Dear Nigel,
We thank you and the reviewers for the constructive feedback on our manuscript. We have carefully addressed all reviewer comments (see responses below) and made the necessary changes in the manuscript and R code.
Best regards,
Rahel

Responses to reviewers

Reviewer 1

[... summary of ms by reviewer omitted…]

The study has no fundamental critical flaws that invalidate the conclusions of the model estimating the 'fecundity' from imperfect ground observations, but I believe that the population projection has a critical flaw that could render the projections (Fig. 4A) biologically unrealistic. There are several aspects how the study could be improved to make it more widely applicable:

1) The population projection is based on a matrix population model for red kites, and the authors adopt the structure and some demographic parameters from an existing published model (Sergio et al. 2021) to explore how much future population sizes would differ between observed and actual fecundities. However, the 'fecundity' used in a population model is not the same as the 'fecundity' calculated in the model presented here to account for observation errors: the authors exclude all nests with 0 chicks, and for a population process all the females that either did not breed or did not succeed (and thus had 0 fledglings in a given year) need to be incorporated – but it is not clear how the 'fecundity' presented in this manuscript (which is based on the subset of nests with >0 chicks) can adequately reflect the number of female offspring produced by a breeding female every year.

According to the provided R code the number of nestlings produced is simply a random multinomial draw of the number of breeders and the probability vector of nests with 1,2,3, or 4 fledglings (which sum to 1) – the code therefore does not include any opportunity for a breeder to fail to raise offspring (which is biologically unrealistic).

Note that the fecundity or productivity used in other Red Kite population models is generally <2 (Mammen et al. 2017, Katzenberger et al. 2021, Sergio et al. 2021, Pfeiffer and Schaub 2023), so the projections shown in this paper likely exaggerate the actual population trajectory that could be expected from a red kite population.

Response: The number of young per nest (fecundity in our ms) fits in the general range of the number of eggs reported in a review article by Sergio et al. 2019 (from 1.9 to 3.2). As we explain in l 188ff, when nestlings are counted, they have not yet fledged, therefore we used the reported range of number of eggs, rather than fledglings.

2) This may not matter for the relative comparison of observed vs. estimated ‘fecundity’, but given that the discrepancy is larger for the rapidly growing population I suspect that this mis-specified ‘fecundity’ may therefore also slightly exaggerate the effect of accounting for observation error. This error could be easily fixed if the fledged brood size (termed ‘fecundity’ in this manuscript) is multiplied by a probability to fledge any young at all (0.81 based on (Mammen et al. 2017)) to arrive at a more realistic value for productivity in the population projection.

Response: As the reviewer correctly notes, the simulation in the previous version only accounted for the proportion of non-breeding adults, but not the proportion of attempted breeders who fail. We have incorporated the latter into an updated simulation. White that reduced the growth in the growing population, the general conclusions remained the same (see new Table 1, Figure 4 and code).

3) The model appears very complex, and is – in fact – too complex for the available data. This leads to a number of practical constraints that need to be imposed for the model to be identifiable (and an entire section on ‘problematic datasets’ that needed to be excluded, L. 263-284). Some of these constraints, while understandable, limit the utility of the model for a broader application beyond the presented case study. In particular, the constraint that all nests contain at least 1 chick is in my opinion a severe shortcoming for the broader applicability of the model. Although valid for the example study (where all nests are known a priori to contain at least one chick), the general case in nest monitoring will often include nests that have failed and contain 0 chicks, and the model should be able to incorporate such cases. Extensions such as those discussed (L. 450-458) are in my opinion not very promising, and the more commonly encountered scenarios in any monitoring (nest is empty and nothing can be seen, or not seeing any fledglings where the deviation can be >1) would be more valuable to be included.

Response: The modeling framework is actually not limited to having 1+ nestling (true or observed). When implementing the case study, we parameterized the model in this manner because our example data set contained no true or observed 0’s. But we realized, upon re-reading the Methods, that we gave the impression that this is generally true of nestling count data sets. We have now (a) generalized the model development section (l 134ff) to state that z can be 0; and (b) moved the specifics of how and why we parameterized our estimation simulation without z=0 to the ‘Simulation - parameter estimation’ section (l185ff). Further, we have added a short Limitations section to the manuscript in order to address several reviewer comments; here, we come back to our model evaluation not including 0 categories for z and y and what that means for the applicability of the model to real world datasets (l 508ff).

