A response to Butterworth: FISHERIES/2020/AUG/SWG-PEL/82

Richard B. Sherley[1,2]

1. Centre for Ecology and Conservation, College of Life and Environmental Sciences, University of Exeter, Penryn Campus, Cornwall, TR10 9FE, United Kingdom.
2. FitzPatrick Institute of African Ornithology, DST-NRF Centre of Excellence, University of Cape Town, Rondebosch 7701, South Africa.

Email: r.sherley@exeter.ac.uk

Note: For readers' ease, responses have, in the main, been inserted at appropriate points in the original document below in blue. Italics are used for direct quotes. The original figures and tables from FISHERIES/2020/JUL/SWG-PEL/53REV that aren't directly referenced in my replies are omitted here to keep the file size small.

Primary overview comments

Sherley's document below, as it states, provides a response (in commendable detail) to some suggestions made by the 2019 International Review Panel regarding the selection of random effects structures for models to estimate the closure effect from the island closure experiment which Sherley and colleagues have submitted previously. That goes to the question of how best such models might remove the effects of non-independence (or pseudo-replication) in the individual measurement data they use to prevent their providing negatively biased estimates of the standard errors of these closure effects.

However, the document fails to address the more basic question of whether, even if perhaps such removal may be achieved, the use of such individual data can provide improved (lower standard error) estimates of such precision compared to those based on annually aggregated values of the corresponding response variables. This is the issue raised, for example, in the last section of FISHERIES/2020/JAN/SWG-PEL/08, where a limiting case example is used to suggest that this may not be so.

The Annex added to this document provides a mathematical-statistical demonstration that ***this is indeed not so****. Thus even if the random effects approach to making use of individual data can fully account of their non-independence, and hence prevent this from negatively biasing estimates of the standard error of the island closure effect, the resultant estimates could not have better precision than those provided by corresponding models based on annually aggregated values of measurements of the response variables.*

The underlying reason for this is the absence of any inter-year linkage of the data sources available to provide the response variables in these analyses. Sherley confirmed at the SWG meeting on 30 July when this document was discussed that in all the instances examined, there was no such connection: for chick condition and for chick survival there is no linkage between parents or nests from one year to the next, and similarly for foraging length (maximum distance travelled) the birds used to obtain the data are not linked inter-annually. This means that the observations from one year to the next are independent, which in turn leads to a diagonal structure in terms of annual sub-matrix blocks in the variance-covariance matrix and its inverse which are used for the closure effect estimation, and that in turn leads to the key result of the Annex.

Explained less formally:

a) Unlike the case of individual linkage, which for example enables a paired-t test to have more discriminatory power than a comparison based on means only, in this *instance annual means contain all the information content of pertinence to the key effect being estimated.*

b) Estimation of the island closure effect is one relating to inter-year (not intra-year) variation, so that use of individual data in the estimation is unable to improve estimation precision for this parameter.

At the 30 July meeting, Sherley stated that in his view the estimates from analyses based on annually aggregated data (such as the approach developed in collaboration with the International Panel, and (re-)implemented in FISHERIES/2020/JAN/SWG-PEL/09) should not be considered (or words to similar effect) because the ratio of degrees of freedom to number of parameter values estimated is too small. The analyses in the Annex show that the individual based approach (applied in such a way that there is adequate correction for non-independence/pseudo-replication effects) cannot improve on this precision. From this it therefore follows that Sherley's results below should also not be considered. I would not, however, agree that this is an adequate reason not the consider the results of FISHERIES/2020/JAN/SWG-PEL/09. Admittedly the nature of the island closure experiment is such that the number of degrees of freedom is limited, with effectively each extra year adding two more data points and one additional (year-effect) estimable parameter only, and hence only one further degree of freedom. Nevertheless, the results for standard errors of the closure effect estimates in Table 2 of FISHERIES/2020/JAN/SWG-PEL/09 do show that meaningful results may be obtained from the annually aggregated data that have become available from this experiment.

R1. It is good to finally see the acknowledgement here and in FISHERIES/2020/JAN/SWG-PEL/08 (Butterworth 2020a) that the models used in FISHERIES/2020/JAN/SWG-PEL/09 (Ross-Gillespie and Butterworth 2020a) and similar documents outlining the MARAM approach are limited in the number of degrees of freedom relative to the number of parameters estimated – and thus essentially in their statistical power (see also Sherley 2020a, FISHERIES/2020/AUG/SWG-PEL/83).

Indeed, it has long been Butterworth's contention that using the annually aggregated means and then fitting a statistical model to those means ("statistics on statistics", or a two-stage approach) is the best approach available because it is standard practise in fisheries management, for example when incorporating CPUE data into stock assessment models using GLM standardisation. However, Mark Maunder, head of the Stock Assessment Program at the Inter-American Tropical Tuna Commission (IATTC), suggested back in 1998 that such a two-step procedure "*has a number of disadvantages, which include: (i) information is lost in the standardization process, (ii) assumptions in the standardization procedure are often inconsistent with those of the population-dynamics model, (iii) the error structure of the standardized index is difficult to determine, (iv) uncertainty may not be adequately transferred from the standardization procedure into the fitting of the population-dynamics model, and (v) the separation of the analyses may reduce the ability to diagnose any lack of fit*" (Maunder 2001). And Maunder (2001) also showed, using a simulation analysis, that an integrated approach (one-step) leads to narrower confidence intervals that more accurately represent the uncertainty associated with the parameter estimates (Maunder 2001). So, it would seem that it has been known for over a decade in fisheries management that fitting to the observations directly and not some aggregated index is preferable when possible.

Ecologists have known this too for more than a decade (Besbeas and Freeman 2006). It is also clear from a number of disciplines that the use of parsimonious mixed models containing random effect supported by the data improves the balance between a Type I error and statistical power (e.g. Matuschek et al. 2017, Bates et al. 2018, Silk et al. 2020), and so allows exactly this kind of one-step approach in FISHERIES/2020/JUL/SWG-PEL/53REV (Sherley 2020b). Additionally, as I pointed out already back in 2016

(MARAM/IWS/DEC16/Peng Clos/P4, Sherley 2016a), mixed-effects models have even been advocated for and used in fisheries management for more than a decade (Venables & Dichmont 2004, Punt et al. 2006, Thorson & Minto 2015, Thorson et al. 2016), including by Butterworth himself (Brandão et al. 2004).

The individual-data-based approach can thus at best equal the aggregated data approach in terms of precision of the estimates of island closure effects, but only provided that there is complete adjustment for the non-independence effects through the use of random effects models. This requires first that the random effects structure is appropriately chosen, as the document below addresses; but even if that is the case, as the 2019 Panel stated, in natural experiments such as the island closure experiment, it remains only "a working hypothesis that including random effects chosen using model selection methods will appropriately account for the pseudo-replication". Thus, even if best practice is used to select the random effects structure, this provides no guarantee that the closure effect standard error estimates arising will not be negatively biased, and to an unknown extent. Hence, why consider the results from these models, when the aggregated approach already accounts for within-year data non-independence without raising this concern?

Overview summary

The individual-data-based estimates of the closure effect are indicated by the Annex to be unable in principle to provide any improvement on annually aggregated analyses. To the extent that they might appear to do so, no guarantee can be provided that this appearance is not a consequence of a failure of the random effects approach used to account for all sources of non-independence in the data.

Why therefore proceed further with any comments on the document below, since its results cannot be used as a basis for decisions regarding the implications of the island closure experiments? This is only because other useful discussion points arise co-incidentally from this text, which may well assist in further analyses (e.g. standardisation of annual aggregate measurements for co-variates), consideration and interpretation of results for this experiment.

R2. As outlined in FISHERIES/2020/AUG/SWG-PEL/83, Butterworth seems to have directly conflicting issues with the analysis presented in FISHERIES/2020/JUL/SWG-PEL/53REV. On the one hand, he says that the results cannot be used as a basis for decisions because they might produce negatively biased estimates of the standard errors of closure effect compared to the aggregated data approach (i.e. the standard errors are smaller than obtained from the aggregated data approach). On the other hand, he also contends that the results presented in FISHERIES/2020/JUL/SWG-PEL/53REV should not be considered further because that document has NOT shown that they can provide improved estimates of such precision (i.e. smaller standard errors than those obtained from the aggregated data approach).

The two cannot both be true.

It is worth noting, once again, that the 2019 panel recommended: "*Given the nature of the experiment, use of individual data is to be preferred. However, this is only the case if an appropriate random effects structure is chosen*" [my emphasis]. And Butterworth admits above that FISHERIES/2020/JUL/SWG-PEL/53REV did indeed provide results with appropriate random effects structure: *"This requires first that the random effects structure is appropriately chosen, as the document below addresses".*

Which of the following two options do we adopt?

Is FISHERIES/2020/JUL/SWG-PEL/53REV to be disregarded because Butterworth feels it cannot "*guarantee that the closure effect standard error estimates arising will not be negatively* *biased*" – in other words, because he maintains that it DOES yield smaller standard errors than those obtained from the aggregated data approach?

Or are we expected to accept the complete opposite contention that FISHERIES/2020/JUL/SWG-PEL/53REV should be disregarded because its approach DOES NOT produce smaller standard errors than those obtained from the aggregated data approach?

