ICES WGMG REPORT 2007

ICES RESOURCE MANAGEMENT COMMITTEE

ICES CM 2007/RMC:04

Ref. ACFM

REPORT OF THE WORKING GROUP ON METHODS OF FISH STOCK ASSESSMENTS (WGMG)

13–22 MARCH 2007

WOODS HOLE, USA

ICES International Council for the Exploration of the Sea

CIEM Conseil International pour l'Exploration de la Mer

International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

H. C. Andersens Boulevard 44–46
DK–1553 Copenhagen V
Denmark
Telephone (+45) 33 38 67 00
Telefax (+45) 33 93 42 15
www.ices.dk
info@ices.dk

Recommended format for purposes of citation:

ICES. 2007. Report of the Working Group on Methods of Fish Stock Assessments (WGMG), 13–22 March 2007, Woods Hole, USA. ICES CM 2007/RMC:04. 146 pp.
For permission to reproduce material from this publication, please apply to the General Secretary. https://doi.org/10.17895/ices.pub.9039

Contents

Executive summary

The purpose of the Working Group on Methods of Fish Stock Assessments (WGMG) is to develop and critically evaluate the models and software code used in assessments, forecasts and management simulations, and to suggest ways in which these might be improved. WGMG meets to address particular concerns raised by ACFM and the Resource Management Committee of ICES. The issues covered by each meeting are a function both of the Terms of Reference, and of the interests and expertise of the participants.

The 2007 meeting of WGMG was held at the Northeast Fisheries Science Center (NEFSC), NOAA, Woods Hole, USA. The principal reason for this was to draw on existing expertise at NEFSC on detecting and accounting for retrospective bias in fish stock assessments. The ToRs for the meeting were extremely wide, and covered many problems currently encountered in fisheries assessment and management science. With the time available WGMG could not address all the ToRs, so following an opening series of presentations of previous and current work, the group was divided into three subgroups to work on more focussed issues.

Subgroup A looked at methods for running management strategy evaluations (MSEs), and started designing simulations to assess how management advice might be affected by errors in assessments (in particular, retrospective bias). Subgroup B investigated ways in which the uncertainty in outputs from assessment models could be estimated. As a starting point, this was done by comparing Bayesian and bootstrap estimates of uncertainty arising from a comparatively simple surplus production model. Subgroup C looked further into the problem of retrospective assessment bias; that is, where each successive annual assessment substantially alters the perception of historical stock in a systematic way (either consistently increasing or decreasing it).

Subgroup A reached three main conclusions. Firstly, WGMG is not yet in a position to answer the questions of whether and how management should proceed in the presence of retrospective bias. The presence of such bias should lead to more cautious management, but how to implement this and how cautious such management should be is less clear. This is due principally to the complexity of programming management-strategy evaluations, but answers to these questions are certainly feasible using current approaches. Secondly, any management-strategy evaluation toolbox must allow for assessments to be run "live" as part of the evaluation loop. And thirdly, managers will get management decisions wrong if these are based on biased advice. This last point may seem obvious, but the analyses presented by Subgroup A highlights the issue with great clarity.

Assessments will always have problems of one sort or another, and it is important that MSEs are able to accommodate this fact. The function of WGMG in this regard is then to provide methods to do this. This endeavour therefore links the work of all three Subgroups.

Subgroup B provided important advances in the implementation of MCMC algorithms for model fitting, and went a considerable distance in generating comparisons of uncertainty estimates from bootstrap and Bayesian methods, with observation and/or process error, using a number of different datasets with different problems (one-way trips, under-reporting and changes in survey catchability). They were able to explore the varying reactions of models to these situations, but firm conclusions remained elusive due to considerable problems in software coding. The Section should be viewed as a strong advance in a work-in-progress. Nonetheless, it seems that not accounting for process errors can lead to a biased view of the true uncertainty in stock estimates based on approximate populations' models. Reliable methods that account for process and measurement errors simultaneously in stock assessment models are not yet available.

Subgroup C used four different techniques to try and detect model mis-specification in six simulated datasets. The techniques were:

1) Pre-screening of data inputs to assessment models.
2) Local influence diagnostics (LIDs).
3) The ADAPT approach with year effects in survey catchability (SPA YE).
4) The ADAPT approach with year effects in a catch multiplier (B-ADAPT)

In the case of LIDs, the method was used to try and ascertain the cause of retrospective bias *directly*; the use of the other methods was restricted to an evaluation of which model mis-specification had been applied (and when), without a concomitant analysis of the effect on retrospective bias (although this would be the next step). Pre-screening techniques can only be used to identify large changes in survey catchability. Similarly, the SPA YE model can only improve assessments when mis-specification of survey catchability is known to be the problem; and the B-ADAPT model performs best when errors in catch are the true source of mis-specification. LIDs suggested that survey catchability changes were responsible for retrospective bias in *all* simulations, even those in which the true cause was under-reporting and/or changes in natural mortality. In addition, correcting assessments using LIDs often removed retrospective bias but resulted in an incorrect assessment. These LIDs cannot therefore be considered reliable indicators for such problems, although they may still have utility when the VPA mis-specification is known to be small. However, a more positive result was that the diagnostics could more reliably detect the timing and direction of the problem when the source was known (e.g. M or survey catchability), especially in the more converged part of the VPA, Such models and diagnostics will perform best when used in combination with a) each other, and b) (more importantly) external information about the likely source of mis-specification.

Finally, analyses of different approaches to calculating a representative average F estimate for a given year were not able to determine any particular method that consistently performed well. Sensitivity of management advice to the method used needs to be evaluated on a case-by-case basis.

The main recommendations from the 2007 meeting of WGMG are summarised above. Of most direct relevance to this year's assessment Working Groups are the conclusions from Subgroup C, regarding testing for and correcting retrospective bias. The work of the other two Subgroups is at an earlier stage, but strong foundations for further work have been laid and plans are in train to continue collaborations. In addition, it was agreed that WGMG was an appropriate forum within which to carry forward certain aspects of size-based analyses; specifically, an exploration of the biases inherent in assuming size-based processes are age-based.

1 Introduction

1.1 Terms of reference (ToRs)

The **Working Group on Methods of Fish Stock Assessments** [WGMG] (Chair: Coby Needle, UK) met in Woods Hole, USA, from 13–22 March 2007 to:

a) investigate further, and test, the sensitivities of stock assessment methods to known data problems with particular reference to the retrospective problem;

b) operationalize methods to include discard data in stock assessments;

c) review developments in fisheries-independent (e.g. survey-based) assessment tools;

d) evaluate the current state of operational evaluation tools for fisheries management options;

e) provide guidance on incorporation in assessments of estimates of variance in input data; and

f) provide guidance to assessment Working Groups on the inclusion of variable weights and maturities in assessments, predictions and management simulations.

WGMG will report by 15 May 2007 for the attention of the Resource Management Committee and ACFM.

In addition to the ToRs, WGMG was asked to address a number of special requests arising from other Expert Groups within ICES. These were as follows, where the ICES Group(s) making the request is given in parentheses.

1) Bias correction in North Sea sandeel forecasts (WGNSSK).

2) John Simmonds' method of determining breakpoints (WKREF).

3) Possible future WGMG involvement in length-based analyses (SGASAM).

4) Evaluation of standard ICES PA advice rule (WKREF/AMAWGC).

5) Quality indicators – summarising results from successive meetings including forecasts (AMAWGC).

6) Survey variances. ICES DATRAS now includes these, for example, and WGMG were asked to consider how best to use them (AMAWGC).

7) A review of the weights-at-age derivation procedures, both for the assessment and forecast, so that appropriate guidance can be given to WG members in the preparation and use of such data and techniques (WGSSDS).

8) An investigation of the appropriate inclusion of varying maturity data in assessments, given that the EU Data Collection Regulation will lead to the provision of such data (WGSSDS).

9) The wider implications of declining abundance of key species to the assessment process (WGSSDS).

10) Sensitivity of estimation of stock-recruit breakpoints to additional data points (WKREF).

1.2 Report structure

The meeting began with a number of presentations on topics related to the ToRs and special requests. The author of each presentation and/or paper was asked to provide an abstract, and a rapporteur was appointed for the discussion sessions that followed each presentation. These abstracts and discussion summaries are given in Section 2 of the report.

Following the presentations, the WG was divided into three subgroups with the following broad remits:

- Subgroup A: Management strategy evaluations.
- Subgroup B: Uncertainty and variance in assessments.
- Subgroup C: Detecting and dealing with retrospective bias in assessments.

A Chair was appointed for each subgroup, and daily plenary sessions of the whole WG were held in which the progress of each subgroup was presented and suggestions made for further analyses. In this report, the work of subgroup A is covered in Section 3, that of subgroup B in Section 4, and that of subgroup C in Section 5. One other analysis that did not fit into the subgroup structure is included in Section 6. Overall conclusions are given in Section 7, with references listed in Section 8.

ToR a) is addressed principally by subgroup C in Section 5, although subgroup A (management strategy evaluations) also explored the influence of retrospective bias on the ability of managers to control stock dynamics (Section 3). WP 2 (Section 2) was presented as the response to ToR b), which however was not explored further within the subgroup structure. Similarly, WPs 1 and 9 addressed ToR c) and are summarised in Section 2; the subgroups did not focus further on survey-based assessments. Section 3 includes a review of the current status of operational tools for management strategy evaluations (ToR d), specifically focussing on FLR, F-PRESS, PROST and POPSIM. An interpretation of ToR e) is addressed in Section 4, following the work in subgroup B comparing Bayesian and bootstrap approaches to incorporating uncertainty. ToR f) was not directly considered in great detail, although one presentation (WP 13; Section 2) suggested that changes in weights-at-age may be one source of bias and noise in assessments of Northern Shelf saithe.

There were a large number of special requests (SR) submitted to this year's meeting of WGMG, and it proved to be impractical to address them all. The requests which were considered at least in part were SR 3 (plenary discussion, and discussion following WPs 11 and 17; Section 2), and SR 6 (indirectly in Section 4 via consideration of variance incorporation in general). An email was received from Martin Pastoors (chair of ACFM) late in the meeting which contained analyses pertinent to SR 5, but in the available time the WG was not able to evaluate this work. The remaining SRs were not addressed, and are not mentioned further in this report. The more general issues of the number of such requests directed to WGMG, whether WGMG should be expected to address them all, and what the purpose of the meeting should be, are considered in Section 7. Section 8 lists references cited in the text.

2 Working papers and presentation

2.1 WP1

John Cotter, Rob Fryer, Coby Needle, Dankert Skagen, Maria-Teresa Spedicato, Verena Trenkel. A review of fishery-independent assessment models, and initial evaluation based on simulated data. Edited by Benoit Mesnil.

2.1.1 Abstract

Large uncertainties in the catch data (official landings and discards) are undermining the ability of ICES and other advisory bodies to provide valid management advice based on the conventional approach of analytical assessments. There is thus an urgent need to consider alternative tools that do not depend on long series of precise catches, with their age composition. This WP presents a number of fishery-independent assessment models developed by the EU project FISBOAT (Fishery Independent Survey Based Operational Assessment Tools). It also reports on rudimentary tests based on simulated data, following the same protocol as an evaluation study conducted by the US National Research Council in 1997. It appears that the available survey-based assessment models are able to reliably capture the major signals in biomass and recruitment, although they smooth out transient changes. However, they cannot provide absolute abundance estimates, but only relative values on an arbitrary scale. Their operationalization in ICES would thus require an adaptation of the advisory framework, in terms of nature of the advice and definition of reference points; indeed, this might be needed anyway, if we were more lucid about the myth of VPA estimates being absolute (which they are not). It is also noted that survey-based approaches have the potential to provide much more rapid updates of the state of stocks than catch-based methods.

2.1.2 Summary of discussion

The presentation is a summary progress report on WP3 of EU project FISBOAT (2004–2006). FISBOAT WP3 aims to analyse fishery independent data to provide managers with relevant information. Fisheries independent methods are not affected by corrupted catch data. Hence though there is limited expert knowledge involved in the modelling process, such methods enhance quick delivery of advice.

The report covers tests of fishery-independent assessment methods, using five submitted models (BREM, YCC, SURBA, TSA, and LENSUR) along with the ALADYM data simulator, and based on probing using the NRC simulated data (in the public domain). The aim is to acquire a sense of potential usefulness. The performance evaluation was on the basis of biomass and recruitment rather than mortalities.

In the absence of data uncertainty, all the models performed relatively well. In general, since the models may be viewed as data smoothers (involving few parameters), their performance in terms of general trend evaluations was good. It can be envisaged that the performance of the models may become degraded when the input data is corrupted with noise.

The general conclusions from the WP3 package indicate the need for more surveys, preferably annual survey data, when catch data are unreliable or unavailable. Further that the methods could be sensitive to inconsistencies in the survey because of the reliance on relative indices.

The plenary discussions looked into the question of whether all the models involved could handle

1) multiple surveys
2) conflicting indices
3) other environmental and ecological covariates, e.g. temperature.

With the exception of TSA and SURBA, it was uncertain whether the other models could handle multiple surveys. However, in the case of conflicting indices, most models have methods of incorporating effects of e.g. fleets, by weighting. Hence it is envisaged that conflicting indices will be handled in a similar manner. None of the models, however, incorporate other ecological covariates such as temperature.

A number of such methods currently in use in the USA were also mentioned. These included a catch-free model for goliath grouper (Porch *et al.*, 2006), a survey-based VPA model applied to cod on the Flemish Cap (Murua at al 2006), and an icefish model applied in the CCAMLR area (see Section 3.4).

2.2 WP2

Colin Millar and David Hirst. Estimating discards at age from discard sampling data by using a Bayesian hierarchical model.

2.2.1 Abstract

The work presented summarises extensions to a model for estimating catch at age from market sampling data (Hirst *et al*., 2004) to allow the estimation of discards (and landings) at age with proper account taken in the uncertainty in the estimates. The extended model is still under development but will, like its antecedent, include covariate effects such as season and gear, and will also be able to account for the within-boat correlation. The model uses observations on the length frequency distribution of the population and observations on the age conditional on length distribution (age length keys). If we have an age-given-length model and a proportion-at-age model then we can build up the likelihood of the length distribution and age conditional on length distributions directly. Such models were proposed. A further complicating factor is that samples are collected either from the discards or the landings, this layers a further conditional variable which, with a discarding-conditional-on-length model (discard ogive) we can again directly form likelihoods for all the data components. Several random effects and fixed effects are proposed in the proportion-at-age model and discards-conditional-on-length model to deal with the effects of space, time, fishing gear and year. A complicated and potentially slow algorithm for sampling from this model to estimate total discards-at-age and landings-at-age was proposed. This algorithm requires further work.

2.2.2 Summary of discussion

The presentation led to a considerable amount of in-depth methodological discussion, not all of which is covered here. It was explained that the model is fitted in a Bayesian framework using MCMC techniques. The author pointed out that the data used in the model are not actually length-at-age distributions, because they are derived from length-stratified sampling schemes, but they are treated as such for fitting the model – the distinction appears to be irrelevant for this purpose.

The basis of the approach is that we have f(a|l), which is a frequency distribution of age given length, and we estimate (via an analogue to a linear growth model) the corresponding f(l|a), the frequency distribution of length given age. It is based on a proportion-at-age model which allows for flexibility. However, the discard model is a function of length only – the response of fisheries to economic or regulatory factors is not accounted for. The simulation of fishing trips is also problematic – it can be done to a certain extent for Scottish vessels because there are good estimates of what a realistic landed yield would be, but for other countries it might be difficult.

