Malpas et al. 2013; Smart et al. 2014; Franks et al. 2018; Ausden et al. 2019). However, a central issue is whether the conservation actions result in positive benefits to the target populations: is the population performance better than would have occurred in the absence of these interventions? We test this by comparing breeding trends on the reserves with matched counterfactual trends that represent how the trends might have developed in the absence of reserve-based conservation

6

interventions. This is, to our knowledge, the first time post-hoc evaluation of conservation interventions using quasi-experimental After-Control-Impact (ACI) analyses has been carried out for conservation interventions in the UK (We are using trends post intervention thus "After" and matching reserve trends to counterfactual controls therefore "Control – Impact").

Methods

Data

We used bird counts from RSPB lowland wet grassland reserves and from the UK Breeding Bird Survey (Harris et al. 2019) for the period 1994 – 2018. The RSPB manages more than 200 reserves across the UK, with 47 of these containing lowland wet grassland (see map S1). Most of these reserves are in England (35), with the rest in Scotland (7), Wales (3) and Northern Ireland (2). We chose the habitat type lowland wet grassland – periodically flooded grasslands below c. 250m altitude (Jefferson & Grice 1998) – because this is a habitat in which considerable resources have been invested in habitat restoration and creation in recent decades. The area of lowland wet grassland on individual reserves varies from 18 ha to 1,300 ha (mean site area = 95 ha; SD = 144). Some reserves consist of two or more non-contiguous blocks of lowland wet grassland habitat, which we refer to as 'sites'. For this study, we used a total of 101 sites within the 47 reserves. We treated new acquisitions of land as separate sites. RSPB reserves are managed in accordance with the biological requirements of priority species selected for that reserve. The number of breeding pairs of priority bird species are counted three times annually at each site using standard methods described in Gilbert et al. (1998) (Supporting Information).

The focal wetland species were Garganey (*Anas querquedula)*, Shoveler (*A. clypeata*), Black-tailed Godwit (*Limosa limosa)*, Lapwing, Curlew (*Numenius arquata*), Snipe (*Gallinago gallinago*), Redshank (*Tringa totanus*) and Yellow Wagtail (*Motacilla flava*), with analysis concentrating on the latter five abundant species. These species were chosen for practical reasons: first, populations breed on reserves; second, they are currently RSPB priority species and have been monitored both on reserves and in the wider countryside (see below); and third, most importantly, conservation interventions are designed closely to match their biological breeding requirements, making the number of breeding birds a natural response to the conservation type evaluated in this study (Supporting Information).

In the case of Snipe and Yellow Wagtail, a large proportion of their breeding reserve population (59% and 90%, respectively, at the start of the period analysed) occurred at a single reserve, the Ouse Washes in Norfolk/Cambridgeshire. This site is atypical because breeding birds are sometimes disrupted by flooding during the breeding season, as the site is designed to temporarily store floodwater. This flooding is outside the control of the reserve management and has been shown to explain population declines for Black-tailed Godwit (Ratcliffe et al. 2005). We therefore carried out analyses both with and without the Ouse Washes for Snipe and Yellow Wagtail.

We obtained matching data to compute counterfactual population trends from the UK Breeding Bird Survey (BBS), managed by the British Trust for Ornithology. This scheme was started in 1994, and monitors changes in the national breeding trends of more than a hundred common and widespread bird species (Gregory et al. 2000; Harris et al. 2019). Surveying is performed

within 1x1 km grids, each consisting of 10 transects. Habitat is recorded in a separate visit prior to two annual bird counts (Supporting Information). We used the habitat data recorded in the BBS and altitude data from the OS terrain 50 data set and the USGS EROS Archive – Digital Elevation (SRTM) 1 Arc-Second Global to calculate mean altitude levels.

We selected observations from lowland wet grassland sites and target species to create one reserve sample (i.e. treated sites) and matched the BBS data exactly on covariates affecting reserve selection and breeding trend (Table 1) to create the counterfactual sample (i.e. the control sites) for each species. We call this our benchmark counterfactual, as opposed to two other variants introduced to test sensitivity of the results (see below and Table 1). The counterfactuals were created by selecting observations from BBS grids containing certain habitats (Table 1) as we believe these are the best approximations of how reserve land would have developed without reserve conservation. We did not set a minimum proportion of the selected habitats, nor the exact mix of habitats that a grid had to contain in order to be included in the counterfactual sample. In the BBS, birds are counted in transect of 200 meters and habitat is determined similarly. This also means that both bird numbers and habitat distinction come with some uncertainty regarding exactly where habitat changes and birds are observed. To account for this uncertainty, we operated on 1 km grid level. Furthermore, the counterfactuals must reflect how reserves would have developed without conservation which are not necessarily into wet grasslands. We used a Directed Acyclic Graph (DAG) to present our hypothesis for how wetland conservation affects breeding trends and to select matching covariates (Figure 1) (Pearl 2010; Stuart 2010; Pearl & Mackenzie 2018; VanderWeele 2019; Hernan & Robins 2020).

9

We excluded counts from the matched control sample if they originated from grids spatially overlapping with the chosen reserves (see Stable Unit Treatment Value Assumption in Rubin 1980). Transect counts were summarised for each grid, excluding transect counts with >10 individuals as birds on passage, as it is not likely that the study species breed in such high densities (Field & Gregory 1999). The maximum annual grid count for each species was used and grids that were only surveyed once were excluded. Furthermore, to avoid uncertain trend estimates, we excluded all BBS species that were observed in less than 30 grids annually (Newson et al. 2009). Pre-analysis data manipulation and graphics were done using the tidyverse packages (Wickham et al. 2019) and DAGs using the dagitty package (Textor et al. 2016). All analysis, visualisation and manipulation were implemented using R version 3.5.1 (R Core team 2019). All code used in this study is available at: https://github.com/seanjellesmark/lwg.