Given the above contradictions, and since the 2019 panel stated that "*results presented to the Workshop suggest that estimates of closure parameters using models fitted to aggregated and individual data had similar standard errors*" and "*Given the nature of the experiment, use of individual data is to be preferred"* (Die et al. 2019), **I repeat my suggestion that we seek guidance from the panel on whether FISHERIES/2020/JUL/SWG-PEL/53REV and MARAM/IWS/2019/PENG/P4 can be used *as a basis for decisions.***

I also repeat the key conclusion of FISHERIES/2020/AUG/SWG-PEL/83 and FISHERIES/2020/JUL/SWG-PEL/53REV, that we focus on the fact that "we have now iterated to a place where two independent sets of analyses agree that biologically meaningful effects of fishing around African penguin breeding colonies are apparent and, importantly, that some of those effects are on variables (chick survival, fledging success) that contribute to the demographic process". See also Table 1 in FISHERIES/2020/SEP/SWG-PEL/87 (Sherley 2020c), which further supports this point.

1. Introduction

The Panel at the 2019 International Stock Assessment Workshop made the following key recommendations regarding the work outlined in MARAM/IWS/2019/PENG/P4 (hereafter Sherley et al. 2019).

- Model selection methods should be applied to select an appropriate random effects structure;
- Use of a Bayesian fitting process combined with the Widely Applicable Information Criterion (WAIC) comparisons should be used to check model selection;
- The set of covariates to consider in the analysis should be identified by the relevant DEA working group. This requires an understanding of how the individual data are collected;
- Best practise for fitting mixed effects models (e.g., Zuur et al., 2009) should be followed (if this is not the case already). This should include standard residual analysis as well as residual analyses that are tailored to the problem at hand (e.g., temporal, spatial or within-season plots of residuals).

Here, I have refit the three key models in Sherley et al. (2019), namely Eastern Cape Maximum Distance, Western Cape Chick Condition and Western Cape Chick Survival, using model selection and followed the guidelines in Zuur et al. (2009) to choose the best fitting random effect structure. I then present a series of residual analysis for each best fitting model, using standardized residuals for models with Gaussian errors and Pearson residuals for models with Gamma errors (Zuur et al. 2009, 2013).

A potentially important aspect associated with the application of random effects models is whether the quantities assumed to be random are indeed so, and do not exhibit systemic patterns indicative of model mis-specification. It is unclear whether the diagnostic plots presentation in, for example, Figure A4.10 are addressing this matter, and if so how.

R3. The diagnostic plots presented are those asked for by the 2019 panel: "*This should include standard residual analysis as well as residual analyses that are tailored to the problem at hand (e.g., temporal, spatial or within-season plots of residuals)*" (Die et al. 2019). The residual plots presented show no evidence of deviation from the assumptions underlying the use of the relevant mixed effects models. Figure A4.10 is a posterior predictive check (not a residual plot) and is not used to diagnose whether quantities are random. As we were informed at the 2016 IWS when a plot of this nature was first presented to the panel in MARAM/IWS/DEC16/PENG CLOS/WP4 (Sherley 2016b), the expected distribution in this case is uniform.

I have retained the fixed effects used in Sherley et al. (2019) because these were selected and/or approved by seabird biologists with understanding of how the individual data were collected and that sit on the most relevant DEA working group (the Seabird Task Team of the Top Predator Working Group).

The Panel also suggested developing a simple guide to mixed effects models to broaden the group of scientists capable of discussing these models meaningfully (Die et al. 2019). Although many good simple guides exist (e.g. Harrison et al. 2018), I will attempt to go some way to meeting this Panel recommendation here. Mixed effects model contains two components: a fixed effect component (the explanatory variables) and the random effects. In most cases, the variables we are interested in the effects of are the fixed effects (e.g. the closure effect in our case). But there may also be sources of variation that we wish to control for when we estimate the fixed effects, but that would require a large number of parameters to estimate as fixed effects themselves. For example, the year effect in our case would require 11 parameter estimates if included as a fixed effect. This can lead to models that are over-parameterised (e.g. Robinson and Butterworth 2014). Zuur et al., (2009) provide the example of such a model that has 45 data points and estimates 17 model parameters, saying "the number of parameters used by this model is excessively high". Over-parameterisation can lead to inflated standard errors for parameter estimates (compromising statistical power). Instead, to save on model parameters, we can use random effects.

While a high ratio of data points to estimable parameters is indeed desirable, the nature of the island closure experiment sets a low upper bound to this ratio. As indicated in the Annex, and explained further in the Primary overview comments above, the use of random effects approaches with individual data is unable to ameliorate this problem in estimating the island closure effect.

R4. The 2019 panel would appear to disagree: "*Given the nature of the experiment, use of individual data is to be preferred*" (Die et al. 2019).

To select the appropriate random effect structure, Zuur et al. (2009) recommend a top-down strategy (pp. 121–122) that begins with a model that contains all the fixed components of interest and as many interactions as possible, or this is impractical e.g. due to a large number of explanatory variables or interaction, a selection of covariates that you think are most likely to contribute to the optimal model. For the analysis here, I retained the fixed effect structures from Sherley et al. (2019). This means I retain an 'Island' × 'Closure' interaction in the fixed effect component as this leaves open the possibility of island-specific closure (or fishing) effects, an approach that has been almost unanimously applied for some time (e.g. Robinson 2013, Robinson & Butterworth 2014, Hagen et al. 2014, Sherley et al. 2018). Using these fixed effect model structures, I then find the optimal structure of the random component. As Zuur et al. (2009) put it, "Because we have as many explanatory variables as possible in the fixed component, the random component (hopefully) does not contain any information that we would like to have in the fixed component". In other words, effects we are interested in directly should be in the fixed component, those we only wish to account for should be random effects. In the context of the Island Closures experiment, this means that Island would be best placed in the fixed (not random) component of the model (plus, it only used 1 additional degree of freedom in the fixed component. Nevertheless, since models containing Island in the random component were specifically asked for by one participant at the 2019 International Stock Assessment Workshop, I have included some here (but see methods section).

Below I present the results of models based on a Bayesian fitting process and provide the Widely Applicable Information Criterion (WAIC; Watanabe 2010) for each model, as recommended by Die et al. (2019). However, since leave-one-out (LOO) cross-validation is now generally recommended over WAIC for Bayesian model selection (Gelman et al. 2014, Piironen & Vehtari 2017, Vehtari et al. 2017), I base my model selection on Pareto smoothed importance sampling (PSIS) LOO cross-validation (PSIS–LOO; Vehtari et al. 2019a) and provide model averaged results based on stacking of predictive distributions (Yao et al. 2018).

2. Methods

Datasets and the basic analytical approach remain the same as in Sherley et al. (2019), with the exception that the chick survival models (section 3.3) which now use data from 2008–2018 and are now implemented using a log-normal hazard function rather than an exponential hazard function (initial analysis indicated this provided a slightly better model fit, see Appendix 1 for details).

In this last respect, see the comments entered underneath Figure A4.13 below.

R5. See response there also.

All models were implemented using Markov Chain Monte Carlo (MCMC) estimation in JAGS (v. 4.3.0; Plummer 2003), with all priors as specified in Sherley et al. (2019). The fixed component of each model is specified in the results section along with the various random-effects structures tested. To implement the WAIC and PSIS–LOO model selection I fit each candidate model using the *jags* function in the 'jagsUI' library (version 1.5.0; Kellner 2018) for R (version 3.5.2; R Core Team 2019) and traced the log-likelihood for each observation using the log density distributions implemented in JAGS (Plummer 2017). I then pass the posterior samples for the log-likelihood for each model to the *waic* and *loo* functions in the 'loo' library (version 2.2.0; Vehtari et al. 2019b) for R to calculate the WAIC and PSIS–LOO value respectively for each model. Finally, stacking weights (Yao et al. 2018) for all the candidate models were calculated by passing the expected log pointwise predictive density (elpd) values for each model observation (the pointwise output from the *loo* function) to the *stacking_weights* function in the 'loo' library (see Vehtari & Gabry 2019). All JAGS models were fit using 3 MCMC chains of 120 000 iterations each, with the first 20 000 iterations discarded as burn-in and a thinning rate of 10, leaving 30 000 samples for inference. Unless otherwise specified, we present means and 95% highest posterior density intervals (HPDI) as the credible intervals. Convergence of all models was checked visually and using Gelman–Rubin diagnostics. All models unambiguously converged (all $\hat{R}$ values ≤ 1.001).

Lastly, we used the best fitting models for each dataset to convert all effect sizes to a percentage effect following the approach in Sherley et al. (2019). We then we combine these six percentage effect posteriors with the four from the datasets that have not been updated here (two islands for Eastern Cape Chick Condition, two islands for Western Cape Maximum Foraging Distance), but that featured in Sherley et al. (2019). We then recalculate the Overall Closure Effect (%) presented in Sherley et al. (2019) based on the updated results presented here. This approach follows that developed recently to combine regional % declines into a weighted global change % for IUCN Red List assessments (Sherley et al. 2020). Here all posterior samples are of length 30 000, so each data set is automatically given equal weighting. We plot this combined distribution, and calculate the mean, median, 95% credible intervals and the percentages of the Overall Closure Effect posterior that are above and below zero and the pre-identified 10% threshold for management action (Cochrane 2016).