Finally, we disagree that the simplification of the model in line 450ff is not promising and retain it to highlight that reparameterization of the model to specific, more parsimonious cases is straightforward.

4) Given the complexity outlined in (2), the authors miss the opportunity to use a very simple and basic correction factor to assess whether overall fecundity could be simply adjusted by this

constant (rather than using a complex hierarchical model) with the same benefit than they claim to derive from the sophisticated model. If, for example, ground observations of the 26 paired calibration nests are divided by the actual number of chicks (from climb counts) in those nests, the resulting correction factor (0.95 from L. 312-313) could then simply be used to 'correct' all remaining ground counts. It would be very useful to include this 'basic correction' scenario in the comparison of population projections because it would be far easier to implement than the hierarchical model (especially for field practitioners).

Response: Using a correction factor for lambda rather than estimating all observation probabilities was the motivating thought behind our attempts to develop a model with a Poisson (rather than multinomial) observation process (see l 417ff). Unfortunately, this simplified model did not work. Of course the reviewer is correct that instead of using a statistical model at all, one can simply calculate the ratio of ground to nest counts and use that to adjust further ground counts. We have included a new Limitations section, in which we also mention this possibility (508ff); further, we provide an error plot for correction factors for the analyses under Question 3 in Appendix 1 (Figure S8). We note, however (and also included that in the manuscript), that the correction factor approach does not allow for the inclusion of covariates that may affect the classification process (or here, the correction factor). Several reviewers (and we ourselves) think that this is an important future development to make this approach more applicable to real world data.

Finally, we chose not to run another population projection for two reasons: a) given the average unbiased correction factor, the population trajectory would look like the true simulated trajectory, and b) our population projections are based on the full distribution of number of nestlings per nest, not the average fecundity lambda; switching to generating the number of nestlings from a Poisson distribution using lambda would require translating lambda into the effective mean of a Poisson distribution that is truncated at 0 and 4. Given that the focus of this manuscript is on the newly developed model, we felt that this was outside its scope.

5) In return, more elaborate methods (such as those suggested in L. 476-480) could be removed from the manuscript.

Response: We feel that we need to retain CMR as a more intensive but also more robust method to estimate fecundity to contextualize the newly developed model.

6) The main benefit of the model is that it estimates the probability of observing the correct number of chicks. However, that probability will vary a lot based on nest location and local topography. While the authors acknowledge this as a 'potential future extension' (L. 469-476, 505-509), I think that this issue is actually more important than the scenario where they estimate detection probability from one population and then estimate 'fecundity' in another population, where 'fecundity' is different (but detection probability is constant). That scenario makes the possibly unrealistic assumption (L. 238 and 467: "assuming the classification process is the same.") that detection probability is constant (or follows the same normal error distribution) between the two populations. Whether that assumption is realistic and likely to be met cannot

be assessed without understanding the factors that affect the accuracy of chick counts. The recommendations to ensure that observation conditions are standardized are valid to a certain point, but in my opinion it is unrealistic to expect that anybody can standardize the “distance from the ground observer’s position to the young” (L. 486) because we cannot tell a red kite how high the tree is they prefer to nest in! Thus, it would be good to present nest-specific variables (nest height, tree density, slope etc.) that likely affect the accuracy of chick counts, and use the 26 nests with known true chick number to explore which factors actually explain any errors in ground counts. That sample size may not be sufficient to construct a linear predictor function in the model and estimate parameters that affect detection probability, but given that these nest-specific factors are likely the reason WHY there is a discrepancy between observed count and true state (=the premise for the entire study in the first place), this should be more prominently included in this manuscript rather than listed as a potential future extension. Including the detection process more formally may also facilitate other simplifications that have so far failed (e.g. L. 413-419 explains that a Poisson distribution did not work – but if the observation error could be modelled as a function of variables this approach may fit the data better?).

Response: We tried fitting such models but as the reviewer suspects, the sample size was too small for robust conclusions about important effects, and the only effect that came out as important was the true brood size - which in practice is not known. We briefly mention this now in the Discussion in l 480f.