Data analysis

We used imputed counts to calculate the species totals used to create both reserve and counterfactual trend indices. Imputed means that if a given site (BBS grid or reserve site) at a given year has been monitored, then the observed count is used, otherwise the missing count is estimated (Supporting Information). Missing population counts were estimated separately for each species x reserve or counterfactual combination using a loglinear model with Poisson error terms, modelling each count as a function of site and year effects (equation 1) using the rtrim package (Bogart et al. 2020). The standard error was adjusted for overdispersion and temporal autocorrelation (Bogart et al. 2020; Pannekoek et al. 2018).

$$\text{Equation 1} \quad LnY_{ij} = \alpha_i + \beta_j$$

Where Y_{ij} is the estimated count for site i at time j, α_i is the average log-count of site i and β_j is the average log-count deviation at time j across all sites.

We used indices to reflect relative changes in breeding pairs through time. The indices were calculated by dividing each annual total imputed count by a reference value which is set as the total count in the first time point (year 1994). Each set of indices was then tested against its counterfactual to examine whether the two sets of indices are different using a Welch two-sample t-test. If any difference could be statistically substantiated (p<0.05), the effect size was assessed as the mean trend of the counterfactual indices subtracted from the corresponding annual reserve indices.

A concern with quasi-experimental inferences is whether the correct variables have been included in the matching process (Stuart 2010). We therefore created two alternative counterfactuals, imposing different matching requirements (Table 1). We created a 'liberal' counterfactual imposing only exact species as a covariate restriction. The liberal counterfactual relaxes the criterion to define "like for like" in control populations but has the potential advantage of increasing the number of control populations. This counterfactual assumes that, on average, the reserve populations would have developed like any other population in the UK. We also created a 'stringent' counterfactual which matches on exact species observations, and a

subset of the habitat types used in the benchmark that is closer to the lowland wet grasslands within RSPB reserves. That is, matching grids were lowland (mean altitude below 250m) and contain transects of either dry grassland, water meadows/grazing marsh, reed swamp, or open marshland. The stringent counterfactual thus assumes that, for each species, the average reserve trend would have developed like that of an average primarily lowland wet habitat regardless of conservation action. The increase in similarity requirements of matching populations comes at the cost of further limiting their numbers, thus potentially reducing the statistical power of the analyses; however, it might better describe the effect of conservation by reducing confounding effects. We assessed whether the results were robust to the counterfactual used by comparing the t-test results from both the liberal and stringent counterfactual (each one tested separately against the reserve indices) to the t-test results of the benchmark counterfactual (benchmark indices tested against reserve indices). We also examined the relationship between site age and changes in breeding counts and whether reserve trends were sensitive to exclusion of sites with large breeding counts (Supporting Information).

Results

Shoveler, Garganey and Black-tailed Godwit were not sufficiently represented in the BBS data to create valid benchmark counterfactuals but showed either stable or increasing trends on reserves (Supporting Information). The distribution of the remaining target species across lowland wet grassland reserve sites varied considerably. Lapwing and Curlew were present on most reserve sites and BBS grids, Yellow wagtail and Redshank were consistently rarer than other species, regardless of the counterfactual approach used (Table 2 and Supporting Information). The BBS grids used for the benchmark counterfactuals consisted primarily of

farmland (45.5%), wet grassland transects (the semi-natural grassland types used in the stringent counterfactual in table 1) (19.9%) and other semi-natural grassland transect (the remaining semi-natural grassland types) (12.7%) whereas the liberal counterfactuals consisted primarily of farmland (67.3%) and other habitat types (24.7%). The stringent counterfactuals consisted primarily of wet grassland transects (27.6%) and farmland (47.4%) (Supporting Information). The largest relative increase in breeding pairs occurred within the first 10 years of reserve creation (Supporting Information).

The breeding indices for Snipe and Yellow Wagtail across all lowland wet grassland reserves could not be statistically distinguished from their benchmark counterfactuals (Snipe: $t = 1.9$, df $= 40$, $p = 0.07$. Yellow Wagtail: $t = -0.3$, df $= 39$, $p = 0.79$). However, when the Ouse Washes was excluded from the reserve dataset (because its spring flooding is known to negate the effect of wetland management), the Snipe indices became more positive than its benchmark counterfactual (Fig. 2; $t = 4$, df $= 47$, $p = 0.0002$ & Supporting Information). The indices for Yellow Wagtail were unchanged by this exclusion (Supporting Information).

Indices of Lapwing ($t = 7.6$, df $= 40$, $p < 0.0001$), Redshank ($t = 9.4$, df $= 45$, $p < 0.0001$) and Curlew ($t = 5.3$, df $= 35$, $p < 0.0001$) were all more positive on reserves. The mean annual trend difference represented an improvement of around 2.4% for Lapwing, 4.5% for Redshank, 1.5% for Snipe (Ouse Washes excluded) and 1.4% for Curlew. Thus, in the period 1994 - 2018 on lowland wet grassland reserves, Snipe populations increased by 36% while the benchmark counterfactual remained stable around index 1, suggesting that conservation interventions on these reserves were responsible for that increase. Curlew populations decreased by 23%

compared with a 55% decline on the benchmark counterfactual, implying a 33% improvement caused by conservation interventions on reserves. Between 1994-2018 Lapwing populations increased by 13%, but the benchmark counterfactual suggests that they would have decreased by 44% without conservation interventions, resulting in a 57% index improvement by conservation. Redshank populations on reserves increased by 51% whereas the benchmark counterfactual shows a decrease of 57% without conservation, attributing a relative improvement of 108% to conservation interventions.