Note that models with ‘Island’ alone as the only random effect were omitted because such a model structure effectively accounts for the exact same source of variation twice (in both the fixed and random component) and, because there are only two islands (two levels of the effect), the models cannot estimate the standard deviation for the Island random effect (because you cannot reliably estimate a variance from two observations). The result is that the models cannot partition the variance between the fixed and random effects, and either would not converge (after 500 000 iterations) or produced biologically impossible estimates for the Island fixed effect (e.g. chick condition estimates spanning the positive to negative 100s, Appendix 2). I noted this potential issue with having ‘Island’ in the random effect during the discussions at the 2019 International Stock Assessment Workshop.

3. Results and Discussion

3.1. Chick Condition, Western Cape

The fixed effect structure used for these models had the following components: ‘Island’ main effect, ‘Closure’ main effect, ‘Island’ × ‘Closure’ interaction, a sardine biomass main effect and an anchovy biomass main effect. The random effects structures tested are shown in Table 1, along with their corresponding WAIC, PSIS–LOO scores and stacking weights. All models are random intercept models.

The best fitting model contained the random ‘Island/Year/Month’ intercept and yielded positive point estimates for the closure effect at Robben Island of 22.9% (HPDI: −4.5–51.1%; mean effect size = 0.07, 95% HPDI: −0.010–0.14) and at Dassen Island of 13.1% (HPDI: −12.4–39.0%; mean effect size = 0.03, 95% HPDI: −0.03–0.10). Although neither effect was credibly different from zero based on the 95% HPDI range, 95.6% of all the posterior estimates for the closure effect were > 0 at Robben Island (down from 99.99% in Sherley et al. 2019). Although not directly comparable to a frequentist p-value, this could be thought of as corresponding to a p-value of 0.044 on a one-tailed test (i.e. assuming *a priori* that fishing cannot improve chick condition). In addition, 81.7% of the posterior estimates exceeded a 10% effect size. Although this is down from 99.8% in Sherley et al. (2019), there is still strong evidence (> 80% probability) from this model for closure effect at Robben Island that exceeds the 10% threshold (Cochrane 2016). Moreover, the evidence for a closure effect at Dassen Island is greater based on M1 (Table 1) than it was in Sherley et al. (2019), with 84.6% of all the posterior estimates being > 0 (up from 45.5% in Sherley et al. 2019) and 57.2% exceeding a 10% effect size (up from 9.4% in Sherley et al. 2019).

Table 1. Model selection results for the candidate models with different random effect structures, tested to assess the impact of the fishing closures on African penguin chick condition at Robben and Dassen Islands. M3 (Year/Month) corresponds to the original model presented in Sherley et al. (2019). Effect sizes marked in bold text are credibly different from zero (≥ 97.5% of the posterior > 0). Models are ranked by PSIS–LOO value (the *smaller* the PSIS–LOO, the better the relative model fit).

Model Number	Random effects structure	WAIC	PSIS–LOO	Stacking weight	Robben Closure effect mean (95% HPDI)	Dassen Closure effect mean (95% HPDI)
M1	Island/Year/Month	10365.9	10366.2	0.946	0.07 (−0.01–0.14)	0.03 (−0.03–0.10)
M3	Year/Month	10680.5	10680.7	0.022	**0.10** **(0.05–0.14)**	−0.002 (−0.05–0.04)
M4	Island/Month	11348.0	11348.0	0.002	**0.10** **(0.08–0.12)**	0.01 (−0.02–0.03)
M6	Month	11449.9	11449.9	0.019	**0.10** **(0.08–0.12)**	−0.002 (−0.03–0.02)

M2	Island/Year	11499.6	11499.6	0.000	0.08 (−0.01–0.16)	0.02 (−0.07–0.10)
M5	Year	11582.6	11582.6	0.012	**0.11** **(0.06–0.15)**	0.01 (−0.03–0.06)
	Model-averaged results				0.07 (−0.01–0.14)	0.03 (−0.03–0.10)

Notes: / denotes nesting of the random effects, thus Island/Year/BirdID = Month nested in Year, nested in Bird Identity. WAIC = Widely Applicable Information Criterion (Watanabe 2010). PSIS–LOO = Pareto smoothed importance sampling, leave-one-out cross-validation (PSIS–LOO; Vehtari et al. 2019a). HPDI = highest posterior density interval.

It is of importance to note that when Island/Year is included in the random effects structure, the 95% HPDI widens appreciably, and to the extent that for Robben Island it is no longer credibly different from zero. This corroborates the concern expressed in FISHERIES/2020/JAN/SWG-PEL/08 that failure to include this interaction was leading to negatively biased estimates of standard errors (i.e. unduly high precision) for the estimates of island closure effects. It should be noted that some earlier analyses of this nature have also not included this interaction in their random effects structure, e.g. MARAM/IWS/DEC19/Peng/P4 included only a Year/Month interaction term; this renders their conclusions questionable.

R6. Butterworth here confuses an interaction (applied to fixed effects) and a nested random effect. Nevertheless, there is still the question of whether Island should simultaneously be included in both the fixed and random components of these models and whether Island, which only has two levels, should be included in the random effect structure at all.

Again, I had requested that the SWG-PEL go back to the 2019 to clarify this, but this request was refuted on procedural grounds. I have sought to clarify this myself (as agreed by the SWG-PEL) with Andre Punt, who said in response to seeing an earlier version of FISHERIES/2020/JUL/SWG-PEL/53REV, "*No problem with not using island as a random effect – I hope that was not my suggestion originally*". **I repeat that request again; the SWG-PEL should officially go back to the 2019 panel and request clarity on this issue.**

Further, although the effect at Robben Island is no longer credibly different from zero at the 95% level, 96% of all the posterior estimates were greater than zero. In other words, it was credibly different from zero at the 94% level. But, rather than considering arbitrary cut offs around p = 0.05, modern day statistics is moving towards communicating uncertainty – see the recent special issue of 43 papers on this topic in the American Statistician, headed with the editorial "Moving to a World Beyond "p < 0.05"" (Wasserstein, Schirm and Lazar 2019). M1 in the table above is the maximal model (the most complex possible random effect structure); maximal models are "generally wasteful and costly in terms of statistical power for testing hypotheses" (Stroup 2012, pg. 185) and maximal models – even when they converge – can result in overparameterization that leads to uninterpretable models (Bates et al. 2018). Furthermore, the maximal model may actually trade-off power for some conservatism beyond the nominal Type I error rate, even in cases where the maximal model matches the generating process exactly (Matuschek et al. 2017). Nevertheless, it presents a > 96% probability (given the data and model structure) of a closure effect at Robben Island. Ignoring this, particularly given that an independent analysis by Ross-Gillespie and Butterworth (2020a, FISHERIES/2020/JAN/SWG-PEL/09) concluded that there was "*a biologically meaningful fishing effect*" on chick condition at Robben Island, using the 2004 to 2018 aggregated data would certainly risk making a Type II error about the impact of the closure.

The best-fitting model in this case was given ~95% of the stacking weights (Table 1), so averaging across all the models based on the stacking weights gives similar results to the best fitting model (Table 1). The model averaged closure effect at Robben Island represented an improvement during closed years of 23.6% (−4.9–51.9%) with 96% of all the posterior estimates > 0 and 82% > 10% (Figure 1). To put this in perspective, to be considered credibly different from zero (and thus bold in Table 1), 97.5% would need to be > 0. For Dassen Island, the corresponding model averaged estimates represented an increase of 13% (−12–39%) with 83% of all the posterior estimates > 0 and 55% > 10% (Figure 1).

The estimates and standard errors of the island closure effect given in Table 2 for the aggregated-data-based analyses of FISHERIES/2020/JAN/SWG-PEL/09 are in log-space, and so correspond roughly to proportions (multiply by 100 for percentages). Adjusting here, and also in similar comparisons below, for the sign change in the convention used, these are 0.14 (se 0.13) for Robben and 0.03 (se 0.14) for Dassen. These PEL/09 values are thus notably different for both the magnitude of the effect itself (smaller) and the associate precision (less).

R7. Actually, there is a difference between the datasets used in FISHERIES/2020/JAN/SWG-PEL/09 and FISHERIES/2020/JUL/SWG-PEL/53REV. FISHERIES/2020/JAN/SWG-PEL/09 appears to be using data on chick condition at Robben Island for 2004–2018 and at Dassen Island for 2008–2018. FISHERIES/2020/JUL/SWG-PEL/53REV is using data for 2008–2018 at both islands. FISHERIES/2020/JUL/SWG-PEL/53REV doesn't use the data for 2004 because it is biased high – the 2004 data at Robben Island was the data set used to create the chick condition index. It was based on a sample "collected between March and September 2004 for 125 chicks that fledged at Robben Island" (Lubbe et al. 2014). Since that 2004 sample only contained chicks that fledged, it is not comparable to the random samples made once the closures experiment started in 2008 and should be omitted. Fortunately, FISHERIES/2020/FEB/SWG-PEL/12 (Ross-Gillespie and Butterworth 2020b) offers us a bridge here; when the 2004 data were excluded from the analysis in FISHERIES/2020/FEB/SWG-PEL/12, the effect size for Robben (adjusting for the sign change as above) was 0.20, or ~20%. Broadly comparable to the ~23% effect reported in FISHERIES/2020/JUL/SWG-PEL/53REV.