A query was raised about how sampling regions were defined. Unlike Norway, where regions are defined on the basis of proximity to ports, in Scotland a different scheme is used (although this was not specified in the discussion). Furthermore, the bootstrap approach used in Norway

does provide similar results, but *only* if there are no missing cells. The way the model homogenises trip autocorrelation was also questioned – this is intentional as the idea is to remove trip-to-trip variation, but might cause underestimation of variance.

The percentage of trips sampled (around 0.1%) was cited as a concern by one of the Canadian participants, who suggested that in Canada 2%, would be considered low. It was pointed out that the Scottish sampling programme on which these analyses are based is actually the most complete in Europe. The model leads to a reduction in the over-stratification previously seen, so increases the relevance of the trips undertaken. It was also emphasised that the purpose of the collation approach is to describe the variance of discard estimates *no matter what* the sampling scheme; the decision of whether to use these estimates in assessments is separate. Finally, it was mentioned that the model does not include weight estimation error.

2.3 WP3

Andrew Campbell. Fisheries projection and evaluation by stochastic simulation (F-PRESS).

2.3.1 Abstract

The F-PRESS (fisheries projection and evaluation by stochastic simulation) model presented to WGMG06 has been further developed and employed in the early stages of the development of a management plan for Western Horse Mackerel.

Improvements to the F-PRESS software have modularised the code, improved the usability of the application and resulted in a (up to) 20 fold increase in performance. Additional graphics routines have been developed for the display of model output, which can also be saved permanently as FLQuant objects, allowing users' familiar with the Fisheries Library in R to use FLR routines for further examination of the simulation output. The software has been compiled into an R package including unit test code, a new revision of the technical documentation, help files, the source code and set-up programs allowing users to install windows GUI applications developed to support the creation and management of F-PRESS input and options files.

F-PRESS simulations for the Western Horse Mackerel stock have been used to demonstrate the important factors that require consideration in the formulation of a management plan for the stock. A number of harvest control rules based on SSB limit and trigger points have been used to demonstrate the relationship between progressively stringent management actions, risk to the SSB falling below a predefined limit point and the variability in yield. Simulations show that the more punitive the harvest control rule action, the higher target yield is available for the same levels of risk to SSB but the more variable the yield becomes. Additionally, the assumption of pulse recruitment with a probability of 1/20 (as appears to be a feature of the stock) significantly reduces the risk to the SSB limit point. Simulations over a range of target yields show that the risk profile changes above target yields of 150kT.

2.3.2 Summary of discussion

There were general questions about the specifics of the Irish Sea cod application, including when SSBs were compared to triggers and limits, how TACs were set, and what was the risk of extinction. The method to model spasmodic recruitment was described. It involved a 1:20 probability of a large recruitment. Otherwise recruitment was drawn from a fitted Ricker curve. A concern was expressed about if the Ricker stock-recruit (SR) function was well estimated. The evaluations were also run using a segmented regression SR model, and that results did not seem too sensitive to the choice of SR model. A question was asked about how population vectors (maturities, weights) were treated. The author described that they were randomized each year in some way, but this could be further refined.

There was concern about *when* stock size was measured for the management plan. The model did not measure stock size after the proposed TAC management action, although it was felt that this was a better approach. The model based TAC on current year stock size which did not directly account for the impact of future proposed fishing. It was mentioned that the "cod recovery plan" specified when biomass should be measured. A dialogue needs to occur between scientists and managers to specify these types of issues. The FPRESS model does not currently include a "live" assessment implementation, which could limit its utility in evaluating management plans.

2.4 WP4

Carmen Fernández. Bayesian methods in fisheries research.

2.4.1 Abstract

This presentation provided an overview of Bayesian statistics and how these methods could be usefully applied in fisheries research. The talk started with a description of the main aspects of Bayesian statistics: the incorporation of prior information into the analysis, and how this is combined with the information coming from the data (encapsulated in the likelihood function) to obtain the posterior distribution, which reflects the knowledge (and uncertainty) that is available after the analysis has been conducted. It was explained that this is a joint distribution on all model unknowns, how to deal with parameter transformations, model uncertainty, and predictions or projections (always probabilistic) of future population trends under different management scenarios. This is a general methodology that can be applied to a wide variety of problems.

Then attention focused on Bayesian population dynamics modelling in the context of fish stock assessment. Several Bayesian hierarchical models were developed all set in the context of state-space models. State process equations model how population abundances evolve in time. In the cases presented, population dynamics were assumed to be deterministic. State equations relate stochastically the observations (survey indices or CPUEs of commercial fleets) to the underlying population abundances. Observation equations provide a framework for the incorporation of the uncertainties associated with the survey indices or catches into the model. Prior distributions should be specified for all unknown model parameters.

Two types of models were considered. The first one had many common features with XSA: it started from (a prior distribution on) survivors and worked backwards in time using cohort analysis assuming the commercial catch is known without error. Tuning indices were incorporated via the observation equations. Results for the particular stock considered were very similar to the XSA results, but an entire posterior distribution was provided, hence there are immediate measures of the uncertainties associated with each of the estimates. Next, catch error models were considered. Starting from prior distributions on yearly recruitment and abundances in the first year of the study, the population is projected forward in time using total mortalities. Independent prior distributions are set on year- and age-specific fishing mortalities (without separability assumptions). Observation equations for the tuning indices as well as for the commercial catch estimates are considered. This permits the incorporation of uncertainty associated with the catch estimates as well as the indices. It was noted that results could be sensitive to the assumed uncertainty associated with each of the inputs (the different tuning indices and the commercial catch estimates) and hence the importance of quantifying these uncertainties and incorporating them in the model was highlighted.

2.4.2 Summary of discussion

Throughout much of this paper, Carmen compared the Bayesian approach to a VPA analysis with the XSA approach.

She noted that priors are weighting factors that are integrated with the likelihood from each model to get the Bayesian model average posterior. To get proper model output it is important to have good quality data rather than a great deal of data. Since the data are used to create the likelihood and are integrated with the priors, it is important to test the *sensitivity* of the priors to the data, along with the sensitivity of the posteriors to the priors.

It would be possible to compare uncertainties using XSA and Bayesian bootstrapping. This would allow one to test the sensitivities of the priors. It would also be nice to look at retrospective estimates using this method.

Carmen indicated that she has not completed the sensitivity testing with her priors.

Coby Needle wondered whether the Bayesian approach provided an appropriate way of accounting for uncertainty and whether it provides useful information for managers. If you have a data poor situation then the priors will be very influential. XSA provides a bootstrapping means of determining uncertainty.

Chris Darby was worried that people were comparing XSA and Bayesian methods. If the answers were similar then people would probably think that the answers were right. It would be more proper to take the comparison further and look at variation in catchability.

Noel Cadigan wondered whether managers would be happy if, for data poor situations, we based our advice on priors. He wondered whether it would be possible to derive meaningful results if the priors were not meaningful to begin with. We require good data in order to obtain valid results using either Bayesian or frequentist methods. Without proper priors, the analysis is basically a frequentist approach and that the MCMC methods are not used solely by Bayesian analyses. He noted that MCMC methods are not unique to Bayesian modellers; frequentists also use MCMC methods in resampling. Regardless of whether one uses Bayesian or frequentist statistics, the objectives for the analyses have to be clearly laid out; the authors have to fully describe what they mean by "probability of outcome".

Coby Needle asked whether the Bayesian version of XSA took shrinkage into account. Carmen was not sure how to account for shrinkage. Noel pointed out that it should be easy to come up with shrinkage on F at the last age in the last year.

Yuri Kovalev noted that Carmen did not go through the second example in which there would be Bayesian model averaging. He felt that model averaging would have been inappropriate because the model would have been over parameterized.

Liz Brooks wondered whether bootstrapping could be used in looking for time trends in residuals. If there were trends in the residuals then you would be introducing biases into the analyses and the confidence intervals would not be meaningful.

Chris Darby noted that Carmen's priors were log normal while bootstrapping indicated that the data were not necessarily log normally distributed. Even though Bayesians often use log normal assumptions they may be poor assumptions.

The log normal distribution for the priors was skewed toward low values. However, Carmen noted that the priors are supposed to be uninformative and are broader than the posterior distributions. The right hand tails were thick.

2.5 WP5

J. Dowden, N. Cadigan, J. Morgan and J. Brattey. Improved estimation and forecasts of stock maturities using generalized linear mixed effects models.

2.5.1 Abstract

Annual biological sampling programs produce data on the number of fish examined, and the number found to be mature, for a wide range of age classes in the stock. A common model used with such data to estimate the proportion mature-at-age (maturities) is logistic regression. This is a generalized linear model with a logit link. The most appropriate way to produce such estimates is by cohort; however, there are problems with this approach. Data are updated annually for unfinished (e.g. recent) cohorts and this can result in substantial changes from year to year in the estimated cohort maturities. For example, the estimated maturity at age 5 for the 1998 cohort based on data up to 2004 can be quite different than the estimate based on data up to 2003.

Often the annual trends in cohort maturities are fairly smooth. The purpose of this paper is to explore how to utilize the autocorrelation structure in cohort maturities using a GLIM with autocorrelated random cohort effects to improve the estimation of maturities, particularly for unfinished cohorts. We apply the method to a case study involving Atlantic cod (*Gadus morhua*) in NAFO Subdivision 3Ps. Fisheries managers often consider changes in SSB in stock projections for different future management scenarios, which requires that maturities be forecasted in the next several years (or more) to compute SSB's. We also investigate if the approach can improve forecasted maturities.

2.5.2 Summary of discussion

Discussion points on this presentation fell into four categories: size effects, implementation and interpretation of year effects, patterns in maturation over time and model residuals, and skipped spawning events.

Regarding size effects, several questions pointed to maturity most likely being a function of length, and inquired whether the observed changes in maturity could be explained by different growth rates (related, perhaps, to cohort density). Alternatively, aging effects could contribute to the noise. The author noted that, if year effects are real population effects (synchronicity in the decision among immature fish in all or most cohorts to become mature, due to varying environmental conditions) then this reduces the predictive capability of the model. WGMG noted that was that the synchronicity could still be due to growth effects, with cohorts with individuals of overlapping sizes deciding to mature at the same time. This would give the appearance of year effects.

An assumption was made in the method that maturity was a monotonically increasing function, i.e. once a fish decides to become mature, it is an irreversible decision and maturity continues to increase with age. The issue of skipped spawning was raised, suggesting that such events would violate the monotonically increasing assumption. The presenter acknowledged that it is implicitly assumed that no skipped spawning occurs. This is one reason why the method does not address the deeper problem of using SSB as a proxy for realised egg production. Also, simply putting a smoother through the time series would violate the monotonicity in maturity.

There were a variety of questions on the implementation and interpretation of random year effects. At present, the model only incorporates a random effect in the intercept of the maturity function. WGMG questioned why it wasn't also considered for the slope as well. The presenter responded that there is no reason why it couldn't be considered; however, for this work, it was not. The meaning of a year effect was questioned, and the author responded that

there was no mechanism implied, it could be in the population or it could be in the sampling. This was followed by a comment that the source of a year effect would be important for both retrospective and predictive understanding of SSB. If the main source is sampling error, the model could be simply tracking noise and it would be wise to use a smoothing model rather than raw data, or one could try to sample the same area for all time periods. Regarding forecasting, and whether year effects were used in predicting future states, the author responded that no year effects would be included in forecast years. A question was asked whether the predicted year effects line up with the deviance residuals on the slide "a problem: year effects." The author responded that it lines up a bit, and that is what bothers him about the approach—namely, whether year effects are confounded with cohort variability, although maybe there is potential to separate the two if year effects are similar for cohorts within a given year. One participant inquired whether cohort errors were assumed random and correlated, and the presenter responded that they were but were not over-dispersed; the hypothesis is that year effects are correlated but that they are not the source of the over-dispersion.

It was noted that the pattern in maturity rates by age consistently increased over the time series, and WGMG questioned whether this could correspond to increased exploitation rates. The implication is that fish are maturing earlier in response to exploitation reducing density. Alternatively, the growth characteristics of early-maturing fish could be making them more vulnerable to capture, and hence appear more frequently in catch and survey data. A related question was whether natural mortality had remained constant over time. The presenter responded that he did not know, but that one would expect the decision to mature earlier would invoke a trade-off with higher natural mortality.

Minor points:

- The presenter noted that on the slides "Mixed Effects Model" and "Random Year Effects (YE)" there should be no 'c' subscript;
- On the "FE Cross-validation slide", the pattern shows larger residuals in middle ages; this seemed counter-intuitive to several in the audience, and the presenter suggested that he would need to look closer at how the student responsible for the slide had generated the plot
- The fixed and year effect model differences look trivial on the plot "observed and predicted proportion mature"; the author responded that it might be better to show those plots by age rather than for all ages, because there are not a lot of cases where the year effect model is fitting better (referring to "total fit on all ages" slide where FE vs YE chi-square residuals are plotted against the 45◦ line)
- Clarification was provided for the cross-validation approach; it was essentially a jack-knife procedure, where one age was removed then the fits to all of the remaining ages provided the model prediction for the age that was removed. This approach was limited to ages 4–8, the "dynamic age range of the maturity ogive' because the presenter wanted to avoid ages where the maturities were close to 0 or close to 1.

2.6 WP6

Alan Seaver. NOAA Fisheries Toolbox Version 2.10.

2.6.1 Abstract

The NOAA Fisheries Toolbox (NFT) is a collection of programs for use in fishery stock assessment. NFT represents a major revision in the design of the fisheries toolbox concept wherein the graphical interface and the calculation engine are independent. In earlier implementations (i.e. FACT and WHAT), the toolbox consisted of a single program with many subordinate models accessed from a single graphical interface. In the new toolbox design each model is an independent application. This creates a more robust and expandable

framework, eliminates critical dependencies among models, and allows for distributed development of models at various research sites.

Communication between calculation engine and graphical interface is through ASCII text input and output files. This approach has a number of advantages for development, testing, and implementation of assessment models. This approach assures the preservation of input integrity. Calculation issues can be kept separate from graphical issues, since the calculation module can be developed independent of the graphical interface, and the development can be done in different offices. The NFT website is http://nft.nefsc.noaa.gov, which currently requires a login (nft) and password (nifty).

2.6.2 Summary of discussion

There was a suggestion to modify the consumer reports model to allow for lagging of time series. The EU Fisheries Library for R (FLR) approach is a collection of libraries and packages for R that has a similar purpose, however, a lot is still in testing mode and a lot of training is required because one needs to know a lot about R in order to use it. An advantage of the FLR approach is flexibility. NFT is a managed package and so has version control, testing, and quality control but has the disadvantage of not being as flexible. For this reason, NFT is developing R interface to output results directly to R to allow for this. There can be issues of testing with any sort of toolbox, and there needs to be more rigorous in testing of all models. FLR has test datasets with each unit that do automatic testing when making any changes to the module. The NFT is not designed to stifle creativity, but rather to allow easy access to a wide range of fishery stock assessment methods. The use of population simulators linked with the different models allows for case-specific testing of assessments.

2.7 WP7

Chris Legault, Bob Mohn and Larry Jacobson. A quick overview of retrospective analyses from NEFSC.