Regardless of which counterfactual we compared with, we found the reserve indices to be more positive for the four wader species and similar for Yellow Wagtail (Figure 3). The difference between the Curlew reserve indices and its liberal counterfactual became less pronounced (Fig. 3; t = 2.4, df = 39, p = 0.02) than when the reserve indices were compared with the benchmark scenario, whereas the reserve indices differed more from their stringent counterfactuals for both Curlew (t = 5.1, df = 32, p < 0.0001) and Snipe (t = 10.2, df = 48, p < 0.0001).

Discussion

We used a quasi-experimental approach to demonstrate how long-term population monitoring data can be used to evaluate the impact of conservation. We found that lowland wet grassland conservation has benefitted Lapwing, Redshank and Curlew populations and, if an atypical site is excluded, that it also benefitted Snipe. We were not able to compare breeding populations of three other species (Black-tailed Godwit, Garganey and Shoveler) because they were too rare outside of nature reserves, although they showed either stable or increasing trends on reserves. Based on the benchmark counterfactual trends, Snipe (Ouse Washes excluded), Lapwing and Redshank populations all increased on reserves, but would have decreased or remained stable

14

without this conservation, while Curlew populations decreased much less on reserves than they would otherwise have done. For the four wader species, we found the reserve indices to be higher than their counterfactuals regardless of which counterfactual they were compared with, showing strong positive effects of reserve conservation in all cases. However, different counterfactuals can produce different results, here illustrated by the different counterfactual trends within each species (Figure 3). The effect of reserve conservation became less pronounced for Curlew under the liberal counterfactual, suggesting that this species may be faring slightly better in habitats other than wet grassland. Nevertheless, the differences in the three counterfactual trends for Curlew were small (Figure 3). Overall, our findings concur with others (Ausden et al. 2019; Verhulst et al. 2007) in substantiating the positive effects of conservation actions on target breeding wetland bird populations.

The target wader species in our study should theoretically benefit from lowland wet grassland conservation, but not necessarily in equal measure. European grassland-breeding waders display species-specific responses to different types of grassland conservation (Franks et al. 2018). Wetland conservation management incorporates a range of different intervention types – from the conversion of, for example, ex-arable land to grassland, to changes in hydrology and grazing/mowing regimes. The degree to which each intervention type provides suitable conditions for the different study species may therefore differ. For example, Ausden et al. (2019) suggested that limiting livestock grazing in spring, which aims to reduce trampling of waders' nests, could also reduce habitat suitability for Yellow Wagtail, because they often feed in close association with domestic livestock. While Yellow Wagtail breed in wetland habitats, it has not been a priority species until recently and has not been actively targeted by management. This species is also the only long-distance migrant among the study species, and changes on its

15

wintering grounds in Africa and migration paths may also affect its breeding population (Wood 1992, Newton 2006), thereby rendering conservation efforts in the breeding range less effective or redundant.

There are also multiple reserve specific conditions which we have not accounted for in this analysis. For example, because of improved breeding conditions, new sites recruit breeding pairs faster than older reserve sites (Supporting Information). Further research is needed to explore why reserve effects differ across study species (e.g. the declining reserve trend for Curlew in contrast to the increasing reserve trends for Redshank, Lapwing and Snipe), and in particular how population responses relate to site-specific interventions, reserve age and size, and finer-scale abiotic and habitat covariates.

We created separate reserve and counterfactual indices for each species based on the total annual number of breeding pairs. Because of the method used, a large decline on one reserve and stable or slightly increasing breeding numbers in all other reserves could still produce a decreasing trend, if the total number of breeding birds declined overall. This can potentially mask the individual reserves' conservation success, as illustrated when excluding the Ouse Washes from the analysis of Snipe populations. However, our results were largely robust to exclusion of sites with large proportions of breeding numbers (Supporting Information).

The method we used provides several benefits over other evaluation methods for conservation impacts. It allows the use of population monitoring datasets to emulate a robust ex-post study design. The interpretation of the results is intuitive (diverging lines on the graphs in Figure 2 mean that the observed scenario differs from its counterfactual), and results are easily

16

communicated to an audience without statistical knowledge. Although our method is marginally more complex than study designs such as the "After", it does not require more resources. European monitoring data, such as the BBS data, are often freely available.

This method also allows a more detailed analysis of impacts than other study designs. For example, using the "After" evaluation methodology, which examines the reserve trend after the establishment of the reserve exclusively, Redshank and Snipe would be the only species with a clear increasing trend. Assessing whether reserve conservation works exclusively based on whether a population trend is increasing implicitly assumes that the population would remain stable in the absence of conservation, which is far from the reality of ongoing population declines outside reserves (Harris et al. 2019). If the assessment had been done using a classical land-use "Control-Impact" study design, where the number of birds in each reserve would have been counted at one point in time, we would be able to compare densities but not trends. Our method (After-Control-Impact) ex-post compares trends and depicts the dynamic development of populations through time, whereas Control-Impact studies provide only a temporal snapshot. The dynamic element is advantageous as it allows identification of divergent mechanisms through time and shows visually how adding new reserves affects the overall reserve trend.