Also, the comment on the differences in the associated precision is confusing, since standard errors are not presented here, and Butterworth does not offer 95% confidence intervals for the effect sizes in Table 2 of FISHERIES/2020/JAN/SWG-PEL/09. So, how is he comparing the precision directly?

Direct, like-for-like comparisons in Sherley (2020d), FISHERIES/2020/SEP/SWG-PEL/86, show that fits to the aggregated data do not consistently yield smaller effect sizes (as suggested above), but do indeed provide consistently more precise estimates (as would be expected from models with greater statistical power).

A notable difference between the best fitting model (M1, Table 1) and the model presented in Sherley et al. (2019; M3, Table 1) is also that the estimate for the 'Island' × 'Closure' interaction was not credible different from zero in the former, but is in the latter (Figure 2). Proceeding to simplify the fixed effects in M1 by dropping the 'Island' × 'Closure' interaction first yields a 'Closure' main effect of 0.047 (−0.0018–0.095) with 97.1% of the posterior estimates > 0, which corresponds to an 18% increase at Dassen Island and a 15% increase at Robben Island. The PSIS–LOO of this simplified model = 10366.0, thus there is no evidence from either classical Bayesian inference or model selection to retain the interaction in the model. Simplifying this model further, to drop the 'Island' main effect (which also overlapped zero) as well yields a 'Closure' main effect of 0.046 (−0.0005–0.094) with 97.33% of the posterior estimates > 0, which corresponds to a 17% increase in chick condition during 'Closed' years at both islands. The PSIS–LOO of this simplified model = 10367.0, so again this model would be considered the more parsimonious. See Appendix 3 for the candidate set of models tested and model selection results for the fixed component of the model.

Model selection must also be informed by relevant external information when available, and not only the analysis of the data from the experiment in question ***alone****. The totality of the estimates of the island closure effect from various sources in Table 2 of*

FISHERIES/2020/JAN/SWG-PEL/09 are strongly suggestive of a real difference in the values for each island. Thus, the results above without an Island/Closure interaction, while of themselves indicative of some average value across the two islands, cannot be used reliably to draw inferences about values for either island separately.

R8. “Everything should be made as simple as possible, but no simpler” – Albert Einstein. A model is by its very nature a simplification of a more complex process. The removal of the interaction based on parsimony does not say that in reality the closure effect does not or cannot differ between the two islands. It simply says that the data as available do not provide evidence to support an effect that differs in both its magnitude and direction in this case. Retaining model parameters that are not doing anything useful (redundant parameters) is a form of over-fitting, which costs explanatory power. The goal of model selection “is to identify the most parsimonious model that can be assumed to have generated the data” (Matuschek et al. 2017). Both sets of results are reported in the interests of full transparency.

Finally, something to note is that the inclusion of ‘Island’ in the random component of these models consistently decreased the precision of the ‘Island’ fixed effect in all models but not the ‘Closure’ fixed effect or the ‘Island’ × ‘Closure’ interaction (Figure 2). This hints that, even with the ‘Island’ random effect nested, the models are struggling to partition the variance due to the different island between the fixed and random effects. Further advice should be sought from the 2019 Panel on whether ‘Island’ should be included in both the fixed and random effects based on this observation (also see Appendix 2) and the statement in the introduction from Zuur et al. (2009).

To conclude here, although the effect is not unambiguous (as it was in Sherley et al. 2019), the results from M1 (Table 1) provide strong evidence for a meaningful ‘Closure’ effect at Robben Island (>95% of all the posterior > 0) and simplifying this model strongly suggests an overall ‘Closure’ effect (i.e. improvements at both Dassen and Robben Island) on chick condition on the Western Cape (>97% of all the posterior > 0). Just as it has been argued that because of the random effect structure used, Sherley et al. (2019) made a type I error in concluding that a ‘Closure’ effect was apparent on chick condition at Robben Island (Butterworth and Ross-Gillespie 2019), I suggest that ignoring the evidence presented here strongly risks making a Type II error. Updating the chick condition analysis to include 2019 data should be seen as a priority.

3.2. Maximum distance travelled, Eastern Cape

The fixed effect structure used for these models had the following components: ‘Island’ main effect, ‘Closure’ main effect, ‘Island’ × ‘Closure’ interaction, a sardine biomass main effect, an anchovy biomass main effect and a brood mass main effect (implemented as an imputed z-score standardised total brood mass (see Sherley et al. 2019). The random effects structures tested are shown in Table 2, along with their corresponding WAIC, PSIS–LOO scores and stacking weights. All models are random intercept models.

For this dataset, there were four well supported models (M6, M3, M2 and M1, Table 2), but the estimated closure effect sizes and 95% HPDI were essentially identical in all four models (Table 2). The results of all four were unchanged from the result reported in Sherley et al. (2019). They were also unambiguous, with 100% of the posterior indicating a positive effect of the Closure (max. distance decreasing in closed years) in all four models and 99.99% of the posteriors for the effect size at St. Croix exceeding a 10% effect size. Averaging across the models, according to the stacking weights assigned to each model, the effect size at St Croix Island was −0.33 (−0.49–−0.19; Figure 3), corresponding to breeding penguins foraging 28% (14–44%) closer to St. Croix Island during ‘Closed’ years, or a mean maximum distance travelled of 20.9 (17.1–24.4) km at St. Croix during ‘Closed’ years and 29.4 (26.5–32.8) km during 'Open' years. In all models, the closure effect at Bird Island was not credibly different from zero, as reported in Sherley et al. (2019).

In conclusion the inference here remains unchanged from that presented in Sherley et al. (2019) for Maximum distance travelled at the Eastern Cape islands.

A general concern about these results is their sensitivity to the metric used to reflect the response. In Table 2 of FISHERIES/2020/JAN/SWG-PEL/09, when Robben and Dassen are considered, although the estimates for maximum distance travelled are similarly small (albethey of opposite sign), for the other metrics of forage length and forage duration they are all of opposite sign (i.e. the effect of closure is negative, not positive, for the penguins), and substantially so for Dassen Island.

R9. I'm not clear on what the point being made here is. The above talks about results in FISHERIES/2020/JAN/SWG-PEL/09 for foraging data at Robben and Dassen Island, but this comment comes after a section and figures presenting results on foraging data at St Croix and Bird, thus seems irrelevant or perhaps misplaced?

3.3. Chick Survival, Western Cape

The fixed effect structure used for these models had the following components: 'Island' main effect, 'Closure' main effect, a sardine biomass main effect, and an anchovy biomass main effect (see Sherley et al. 2019). The shared frailty term structures (akin to random intercepts) tested are shown in Table 3, along with their corresponding WAIC, PSIS–LOO scores and stacking weights.

The best fitting model contained the random 'Island/Year/Month' intercept and yielded an estimated Closure effect size of 0.38 (HPDI: 0.21–0.58). This corresponds to an improvement in survival of 10.3% (5.4–15.2%) at Robben Island and 10.6% (5.2–16.2%) at Dassen Island when the closure was in place. And, as with Max. foraging distance at the Easern Cape islands, this effect was unambiguous, with 100% of the posterior indicating a positive effect of the Closure. Respectively, 53% and 57% of the posterior distribution exceeded the 10% threshold for management action at Robben Island and Dassen Island.

The estimates and standard errors of the island closure effect given in Table 2 for the aggregated-data-based analyses of FISHERIES/2020/JAN/SWG-PEL/09 are 0.04 (se 0.11) for Robben and 0.13 (se 0.10) for Dassen. Similar to results for chick condition then, these PEL/09 values are thus notably different for both the effect itself (smaller) for Dassen and the associated precision (less) for both islands. Part of the improved precision here is arising from assuming that the closure parameter is the same for both islands – see comment immediately preceding Figure 1 above which questions the appropriateness of this assumption.

R10. Again, there is a difference between the datasets used in FISHERIES/2020/JAN/SWG-PEL/09 and FISHERIES/2020/JUL/SWG-PEL/53REV. FISHERIES/2020/JAN/SWG-PEL/09 used chick survival data from 2001–2015 for Robben Island and 2008–2015 for Dassen Island (I assume the latter, it is not specified in FISHERIES/2020/JAN/SWG-PEL/09, but data on chick survival are not available before 2008 at Dassen Island). FISHERIES/2020/JUL/SWG-PEL/53REV used data for 2008–2018 for both Robben Island and Dassen Island. Therefore, in this case the estimates are not directly comparable in the way suggested above.

Averaging across all the models based on the stacking weights (Table 3) gives similar results to the best fitting model. The model averaged 'Closure' effect estimate was 0.36 (HPDI: 0.18–0.54), with 99.9% of the posterior indicating a positive effect (Figure 4). At Dassen Island, chick survival improved by 10.1% (0.2–19.7%) with 97.7% of the posterior estimates > 0 and 51% > 10%. At Robben Island, chick survival also improved by 9.9% (1.1–18.2%) with 97.9% of all the posterior estimates > 0 and 49% > 10%.

In conclusion, the closure effect is credible, even when taking the uncertainty of all model structures into account. And, at present the inference here remains unchanged from that presented in Sherley et al. (2018, 2019) for chick survival at the Western Cape islands, even with 3 more years of data added to the analysis.