2.7.1 Abstract

Analyses of retrospective patterns at the Northeast Fisheries Science Center, in conjunction with our colleague at DFO (Bob Mohn), have focused on simulation studies. The ability to produce datasets that exhibit a retrospective pattern when assessed with standard tools has been demonstrated. Large changes in the time series of data are required to produce retrospective patterns similar to those seen in actual assessments. Retrospective patterns in simulated data have been caused by changes in simulated survey catchability, natural mortality rate, and under-reporting of catch. Two metrics have been used to measure the retrospective pattern; ρ_{tip} and ρ_{path} which both compare the results of assessments with truncated time series with the full time series. The local influence surface (LIS) approach of Cadigan and Farrell (2002, 2004) has been applied to fix the retrospective patterns. However, when the three by three cross of retrospective source and fix was analyzed, all nine cases had the retrospective pattern removed, but none were corrected to the underlying truth. Furthermore, the correct source of the retrospective pattern could not be identified. Thus, even though the combination of ρ_{path} and LIS is able to remove a retrospective pattern, it is not recommended for use in assessments because the fix does not correct the results to the underlying truth. Another set of analyses demonstrated the dependence of the LIS using ρ_{path} on the number of years removed in the retrospective calculations as opposed to the timing of source of the retrospective pattern. Finally, the inclusion of only random noise can cause some retrospective patterns, but when a simulator was set up to mimic a specific stock assessment the observed level of retrospective pattern could not be produced by noise alone. These case specific simulations are recommended to allow determination of whether or not random noise could be the source of an observed retrospective pattern.

2.7.2 Summary of discussion

The contrasts and similarities between the local influence of various parameters or data sources on the retrospective statistic and sensitivity analysis and even likelihood profiles was noted and commented upon by several in the WG. Noel Cadigan and Chris Legault emphasized that the goal of the local influence analysis is to correct mis-specified input. Coby Needle noted that the retrospective pattern can arise from differing signals provided by different data or inputs. Chris Darby cautioned that for long time series recent assessments incorporating recent catch information can provide poorer inferences on the system (fishery and population) than historic assessments because of more recently developing issues such as misreporting or underreporting of catches. Tim Miller noted that there may be substantial variation of the VPA and local influence results from time series to time series and that it is important to consider whether the process provides biased estimates of important attributes on average over time series. Noel suggested that exploring the retrospective patterns of attributes after corrections are applied would be worthwhile. Noel Cadigan also suggested that other metrics of retrospective pattern might be better because patterns may exist when the currently used metrics imply no pattern. It was pointed out by several participants that retrospective bias is a *symptom* of problems, not a diagnostic.

2.8 WP8

Chris Darby. BADAPT: a modification of Adapt used to estimate unallocated mortality

2.8.1 Abstract

Darby (2004, 2005) modified the approach of Gavaris and Van Eeckhaute (1998) to estimate removals of North Sea and Irish Sea cod. VPA models fitted to the catch at age and research survey data, under an assumption of unbiased catch data, indicated a mismatch between population abundance derived from the catch-at-age and CPUE data from two research survey series, identified by a step in the times series of log-catchability residuals. If the assumption is made that historic catch at age data were unbiased and that survey catchability is constant, a year effect in the form of a multiplier on reported catches could be estimated. The time series of estimates of adjusted total catch were consistent with anecdotal reports and information supplied to the working group on the level of unrecorded landings.

Whilst unrecorded landings are considered to have been significant for the two stocks, the estimated unassigned removals could not uniquely be attributed to under-reporting bias as similar effects could result from a trend in natural mortality and/or discarding and survey catchability. The model was reviewed and used to provide management advice by the North Sea and Skagerrak Demersal Working Group (ICES, 2004c, 2005e) and the Working Group on Northern Shelf Demersal Stocks (ICES 2005c, ICES, 2006c).

Subsequent testing of the model using simulated data has shown that the model provides an approach for adjusting catch at age data when unaccounted removals (un-recorded discarding, under reported catch, additional natural mortality) affect the stock.

2.8.2 Summary of discussion

First, one should have information to find out what aspect of the input data is the most likely to need changing. These could include catch data, survey indices (due e.g. to varying catchabilities over time), natural mortality assumptions, etc. The presentation did not use ρ statistics values as a measure of retrospective pattern.

Earlier work tried to estimate a separate natural mortality parameter for every year, but found out that not all the model parameters could be estimated in that case. The solution that was found then was to fix recruitment in one year. Once this was done, the remaining model parameters could be estimated.

The case study considered here is the retrospective pattern in the North Sea cod assessment. It was thought that there could have been unallocated removals in recent years, causing this bias. A trend in survey log-catchability residuals could also be observed, with higher values in the last few years. It was felt that the catch data were appropriate up to a certain year, but needed year-specific factor multipliers thereafter. These (age-independent) catch multipliers can be estimated in the BADAPT setting, after having chosen the year up to which the catches do not need the multiplying factor (this year was finally chosen on the basis of the results obtained under different values for it, seeing at which point the multiplier factors starting to be fitted to values different from 1). The key point is that BADAPT assumes that surveys are correct, and scales catch data to agree with survey-derived population trends.

First, catch was estimated with no smoothing, and this led to estimates of unallocated removals. Then, smoothing parameters for catch or F were considered, obtaining improved estimates.

1000 simulated survey datasets were considered, and it was observed that the method worked well, without need for smoothing. The residual pattern of the log-catchability residuals was corrected.

For North Sea cod, estimates of bias in catches were presented. For Irish Sea cod, where direct estimates of missing catch components exist, the method was seen to produce good estimates.

The procedure was used to provide advice, and it was noted that it was difficult for managers to handle estimates of misreporting and uncertainty estimates. A possibility is to look at multipliers of fishing mortality and examine how these would affect prediction in subsequent years.

It was discussed that it would be interesting to try Chris Legault's simulated datasets with this method.

2.9 WP9

Coby Needle. Summary of SURBA 3.0.

2.9.1 Abstract

SURBA is a simple separable model of mortality in which parameter estimation is based on research-vessel survey indices only. It is based on the RCRV1A model first presented by Robin Cook (1997, 2004), and has been considerably developed since (Beare *et al*., 2005; Needle 2003, 2004a, 2004b, 2005). Although it has been investigated during the previous two WGMG meetings, it was thought to be worthwhile to present it here as it was not familiar to the American participants.

The presentation summarised the data required by the model, the methods used within it, the parameter estimation approach, and some examples of graphical output produced by the Windows GUI available for the program. SURBA is being developed on an irregular basis (both independently and under the FISBOAT project, see WP 1), and the presentation finished with a summary of the key issues to be addressed in the near future.

2.9.2 Summary of discussion

SURBA 3.0 is a multi-parameter, statistical survey-at-age model based on original ideas in RCRVIA (Cook 1997). The modelling depends on exponential cohort decline, separable mortality and abundance at age data. There is an assumption of proportionate relation between abundance and survey index. The model performance index is based on age-structured indices, biomass indices and a penalty term to smooth out year effects. The model assumes no stock recruitment referencing and has the same file structure as used for XSA and related programs.

SURBA 3.0 provides an option to scan over run-settings and limited sensitivity analysis.

The model has analytical uncertainty estimation of total mortality and recruitment. Uncertainty in SSB is absent, due mainly to coding problems. Retrospective analyses are conducted back to time corresponding to half the earliest survey available.

The discussion centred on limitations of SURBA 3.0 as a model tool and as a source of information on which to base management advice. These are particularly apparent when the model is applied to flatfish (for which catchability by age is generally dome-shaped), especially with respect to uncertainty. A typical uncertainty range for total mortality, Z, for such stocks can be -100<=Z<=100, and this problem needs to be addressed.

The main question raised is about how catchability is determined. The discussion indicated that if catchability is fixed with time, it does not matter as much as the trend. However, if the catchability is pitched relative to a management parameter, e.g. Blim, changing the catchability at a certain age might move the reference point. Estimating catchability on the basis of survey alone is impossible. The effect of variable catchability could also explain the overestimation of (for example) the uncertainty bound in total mortality. The overestimation may be linked to the non-existence of trend or changes in the mortality.

Application of SURBA 3.0 to management advice is difficult because the modelling results are relative rather than absolute indices. However, trend-based management is tractable using such an approach.

Future work will include aspects such as FLSURBA (an implementation in FLR), extending the analytical uncertainty estimation, and improving the scanning procedure and component weighting (inverse-variance reweighting).

2.10 WP10

a) *David Orr. Using an empirical traffic light procedure for monitoring and forecasting in the Gulf of St Lawrence snow crab fishery.*

b) *David Orr. Northern Shrimp (Pandalus borealis) off Baffin Island, Labrador and Northeastern Newfoundland.*

c) *E. Colbourne, J. Craig, C. Fitzpatrick, D. Senciall, P. Stead and W. Bailey. Northwest Atlantic climatic update for 2006.*

2.10.1 Abstract

The traffic light approach was introduced as a means of presenting and summarising changes in stock or environmental status. Three examples of performance reports were presented:

1) Southern Gulf of St. Lawrence Snow Crab;

2) Northern Shrimp off the eastern coasts of Labrador and Newfoundland; and

3) Climatic conditions as presented by the Atlantic Zonal Monitoring Program (AZMP).

Each set of performance reports made use of their own metrics for determining within parameter changes in colour. The Snow Crab researchers divided the data into three equal portions; colour within the northern shrimp data was determined by whether it was above, within or below the 95% confidence intervals around the long term mean for that variable, while the climatic data made use of z-transformed deviations around the long term mean. The Snow Crab and shrimp reports made use of three simple colours; red, yellow and green. Climatic data were presented on a colour scale from dark blue for strong negative deviations to light blue and pink for minor deviations and dark red for strong positive deviations. Regardless of the metric used, the trends within several parameters could be clearly presented on a single page. Through usage of the appropriate time lags, it was possible to create a

forecast mode. Such reports need not be limited to data poor situations; the method is equally suited when there are rich long term datasets and can include model output.

However, the method does not include an objective means of determining weights for individual parameters. A weighting scheme is important because not all variables should have equal importance. For instance, trends in long term fishery independent biomass and recruitment indices are critical to overall status and could merit a higher score relative to trends in fishery dependent indices which are less reliable as stock indicators.

Additionally, it is important that one takes care when choosing the appropriate parameters. It is possible to bias the method by loading the report with positive or negative parameters.

Since the goal is to maximize objectivity in the assessment process; the assessment biologist must produce an objective scoring of overall stock status. This final scoring system has to be easily interpreted by fishery managers.

Even though not demonstrated within the presentation, it was pointed out that the NOAA Fisheries Toolbox (WP6) includes a visual report that is similar to the Traffic Light performance reports.

The fact that several methods could be used to provide an easily interpreted report showed the flexibility of the presentation method.

2.10.2 Summary of discussion

David Orr confirmed that the Northern Shrimp biomass and abundance estimates are derived from survey results (question from Coby Needle). An ogive analysis and triangulation method is used. Bootstrapping is used to derive uncertainty estimates. A paper is available describing the method. Coby Needle identified this work as a contributor to the ToR c) since the biomasses are estimated from survey information, or ToR d) as it is essentially a management tool. It was then decided that further work on the traffic light approach would be undertaken during the WGMG meeting under ToR d).

Carmen Fernandez queried the selection of the weights used for each of the indices. David Orr explained that an upcoming workshop will discuss this but for the moment personal experience and preference has been used. He agreed with Coby Needle that a sensitivity analysis would be a valuable exercise.

Coby Needle noted that the Northern Shrimp area has not increased with the catches. David Orr commented that the exploitation rate has not changed much from 15% and explained that the exploitation rate here is calculated as the catch over the previous year's fishable biomass (in response to Noel Cadigan).

Coby Needle expressed concerns with continuums of identical indicators. A number of indicators that are just green will produce a green assessment. There are parallels with the reference approach where SSB can be just above or below Blim which can result in significantly different management strategies being pursued. David Orr commented that if there were a number of indicators that were 'just green' then there would also likely be some that were red or orange. Chris Legault drew comparison with the management indicator plots presented earlier and expressed concern with the relative short time series and the dangers of small changes which could be due simply to error resulting in a management regime change. David Orr commented that the fact there is no dynamic range for the northern shrimp is an added complication. He also added that if there are model results then they can also be included as an indicator with a heavy weighting if the confidence in the result is high. Chris Legault made the point that if all that is available is a poor time series then it is not possible to devise a traffic light scheme.

It was suggested that perhaps more structure is required in the scheme e.g. 'very green'.

Noel Cadigan expressed concerns that this approach is trend based and does not provide managers with a tool on which to set TACs. David Orr mentioned that a forecast is available using the lags that are in the system. It was agreed that a management plan would have to be in place which defines the actions that should be taken depending on the traffic light indication, and that this would need to be tested via simulation.

It was mentioned by several contributors that a model formulation is required. David Orr replied that the approach is open to the use of a model but the short time series and lack of dynamic range make this challenging at present.

It was mentioned by several and widely agreed that the choice and weighting of indicators is very important. Inappropriate indicators or multiple similar indicators need to be either eliminated or weighted appropriately.

2.11 WP11

Dvora Hart, Larry Jacobson, Toni Chute, and Alan Seaver. Length-based assessment models.

2.11.1 Abstract

Length- (or size-) based assessment models have primarily been used in cases where age information is lacking, such as in invertebrate assessments where ageing is difficult or impossible. Based on earlier work by Sullivan *et al.* (1990), we developed a statistical catch-at-size (CASA) model for use in the Atlantic sea scallop (*Placopecten magellanicus*) assessment. Our model uses a growth transition matrix from shell increment data to project shell heights forward in time, and uses a maximum likelihood approach to fit survey, landings, and catch-at-size (height or length) data. Interesting aspects of the model include: 1) provision for priors on survey gear sampling efficiency; 2) explicit characterization of measurement errors (in addition to sampling errors) in size data; 3) robust likelihood calculations for noisy survey abundance data; (4) ability to use shell increment data directly, or as tuning data for a growth model in deriving the growth transition matrix; and (5) estimation of per recruit reference points (*i.e.,* F_{MAX}, $F_{0.1}$, and $F_{40\%}$) as model parameters so that variances of status determination ratios (*e.g.* F_{Recent}/F_{MAX}) can be directly computed. Preliminary results, using both actual and simulated data sets, are encouraging and show little or no retrospective bias. The CASA model may be an attractive alternative to age-based assessment models when survey and catch length (or height) data are available.

2.11.2 Summary of discussion

Disadvantages of length-based assessment models were not discussed in detail during the presentation, but acknowledged to exist, particularly difficulties due to changes in growth rates over time. In theory, these changes could be incorporated into the models, but would require data to support them. Future work could explore density dependent changes in growth and their impact on length-based assessments as well as how estimates of F change with changes in growth rates during the time series.

If in the future the WGMG had a TOR for length-based models, some possible topics are:

- comparison of age-based and length-based assessments using real or simulated data, possible examples include the North Sea whiting and Gulf of Maine winter flounder assessments because both have issues related to age estimation;
- examination of situations that cause estimation difficulties for length-based models through simulation;
- the impact of fishing, especially targeting of size classes, on estimates of growth and creation of growth transfer matrices for length-based assessment models;

- the bias inherent in assuming length-based biological processes are age-based, as is commonly done in stock assessments.

The scallop assessment conducted using the CASA (catch at size analysis) approach was preferred to catch survey analysis because the former utilizes the large amount of length frequency data available while the latter has a single cut-point between pre-recruit and recruit stages that is not well matched by the data.

The software program (SAMS) used to generate the datasets used in the scallop simulations is currently not available because it is still undergoing testing and modification. The goal is to incorporate this simulator, along with the CASA model and Gedamke and Hoenig non-equilibrium length-based Z estimator, in the NOAA Fisheries Toolbox once they have been fully tested.

2.12 WP12

Tim Miller, Chris Legault and Paul Rago. An experimental approach to Bigelow calibration.