Matching is increasingly being used in combination with regression techniques to assess the effect of conservation initiatives (Terraube et al. 2020). However, matching alone does not necessarily improve effect inferences and because of reductions in sample size, may not have the same power to detect effects as regression techniques (Brazauskas & Logan 2016). The RSPB reserve and BBS datasets used in this study covered long time periods (>20 years) and

17

included breeding bird counts derived from robust study designs. Such datasets are not common, and a quasi-experimental evaluation design like ours will not necessarily be applicable or appropriate elsewhere (see Walker et al. 2018 for alternative impact evaluation using BBS monitoring data). Furthermore, for matching to be appropriate, it requires a clear theory of how the treatment changes the outcome (see Figure 1), and careful selection of matching variables and methods accordingly (Schleicher et al. 2019). Using exact matching, we were able to retain sufficiently large sample sizes to run the loglinear models for five out of eight species. Other quasi-experimental designs with fewer data or higher covariate complexity (higher number of covariates or continuous covariates) will either be impractical or require other matching methods (Iacus et al. 2019).

Reserves and BBS grids are surveyed using different survey protocols. Some of these differences could potentially lead to larger uncertainty and year-on-year variance:however, we do not believe this is the case. Each grid or site is surveyed with consistent effort each year, which means that a potential bias is also consistent and accounted for by using indices. Additionally, the counterfactuals created from the BBS are generally based on a relatively large number of annual observations. For further discussion see Supporting Information.

One way to create credible counterfactuals is through well-monitored control areas. This should reduce the likelihood of a mis-specified control group and enhance the credibility of the inference, but in order to make this possible, monitoring of control sites must be a priority, with a further emphasis on consistent survey methodology. This may be difficult for the reasons

18

described in the Introduction. The results of this study nonetheless suggest that dedicated conservation efforts have benefited target lowland wet grassland bird species and that monitoring programmes can be used to evaluate the impact of conservation interventions, by creating credible counterfactuals through matching approaches.

Acknowledgments:

Literature cited

Acreman M C., Mountford J O, Gowing DJ, Stratford C, Durrel S, Duenas M A., Rispen, E., Youngs E, & Moy IL. 2010. Restoration of wet grasslands through re-instatement of surface grips. Extended Final report to Department for Environment, Food and Rural Affairs. Centre for Ecology and Hydrology, Wallingford.

Adams V M, Barnes M, & Pressey R L. 2019. Shortfalls in Conservation Evidence: Moving from Ecological Effects of Interventions to Policy Evaluation. One Earth, **1**(1), 62–75. https://doi.org/10.1016/j.oneear.2019.08.017

Ausden M & Hirons G J M. 2002. Grassland nature reserves for breeding wading birds in England and the implications for the ESA agri-environment scheme. Biological Conservation. https://doi.org/10.1016/S0006-3207(01)00254-3

Ausden M & Bolton M. 2012. Breeding waders on wet grassland: Factors influencing habitat suitability. In R. Fuller (Ed.), Birds and Habitat: Relationships in Changing Landscapes (Ecological Reviews, pp. 278-306). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139021654.014

Ausden M, Hirons G, & White G. 2019. Wetland restoration by the RSPB – what has it achieved for birds? British Birds, **112** (June), 315–336.

Baylis K, Honey-Rosés J, Börner J, Corbera E, Ezzine-de-Blas D, Ferraro P J, Lapeyre R, Persson U M, Pfaff A & Wunder, S. 2016. Mainstreaming Impact Evaluation in Nature Conservation. Conservation Letters, **9**(1), 58–64. https://doi.org/10.1111/conl.12180

Brazauskas, R & Logan B R. 2016. Observational Studies: Matching or Regression? **25**(3), 289–313. https://doi.org/110.1016/j.bbi.2017.04.008

Boatman N D, Parry H R, Bishop J D & Cuthbertson A G S. 2007. Impacts of agricultural change on farmland biodiversity in the UK. Environmental Science and Technology **25**. https://doi.org/10.1039/9781847557650-00001

Bogart P, van der Loo M & Pannekoek, J. 2020. rtrim: Trends and Indices for Monitoring Data. R package version 2.1.1.https://CRAN.R-project.org/package=rtrim

Bull JW, Strange N, Smith R J & Gordon A. 2020, Reconciling multiple counterfactuals when evaluating biodiversity conservation impact in social-ecological systems. Conservation Biology. Accepted Author Manuscript. doi:10.1111/cobi.13570

Christie A P, Amano T, Martin P A, Shackelford G E, Simmons B I & Sutherland W J. 2019. Simple study designs in ecology produce inaccurate estimates of biodiversity responses. Journal of Applied Ecology **5**: 2742–2754. https://doi.org/10.1111/1365-2664.13499

Crick H Q P. 1992. A bird-habitat coding system for use in Britain and Ireland incorporating aspects of land-management and human activity, Bird Study **39**:1, 1-12, doi: 10.1080/00063659209477092

Colhoun K, Mawhinney K, McLaughlin M, Barnett C, McDevitt A M, Bradbury R B & Peach W. 2017. Agri-environment scheme enhances breeding populations of some priority farmland birds

in Northern Ireland. Bird Study **64**(4), 545–556.

https://doi.org/10.1080/00063657.2017.1415296

De Palma A, Sanchez-Ortiz K., Martin P A, Chadwick A, Gilbert G, Bates A. E, Börger L, Contu S, Hill S L L & Purvis A. 2018. Challenges With Inferring How Land-Use Affects Terrestrial Biodiversity: Study Design, Time, Space and Synthesis. In Advances in Ecological Research **58**(2). https://doi.org/10.1016/bs.aecr.2017.12.004

Department for Environment, Food and Rural Affairs, UK. 2019. UK Biodiversity Indicators 2019.