3.4. Overall Closure Effect

The Overall Closure Effect (%) is calculated by combining the posterior distributions for the percentage effect sizes for all ten penguin responses (Figure 5) into a single vector in R using the *c* (combine) function. The posterior distributions are combined with equal weighting (30 000 posterior samples each) given to each penguin response. This is akin to the approach developed recently to combine regional % declines into a weighted global change % for IUCN Red List assessments (Sherley et al. 2020). We plot this combined distribution, with all negative posterior samples (≤ 0) shown in red, all positive posterior samples (>0) shown in green (Figure 5), and calculate the mean, median, 95% HPDI and the percentages of the Overall Closure Effect posterior that are above both zero and the 10% threshold for management action (Cochrane 2016).

How best to combine the results across both different response variables and islands is not an easy question. First though, for reasons given above, combination across islands is inappropriate (and as above for chick survival where the analysis itself assumes the effect to be island independent). Further, a formal "statistical equal weighting" is questionable. For St Croix, the distributions for condition index and maximum distance are entirely of opposite sign, i.e. completely non-overlapping, so that no probabilistic interpretation can be placed on the combined result presented above – rather if some real impact of closure on penguins is indeed occurring, at least one of the two models assumed for these two analyses must be wrong. There is no simple way to address this matter; discussions are first needed about the relative reliability of the response variables as measures of forage fish availability on penguin reproductive success, together with some quantification of the likely impact of each on penguin dynamics.

R11. This is purely an opinion-based statement. No evidence is offered for why a statistical equal weighting would be deemed questionable, nor why effects cannot be considered across islands. In addition, the objection is directly contrary to the approach advocated by Butterworth for many years; he has repeatedly advocated for a simple tallying-up of positives and negatives approach (which really does give all models statistical equal weighting). For example: "*These results (see Table 3.10 of Robinson, 2013) indicate that for 144 GLMs conducted across six monitoring indices, 16 indicate statistically significant positive effects at the 5% level for aspects related to penguin reproductive success, while none indicate similarly significant negative effects. Overall some 80% of the GLMs indicate positive (though not always statistically significant) effects*" (Butterworth 2014, FISHERIES/2014/APR/SWG-PEL/ICTT/16). Moreover, it a misrepresentation to suggest say that FISHERIES/2020/JUL/SWG-PEL/53REV suggests that the overall closure effect gives all models "statistical equal weighting" since the phrase "statistical equal weighting" does not actually occur in FISHERIES/2020/JUL/SWG-PEL/53REV. What FISHERIES/2020/JUL/SWG-PEL/53REV says is that all posterior distributions are sampled with equal weighting. It is not entirely accurate to say that the overall closure effect gives all models "statistical equal weighting" since it automatically and inherently incorporates the uncertainty in each model output – thus, not all models contribute entirely equally, but do so in proportion to the level of uncertainty therein.

3.5. Conclusions and next steps

It has been argued that the effects presented in Sherley et al. (2019) were not robust because the allegedly poorly chosen random effect structure resulted in fixed effect estimates that were overly precise (Butterworth and Ross-Gillespie 2019). However, the results presented here suggest little meaningful change in inference whether or not Island was included as a higher-level random effect.

This comment is completely at variance with the results shown earlier in Table 1 (see also the comments made immediately below that). Quite clearly results for precision are distinctly non-robust to decisions made about which factors to include in the random effects considered in that case. Consequently, results reported in earlier papers have been incorrect because this selection process was not carried out appropriately.

R12. Again, the above is a purely opinion-based statement and not at all in keeping with the results presented in FISHERIES/2020/JUL/SWG-PEL/53REV. Again, this also ignores the question of whether Island should simultaneously be included in both the fixed and random components of these models and whether Island, which only has two levels, should be included in the random effect structure at all. I repeat my request that **the SWG-PEL should officially go back to the 2019 panel and request their input on this issue and whether FISHERIES/2020/JUL/SWG-PEL/53REV meets their recommendations from December 2019.**

This result has also just been supported by a new simulation study on the robustness of linear mixed-effects models which concluded that "missing random effect predictors had little effect on the fixed effect estimates but had systematic effects on the estimates of random effects. The variance due to unmodelled higher level predictors [like Island in our case] was almost completely absorbed by the nested random effect variance of interest" (Schielzeth et al. 2020).

This study provides results similar to those reported in FISHERIES/2016/AUG/SWG-PEL/65 (MARAM/IWS/DEC16/Peng Clos/P3) which showed that the appropriate adjustment using a random effects approach of estimates of precision for non-independent data were reasonably robust to choice of the unknown time scale at which that effect modelled was operating. Whatever, the Schielzeth et al. study is not pertinent to the key aspect of the analyses in question here, because the models which it considers do not include year-independence in a diagonal sub-matrix block structure as pertinent here to the estimation of the parameter of interest – the closure effect (see Annex and the Primary overview comments above).

As a consequence of our updated results, it has now been shown that meaningful, unbiased, positive effects of the fishing closures are apparent in all three of the following cases:

a) When the observation level (disaggregated) data were modelled using random effects that were selected *a priori* by the analyst to adequately reflect the sampling structure of the data (Sherley et al. 2018, 2019);
b) When the observation level (disaggregated) data were modelled using random effects that were selected using model selection (this document), as advised by the 2019 Panel (Die et al. 2019);
c) When the aggregated data (annual means) are modelled appropriately (e.g. Sherley and Winker 2019) to avoid over-parameterization (as in Robinson 2013, Robinson and Butterworth 2014), even when those models account for the differing precision of each annual estimate (Sherley et al. 2015).

Sherley and Winker (2019) is not included in the references listed. Whatever, these conclusions do not follow, for the reasons elaborated in the comments above.

R13. Apologies: Sherley and Winker (2019) is MARAM/IWS/2019/PENG/WP3, which demonstrated that the inference about the Robben Island chick condition effect and the St. Croix max distance effect remains the same whether fitting to aggregated or disaggregated data.

Moreover, the most recent update of the MARAM power analysis (Ross-Gillespie and Butterworth 2020) also finds "evidence in the current data of a biologically meaningful fishing effect" on chick condition at Robben Island, fledging success at Robben Island (chick survival measures a component of fledging success), and chick survival at Dassen Island. Those results from Ross-Gillespie and Butterworth (2020) concur with results presented here in section 3.1 and 3.3.

As indicated in many of the comments above, there are many and important differences in the results in Table 2 of Ross-Gillespie and Butterworth (2020) - FISHERIES/2020/JAN/SWG-PEL/09 – to those reported in this document. Broad indications are that the results for the island closure effect in this document tend to be of larger magnitude and are indicated to be of greater precision – for reasons discussed above, this last result may well be a reflection of the individual-based approach used in this document having failed to account adequately for pseudo-replication effects, and hence leading to incorrect conclusions as to whether certain probability thresholds have been met.

R14. Again, it is worth considering that the 2019 panel appeared to disagree with this statement:
"*results presented to the Workshop suggest that estimates of closure parameters using models fitted to aggregated and individual data had similar standard errors*" (Die et al. 2019). I repeat my request that **the SWG-PEL should officially go back to the 2019 panel and request their input on this issue and whether FISHERIES/2020/JUL/SWG-PEL/53REV meets their recommendations from December 2019.**

Moreover, this comment once again seems to be completely contradictory to concerns expressed above in Butterworth's overview comments that "*the individual-data-based estimates of the closure effect are indicated by the Annex to be unable in principle to provide any improvement on annually aggregated analyses*" in terms of precision. Again, I repeat: on the one hand, Butterworth says that the results in FISHERIES/2020/JUL/SWG-PEL/53REV cannot be used as a basis for decisions because they might produce negatively biased estimates of the standard errors of closure effect compared to the aggregated data approach (i.e. the standard errors are smaller than obtained from the aggregated data approach). On the other hand, he also contends that the results presented in FISHERIES/2020/JUL/SWG-PEL/53REV should not be considered further because that document and his Annex indicate that they are not capable of providing improved estimates of such precision (i.e. smaller standard errors than those obtained from the aggregated data approach).

Again, both cannot simultaneously be true.

Plus, Ross-Gillespie and Butterworth (2020) also find that "a biologically meaningful fishing effect is likely to be detected" if the experiment continues for 2–5 years (using a dataset that ended in 2015) for chick survival at Robben Island. This broadly concurs with the results presented here in section 3.3 as well. In other words, **we have now iterated to a place where two independent sets of analyses agree that biologically meaningful effects of fishing around African penguin breeding colonies are apparent** and, importantly, that some of those effects are on variables (chick survival, fledging success) that contribute to the demographic process.