7) In many raptor monitoring projects, nests are visited more than once. Especially when the goal is to estimate fecundity, nests are often visited several times during a breeding season to ascertain the number of fledglings (Steenhof and Newton 2007). The authors point out in L. 443-448 that a Poisson-Poisson N-mixture model could work well for such data. While it may be beyond the scope of the current manuscript to demonstrate how well such a model would perform with repeated nest monitoring data, it would be good to add some thoughts to this part of the Discussion how the closure assumption could be met in these models: nests can fail at any time (although the probability often decreases as chicks get older), and repeated visits may therefore encounter an altered true state of the number of chicks in the nest (which may violate a core assumption of N-mixture models?).

Response: We agree that examining model performance based on repeated monitoring of nests is beyond the scope of our paper and also find that adding detailed discussion on how closure could be achieved in such a sampling scenario is not within the context of this work.

8) In the current case study observations are made at a single point in time, which is a typical process in basic monitoring schemes that count the number of almost fully-grown chicks just prior to fledging (Steenhof and Newton 2007). However, such observations would almost always include 0 counts (where nests had failed), which is not envisioned in the current study (see point (2) above), and a better explanation under what monitoring scheme the described model could actually be applied would be useful (unless it is expanded to allow for nest failures and 0 counts).

Response: Please see our response above regarding allowing for nests with 0 counts/fledglings.

9) The opportunity to use drones for accurate nest surveys is only briefly mentioned in the manuscript (L. 76) but given the cheap and ubiquitous availability of drones and expanding literature on how to use drones for wildlife monitoring (Chabot and Francis 2016, Brack et al. 2018, Hodgson et al. 2018, Edney and Wood 2021, Zink et al. 2023) I think it would be reasonable to suggest that using a drone could achieve the same certainty as a climber with less effort, potentially eliminating the need for an observation-correction or at least allowing for a much larger sample size.

Response: We added a reference to drones as a potential substitute for climbing nest trees in line 490ff.

10) The Swiss Ornithological Institute (where I work) could contribute much more data for the extension or evaluation of those models. Please contact us if you would like more data from nests where the number of chicks was validated by either drone or climb counts.

Response: Thanks for this offer, and we may reach out to you for future studies.

Reviewer 2

1. I have mixed comments about this work. On one hand the method is interesting. It has been described clearly and it has several possible applications. It is not completely novel though. It roots in N-mixture models and occupancy models accounting for false positives. Differently from this class of models, the one presented here does not need repeated measures, but it does need some true values to account for the uncertainty in the dataset. It is also interesting the insight on parameter redundancy and the use of the calibration methods with unbiased data. Simulations are used to prove model reliability and to answer specific questions on model performance.

Response: We are glad the reviewer sees value in our approach and findings.

2. On the other hand, I think the case study, i.e the illustration of the method, is weak. The application is based on a small dataset. The bias of considering unadjusted counts is negligible, especially in long-lived species, which fecundity has a low elasticity on population growth rate. I consider a bias of 1 offspring in either direction (one less or one more) as negligible. Especially because it is more pronounced in larger broods and because positive biases seem to compensate negative ones. A smaller bias on adult survival probability would have a more serious impact on population dynamics. This is even less important if one considered that about ¾ of the nests were not affected by the bias (19/26).

Response: We have added a Limitations section to the manuscript (l 508ff) in which we discuss the weak bias we used in our simulation. Generally, we do not think this invalidates the approach, but recognize that it is probably more worthwhile to use this approach in situations where that bias is more pronounced. Further, we found in our simulations that for an exponentially growing population, even this small bias in fecundity had implications for the population trajectory. This does not negate that a change to adult survival may have an even stronger impact. But our analysis is not concerned with the relative importance of fecundity vs survival, it is only concerned with potential implications of using biased fecundity estimates in population projections.

3. The part on population projections is also on a shaking ground. First, the consequence of a lower fertility can be calculated analytically using elasticity or sensitivity of the fecundity, which is available in the literature or in the COMADRE dataset, for the same and/or very close species.

Response: While we agree these approaches are useful for evaluating elasticity or sensitivity of population projections to variation in fecundity, that does not invalidate our numerical simulation approach.

4. Second it assumes that i) biases do not depend on observers or habitat and ii) are constant over time.

Response: The likely dependence of observation bias on survey conditions is discussed and extension to model such variation is suggested in lines 474ff.

5. Moreover, the bias is more important when the population is increasing, a situation of least concern in recovering populations.

Response: Growing populations are also of interest in population ecology and conservation science.