Eaton M A, Gibbons D W, Evans A, Gregory R D, Brown A F, Noble D G, Musgrove A J, Hearn R D & Aebischer N J. 2015. Birds of conservation concern 4: the population status of birds in the united kingdom, channel islands and isle of Man. British Birds, **108**(6), 708–746.

Eglington S M, Gill J A, Bolton M, Smart M A, Sutherland W J & Watkinson A R. 2008. Restoration of wet features for breeding waders on lowland grassland. Journal of Applied Ecology, **45**, 305–314. https://doi.org/10.1111/j.1365-2664.2007.0

Ferraro P J. 2009. Counterfactual thinking and impact evaluation in environmental policy. New Directions for Evaluation, **122**, 75- 84. https://doi.org/10.1002/ev.297

Ferraro P J, McIntosh C & Ospina M. 2007. The effectiveness of the US endangered species act: An econometric analysis using matching methods. Journal of Environmental Economics and Management, **54**(3), 245–261. https://doi.org/10.1016/j.jeem.2007.01.002

Ferraro P J & Pattanayak S K. 2006. Money for Nothing? A Call for Empirical Evaluation of Biodiversity Conservation Investments, **4**(4), 482–488. https://doi.org/10.1371/journal.pbio.0040105

Ferraro P J & Pressey R L. 2015. Measuring the difference made by conservation initiatives: Protected areas and their environmental and social impacts. Philosophical Transactions of the Royal Society B: Biological Sciences, **370**(1681), 4–8. https://doi.org/10.1098/rstb.2014.0270

Field, R H. & Gregory R D. 1999. Measuring population changes from the Breeding Bird Survey. BTO Research Report No. 217. The Nunery, Thetford, Norfolk.

Franks S E, Roodbergen M, Teunissen W, Carrington Cotton A & Pearce-Higgins J W. 2018. Evaluating the effectiveness of conservation measures for European grassland-breeding waders. Ecology and Evolution, **8**(21), 10555–10568. https://doi.org/10.1002/ece3.4532

Geldmann J, Barnes M, Coad L, Craigie I D, Hockings M & Burgess N D. 2013. Effectiveness of terrestrial protected areas in reducing habitat loss and population declines. Biological Conservation, 161, 230–238. https://doi.org/10.1016/j.biocon.2013.02.018

Geldmann J, Manica A, Burgess N D, Coad L & Balmford A. 2019. A global-level assessment of the effectiveness of protected areas at resisting anthropogenic pressures. Proceedings of the National Academy of Sciences of the United States of America, **116**(46), 23209–23215. https://doi.org/10.1073/pnas.1908221116

Gilbert G, Gibbons D W, Evans J. 1998. Bird Monitoring Methods. RSPB, Sandy.

Gregory R, Baillie S, Bashford R. 2000. Monitoring breeding birds in the United Kingdom. Bird Census News **13**, 101–112.

Harris S J, Massimino D, Eaton M A, Gillings S, Noble D G, Balmer D E, Pearce-Higgins J W & Woodcock P. 2019. The Breeding Bird Survey 2018. BTO Research Report 717. British Trust for Ornithology, Thetford.

Hayhow D B, et al. 2017. The state of the UK's birds 2017. The RSPB, BTO, WWT, DAERA, JNCC, NE and NRW, Sandy, Bedfordshire

Hayhow D B, et al. 2019. The State of Nature 2019. The State of Nature partnership.

Hernan M A, Robins J M. 2020. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.

Hoffmann M, Duckworth J W, Holmes K, Mallon D P, Rodrigues A S L & Stuart S N. 2015. The difference conservation makes to extinction risk of the world's ungulates. Conservation Biology, **29**(5), 1303–1313. https://doi.org/10.1111/cobi.12519

Hoffmann M, et al. 2010. The impact of conservation on the status of the world's vertebrates. Science, **330**(6010), 1503–1509. https://doi.org/10.1126/science.1194442

Iacus S M, King G & Porro G. 2019. A theory of statistical inference for matching methods in causal research. Polit. Anal. **27**, 46–68.

IPBES (2019): Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. S Díaz, et al (eds.). IPBES secretariat, Bonn, Germany. 56 pages.

Jefferson R G & Grice PV. 1998. The conservation of lowland wet grassland in England. In C.B. Joyce & P.M. Wade (eds) European Wet Grasslands: Biodiversity, Management and Restoration: 31– 48. Chichester: Wiley

Joppa L & Pfaff A. 2010. Reassessing the forest impacts of protection. Annals of the New York Academy of Sciences, **1185**(1), 135–149. https://doi.org/10.1111/j.1749-6632.2009.05162.x

Malpas L R, Kennerley R J, Hirons G J M, Sheldon R D, Ausden M, Gilbert J C. & Smart J. 2013. The use of predator-exclusion fencing as a management tool improves the breeding success of waders on lowland wet grassland. J. Nat. Conserv. **21**, 37–47 (2013). https://doi.org/10.1016/j.jnc.2012.09.002

Margoluis R, Stem C, Salafsky N & Brown M. 2009. Design alternatives for evaluating the impact of conservation projects. New Directions for Evaluation, **122**, 85–96. https://doi.org/10.1002/ev.298

Newton I. 2006. Can conditions experienced during migration limit population levels of birds. J Ornithol. **147**, 146-166. Doi: 10.1007/s10336-006-0058-4.

Pannekoek J, Bogaart P, van der Loo M. 2018. Models and statistical methods in rtrim. Statistic Netherlands, Haga

Pearl J. 2010. 3. The Foundations of Causal Inference. Sociological Methodology, **40**(1), 75–149. https://doi.org/10.1111/j.1467-9531.2010.01228.x

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Pearl J & Mackenzie D. 2018. The Book of Why. Basic Books, New York.