Estimation of chick survival is particularly valuable, as it is a parameter relating directly to penguin dynamics, and hence simplifies relating any estimate of the closure effect to its consequences for penguin population growth rate, However chick survival is itself a component of fledging success, which overall is even more consequential for those dynamics. A concern is that Table 2 of FISHERIES/2020/JAN/SWG-PEL/09 indicates that although closure has a strong positive effect on chick survival at Dassen island, the impact on fledging success there is not small and in the ***reverse*** *direction. There may therefore be negative correlation effects present here (e.g. if conditions to promote successful hatching of eggs are* *good, a greater proportion of comparatively weaker chicks hatched may follow, leading to worse chick survival). This needs further consideration before drawing conclusions about the implications from the chick survival estimates in isolation.*

R15. First, unfortunately, Butterworth is trying to compare apples with oranges. The chick survival dataset in FISHERIES/2020/JAN/SWG-PEL/09 spans 2008–2015 for Dassen Island. The fledging success dataset in FISHERIES/2020/JAN/SWG-PEL/09 spans 1995–1999 and then 2008–2015 (with a gap from 2000 to 2007) at Dassen Island. First, it is difficult to be confident in directly comparing data from the 1990s with data collected from 2008 onwards in this context because there is strong evidence that the ecosystem, the availability of forage fish resources to fisherman and predators, and penguin population dynamics have changed markedly over this timeframe (e.g. van der Lingen et al. 2005, Roy et al. 2007, Robinson et al. 2013, Crawford et al. 2019). We cannot be sure that the trend in the opposite direction is not a consequence of these differences in the state of the ecosystem.

Second, FISHERIES/2020/JAN/SWG-PEL/09 indicates that the experiment would need to continue for more than 10 years before a biologically meaningful fishing effect is likely to be detected for fledging success at Dassen Island. In other words, the fledging success effect at Dassen Island isn't meaningfully different from zero. On the other hand, the chick survival dataset for Dassen Island already provides evidence of a biologically meaningful fishing effect. Thus, the two do not offer equally strong opposing evidence. The chick survival dataset already has enough power to show a biologically meaningful effect, while the analysis of the fledging success dataset reveals that this dataset is still currently too noisy to offer useful information. This is not a case of a statistically significant effect in one direction versus a statistically significant effect in the other; it is a clear effect in the positive direction versus a noisy, but tentative trend in the other.

Third, the above ignores the fact that at Robben Island the fledging success dataset which also already enough power in FISHERIES/2020/JAN/SWG-PEL/09 analysis to provide evidence of a biologically meaningful fishing effect is in the same direction as the chick survival effects at both Robben Island and Dassen Island reported in FISHERIES/2020/JAN/SWG-PEL/09 and FISHERIES/2020/JUL/SWG-PEL/53REV.

Fourth, and most importantly, if we actually do a like for like (apples with apples) comparison between chick survival and fledging success, we find they are positively correlated with one another. The below uses chick survival estimates from OLSPS (2020)

Robben Island 2001 to 2015: Pearson's product-moment correlation, $r = 0.981$, $t_{13} = 18.37$, $p < 0.001$.

Dassen Island 2008 to 2015: Pearson's product-moment correlation, $r = 0.818$, $t_6 = 3.48$, $p < 0.013$.

And using the data underpinning FISHERIES/2020/JUL/SWG-PEL/53REV (note fledging success has not been calculated for 2016–2018):

Robben Island 2008 to 2015: Pearson's product-moment correlation, $r = 0.93$, $t_6 = 6.65$, $p < 0.001$.

Dassen Island 2008 to 2015: Pearson's product-moment correlation, $r = 0.817$, $t_6 = 3.47$, $p < 0.013$.

Thus, Butterworth's concerns about negative correlation effects are unfounded.

I propose the next steps should be:

i) Seek guidance from the 2019 Panel on the sense of including 'Island' in both the random and fixed components of these models (given the results herein and recommendations in Zuur et al. 2009).

ii) Based on the advice pertaining to i), if necessary, undertake model selection as presented here for the other variables examined in Sherley et al. (2019), namely chick condition in the Eastern Cape and maximum foraging distance in the Western Cape.

iii) Based on the advice pertaining to i), if necessary, recalculate the Overall Closure Effect presented in Sherley et al. (2019) using these updated model structures.

iv) Seek written confirmation from the 2019 International Stock Assessment Workshop Panel that the results herein [and if necessary the outcomes of steps i) to iii)] satisfy the recommendations in Die et al. (2019) and, if so, make a final management recommendation regarding the Island Closures Experiment to the Department of Environment, Forestry and Fisheries by December 2020.

Some final comments

1) *It is always useful to have more than one model applied to data, if only to check the robustness of key results. However, if a new model is introduced, standard practice is for the modeller to "build a bridge" between the default model and the new one to reveal which changes are most responsible for any changes in results. This is to ease identification of where discussion on the appropriateness of assumptions is best focussed.*

   In this case, the model of this document differs from the default model that has been adopted for these analyses in collaboration with the International Panel during previous meetings, where this default provides the basis for the computations of FISHERIES/2020/JAN/SWG-PEL/09. All of the structure of the model of this document (e.g. including setting up in normal rather than in log space), the data used (and whether or not these are individual-based or annual aggregates), and the estimation approach differ in this instance. This renders comparison of results problematic, and it would be helpful for future iterations of this document to build that bridge,

R 16. This comment strikes me as being deliberately disingenuous. I have – on more than one occasion – submitted documents to the SWG-PEL or IWS that attempt to bridge across the differences in results that come from using the aggregated vs disaggregated approach. For example, MARAM/IWS/DEC16/Peng Clos/P4 (Sherley 2016a) for the 2016 IWS and MARAM/IWS/2019/PENG/WP3 (Sherley and Winker 2019) for the 2019 panel. Moreover, as Butterworth outlines above, and as I confirmed in one of my responses, the bridge actually has been built in the % effect sizes presented in FISHERIES/2020/JUL/SWG-PEL/53REV. To quote from FISHERIES/2020/AUG/SWG-PEL/82 (Butterworth 2020b) "*The estimates and standard errors of the island closure effect given in Table 2 for the aggregated-data-based analyses of FISHERIES/2020/JAN/SWG-PEL/09 are in log-space, and so correspond roughly to proportions (multiply by 100 for percentages)*". And my response above: "when the 2004 data were excluded from the analysis in FISHERIES/2020/FEB/SWG-PEL/12, the effect size for Robben was (adjusting for the sign change as above) was 0.20, or ~20%. Broadly comparable to the ~23% effect reported in FISHERIES/2020/JUL/SWG-PEL/53REV". So, actually I have already made the effort to build that bridge.

2) *The motivation for suggesting consideration of other than island closure values which are island specific, and the inclusion of island-related interactions (with year – whether or not as random effects), is unclear. The reasons for the need for the first have already been explained above. Similarly, the results of Table 1 seemingly make clear that the inclusion of Island/Year interactions (elsewhere termed "process errors") are essential.*

After all, the basis underlying the whole experiment is that closure and year effects are confounded, but that their distinction is made possible by considering two nearby islands (e.g. the Dassen-Robben pair) for which the year effect (related to forage fish densities) is likely to be similar because they are not that far apart. But of course the differential effect between the two islands will vary from year to year, necessitating the presence of this process error term in the model equations. Furthermore, this term is quantitatively important, as shown by the analyses of MARAM/IWS/DEC15/PengD/P2 which indicated that this process error is far larger than the observation errors remaining after averaging over the number of island-specific measurements typically available each year. It may be that improved standardisation for covariates might further reduce the extent of that observation error, but there have been no indications to date that that might be the case, and in any case this seems unlikely given the process error dominance which is already apparent in the results from the experiment.

R17. See responses above (and peer-reviewed literature cited) about model selection. Also, Butterworth seems in this comment to be confusing the 'Island' × 'Closure' interaction in the fixed effects and the Island/Year nested structure in the hierarchical random effects. Any simplification in this document where the 'Island' × 'Closure' interaction was dropped from the fixed effects was a simplification of M1 (e.g. Table 1, still in this document), and M1 contained the maximal random effect structure (i.e. Island/Year/Month). So, in fact, it contained what Butterworth describes as the process error term. I am thus not clear what point is being made here.

Note: The author acknowledges helpful discussions with Anabela Brandão and Andrea Ross-Gillespie in developing this response, and from the latter in producing it; however, this should not be taken to imply that these persons necessarily concur with all the views expressed in this response.