2.12.1 Abstract

When survey vessels are decommissioned and replaced with new ones, a need arises to make catch rates for each vessel comparable. Paired haul and area-based methods have been used to estimate conversion factors (Pelletier, 1998). Lewy *et al*. (2004) presents a useful paired-haul design that can be used when the old vessel must make the first tow in the sequence. This is important for an upcoming switch in vessels for the Northeast Fisheries Science Center because it is thought imperative that this year's survey be directly comparable with those of previous years. The two tow sequences that must be performed for application of the Lewy *et al.* (2004) design are the decommissioned vessel twice or the decommissioned vessel then new vessel. Given that there are two types of sequences that can be performed at each station, a total number of stations for the experiment and the costs (in time or other resources) of the two types of sequences, the optimal allocation of stations to each sequence type is a function of the disturbance of the starting density at the station and the ratio of the catchabilities of the two vessels. We presented the relationship between the allocation and the resulting coefficient of variation to these parameters and the costs of the two types of sequences for the case where there are either one or two new survey types and emphasize the need for preliminary studies to determine disturbance and catchabilities we should expect for the full experiment.

2.12.2 Summary of discussion

The WG questioned whether this rather complicated approach gives any benefit over simple paired trawls. However, paired trawls are not practical in this case – for example, the new vessel must be towed into the current, and is thus not always compatible with the old vessel. The method presented is an attempt to try and circumvent difficulties such as this. Similar work as been done in Europe and Canada, and a number of potentially relevant papers were mentioned although there is generally no clear solution to this question.

2.13 WP13

Chris Darby. Growth of saithe in ICES Sub-Areas IV and VI.

2.13.1 Abstract

Abstract not provided.

2.13.2 Summary of discussion

The question posed was whether the substantial decrease in weights-at-age in recent years for the North Sea saithe stock has affected the catchability of these fish, and hence caused the

retrospective pattern observed in the assessment. It was noted that the Faeroe saithe data had similar features (e.g. the time trajectory for mean weights at age 6 show similar trends for the two stocks). Estimates of $F_{0.1}$ on a year-to-year basis (using year-specific weight-at-age and exploitation patterns) is highly variable, and poses questions about how to manage the stock using long-term F targets, particularly in the context of the EU-Norway management plan, which is up for renewal this year (also for the NEA stock of saithe).

It is possible that with weights-at-age effects, retrospective patterns won't be seen in the assessments because these are essentially numbers-based, but retrospective patterns may well be seen in forecasts because weights-at-age are used for these. However, if changes in weights-at-age over time affect selectivity, or behave as proxies for changes in selectivity, then these weights-at-age changes may well cause retrospective patterns in the assessment. It was pointed out that weights-at-age for this stock are based on length-at-age, transformed via a fixed (time-invariant) weight-length relationship. Changes in feeding and condition are therefore not directly addressed.

There are other possible causes of retrospective patterns for NS saithe, such as the absence of discard data in the assessment for the period when discarding took place (when saithe were of lower commercial value historically) – i.e. discarding practices may have changed over time.

2.14 WP14

Noel Cadigan. Local influence diagnostics for the retrospective problem in VPA.

2.14.1 Abstract

Local influence diagnostics (LIDs) are metrics that describe the effect of small perturbations of model components on important model results. They can be used to find changes in model inputs that have large effects on outputs, and they can be used to find changes to VPA inputs or assumptions that remove retrospective patterns. We postulate that retrospective patterns are another manifestation of residual patterns, and we propose that a good output statistic to examine in diagnosing the source of retrospective patterns is a measure of the size of the residual problem. Retrospective patterns are almost always associated with time trends in residuals, so we examine the mean square average annual residual (MSAAE) as a measure of the size of the residual problem. We examine perturbations to VPA catches, natural mortalities, survey catchabilities, and estimation weights to find changes in these inputs that remove residual patterns and also retrospective patterns. We applied the approach to six simulated case studies with mis-specifications that resulted in retrospective patterns. Unfortunately in all six cases the diagnostics suggested that smaller changes in survey catchabilities could reduce the residual patterns than could changes in the other components. This was independent of the real source of the problem and suggests that the approach cannot be used to diagnose the source of the problem. In addition, bootstrap results suggested that the four perturbation schemes were not directly comparable. For example, a multiplicative perturbation to catchability appeared to be a larger perturbation than a multiplicative perturbation to catches. However, a more positive result was that the diagnostics could more reliably detect the timing of the problem.

2.14.2 Summary of discussion

The discussion on WP14 covered the utility of the MSAAE objective function, and issues related to the basic design of the LID approach.

The utility of MSAAE was questioned due to the fact that it did not, like Mohn's rho, take into account any trends in time. An appropriate solution was not provided; however, an objective function that minimised autocorrelation in the residuals in some way was suggested as a possible direction.

An alternative approach to LID was put forward where the model is altered to accommodate suspected problems in the data; however, the author suggested that this was more of a solution to the problem and not a method of finding the problem. This particular application of LIDs was intended to find changes in the data that would reduce some measure of retrospective bias to zero, based on an underlying dynamics model which was believed to be true. The changes that were selected would hopefully inform on the source of the retrospective problem.

A problem with the approach was then highlighted by the author: the LID approach seeks to find the smallest perturbations to the data that reduce the objective to zero. There will of coarse be many changes to the data that will result in the objective being zero, but LID finds the most local change, thus if the true misspecification is not local then LID will not find the cause if it can be explained by smaller perturbations of some other data points. It was suggested here that one might try and bound the problem so that only realistic changes to data be allowed as it was noted that some results had indicated unfeasible changes as causes of the retrospective bias.

It was concluded that LIDs using MSAAE are good at finding the timing of events in the data that cause retrospective bias, but that the current state of knowledge does not allow us to identify the source of the retrospective bias. It was also suggested that a corresponding analysis of a simpler model might be beneficial, if said model was amenable to analysis by analytic means.

2.15 WP15

Chris Legault. Treatment of null values.

2.15.1 Abstract

Many stock assessment models use time series with assumed lognormal error that can have an observation of zero. For example, consider a research survey which does not encounter any age 1 individuals in 1995. This does not mean that there were no fish in that cohort, but rather that abundance, availability, and gear selectivity combined to produce an observation of zero. The standard procedure at the Northeast Fisheries Science Center (NEFSC) is to treat these null values as missing in stock assessments. Others advocate replacing the null with a small positive value c. Often $c = 1$, $c = 0.01$, or a rule is used which sets c equal to one sixth the smallest non-zero value in the series and adds this value to the entire time series. Treating these null values as missing avoids the issue of which value to use for c but could potentially bias the assessment by not providing information to the model. When there are multiple null values in a time series, replacing them all with a single value provides not just information about magnitude but also trend, which may or may not match the true trend in the population.

Simulations were conducted with survey values below an arbitrary level set to zero and these zeros then treated as missing, replaced with 0.01, or replaced using the one sixth rule. The $c = 0.01$ case performed poorly with highly biased estimates of N and F. Treating the zeros as missing performed best, with the one-sixth rule performing only slightly worse. A second set of simulations was conducted which added a second set of tuning indices which did not follow the true population trend under the assumption that this would be a harder test for the missing case. The results from this second set of simulations were hard to interpret. Treating the values as missing produced results closer to the case when all the values were used, but further from the underlying truth than the one-sixth case. The $c = 0.01$ case was quite highly biased for many more years than the missing case or one-sixth case. Given the inclusion of the biased indices, it is not clear whether the results using all the data or the underlying truth should be the basis for comparison.

The solution to the problem appears to be use the of a different error structure that allows for null values. Simulation testing will be required to demonstrate that such an alternative error structure is robust to outliers.

2.15.2 Summary of discussion

WGMG considered the use of replacing values with probabilities as part of the observation equation. Delta approaches were suggested as a possibility, whereby the probability of positive values is estimated separately from the magnitude of the positive values. Noel Cadigan mentioned that he encountered problems when trying this approach. Noel recalled the application of a quasi-likelihood function with a quadratic term as an approach to potentially reconcile this problem. This approach allows for null values and is close to the lognormal error distribution for positive observations. However, this approach requires the estimation of an additional parameter. The standard procedure in ICES is to treat null values in survey time series as missing. WGMG concluded that there is no simple solution to this problem, owing to the fact the log-normal distribution is an inappropriate error structure for these types of data series. WGMG stated that one should not change data to fit the model, but rather change the model to fit the data.

2.16 WP16

Coby Needle. HCR evaluation for North Sea haddock.

2.16.1 Abstract

The plan agreed between the European Union and Norway for the management of the North Sea haddock fishery was reviewed during 2006. This presentation summarised the scientific analyses carried out as part of this review. The three-step evaluation loop stipulated by ICES, 2006b, was presented, along with its application to the management plan in question which was shown to be logically incomplete. The modelling framework was discussed, based on R code with FLR objects and functions, and the conclusions of the analysis were summarised: namely, that a) a target fishing mortality of 0.3 with a 15% limit on interannual variation in TACs leads to a low risk to Blim (around 5%), and b) increasing the target fishing mortality above 0.3 leads to an increased risk. The agreed revision to the plan includes a sliding-*F* harvest control rule, which has not yet been evaluated. It was also emphasised that some of the model assumptions (regarding growth in particular) are somewhat crude. Further work will focus on the issues, along with optimisation of code to permit more extensive evaluations.

2.16.2 Summary of discussion

The WG asked about how the mean F was calculated - it is a straight average over ages 2 to 4. The author indicated that the stock assessment was run as part of the HCR simulations, and this was why *F*s, biomass, and recruits changed in the HCR iteration presented. Coby indicated that the assessment mean *F* varied about the target because of the 15% rule on changes in TAC, and because the mean *F* was for a forecast and the subsequently observed *F* would differ somewhat from the target. A question was also asked about how discards were handled; these are modelled as a fixed proportion at age, which of course is a simplification. The stochastic output presented from the HCR was both actual simulated population numbers *and* assessment estimates. Did the rule specify action when stock was below B_{lim} or B_{Pa}? The author answered no, because the action to be taken in these circumstances had not been specified in the management plan, and suggested there should be some consideration for this in the control rule.

2.17 WP17

Richard Methot. Stock Synthesis 2: integrated analysis of fishery and survey size, age, and abundance information for stock assessment.

2.17.1 Abstract

Stock Synthesis 2 (SS2) provides a statistical framework for calibration of a population dynamics model using a diversity of fishery and survey data. SS2 is designed to accommodate both age and size structure, and multiple stock sub-areas. Selectivity can be cast as age specific only, size-specific in the observations only, or size-specific with the ability to capture the major effect of size-specific survivorship. Growth is modelled using *morphs*, discrete population subunits that share growth characteristic (equivalent to super-individuals in a Lagrangian sense). The overall model contains subcomponents which simulate the population dynamics of the stock and fisheries, derive the expected values for the various observed data, and quantify the magnitude of difference between observed and expected data. Parameters are searched for which will maximize the goodness-of-fit. A management layer is also included in the model allowing uncertainty in estimated parameters to be propagated to the management quantities, thus facilitating a description of the risk of various possible management scenarios. The structure of SS2 allows for building of simple to complex models depending upon the data available. SS2 is available in the NOAA Fisheries Toolbox for the first time with a graphical user interface.

2.17.2 Summary of discussion

There was a general discussion regarding the use of length based models and whether it would be worthwhile for WGMG to consider them given that within ICES there is a distinction between the length and age based approaches with most people not having worked with length based methods. If they were to be considered by WGMG then the group would have to be broadened.

The author offered an opinion that the lines between the two are blurred. SS2 is basically an age-based model set up to run in a length-based manner, and has much in common with simulation approaches such as Gadget. It is extremely flexible and can be run without any age data. However, it may not be necessary to go to a length base approach if good age information is available. Much depends on the quality of the available data.

The model is part of the NOAA Fisheries Toolbox demonstrated earlier. It is constantly being updated and the demonstrated model (2007 updates) will be included in the toolbox in the near future.

There is a challenge with using the model's movement dynamics implementation with an increased number of parameters. The inclusion of tagging data can be used to help here but often it is necessary to make some assumptions about selectivity.

The time step of the model is flexible with a capability to include multiple seasons to incorporate differences in timing between surveys and fishing.

The point was made that the weighting of each of the data sources is important. The author described an approach whereby externally estimated variance is used to determine the weights. However, an iterative weighting approach has also been adopted. Variances are adjusted based on initial model runs with emphasis factors available to modify the weighting.

The recruitment realisation in the model is defined in such a way as to include process error. The growth specification is less advanced in this respect. Variability in a general sense can be modelled through priors, as model parameters are estimated in a Bayesian MCMC framework via ADModelBuilder.

The view was widely expressed that in order to use this model a significant amount of training would be required. Chris Legault mentioned that Richard Methot gives regular training courses and it was perhaps suggested that should WGMG look further into these types of models perhaps a course could be arranged.

It was suggested that North Sea whiting would be an ideal candidate for this approach as traditional methods don't have the necessary spatial and length-based components.

2.18 WP18

Chris Legault. To weight or not to weight... that is the F.

2.18.1 Abstract

Estimates of F at age from virtual population analysis can be highly variable for "fully selected" ages due to the assumption that catch at age is known without error (or other model misspecifications). Managers desire a single value of *F* for each year to compare against reference points. The annual values for a specified range of ages could be calculated either as an arithmetic average or as a weighted average with the weights supplied by population abundance (*N*), biomass (*B*), or catch (*C*). Two simulation approaches were used to address the question of which of these four methods produced the least biased and smallest confidence intervals. The first simulation approach used a top-down approach of applying a gamma error to each *F*-at-age and deriving *N*, *B*, and *C* from an initial population structure. The second simulation approach used the NFT PopSim program (see WP6) to create datasets with noise in catch at age and survey indices which were supplied to a VPA to estimate the *F*-at-age. The two approaches produced contradictory results. In the first simulation approach, unweighted *F* had much lower confidence intervals than *N*-, *B*-, or *C*-weighted *F*. In the second simulation approach, *N*- and *B*-weighted *F* had lower bias and confidence intervals than unweighted *F* or *C*-weighted *F*. Given these results, there is not a clear indication of which method is preferred.

During the meeting Noel Cadigan suggested an alternative method to estimate the annual *F*: subtract *M* from the estimate of *Z* calculated from the sum over cohorts from two successive years. This approach is similar to a ratio estimate using the sum of values in both the numerator and denominator instead of averaging a number of ratios. Application of this approach to a number of test datasets created using the second simulation approach resulted in situations where this method was less biased, as biased, or more biased than the *N*-weighted estimates of *F*. Given this range of responses, there is still not a clear indication of which method is preferred.

2.18.2 Summary of discussion

Several possible methods for calculating the F on fully selected ages were proposed and discussed. The methods included:

- Arithmetic average
- N weighted average
- B weighted average
- C weighted average

Straight average weighting can be a problem when there are very small cohorts. A number of other approaches have been mentioned elsewhere by other scientists. In his ADAPT software, Gavaris uses a population weighted estimate so that the F tracks strong cohorts. Steffansson also favours this approach. In addition, Shepherd (1983) found a method that derived from engineering. Regarding strong cohorts, it was suggested that the mean F in that situation might be irrelevant to what the population is experiencing, whereas if there were more consistent cohorts, then population weighting might not make a difference.

An additional method was suggested, a "survivor-weighted" value, where F is estimated from the log of annual population size averaged over all ages (which would give total mortality, Z) minus natural mortality (M):

$$F = \frac{\ln\left(\frac{1}{a-1}\sum_{a=2}^{A} N_{a,y}\right)}{\ln\left(\frac{1}{a-1}\sum_{a=1}^{A-1} N_{a,y-1}\right)} - \frac{1}{a-1}\sum_{a=1}^{A-1} M_{a,y-1}$$

This method was subsequently tested against the other methods on simulated data sets. When compared to an N-weighted measure, this survivor-weighted F performed similarly/better/worse, depending on the data set. There was no method that consistently outperformed the others, and the question was not resolved during this meeting. As a point of interest, the ICES standard it to use a straight unweighted F.