Ratcliffe N, Schmitt S & Whiffin M. 2005. Sink or swim? Viability of a black-tailed godwit population in relation to flooding. Journal of Applied Ecology, **42**(5), 834–843. https://doi.org/10.1111/j.1365-2664.2005.01076.x

RSPB. 2010. Which species need help? Royal Society for the Protection of Birds, United Kingdom. Available from https://www.rspb.org.uk/our-work/conservation/conservation-and-sustainability/safeguarding-species/which-need-help/ (accessed April 2020)

Rubin R D. 1980. Discussion of "Randomization Analysis of Experimental Data in the Fisher Randomization Test," by D. Basu, Journal of the American Statistical Association, **75**, 591–593

Schleicher J, Eklund J, Barnes M, Geldmann J, Oldekop J A & Jones J P G. 2019. Statistical matching for conservation science. Conservation Biology, **34**, 538-549. doi:10.1111/cobi.13448

Sills E O, de Sassi C, Jagger P, Lawlor K, Miteva D A, Pattanayak S K and Sunderlin W D. 2017. Building the evidence base for REDD+: study design and methods for evaluating the impacts of conservation interventions on local well-being Glob. Environ. Change **43**, 148–60

Rubin DB: Randomization analysis of experimental data: the fisher randomization test. J Am Stat Assoc 1980, 75:591-593

Smart J, Wotton S R, Dillon I A, Cooke A I, Diack I, Drewitt A L, Grice P V & Gregory R D. 2014. Synergies between site protection and agri-environment schemes for the conservation of waders on lowland wet grasslands. Ibis, **156**(3), 576–590. https://doi.org/10.1111/ibi.12153

Smart M & Coutts K. 2004. Footdrain management to enhance habitat for breeding waders on lowland wet grassland at Buckenham and Cantley Marshes, Mid-Yare RSPB Reserve. Conservation Evidence, **1**, 16–19.

Sonter LJ, Barnes M, Matthews JW, Maron M. 2019. Quantifying habitat losses and gains made by U.S. Species Conservation Banks to improve compensation policies and avoid perverse outcomes. Conservation Letters:1–9.

Stuart E A. 2010. Matching Methods for Causal Inference: A Review and a Look Forward. Statistical Science, **25**(1), 1–21. https://doi.org/10.1214/09-sts313

Terraube J, Van doninck J, Helle P, Cabeza M. 2020. Assessing the effectiveness of a national protected area network for carnivore conservation. Nature Communications **11**:1–9. http://dx.doi.org/10.1038/s41467-020-16792-7.

Textor J, van der Zander B, Gilthorpe M S, Liśkiewicz M & Ellison G T. 2016. Robust causal inference using directed acyclic graphs: The R package "dagitty." International Journal of Epidemiology, **45**(6), 1887–1894. https://doi.org/10.1093/ije/dyw341

VanderWeele T J. 2019. Principles of confounder selection. European Journal of Epidemiology, **34**(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

Verhulst J, Kleijn D & Berendse F. 2007. Direct and indirect effects of the most widely implemented Dutch agri-environment schemes on breeding waders. Journal of Applied Ecology, **44**, 70-80

Walker L K, Morris A J, Cristinacce A, Dadam D, Grice P V & Peach W J. 2018. Effects of higher-tier agri-environment scheme on the abundance of priority farmland birds. Animal Conservation, **21**(3), 183–192. https://doi.org/10.1111/acv.12386

Wickham et al. (2019). Welcome to the tidyverse. Journal of Open Source Software, **4**(43), 1686, https://doi.org/10.21105/joss.01686

Wiik E, d'Annunzio R, Pynegar E, Crespo D, Asquith N, Jones J P G. 2019. Experimental evaluation of the impact of a payment for environmental services program on deforestation. Conservation Science and Practice **1**:1–11.

Wilson A M, Ausden M & Milsom T P. 2004. Changes in breeding wader populations on lowland wet grasslands in England and Wales: Causes and potential solutions. Ibis, **146**(suppl. 2), 32–40. https://doi.org/10.1111/j.1474-919X.2004.00371.x

Wilson A, Pendlebury C & Vickery J. 2005. Changes in Lowland Wet Grassland Breeding Wader Number: The Influence of Site Designation. BTO Research Report 365. British Trust for Ornithology, Thetford.

Wood B. 1992. Yellow Wagtail *Motacilla flava* migration from West Africa to Europe: pointers towards a conservation strategy for migrants on passage. Ibis, **134** (suppl 1), 66–76.

Table 1 The variables used in the creation of the benchmark, liberal and stringent counterfactuals.

Benchmark counterfactual	'Liberal' counterfactual	'Stringent' counterfactual
• Altitude < 250m • From 1994-2018 • From the UK • Contains the target species • Grids* containing semi-natural grassland / marsh (chalk downland, grass moor, grass moor mixed with heather, machair other dry grassland, water-meadow/grazing marsh, reed swamp, other open marsh or saltmarsh)	• From 1994-2018 • From the UK • Contains the target species	• Altitude < 250m • From 1994-2018 • From the UK • Contains the target species • Grids* containing semi-natural grassland types more similar to wet grassland (dry grasslands, water meadows/grazing marsh, reed swamp or other open marsh).

* Using habitat data from the Breeding Bird survey. Habitat data were recorded by volunteer surveyors using a standardized habitat recording form, allowing each survey transect two primary habitat types. For more information see Supporting Information.

Table 2 The number of lowland wet grassland reserve sites and Breeding Bird Survey (BBS) grids used to create the reserve and counterfactual indices for each of the study species.