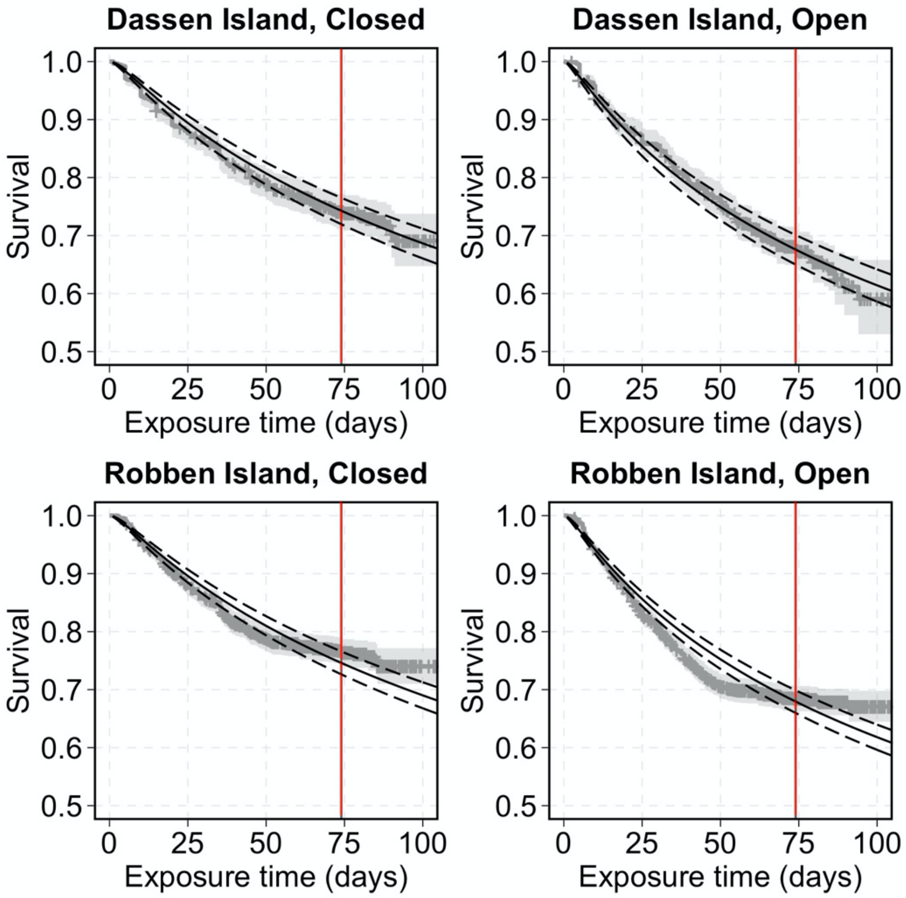

Figure A4.13. Model validation plots for Chick Survival at Dassen Island (top) and Robben Island (bottom) during years that were Closed (left) and Open (right) to fishing. Panels show the comparison of the non-parametric Kaplan-Meier (KM) estimate of survival (grey points, +) and its 95% confidence intervals (grey polygons) and the predicted survival rates (solid black curves) and 95% credible intervals (black dashed curves) based on a model with a log-normal hazard function and no shared frailty term. The vertical red line marks time = 74 days, the age at which the predicted chick survival is compared between islands and closure statuses in the results section of this document and elsewhere (Sherley et al. 2013, 2015, 2018, 2019). Crucially, the predictions from the log-normal model and the KM estimate (which is derived only from the observations) are not credibly different at 74 days, which indicates adequate model fit to predict chick survival at time = 74 days.

The use of an exponential model for survival rather than the "log-normal" model would be simpler and more readily interpreted, and seems attractive given indications (if I am understanding correctly from Sherley's comments during the 30 July meeting) that the differences in results of interest are not large. The particular reason for this is that then the non-equivalence of exposure time and chick age (because of variable commencement of the age at which different chicks are first recorded) does not potentially confound results. But then the marked (and apparently relatively precisely estimated) change in the estimated survival rate at for Robben (but not Dassen) from the KM estimates after some 50 days exposure becomes a concern. To what extent then might these estimates of cumulative survival be confounded by different distributions of the chick age at which this monitoring commences? Some restrictions on the data used for these analyses, for example through elimination of data for chick for which monitoring is known to have started only at a fairly late stage, might be desirable. However, the matter should first be discussed to check whether some prior further diagnostic investigations might provide insight, before perhaps embarking on further onerous data extractions.

R18. Results comparing the log-normal and exponential hazard functions were given already in an appendix of FISHERIES/2020/JUL/SWG-PEL/53REV. It makes no appreciable difference to the inference whether a log-normal or exponential hazard function is used. The log-normal model gives a more parsimonious fit to the data based on model selection (see FISHERIES/2020/JUL/SWG-PEL/53REV) because the log-normal hazard function can be monotonically decreasing based on the mean and standard deviation of survival time on the log scale. It is very unlikely that any parametric survival model will perfectly match the KM curves all the time, but the key point illustrated above is that in all 4 cases above (Dassen Island Closed, Dassen Island Open, Robben Island Closed, Robben Island Open), there is no credible deviation from the KM estimate at 74 days, when the comparisons between islands is made. In other words, whether we compared the means from the KM model, which is purely data-driven (non-parametric), the log-normal model or the exponential model, we would find a credible difference between the open and closed years at both islands. Furthermore, as outlined elsewhere, the comparison at age 74 days was not selected specifically for this study, but dates back to Sherley (2010) and has been applied consistently in Sherley et al. (2012, 2013, 2015, 2018, 2019) and Sherley (2020b). Thus, it cannot be that this choice of the log-normal model was somehow made to accentuate or inflate the chick survival effect.

Moreover, the mean error between the KM model and the LN model is less than 1.5% (Figure R1), meaning that on average, the estimates from the LN are within 1.5% of the KM. And, at Robben Island, where the LN fits the least well of the two islands, the deviation is more positive during the Open years (1.35% mean error and −0.46% at 74 days), than during the Closed years (0.26% mean error and −1.91% at 74 days). In other words, the LN over-estimates survival more on average and at 74 days for the Open years than for the Closed years at Robben Island. This means that the meaningful closure effect detected at Robben Island is in spite of, not caused by, any bias that might exist in the dataset from the different chick ages at which this monitoring commences.

The above notwithstanding, it would be possible (though time-consuming and onerous as noted above) to refit the model to exclude the chicks not monitored from hatching.

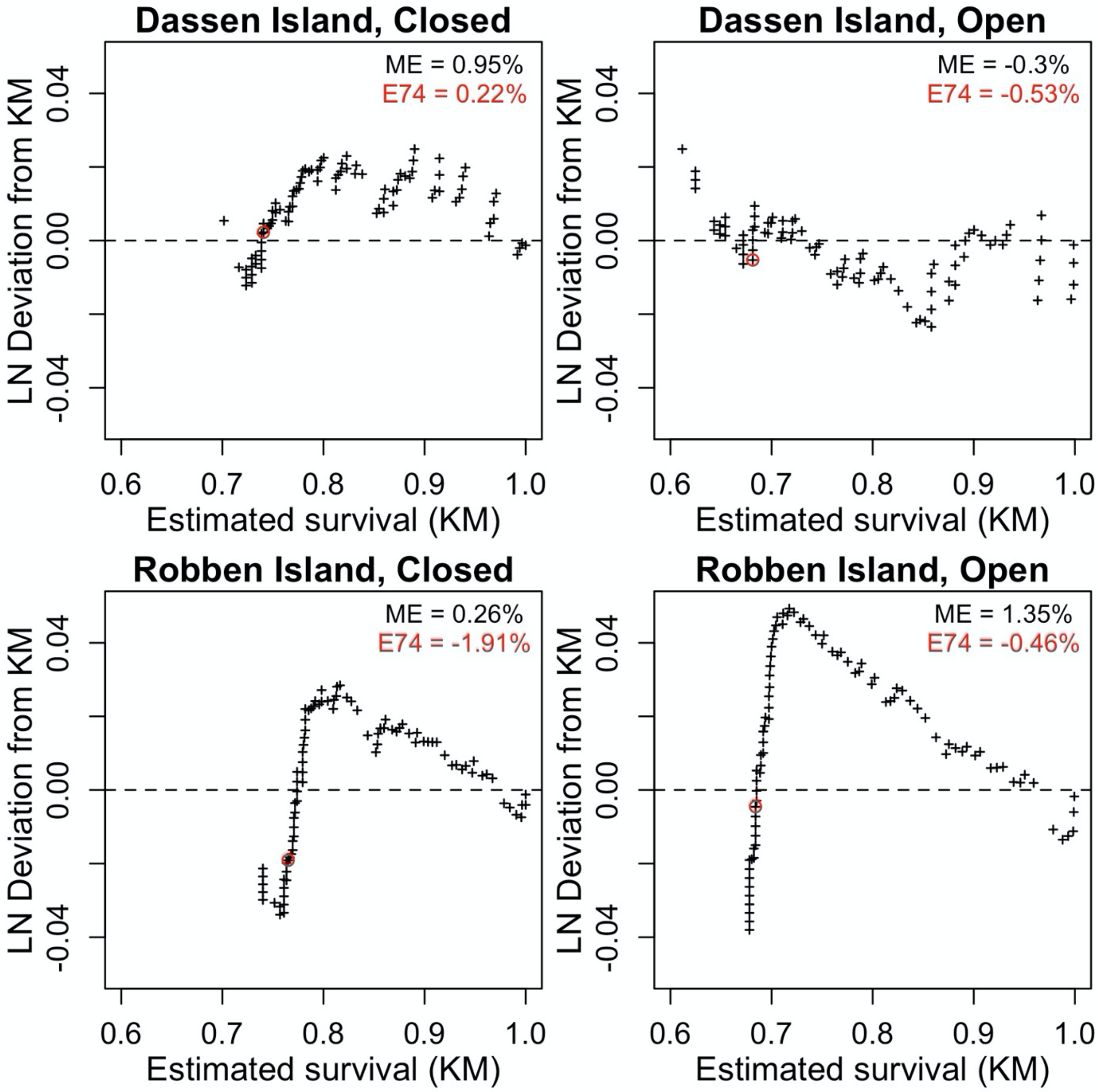

Figure R1: The deviation in the estimate of survival between a model with a log-normal (LN) hazard function and the estimate of survival of the non-parametric Kaplan-Meier (KM) model applied to the Dassen Island and Robben Island data, plotted against the KM survival estimate. The deviation (log-normal estimate – KM estimate) is positive when the LN model tends to over-estimate survival and negative when LN model tends to under-estimate survival. The dashed horizontal line shows a perfect one-to-one match. The red circle shows the deviation at 74 days. ME = the mean error in survival, expressed as a percentage. E74 = the error in survival at 74 days, expressed as a percentage.