6. Finally, a very minor point. If nests (or a subsample of them) are climbed to mark chicks and obtained the calibration data, the bias would concern few nests anyway (here only 27% of them). In species or populations with inaccessible nests model parameters are redundant and the method cannot be applied. This is a minor point because obviously it depends on the study system and it does not influence the model.

Response: We agree that if the nests are inaccessible by a person or drone (see l 76 in Introduction) to obtain a perfect count, then our method cannot be used. We have restated that in the Recommendations for data collection section in l 489.

7. The methods are generally clear, but the simulation part would benefit from more explanations. Possibly a table with dataset characteristics used in each question. For example, in paired datasets with n=10, 25…250 it is not clear if 100% of the datasets is paired or whether the truths concern only an increasing proportion of the datasets.

Response: We clearly define that paired data sets are those in which both ground and climb counts (or uncertain and certain counts) are obtained for each nest, and that combined datasets further contain a set of nests for which only uncertain counts are available (l 181ff). For each Question, we then provide the sample size for the paired and the uncertain data (where appropriate; l 216f, l 231f, l 241f). Moreover, the tables in Appendix 1 contain an overview of all scenarios for all questions, including the number of paired and uncertain-only nests considered. We are unsure what else to add.

Some specific points:

Lines 96. An alternative method, not mentioned here is the double-observer counts (e.g. Nichols et al. 2000 The Auk 117(2):393–408, 2000; Williams et al.,2002 Analysis and Management of Animal Populations. Academic Press. San Diego, CA). In this method there is no need for repeated visits, nor to have a subset of paired counts to make the number of chicks estimable. It is based on the data from two observers that count animals independently.

Response: While we agree that the double-observer method is a single-visit method that can account for false negatives, we are unaware of extensions that allow for false positives. In this manuscript in general and this part of the Introduction in particular, we are concerned with methods that also allow for false positives. Therefore, we have not added in the double observer method here.

Lines 110. Please specify the total number of nests in the dataset. It should be 26 (8 in 2021 and 18 in 2022), is it correct? Do model parameters change between the year? Normally one is interested in a yearly value of fecundity, which stratified the 26 nests even more.

Response: Yes, there are a total of 26 nests, which we added in l 113. We did not stratify the nests by year because this dataset is only used to parameterize the observation process used in the data simulation; we are not attempting an ecologically realistic estimation of fecundity from these data.

Lines 134. This reads as a contradiction. If there is a 1 plus- 1 minus bias in the number of chicks, the classification of an occupied nest based on the observation of at least one chick, is potentially biased as well. The authors inform the reader that the model could be expanded to account for observed empty nests, but I wonder why it has not been done here.

Response: Please see our response to reviewer 1 regarding 0 states and observations.

Lines 263. Due to problematic datasets, it is difficult to have a clear picture of the simulated set of data retained. Would it be possible to add a Table with the simulated vs retained datasets and they characteristics for each question posed?

Response: Please see table S4 in Appendix 1 for a summary of differences in the problematic and retained data. Also see other tables in Appendix 1 for the number of datasets for each *question and scenario that was removed due to missing true categories. If additional information is needed, please provide specific requests for what else to include.*

Line 498. Not always. In short-lived species, only. The population of many long-lived species are not affected by a 5% change in fecundity.

Response: While we agree with the reviewer that in long-lived species adult survival is a stronger determinant of population trajectories than fecundity, we disagree that any changes are necessary in this sentence. Here, we are not talking about sensitivity to changes in fecundity, but we merely list it as one of the parameters determining population dynamics. If one were to build a population model of a long lived species, reproductive success would still be included. Moreover, our sentence clearly states that 'even apparently low bias in fecundity estimates can lead to considerable bias in population projections ***under certain conditions'*** *(ie, not always). This statement is supported by the results of our population projection, where this small bias led to considerable differences in final population size in growing populations.*

In conclusions, I liked the idea and the formulation of the model, but I do not find the illustration very useful. In real dataset, the SE of fecundity measure includes a 5% bias. Stochastic projections that account for parameter uncertainty would probably include the 5% bias in a parameter with low elasticity. I do not see anything fundamentally wrong with this work beside the weak biological realism of the example.

Response: We understand this concern and inserted a Limitations section in the Discussion to address this and some other general issues (l 508ff).