	Reserve sites Number of sites	Benchmark counterfactual Number of BBS grids	Liberal counterfactual Number of BBS grids	Stringent counterfactual Number of BBS grids
Curlew	23	371	2477	267
Lapwing	97	487	3223	380
Redshank	87	140	589	108
Snipe	56	227	1212	147
Yellow wagtail	29	105	1019	82
Total number of sites/grids	101	1377	6507	1071

Figure 1 A Directed Acyclic Graph showing the hypothesized effect of lowland wet grassland conservation on breeding bird trends. Lowland wet grassland conservation is a cause of change in habitat suitability (Habitat, Hydrology, Food availability and Predator pressure (Smart & Coutts, 2004; Verhulst et al. 2007; Eglington et al. 2008; Acreman et al. 2010; Ausden & Bolton, 2012; Smart et al. 2014;) which then causes a change in the breeding trend. Suitability is improved by converting or forming the habitat from other habitat types to grassland, by changing the hydrological conditions using water control structures and land forming, by maintaining a suitable sward through grazing by domestic livestock and mowing; mechanical removal of shrubs and trees to remove perches for avian predators; reducing the impact of predation by controlling or excluding generalist predators. The four yellow circles (Time-

variant factors, Species, Country-variant factors and Habitat potential) represent confounding factors which may affect reserve selection and the breeding trends.

Figure 2 Breeding trends from 1994 - 2018 for the five target species inside reserves (Solid line) and the benchmark counterfactual trends (Dashed line). The Ouse washes reserve was excluded for Snipe. Indices were calculated using imputed counts from loglinear models (as described in the method section). The shaded area delineates the standard error.