Annex

R 19. I have no specific comments on the annex as it seems to cover old ground: Butterworth (2016) argued that there is “nothing to be gained in terms of improved estimation performance by fitting to the individual data for each year rather than to their means”, that “one cannot assume that a random effects estimator will fully correct for non- independence of data; rather it seems likely to yield estimates of standard errors for parameters which are negatively biased to some extent”. These issues have already been dealt with in numerous other places. Moreover, the 2019 IWS panel have already given an opinion on this issue, namely that “*Given the nature of the experiment, use of individual data is to be preferred. However, this is only the case if an appropriate random effects structure is chosen*” Die et al. (2019). Once again, **I repeat my request that the SWG-PEL should officially go back to the 2019 panel and request their input on this issue and whether FISHERIES/2020/JUL/SWG-PEL/53REV meets their December 2019 recommendations.**

References cited in response:

Bates D, Kliegl R, Vasishth S and Baayen 2018. Parsimonious Mixed Models. arXiv:1506.04967v2.

Besbeas P and Freeman SN. 2006. Methods for joint inference from panel survey and demographic data. Ecology 87: 1138–1145.

Brandão A, Butterworth DS, Johnston SJ, Glazer JP. 2004. Using a GLMM to estimate the somatic growth rate trend for male South African west coast rock lobster. Fisheries Research 70: 399–349.

Butterworth DS. 2014. A COMPOSITE PROPOSAL RELATED TO THE PENGUIN COLONY CLOSURE PROGRAMME. Department of Environment, Forestry and Fisheries Report: FISHERIES/2014/APR/SWG-PEL/ICTT/16. Pp. 1–8.

Butterworth D.S. 2016. On the use of aggregated vs individual data in assessment models. Department of Agriculture, Forestry and Fisheries Report No. FISHERIES/2016/NOV/SWG-PEL/65. Pp. 1–6.

Butterworth DS. 2020a. On estimates of the impact of fishing from analyses of the island closure experiment which model individual penguin responses directly. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/JAN/SWG-PEL/08. Pp. 1–2.

Butterworth DS. 2020b. A response to Sherley: FISHERIES/2020/JUL/SWG-PEL/53REV. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/AUG/SWG-PEL/82. Pp. 1–38.

Crawford RJM, Sydeman WJ, Thompson SA, Sherley RB and Makhado AB. 2019. Food habits of an endangered seabird indicate recent poor forage fish availability. ICES Journal of Marine Science 76: 1344–1352.

Die DJ, Punt AE, Tiedemann R, Waples R and Wilberg MJ. 2019. International Review Panel Report for the 2019 International Fisheries Stock Assessment Workshop, 2–5 December 2019, UCT. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/2019/General/5. Pp. 1–18.

Lubbe A, Underhill LG, Waller LJ and Veen J. 2014. A condition index for African penguin *Spheniscus demersus* chicks. African Journal of Marine Science 36: 143-154.

Matuschek H, Kliegl R, Vasishth S, Baayen H and Bates D. 2017. Balancing Type I error and power in linear mixed models. Journal of Memory and Language 94: 305–315.

Maunder MN. 2001. A general framework for integrating the standardization of catch per unit of effort into stock assessment models Canadian Journal of Fisheries and Aquatic Sciences 58: 795–803.

OLSPS. 2020. A simple restructuring of penguin chick survival data to estimate survivorship at Dassen and Robben Islands. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/JAN/SWG-PEL/06. Pp. 1–3.

Punt AE, Hobday DK and Flint R. 2006. Bayesian hierarchical modelling of maturity-at-length for rock lobsters, *Jasus edwardsii*, off Victoria, Australia. Marine and Freshwater Research 57: 503–511.

Robinson WML, Butterworth DS and Plagányi ÉE. 2015. Quantifying the projected impact of the South African sardine fishery on the Robben Island penguin colony. ICES Journal of Marine Science 72: 1822–1833.

Ross-Gillespie A and Butterworth DS. 2020a. Updated implementation of the Algorithm recommended by the Panel for the 2016 International Stock Assessment Workshop for assessing whether or not to continue with the penguin island closure experiment. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/JAN/SWG-PEL/09. Pp. 1–17.

Ross-Gillespie A and Butterworth DS. 2020b. Application of the power analysis algorithm to chick condition data standardised by month. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/FEB/SWG-PEL/12. Pp. 1–18.

Roy C, van der Lingen CD, Coetzee JC and Lutjeharms JRE. 2007. Abrupt environmental shift associated with changes in the distribution of Cape anchovy *Engraulis encrasicolus* spawners in the southern Benguela. African Journal of Marine Science 29: 309–319.

Sherley RB. 2010. Factors influencing the demography of Endangered seabirds at Robben Island, South Africa: Implications and approaches for management and conservation. PhD Thesis, University of Bristol, U.K.

Sherley RB. 2016a. Additional analysis suggested in response to differences in variance estimates between Sherley (2016) and Ross-Gillespie & Butterworth (2016). Department of Environment, Forestry and Fisheries Report: MARAM/IWS/DEC16/Peng Clos/P4. Pp. 1–4.

Sherley RB. 2016b. Updated diagnostic plots for JAGS models on chick condition. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/DEC16/Peng Clos/WP4. Pp. 1–3.

Sherley RB. 2020a. Some comments on FISHERIES/2020/JAN/SWG-PEL/08. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/AUG/SWG-PEL/83. Pp. 1–5.

Sherley RB. 2020b. Revisiting the key results in MARAM/IWS/2019/PENG/P4 in light of the 2019 Panel recommendations. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/JUL/SWG-PEL/53REV. Pp. 1–27.

Sherley RB. 2020c. A reply to Bergh: FISHERIES/2020/AUG/SWG-PEL/84. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/SEP/SWG-PEL/87. Pp. 1–18.

Sherley RB. 2020d. Some observations on comparisons of fitting to the annually aggregated and the individual data, this time using JAGS and for the cases considered in FISHERIES/2020/JUL/SWG-PEL/53REV. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/SEP/SWG-PEL/86. Pp. 1–4.

Sherley RB and Winker H. 2019. Some observations on comparisons of fitting to the annual means and the observation-level data for the cases in MARAM/IWS/DEC19/Peng/P4 that support a positive effect of the island closures experiment on African penguins. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/2019/PENG/WP3. Pp. 1–5.

Sherley RB, Barham BJ, Barham PJ, Leshoro TM and Underhill LG. 2012. Artificial nests enhance the breeding productivity of African Penguins (*Spheniscus demersus*) on Robben Island, South Africa. Emu 112: 97–106.

Sherley RB, Underhill LG, Barham BJ, Barham PJ, Coetzee JC, Crawford RJM, Dyer BM, Leshoro TM and Upfold L. 2013. Influence of local and regional prey availability on breeding performance of African penguins *Spheniscus demersus*. Marine Ecology Progress Series 473: 291–301.

Sherley RB, Winker H, Altwegg R, van der Lingen CD, Votier SC and Crawford RJM 2015. Bottom-up effects of a no-take zone on endangered penguin demographics. Biology Letters 11: 20150237.

Sherley RB, Barham BJ, Barham PJ, Campbell KJ, Crawford RJM, Grigg J, Horswill C, McInnes A, Morris TL, Pichegru L, Steinfurth A, Weller F, Winker H and Votier SC. 2018. Bayesian inference reveals positive but subtle effects of experimental fishery closures on marine predator demographics. Proceedings of the Royal Society B: Biological Sciences 285: 20172443.

Sherley RB, Barham BJ, Barham PJ, Campbell KJ, Crawford RJM, de Blocq A, Grigg J, Le Guen C, Hagen C, Ludynia K, Makhado AB, McInnes A, Meyer A, Morris T, Pichegru L, Steinfurth A, Upfold L, van Onselen M, Visagie J, Weller F and Winker H. 2019. A Bayesian approach to understand the overall effect of purse-seine fishing closures around African penguin colonies. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/2019/PENG/P4. Pp. 1–25.

Silk MJ, Harrison XA and Hodgson DJ. 2020. Perils and pitfalls of mixed-effects regression models in biology. PeerJ 8: e9522.

Stroup WW. 2012. *Generalized linear mixed models: Modern concepts, methods and applications*. Boca Raton: CRC Press.

Thorson JT and Minto C. 2015. Mixed effects: a unifying framework for statistical modelling in fisheries biology. ICES Journal of Marine Science 72: 1245–1256.

Thorson JT, Rindorf A, Gao J, Hanselman DH and Winker H. 2016. Density-dependent changes in effective area occupied for sea-bottom-associated marine fishes. Proceedings of the Royal Society B: Biological Sciences 283: 20161853.

van der Lingen CD, Coetzee JC, Demarcq H, Drapeau L, Fairweather TP and Hutchings L. 2005. An eastward shift in the distribution of southern Benguela sardine. GLOBEC International Newsletter 11: 17–22.

Venables WN and Dichmont CM 2004. GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. Fisheries Research 70: 319–337.

Wasserstein RL, Schirm AL and Lazar NA. 2019. Moving to a World Beyond "$p < 0.05$". The American Statistician 73: Suppl. 1, 1–19.