A separate question of how one determines which ages are fully selected was noted.

2.19 WP19

Chris Legault. Bias correction in stock assessments: is it necessary?

2.19.1 Abstract

Stock assessment models have bias due to nonlinear estimation. This can be observed when the point estimate and the mean of a bootstrap differ. Efron (1987) demonstrated the superiority of bias corrected confidence intervals over percentile confidence intervals in simple models based on coverage. This work has been extended to VPA by Gavaris (1999) and Mohn (1999). However, it is not standard practice in most parts of the world to use bias corrected confidence intervals. Why not? The bias correction is a simple statistical modification, but can be overwhelmed by other sources of bias, such as a retrospective pattern. Often the estimated bias is small, causing the bias corrected values to be close to the uncorrected values. An exception is in projections when these small adjustments can grow over time. Conversely, when bias is large, strange results can occur, such as the point estimate being outside the bias-corrected confidence intervals.

2.19.2 Summary of discussion

WGMG noted that bias correction was usually applied to correct percentiles for the provision of confidence statements and probability distributions, but not to expected values. Bias correction of percentiles had been explored by Restrepo *et al*. (2000) who compared the delta method, parametric bootstrap and non-parametric bootstrap, and a Bayesian approach. The analysis quantified the coverage and accuracy of confidence limits of estimated interest parameters (F0.1, SSB and TACF0.1 in year 26) by comparing against simulated truth. A confidence statement was defined as being accurate if the confidence point achieves the desired probability coverage. It was established that bias correction can improve accuracy when it can be applied and that inference statements about F0.1 were more accurate than those for SSB or TAC. WGMG suggested the exploration of quasi-likelihood methods that do not seem to require bias correction. It was also noted that management targets are based on the uncorrected estimates and therefore bias correction may not be needed in order to provide appropriate advice.

2.20 WP20

Chris Jones. Assessment approaches for Southern Ocean resources.

2.20.1 Abstract

A review was presented on assessment and management approaches for three groups of harvested resources in the Southern Ocean: Antarctic krill (*E. superba*); mackerel icefish (*C. gunnari*); and toothfish (*Dissostichus* spp.). Although these resources have unique biological, life history, population, and ecological characteristics, their assessment and management approaches are united under the application of decision rules intended to ensure they meet the objectives and acceptable levels of risk defined under the Convention for the Conservation of Antarctic Marine Living Resources (CCAMLR). These decisions rules are designed to 1) prevent decrease in size of harvested populations below that necessary for stable recruitment; 2) maintain ecological relationships between harvested, dependent and related species; and 3) prevent or minimize risk of changes not reversible over two or three decades. Accordingly, assessment methods must yield advice in relation to long term stock status, must be precautionary, and must consider the needs of predators and dependent species of the harvested resource. Antarctic krill, which are a key prey species in the Antarctic ecosystem, are assessed using the Generalized Yield Model (GYM), an age-structured Monte Carlo population simulation and projection tool that allows for integration across uncertainties in population parameters. Mackerel icefish, which have a complex ecological role and large pulses in year-class strength, are managed using a short term (2-year) projection model. Toothfish, a higher trophic-level species with a greater number of available data sources, are assessed using the C++ Algorithmic Stock Assessment Laboratory (CASAL), a fully integrated modelling framework that allows multiple data sources to be combined into a single assessment. Challenges of using the integrated modelling framework include uncertainties in model structure, reliable estimates of M and steepness, and weighting factors of datasets when they demonstrate conflicting information. Flexibility is given to assessment approaches, based on the unique characteristics of the harvested resource and data availability, to generate advice that strictly adheres to the management objectives.

2.20.2 Summary of discussion

The body that manages all harvested resources in the Southern Ocean (with the exception of seals and whales) is CCAMLR. CCAMLR includes an international commission and a scientific committee, as well as several working groups which provide scientific advice for management of harvested resources.

The author explained that there are procedures for estimating the ages of krill, enabling the use of age-based modelling approaches. There are a variety of markets for the krill products, including boiled tail meat for human consumption, a dried krill meal product, pharmaceutical products, and as feed in the aquaculture industry.

With respect to toothfish assessments using integrated methods (CASAL), the author showed that there were some conflicts between data series in the model. It was commented that some conflicting series appear to line up with the penalty, which may suggest it is suspect. It was pointed out that dome-shaped selectivities have a danger of allowing a large biomass of small fish to develop in the kind of models such as the one presented. The author confirmed that this was a model artefact.

It was mentioned that with sufficient tagging data, then the M could be calculated, however in this case of toothfish, there is currently insufficient tagging data to generate robust estimates of M.

There was discussion that a large IUU (*illegal, unregulated and unreported*) catch will cause issues with a management strategy. This could be a problem in the Indian Ocean sector of the Southern Ocean, but IUU catch is thought to have been eliminated in the Atlantic and Pacific sectors.

The extent of the winter sea ice appears to be a key factor in the recruitment of krill. This effect is not currently included in the GYM model, which is used to generate advice for precautionary catch limits.

The author pointed out that the CASAL model which is used for toothfish assessments is also used to assess several stocks around New Zealand. As with the SS2 model, some training is required for the use of CASAL. CASAL includes the possibility of using tagging data unlike SS2 (although this will have this soon)

The importance of by-catch in the krill fishery was discussed. Although the krill fishery uses very small mesh sizes, there is little by-catch because krill swarms are specifically targeted, and the catch is largely clean. However, this may not be the case with a new method employed by super trawlers known as 'continuous fishing', whereby the contents of the cod end are continuously pumped out. Thus, the trawl gear only requires very occasional retrieval, and regions between krill swarms are also taken by the vessel.

WGMG commented that the efficacy and utility of the management plan under which Southern Ocean fisheries operate could not be determined without a rigorous management strategy evaluation (MSE), which has not yet been attempted.

3 Management strategy evaluation

3.1 Introduction

This Section reports the work of Subgroup A, which was convened in order to address ToR d): "evaluate the current state of operational evaluation tools for fisheries management options." The relevance of this work to WGMG lies in the development and testing of methods and tools to allow for the appropriate evaluation of management plans and strategies. It is not the function of scientists to propose or advocate management plans or targets – that is for managers themselves, along with stakeholders and the wider society. It is, however, incumbent on scientists to advise managers on the likely consequences of this or that management action, and to assist managers in developing plans that have the best likelihood of achieving whatever it is the managers want to achieve – again, exactly what this is not for scientists to decide. Specifically, WGMG should neither propose management plans nor conclude whether this or that plan is more likely to succeed – rather, WGMG should test and develop methods that enable scientists to answer the questions asked of them by managers.

The Section contains summaries of four existing management evaluation tools (namely FLR, F-PRESS, PROST and PopSim), along with a review of assessment and management approaches in an area (the Southern Ocean) with rather different management issues. Rather than simply review existing methods, it was important that the Subgroup used such methods to try and address a number of key questions in current fisheries management science. In order of tractability, the Subgroup listed these (following the presentations given at the start of the meeting) as:

1) Can harvest control rules be devised that can account for the presence of retrospective bias in fisheries stock assessment?
2) Can the traffic-light approach (WP 10) be used as a functional fisheries management tool?
3) Can the effect of varying growth and maturity be allowed for in harvest control rules?
4) Can the models be used to evaluate the current ICES precautionary approach to management advice, used in the absence of agreed management plans?

This Section reports progress in addressing task 1, and to a lesser extent task 2. There was insufficient time, data and expertise to consider task 3. Task 4 is possible in theory, but the decision tree for the ICES precautionary approach is complicated and would take considerable investment in programming time to simulate. WGMG participants intend to address this in the near future.

The software tools available to the Subgroup are not strictly complementary, and were therefore used in different ways. Implementations of FLR, PROST and F-PRESS were used to explore (through simulation) the effect of retrospective bias on the ability of managers to attain management goals. PopSim is a data simulator rather than a management strategy evaluation tool *per se*. As well as generating sample datasets for other methods, it was used in exploratory analyses to a) identify the onset of retrospective bias, and b) ascertain suitable year-ranges over which to calculate average recruitment to be used in subsequent management projections. The traffic-light approach is a way of summarising data which could then be used in a management context; here we have restricted the analysis to traffic-light summaries of the simulated datasets.

Previous work in related fields has been carried out by a number of groups within ICES and elsewhere. The ICES *ad* hoc Group on Long-Term Advice (ICES, 2005b) applied existing methods to evaluate management plans for a number of European stocks. The ICES Workshop on Reference Points (ICES, WKREF 2007, report not yet available) commented on the use of

stochastic simulations to inform the development of management strategies. The ICES Study Group on Management Strategies (ICES, 2005d, 2007) considered specific management plans and the methodology of their development (building on work presented at the 2006 meeting of WGMG), rather than using evaluations to investigate the effects of specific assessment problems. As a final example, a subgroup of the EU STECF met during the first week of WGMG to look at evaluating harvest control rules using a number of different tools: they concluded that FLR is the only widely-available toolbox that allows the use of "live" assessments within the evaluation loop, but also that setting up such an evaluation is far from straightforward.

The work of these and other groups will not be reviewed in detail in this report, but the reader should be aware of a wider body of work addressing management-strategy evaluation issues.

3.2 Methods

3.2.1 Data simulation

Datasets for analyses in the remainder of Section 3 were created using a modification of the program PopSim of the NOAA Fisheries Toolbox (WP6). PopSim is a general data simulator designed to allow users to create populations with known underlying parameters and error structure. The resulting stochastic realizations can be used to automatically generate multiple datasets for input to a number of programs in the NOAA Fisheries Toolbox: VPA/ADAPT, ASAP, ASPIC, and CSA. It is straightforward to convert datasets into formats suitable for other widely-used assessment programs. PopSim is designed to allow users to rapidly compare the relative merits of alternative modelling approaches. This simulator should have general utility for examining tradeoffs among model dimensionality, degree of fit and generality. The graphic user interface allows for complete graphical analysis of input and output data.

The version of PopSim used to create the test datasets for WGMG reduces some of the available options, but includes the ability to change a number of parameters that are normally fixed, for example survey catchability and natural mortality rate. These changes in parameters can cause retrospective problems, and thus the modified version of the program is called RetroPop.

In RetroPop, the user provides an estimate of the initial population structure and a time series of recruitment values. The population is length and age based with fishery selectivity following a logistic curve over lengths. The von Bertalanffy growth parameters Linf, K, and t0 are provided along with a measure of the spread of length at age. A length-weight relationship is provided to convert catches in number to total catch weight. Given the initial conditions, a time series of natural mortality rates, and a time series of fully selected fishing mortality rates, simple forward projection methods are used to derive catches and abundance indices. The catch at age is built each year based on a random sampling of lengths from the true catch and a random sample of ages selected from these lengths to create an age-length key. Thus, catch at age can be estimated with low or high uncertainty depending on the level of sampling for both lengths and ages. An aging error matrix can be included, if desired. Indices are computed from the true population using a catchability coefficient. Indices can be collected at the start or mid-point of the year, in numbers or biomass, and for specific ages or over an age range. Observation error is then added to the indices specified by a user-supplied lognormal coefficient of variation.

The datasets created for the management strategy evaluation work (Section 3) contained large changes in parameters to create the retrospective patterns but relatively low levels of uncertainty in catch at age and survey indices. The datasets covered years 1970–2004 with tuning indices available at the start of 2005 as well. The population had 5 true ages and an age 6 plus group. Selectivity was nearly full for ages 1 and 2 and full (= 1.0) for ages 3–6+.

Weight at age was calculated from the observed catch at length using a constant length-weight relationship. All four datasets had high fishing mortality rates with an increasing trend from 1970 to 1994 of approximately 0.4 to 1.0 followed by a sudden decline to approximately 0.5 for 1996–2004. Recruitment varied around 3 million fish for most of the time series, but reduced to lower levels of around 1 million fish for years 1996–2004. These factors combined to cause a large drop in catch in 1996 from a relatively constant level during 1970 to 1995.

There were three main test datasets created for the management strategy evaluation subgroup. These three datasets all generated retrospective patterns in the subsequent VPA output. The source of retrospective bias was different for each dataset:

- testdata1: three-fold increase in survey catchability (q) for all ages 1995–2004;
- testdata2: reported catch multiplied by 1/3 at all ages for years 1995–2004 (that is, 66% misreporting for these years – true catch remained unchanged);
- testdata3: M in the simulator increased from 0.2 to 0.6 for years 1995–2004 but the assumed M for the assessments remained constant at 0.2

Each dataset was provided as both the input and output for the NOAA Fisheries Toolbox VPA/Adapt program along with a file documenting the underlying truth. There was a low level of noise in the catch at age due to sampling of the length frequencies at a rate of 50 fish per metric ton of landings but aging all the length sampled fish for age. A 20% coefficient of variation was applied to each of the six indices, one for each age. Due to these relatively small levels of noise in the dataset, estimates would be quite close to the true values if the perturbations of survey q, catch, or M had not occurred. It is the inclusion of the changes in survey q, catch, or M reporting that *cause* the retrospective patterns, not any feature of the F or R time series.

3.2.2 Available methods and required modifications

FLR

FLR (FLR Team 2006) consists of a library of data objects and methods for the R statistical programming package (R Development Core Team 2005). The package has been created largely under the auspices of a suite of European Union projects (EFIMAS, FISBOAT, COMMIT) with the express intention of providing an operating model for evaluating fisheries management strategies. It includes methods for summarising, manipulating and generating fisheries data, running assessments and forecasts, and simulating fisheries management under a variety of harvest control rules. As well as management-strategy evaluation (MSE), it is also increasingly widely used as a stock assessment tool within ICES assessment working groups.

The main advantage of FLR in the context of the current analysis that it is extremely flexible. Essentially the model to be simulated must be programmed by the user, who can therefore (in theory) make the simulation proceed exactly as intended. While there are methods within the library which can be viewed as "black boxes", the entire structure is open-source and it is possible to code simulations from scratch (this is what was done for the analyses discussed below). As the library contains stock assessment methods, it is also possible to run management simulations that include live stock assessments (rather than simple perturbations of underlying abundance), thus enabling a more direct simulation of the assessment-advice-management process. The flexibility of FLR is also its main disadvantage. Simulation code can rapidly become complicated and difficult to debug, and the development cycle of an FLR simulation can be long. The result of this for the work of WGMG was that only one case study could be analysed using the FLR approach.

The basis for the work discussed below was the harvest-control rule evaluation code developed in 2006 and applied to the management plan for North Sea haddock (WP 16). This had to be modified before it could be applied to the simulated datasets: for haddock, the result

of the first assessment carried out in the management simulation was viewed as the "truth" against which subsequent developments were evaluated, while for the simulations in WGMG the actual true population was available and had to be used as the basis.

The FLR simulation model implemented in this meeting can be summarised briefly as follows: further details on the application to North Sea haddock can be found in Needle (2006a, 2006b).

1) Historical assessment data are read in, along with the underlying true data (from the historical simulation described in Section 3.3.1). If the current assessment year is *y*, then assessment data are available up to and including year *y*-1.
2) A biological population simulation is carried out for the first year *y*. This generates recruitment, growth, and mortality, and results in values for abundance and biomass. In this first year, fishing mortality is assumed to be equal to a three-year historical average: in subsequent simulation years, fishing mortality or yield is determined by application of management decisions from previous years. Recruitment is fixed to a 10-year geometric mean of the last historical years.
3) A stock assessment (using FLXSA in this case) and associated short-term forecast are carried out, based on data up to and including year *y*-1.
4) Management decisions for the following year *y*+1 are then determined, using information from the assessment and forecast and following the selected HCR. In the analyses presented here, the HCR is a simple target *F* of 0.3. The forecast tells the model what the intended landings yield for the year *y*+1 should be in order to achieve this target *F*. As we are simulating catch-based management, the intended yield in year *y*+1 will be taken unless the population is too small to permit it. However, taking such a yield may not result in the target *F*; this would only occur if the population forecast was precisely correct, which is very unlikely.
5) Steps 3–5 are repeated for the number of years required in the simulation, which in this case is 25.