Figure 3 Reserve and counterfactual breeding trends from 1994 - 2018 using the liberal, benchmark and the stringent matching settings as in earlier analysis. The Ouse washes reserve was excluded for Snipe. Indices were calculated using imputed counts from loglinear models (as described in the method section). To ease visualisation, the standard errors have been excluded (See Supporting Information for figure with SE).
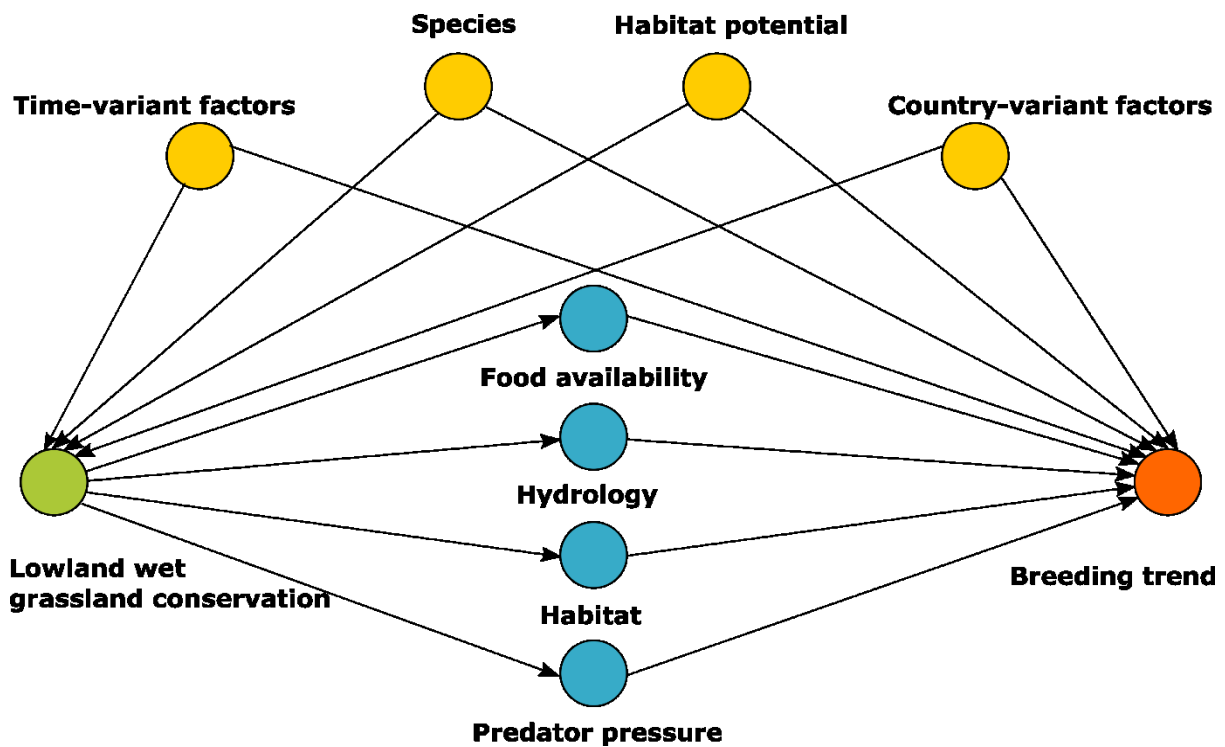
Figure 1 A Directed Acyclic Graph showing the hypothesized effect of lowland wet grassland conservation on breeding bird trends. Lowland wet grassland conservation is a cause of change in habitat suitability (Habitat, Hydrology, Food availability and Predator pressure (Smart & Coutts, 2004; Verhulst et al. 2007; Eglington et al. 2008; Acreman et al. 2010; Ausden & Bolton, 2012; Smart et al. 2014;) which then causes a change in the breeding trend. Suitability is improved by converting or forming the habitat from other habitat types to grassland, by changing the hydrological conditions using water control structures and land forming, by maintaining a suitable sward through grazing by domestic livestock and mowing; mechanical removal of shrubs and trees to remove perches for avian predators; reducing the impact of predation by controlling or excluding generalist predators. The four yellow circles (Time-variant factors, Species, Country-variant factors and Habitat potential) represent confounding factors which may affect reserve selection and the breeding trends.
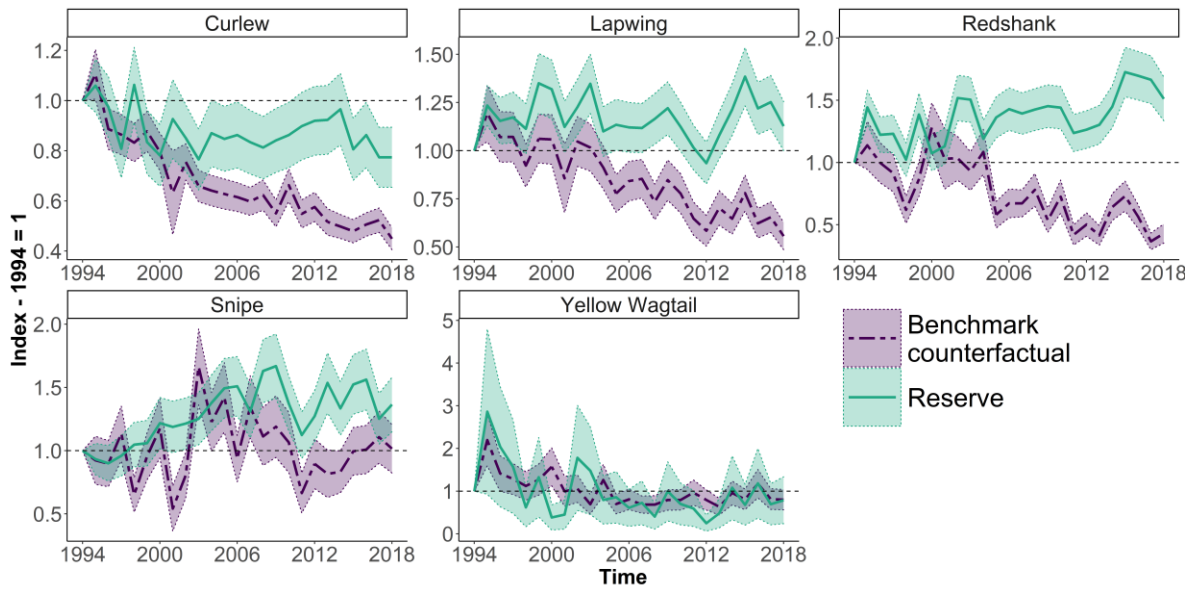
Figure 2 Breeding trends from 1994 - 2018 for the five target species inside reserves (Solid line) and the benchmark counterfactual trends (Dashed line). The Ouse washes reserve was excluded for Snipe. Indices were calculated using imputed counts from loglinear models (as described in the method section). The shaded area delineates the standard error.
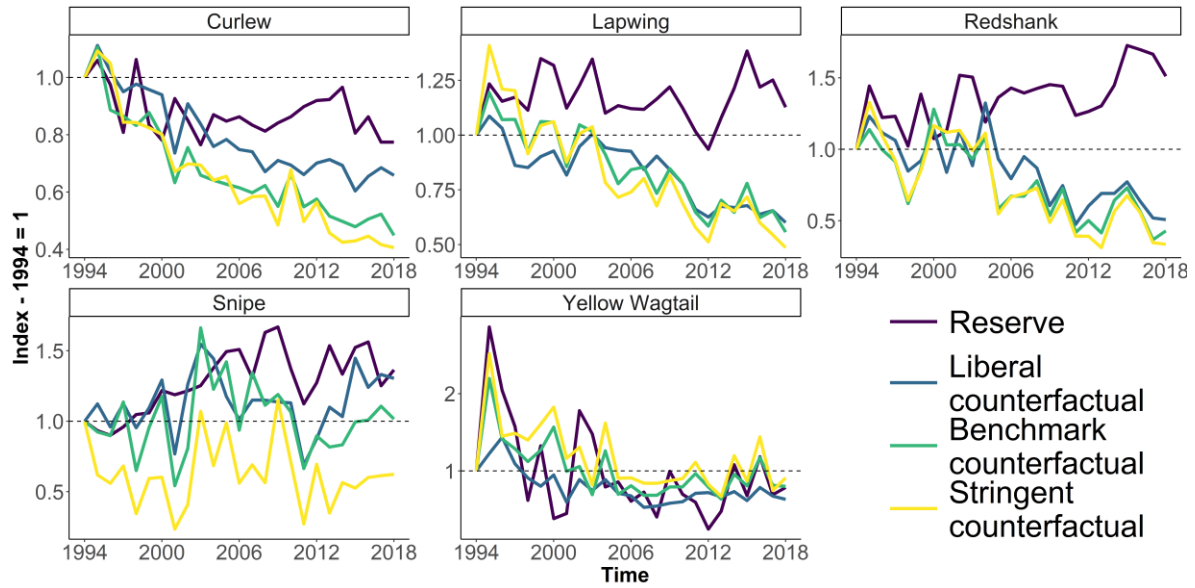
Figure 3 Reserve and counterfactual breeding trends from 1994 - 2018 using the liberal,

benchmark and the stringent matching settings as in earlier analysis. The Ouse washes reserve

was excluded for Snipe. Indices were calculated using imputed counts

from loglinear models (as described in the method section). To ease visualisation,

the standard errors have been excluded (See Supporting Information for figure with SE).