Normally this loop would be repeated many times with different stochastic realisations of the recruitment time-series, which would then enable an estimate of the risk of biomass falling below a given low level (for example). However, in this case recruitment in the future is fixed to a geometric mean of the 1995–2004 period, so the stochastic approach is rather redundant.

PROST

A number of exploratory runs to evaluate an HCR for the simulated stocks (see Section 3.3.1) were carried out using PROST. In these evaluations the corresponding data from assessments were treated as "known to an assessment group". The results of these evaluations were compared to the underlying true dynamics of the population.

Three examples of assessments using simulated data sets were explored (Section 3.3.1):

a) the assessment based on data where the shift of survey catchability was simulated;
b) the assessment based on data where the natural mortality has changed from 0.2 in the beginning of time series to 0.6 for the last ten years;
c) the assessment based on data where the catches were underreported by 1/3 since 1995.

Software used

The PROST software was developed and used for evaluation of the harvest control rules for Northeast Arctic cod and haddock carried out by ICES in 2004–2006 (ICES, 2004a, 2005a, 2006a). The implementation in which the simulations were carried out using stochastic

projections is mentioned by SGMAS (ICES, 2005d, 2007) as a tool for harvest control rule evaluation. The software and users guide are available on the ICES web site (www.ices.dk).

Model parameters

The inputs for the population model were taken from the example data sets and simplified to represent better what would actually be available to a stock assessment working group. The weight at age in the stock and in the catch, maturity ogive, and F selection pattern were averaged for all time series and used in the projections as constant values. The recruitment was modelled as a parameter independent of SSB and equal to the geometric mean of the whole time series. Inclusion of uncertainty in the population model was done assuming random variation (normally distributed with CV=0.5) in recruitment and normally distributed errors in F with CV=0.3.

PROST does not include a "live" assessment model. The problem of retrospectively underestimating *F* in the final assessment was therefore taken into account in the HCR evaluation by imposing a bias in the *F* "estimates" available to managers. The level of bias was estimated as the average deviation of final points in annual assessments from the assessment in the final year. The output from retrospective runs on the simulated data sets were used for estimating the level of bias and CV for *F* in the final year (Figures 3.1 to 3.3).

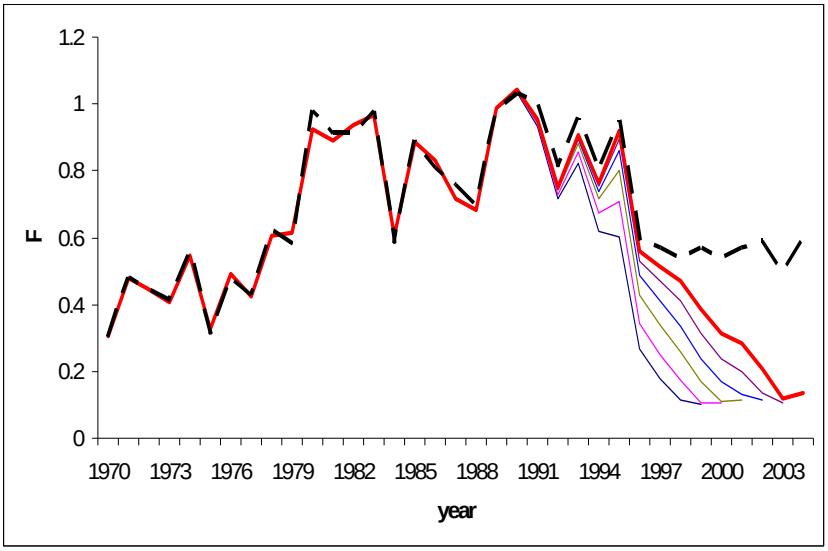

Figure 3.1. The fishing mortality from retrospective assessment runs on simulated data, where a systematic shift in survey catchability occurred in years 1995–2004; the underling true F is plotted as a dotted thick line.

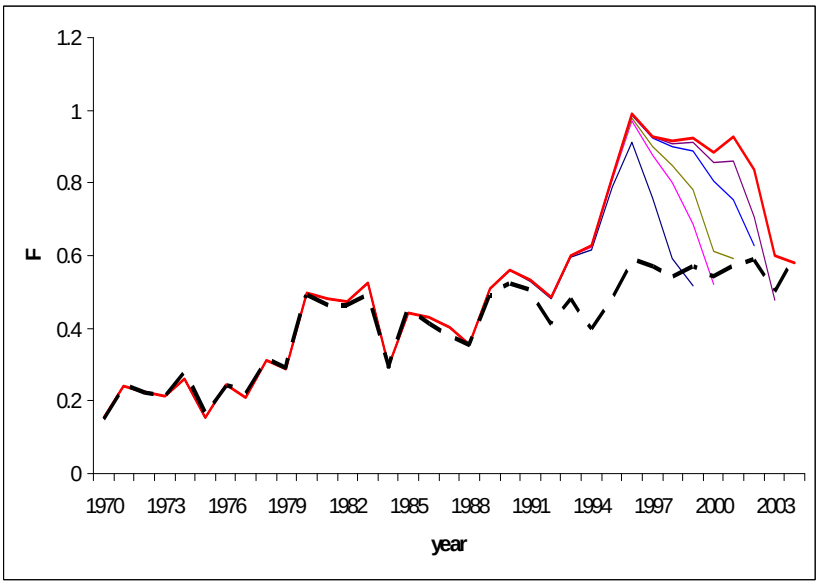

Figure 3.2. The fishing mortality from retrospective assessment runs on simulated data, with a systematic shift in natural mortality from 0.2 to 0.6 in years 1995–2004; the underling true F is plotted as a dotted thick line.

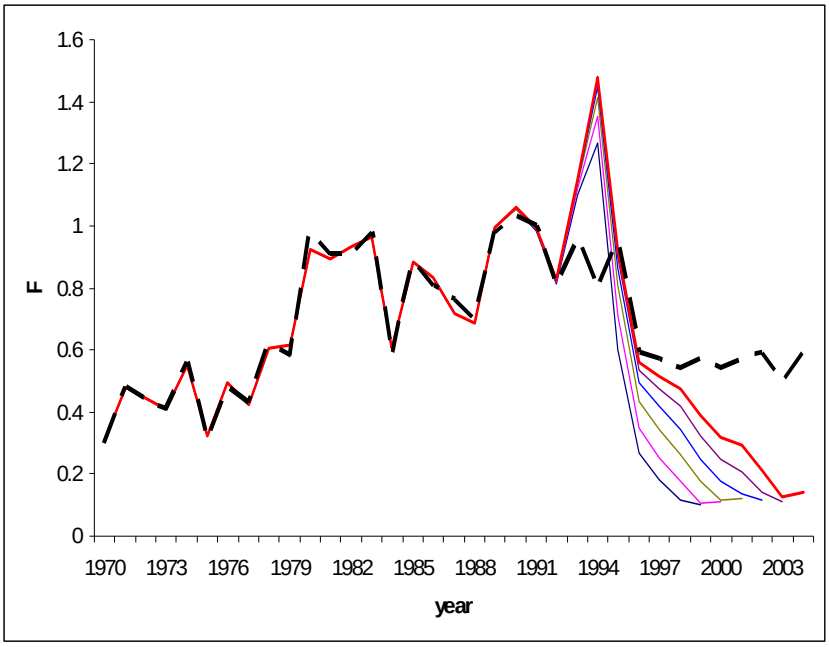

Figure 3.3. The fishing mortality from retrospective assessment runs based on simulated data, where catch was underreported by 1/3 in years 1995–2004; the underling true F is plotted as a dotted thick line.

Projections with constant $F = 0.3$ were tested as a HCR. Predicted distributions of SSB were compared with a hypothetical B_{lim}, where the stock was assumed to be equal to 1.5 million tonnes.

In all cases the model was run for 50 years forward, and the last 20 years were used for evaluating results. In all cases 500 iterations were carried out.

F-PRESS

The F-PRESS (Fisheries Projects & Evaluation by Stochastic Simulation) model is designed as a stochastic simulation tool for evaluating fisheries management strategies and developing management advice. It is designed as a population projection model with the following characteristics and limitations:

- stochastic;
- single species;
- non-spatial;
- age-structured population;
- exponential mortality;
- F or TAC controlled fishery;
- various recruitment models; and
- various harvest control strategies.

The model is designed specifically to carry out investigation of management strategies, specifically harvest control rules. It is not designed to produce quantitative predictions of stock numbers.

Stochasticity is introduced by randomising the various parameters via a supplied coefficient of variation (CV) value. Non-zero CVs and bias can be specified to account for variability in stock and environmental dynamics, assessment errors, TAC non-compliance and data errors.

The design of F-PRESS (deliberately) avoids a complex "assessment feedback" model so that all bias and noise introduced in the assessment process can be qualitatively controlled. During the model realisation of the assessment process the current virtual SSB or *F* value is multiplied by the SSB or *F* assessment bias input parameter. The new biased values are then used as the mean of a normal distribution with a CV value as given by the SSB or *F* assessment CV input parameter. New (randomised and biased) SSB and F values are then drawn from the normal distributions.

The newly biased and randomised SSB and *F* values are inputs to the management module, the purpose of which is to apply any harvest control rule that may be in operation. F-PRESS does not perform a forecast in order deterministically calculate the SSB or F in future years as a basis for the management model. Instead, the stochastic approach (via the inclusion of predetermined levels of noise and bias on all the important model components) is the basis of the F-PRESS model.

A consequence of the design of F-PRESS is that it is unable to introduce bias to the assessment process via the assessment process in itself. However, since bias can be included explicitly, F-PRESS can be used to gain insight into the possible performance and robustness to assessment bias of a harvest control rule (management plan). Using a dataset supplied by the working group for population initialization, F-PRESS has been used to conduct preliminary simulations using a harvest control rule that modifies the TAC in order that *F* is maintained at a value of 0.3. Various levels of bias were applied to the assessment of *F* (1, 0.8, 0.6 and 0.4) and annual recruitment was held constant with zero CV. Three scenarios were considered for each level of bias and the projection run for 20 years (1000 iterations):

1) assessment bias is applied in each year;
2) assessment bias is applied in years 5–9;
3) assessment bias is applied in years 5–9, reducing linearly over the period to zero in year 10.

Future work could investigate alternative harvest control strategies. Also, the model could be adapted to incorporate an actual assessment and/or a short term forecast, possibly making use of the FLR libraries. Although such developments would be contrary to the ethos that is behind F-PRESS, they would demonstrate the ease with which the model can be modified and developed.

F-PRESS and a Traffic Light Management Approach

Traffic light management approaches employ a number of stock abundance indicators and devise rules for their weighting and influence on a decision making process. In general it is straightforward to translate such management models into structure of a harvest control rule and so in theory can be tested by models such as F-PRESS. The limiting factor is the indicators. F-PRESS includes biological indicators such as SSB and fishery statistics such as F to be used as harvest control rule indicators. The use of other fishery indicators could possibly be included in the model along with environmental indicators, which could be modelled stochastically as is done for other model parameters.

FSIM, PopSim and VPA-2Box

FSIM has been uploaded to the WGMG SharePoint site. It can also be downloaded from the ICCAT assessment software catalogue: http://www.iccat.es/AssessCatalog.htm. The manual was uploaded separately as a PDF file. Although it was not used in the simulation analysis, it is a complementary package to PopSim (discussed below) and is summarised here for completeness.

FSIM (Goodyear, 2003) is a general purpose fish population simulator designed to simulate many forms of fisheries data routinely collected from real fisheries. Analyses of these

"known" simulated datasets facilitate studies of the robustness of alternative assessment methodologies. The model is also useful for exploring the implications of uncertainty about the dynamics of fish populations, forecasting consequences of management alternatives, and predicting future trends in population sizes and catches for a wide assortment of possible biological attributes under different management alternatives. The simulated population is structured by age, sex, and growth morph, where individuals within a growth morph (cf. WP 17) are subunits of a cohort of individuals spawned during the same reproductive period that share the same growth characteristics. The time step is seasonal up to a maximum of 12 seasons per year (monthly), and growth within the year can be seasonal. Mortality sources are assumed to operate concurrently within the specified season. There are many options for simulating recruitment and generating indices of abundance (fisheries independent or fisheries dependent, where the catch is specified as known with or without error).

This simulation software is very flexible in terms of the processes that can be modelled and the variability that can be incorporated, as well as the "assessment data" that can be generated as output. At present, automatically formatted output files are only generated for ASPIC, so users of other assessment programs need to assemble their own input files. FSIM does not perform assessment analyses; it is strictly a tool for simulating data that can be used to test other assessment software.

PopSim is a data simulator rather than a management strategy evaluation tool *per se*, and was used in exploratory analyses to a) identify the onset of retrospective bias, and b) ascertain suitable year-ranges over which to calculate average recruitment to be used in subsequent management projections.

Three data sets with known errors were simulated using the NMFS Toolbox package PopSim. After assembling input files in the necessary format for VPA-2Box (Porch, 2003), an attempt was made to run VPA-2Box from the NMFS Toolbox GUI. However, there seem to be errors in how the Toolbox version interprets the control file when there are no tagging data (and hence, no tagging input file), and this endeavour was not pursued further. Instead, the command line version of the software was used for exploring and performing assessments on the data sets.

There proved to be substantial cutting and pasting required to format input files to VPA-2Box, and then additionally to take VPA-2Box output and format it for use in the projection software (PRO-2Box, Porch, 2003). Implementing a feedback between management decisions and projections would have been too cumbersome, therefore all projections were fixed at a constant F=0.3 until the year 2014.

3.3 Results

3.3.1 FLR

The FLR analysis was carried out only for the first simulated dataset described in Section 3.3.1, in which retrospective bias was induced by a tripling of survey catchability (this tripling was maintained into the management simulation). The other two datasets (in which bias was caused by changes in misreporting and natural mortality respectively) are also very amenable to consideration in this way, but only following a certain amount of code restructuring for which there was insufficient time during the meeting. This will be addressed intersessionally.

Figure 3.4 summarises the FLR analysis for dataset 1. The retrospective bias causes *F* in the last historical year to be underestimated and SSB to be overestimated. As the simulation moves forward in time, the tripling of the survey catchability is maintained, which in turn maintains the retrospective bias (FLXSA in this case uses the full time-series for tuning, so the distinction between real and tripled survey indices continues to bias the assessment). The increased survey catchability strongly affects the estimates of recruitment, because recruitment

estimates are more driven by survey data than is the case for older ages, and future recruitments and SSB are considerably over-estimated. The model therefore believes that there are more fish in the sea than is really the case, and concludes that mean *F* is falling towards the target *F* in the assessment year for every future assessment that is done. However, in all cases the true mean *F* is much higher than this. The low estimates of *F* lead in turn to relatively high quotas, which maintain mean *F* at a high level. In this case, management action following scientific advice has caused the stock to decline rapidly. It doesn't disappear altogether, but this is only due to the assumption that recruitment is fixed no matter what the spawning biomass.

Figure 3.5 shows log index residuals from three of the years in the forward simulation. The effects of the onset of a change in survey catchability at or about 1995 can clearly be seen.

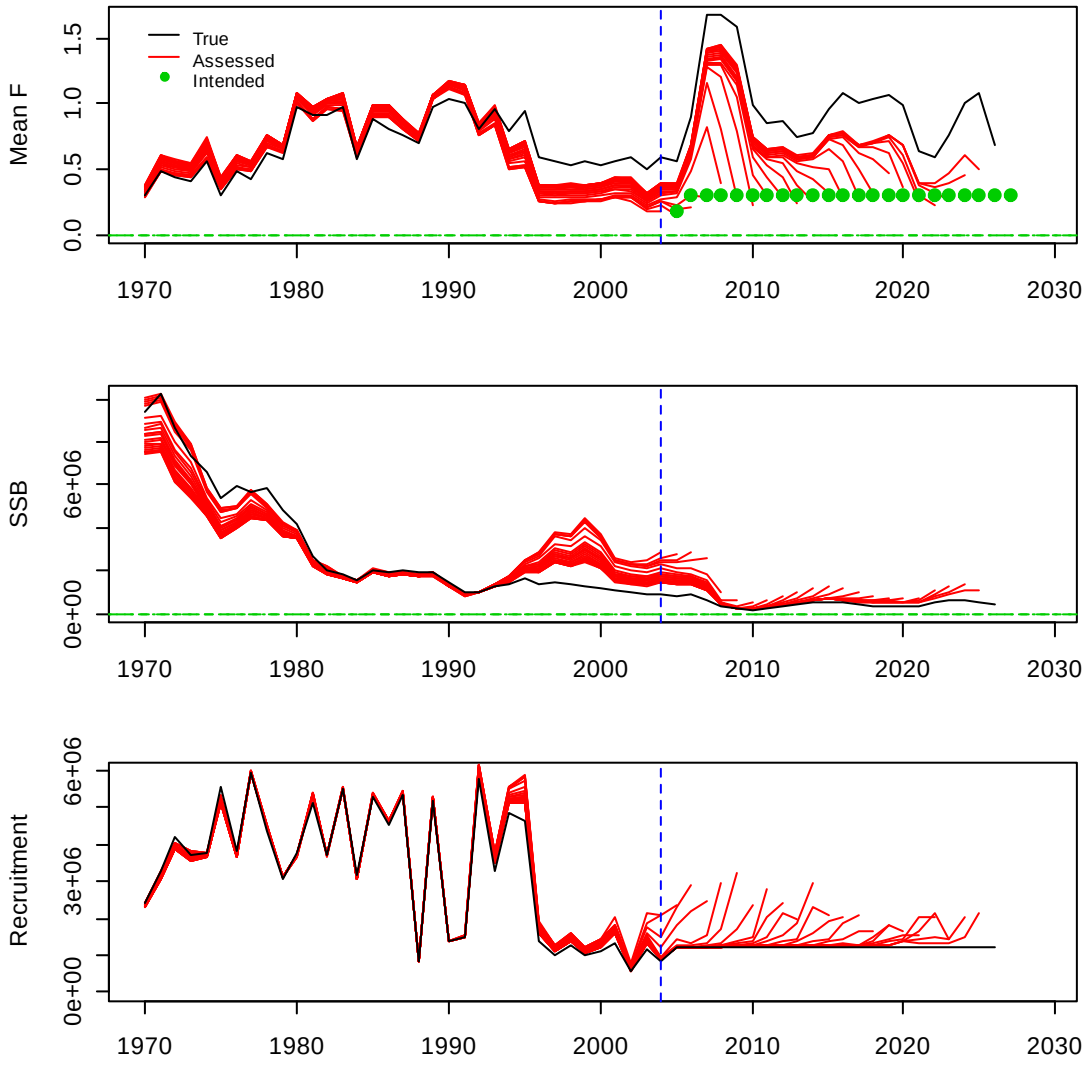

Figure 3.4. Summary plots of the application of FLR to dataset 1. In each plot, the thin black line gives the true values, the red lines are assessment estimates for each year in the forward simulation, the vertical dotted line is the final historical year (2004), and the green points are the intended values. Top: mean *F* over ages 3–5; middle: SSB; bottom: recruitment at age 1.

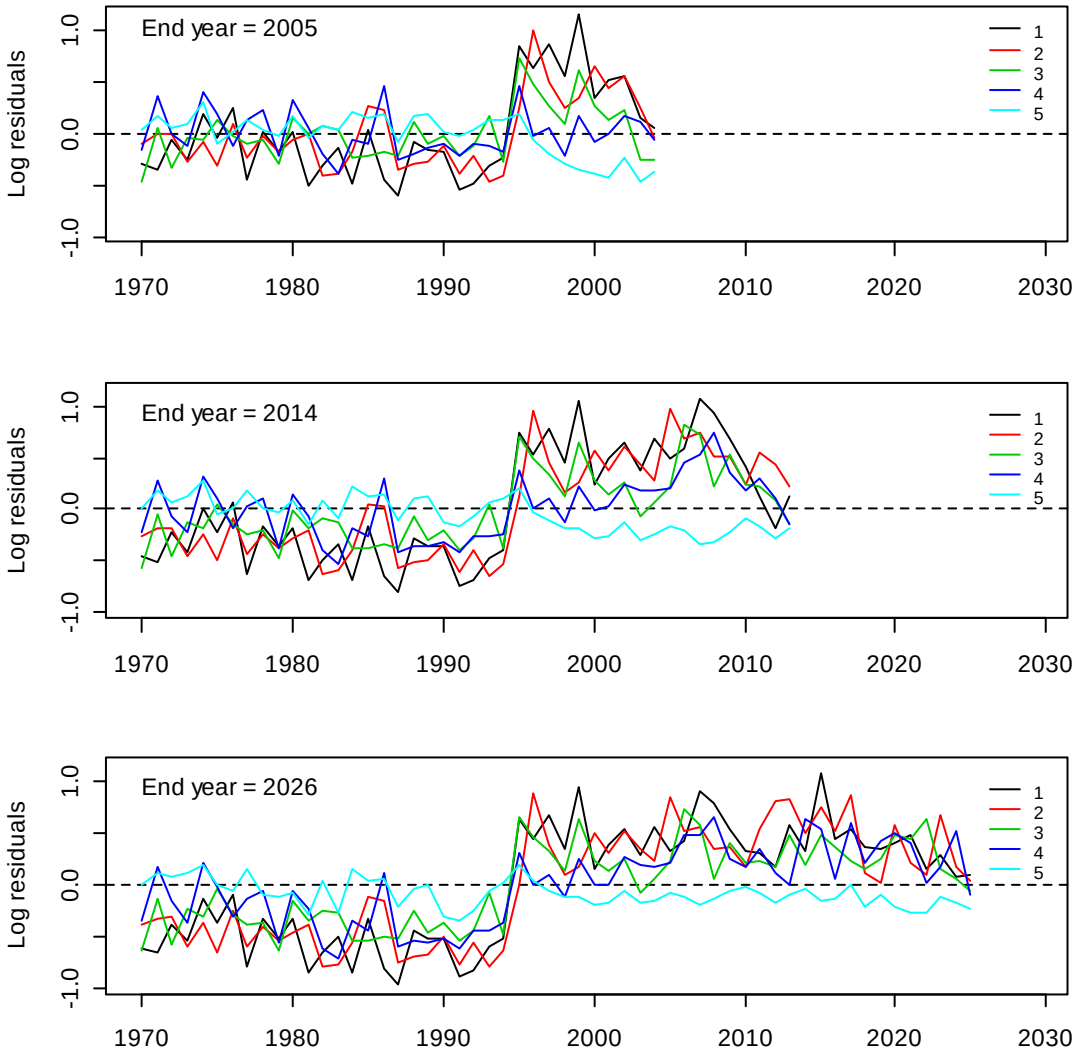

Figure 3.5. Log residuals plots from the application of FLR to dataset 1.

PROST

The results of this HCR evaluation were compared with results of stochastic projections of "real stock" dynamics. In this case, the stochastic models were run with the level of bias in F estimated as the average deviation of final points in annual assessments (retro runs) from the true F.

The description of different runs and results of evaluations are presented in Table 3.1 and Figures 3.6 and 3.7.

Table 3.1. Model parameters used in different runs of PROST and the results of HCR evaluations for simulated data sets.

Run no	Evaluation of HCR based on	Testing data set problem	Modeled bias in F	Natural mortality for all ages	Realised F	Mean catch (kt.)	Mean SSB (kt.)	Prob. SSB<Blim
1	assessment	shift in survey catchability	−0.15	0.2	0.45	1500	3300	0%
2	true dinamic of population	shift in survey catchability	−0.45	0.2	0.75	1180	1410	0.62%
3	assessment	Shift in M	−0.29	0.2	0.59	1350	2160	0.047%
4	true dinamic of population	Shift in M	0.0	**0.6**	0.3	530	1630	29%
5	assessment	25% of catch underreporting	−0.16	0.2	0.46	1520	3230	0%
6	true dinamic of population	25% of catch underreporting	−0.45	0.2	0.75	1180	1410	0.62%

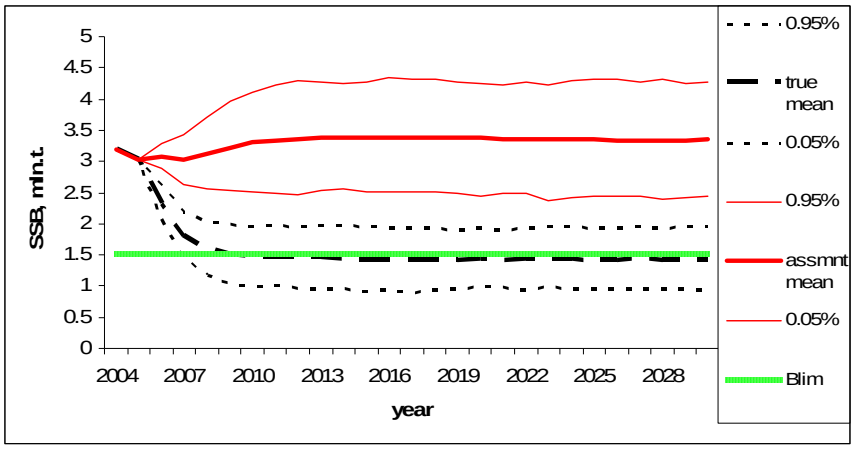

Figure 3.6. The results of stochastic projections of hypothetical population dynamic (data set a) using PROST software based on assessment data (observed assessment bias in F; red lines; 5%, 95% and median) and data from real dynamic of the stock (bias in F; doted black lines; 5%, 95% and median).

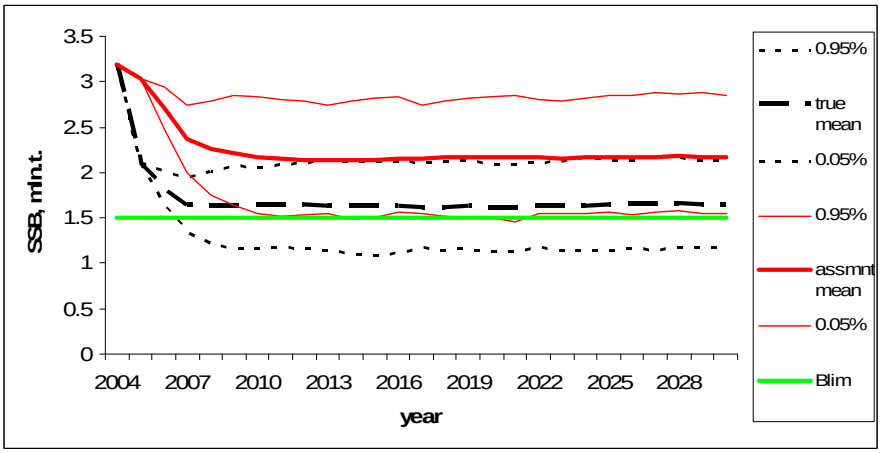

Figure 3.7. The results of stochastic projections of hypothetical population dynamic (data set b) using PROST software based on assessment data (observed assessment bias in F, assuming M=0.2; red lines; 5%, 95% and median) and data from real dynamic of the stock (no bias in F, M=0.6; doted black lines; 5%, 95% and median).

The results of HCR evaluations performed on the data from assessments show that the rule corresponds with the precautionary approach in all cases. The risk that SSB drops below Blim in all runs was lower than 5% (Table 3.1, Figures 3.6 and 3.7). On the other hand, the stochastic projection based on the true stock dynamics shows that if that target level of F (F=0.3) will be used in management, then the mean SSB will be sufficiently lower and the risk of dropping below Blim will be very high. In all simulated data sets, the underestimation of bias in F or the increases in natural mortality could lead to a wrong conclusion, and the HCR would not correspond to the precautionary approach.

Runs 5 and 6 (underreporting of simulated catch) gave almost identical results compared to runs 1 and 2 (shift in survey catchability). The reason for that was that the corresponding levels of bias in F were very close to runs 1 and 2.

The results indicate that if there is a systematic shift in survey catchability or if sufficient underreporting of catches has occurred, an assessment based on this data will have a retrospective problem, and the interpretation of population dynamics and advice for fisheries management could be wrong even if the bias in assessment is taken into account. It may be that considerable underestimation of bias in the assessment estimate of F relative its true dynamic is causing the HCR to fail. The considerable increase in natural mortality, which was not accounted for in the assessment nor in the HCR evaluation, leads to a decrease in stock productivity and substantial underestimation of risk of overfishing.

FPRESS

In general, applying a bias to the F assessment in F-PRESS (in order that the assessed F is less than the actual) results in the harvest control rule permitting a higher TAC than may be precautionary. Whether or not this TAC is sustainable depends on the level of bias, the recruitment and the initial population vector used in the model. Consistently low levels of bias, and to a lesser extent low biases applied for shorter time scales result in increasing yields if the current stock and future recruitment can sustain the elevated fishing mortality. Higher levels of bias however, result in significantly increased risk to the stock even when applied over a short time scale.

The harvest control rule to fish at F=0.3 is a simple and rather inflexible rule. For high assessment biases, model results show that stock is at risk of collapse for consistently high bias. In the situation where the bias is applied for a shorter time period or assumes a linearly reducing behaviour the harvest control rule is more successful. However, variation in yield is high, an undesirable characteristic for successful management and exploitation with fishers likely to seek a less stringent management method. Using more flexible harvest control rules (for example limiting changes in TAC or F to 15% in any one year) is likely to lead to enhanced risk to the stock although such a measure would likely reduce the variability in yield.

The following conclusions can be drawn from the F-PRESS simulation results. Unfortunately time constraints prevent further investigation and model development at the time of writing.

1) The higher the level assessment bias the greater the risk is to stock collapse or any preset biological limits.
2) When assessment bias is high the target fishing mortality must be reduced to maintain a viable fishery.
3) For high levels of assessment bias the risks are reduced when the assessment bias is only applied in years 5–9 and a further improvement when linearly decreasing over years 5–9.
4) Variability in yield increases significantly when the bias is applied over short time scales.
5) A single year of high assessment bias can lead to significant risk of stock collapse.

PopSim and VPA-2Box

VPA analysis

The simulated data sets (Section 3.3.1) spanned the years 1970–2004. VPA assessments were made through 2004, and retrospective patterns were examined for 20 years (terminal assessment years from 1984–2003). Plots of annual F on age 6, spawning stock biomass (SSB), and recruitment were generated in order to examine retrospective patterns to see if they corresponded to the year when a known error occurred (1995, see Figures 3.8 to 3.10). In all three datasets, it was apparent from the SSB and recruitment plots that the retrospective pattern became substantial in years 1995–2004, while the F on age 6 displayed a more variable pattern even before 1995. The nature of the F plots was driven in part by the age class chosen, and the variability in year class strength, among other things. Choosing the proper age class(es) to examine for F, i.e. identifying fully selected ages, and selecting a method to obtain the "full F" was beyond the scope of this exercise (see WP18).

No attempt was made to diagnose the cause of the retrospective pattern, nor to adjust the data in order to remove the retrospective pattern. The sole purpose of the retrospective analysis was to try to identify when the error occurred in the time series. The follow-up analysis tried to identify a sensible level of recruitment to use for making projections, assuming that one had correctly identified a year when the data are first contaminated by the known error. The decision regarding recruitment level rests on whether one believes estimated abundance signals in the years where the retrospective pattern is strong. If one believes the abundance estimates, then recent values might be appropriate for deriving future recruitments. On the other hand, if one were uncertain as to the cause of the retrospective pattern and consequently had less confidence in recent abundance estimates, one might consider basing projections on an earlier time series of abundances or trying to average over the retrospective pattern by including years where confidence was higher.

In all three data sets, a strong retrospective pattern is apparent starting around 1995. The geometric mean recruitment was calculated for a 10 year window prior to the end point of each retrospective vpa analysis (black dashed line in Figure 3.11). In addition, a 10 year moving geometric mean was calculated from each year in the assessment that used data through 2004 (solid red line, Figure 3.11). Each of these recruitment smoothers was compared to the true recruitment time series (green dot-dash line). For data sets one and two, the 10 year moving geometric mean tracked the true recruitments more closely than the retrospective geometric means. This is because the estimated recruitments from the 2004 assessment (blue dashed line) track the true trend and are relatively close to the true recruitments. In the third data set, the true trend is still well tracked by the 2004 assessment estimated recruitment, but they differ by a factor of about two. In this case, both recruitment smoothers suffered from the scaling problem.

The 2004 estimated geometric mean recruitment (years 1995–2004), and the retrospective geometric mean recruitment for years 1990–1999 were compared with the true geometric mean recruitment for 1995–2004 (Table 3.2). For data set 1, the geometric mean recruitment estimated from the 2004 assessment was better than the recruitment level estimated by the retrospective geometric mean, although it was 64% larger than the true value. In the remaining two datasets, both estimates were roughly equivalent, and biased by about the same factor (~ 55–65%).

Projections

Terminal model estimates of NAA for the 2004 assessment were projected at F=0.3 to year 2014. Future recruitments were fixed at either the 10-year geometric mean recruitment as estimated from the 2004 assessment (years 1995–2004) or from the retrospective geometric mean for years 1990–1999. The retrospective geometric mean was explored because it brackets years on either side of 1995, when the retrospective pattern appeared. In addition, the

true underlying numbers-at-age were projected forward at F=0.3 using the true geometric mean recruitment for years 1995–2004. The resulting annual yield and annual spawning biomass trajectories were plotted (Figures 3.12 and 3.13). When q or M tripled in years 1995–2004 (data sets 1 and 2), the projected yields were greater than the true yields. In the third dataset, where reported catch was only 1/3 of true catches, yield projections at both of the estimated recruitment levels underestimated true yield. Similarly, projected SSB was larger than the true value for data sets 1 and 2, and was smaller than the true value for data set 3. Projections made using the recruitment level as estimated by the geometric mean for years 1995–2004 were closer to the true projections for data sets 1 and 2. In data set 3, the estimated abundances were biased low because catch was underreported; this obviously carries forward in the projections, because the estimated geometric mean recruitment is biased low as well (Table 3.2).

Implications for management

Drawing general conclusions from this analysis is not possible, as the results likely pertain to the peculiarities of the individually simulated data sets. Additionally, no management decisions or actions were included in the simulations. This work should be viewed rather as an exploratory analysis which would help to determine the source of the mis-specification problem (analogous to the analyses presented in Section 5).

Future directions

The software used for this exercise was not designed for management strategy evaluations, so if further explorations are to be made, additional coding would be required to facilitate the input/output flow between simulated data, assessment software, and projection software. An additional step in the assessment analysis would be to look at other model diagnostics to determine if the source of error could be identified (see also Section 5). From this, it might be possible to associate different categories of error source with appropriate actions for projection. Simulating datasets with stronger underlying biological processes (e.g. a stock recruit relationship) would make it possible to evaluate the VPA's ability to estimate current status relative to MSY reference points, and thus to test robustness of rebuilding advice under varying sources of uncertainty.

Table 3.2. Relative estimates of geometric mean recruitment (each estimate is scaled by the true geometric mean). Values closer to 1.0 indicate a better estimate.

DATASET	2004 ASSESSMENT 10-YR GEOMETRIC MEAN RECRUITMENT	RETROSPECTIVE GEOMETRIC MEAN RECRUITMENT (1994–2000)
1	1.66	3.06
2	0.54	1.63
3	0.64	0.63

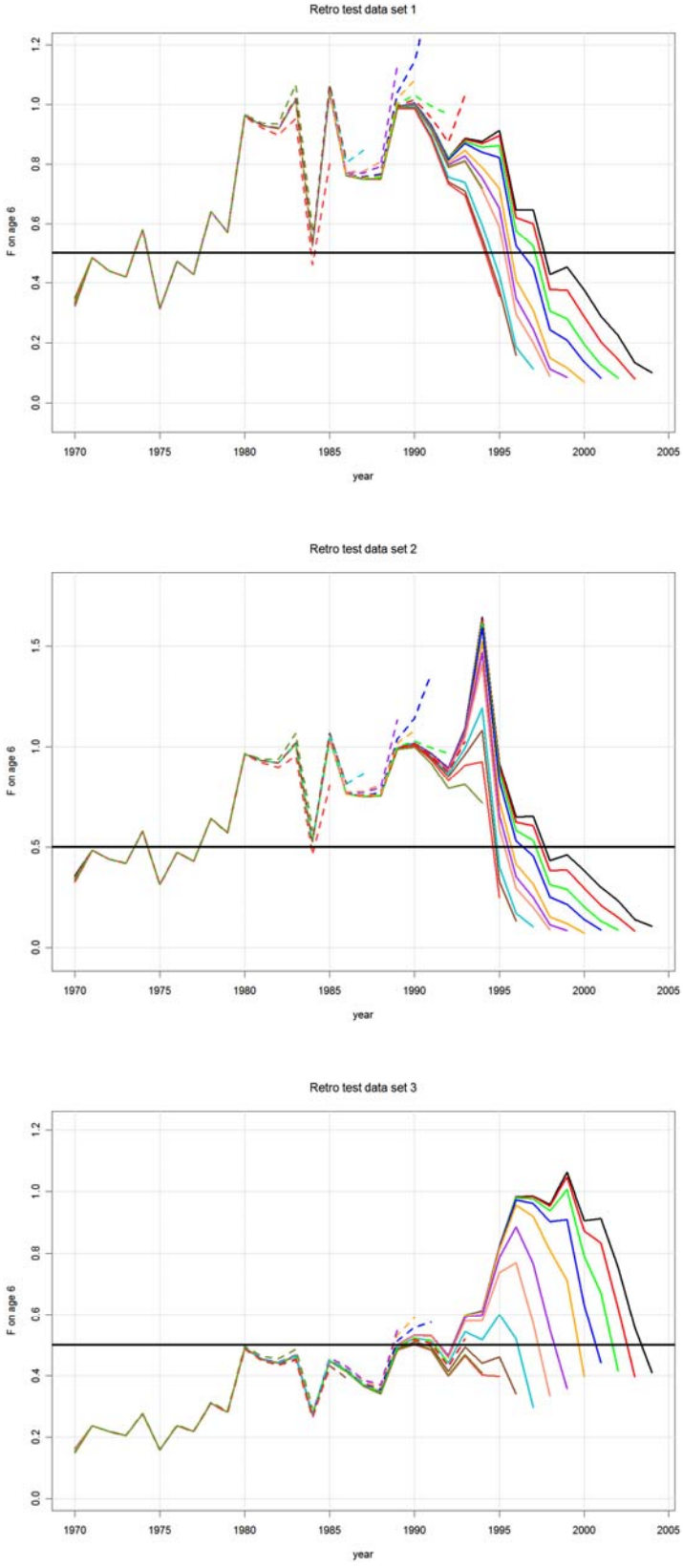

Figure 3.8. Retrospective pattern in annual F on age 6 for data sets 1–3 (top – bottom). True F was 0.5 (solid black line).

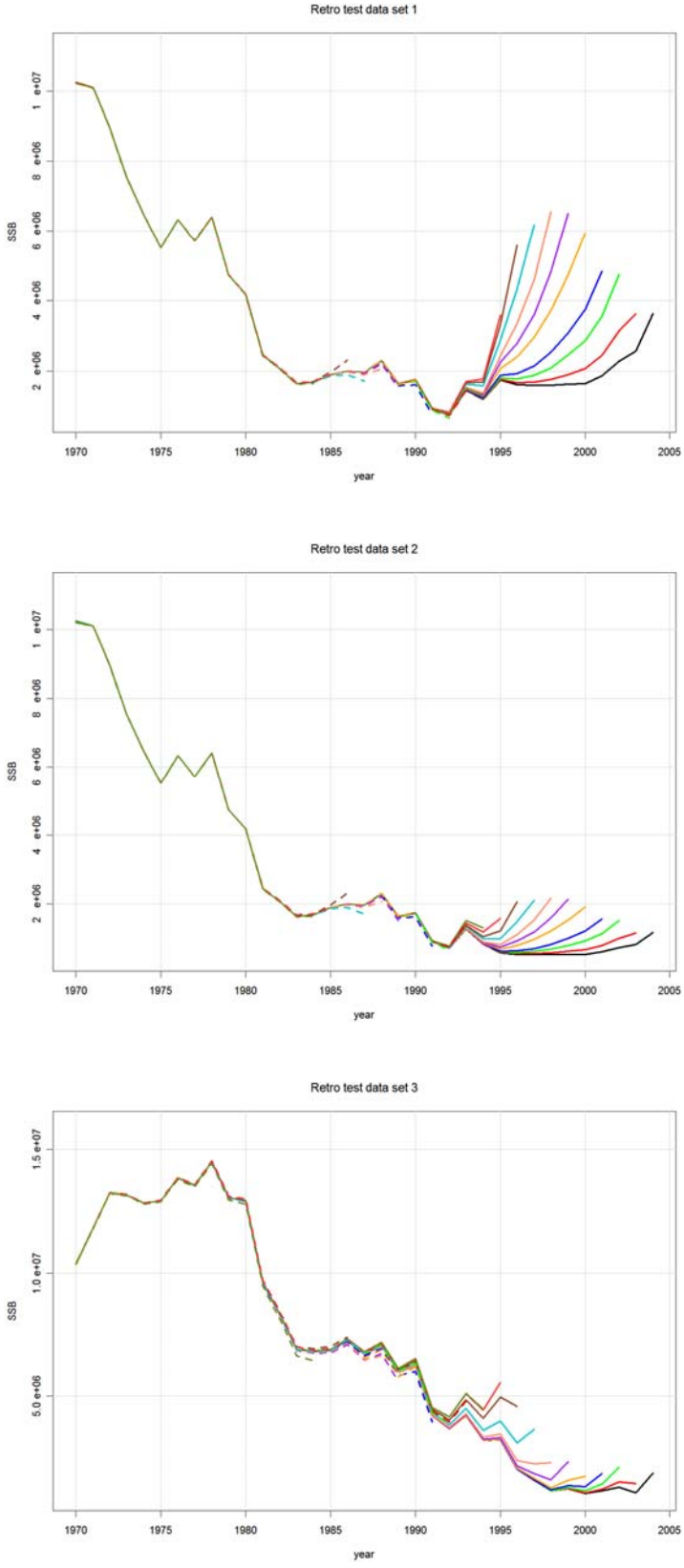

Figure 3.9. Retrospective pattern in annual spawning stock biomass (SSB) for data sets 1–3 (top – bottom).

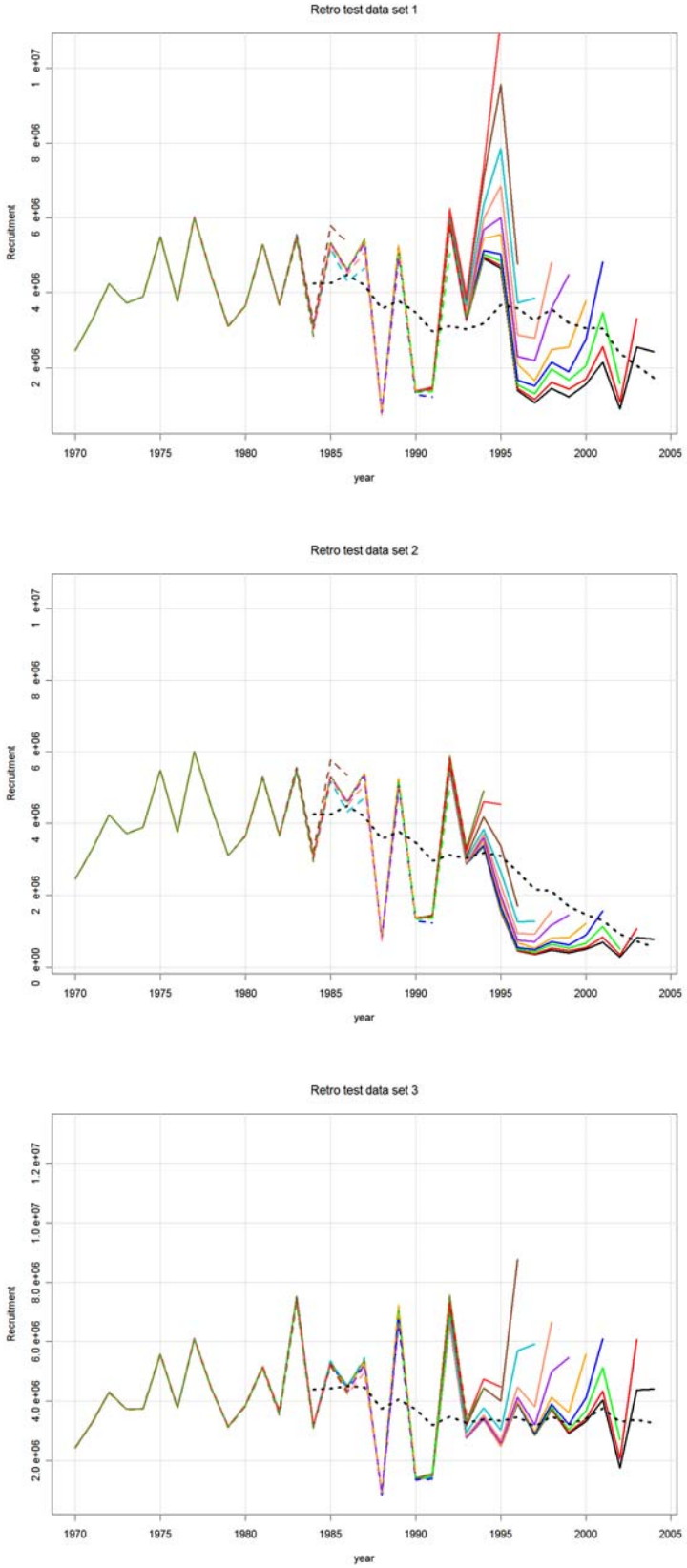

Figure 3.10. Retrospective pattern in annual recruitment for data sets 1–3 (top – bottom). Dashed black line is 10 year geometric mean recruitment estimated for each retrospective VPA.

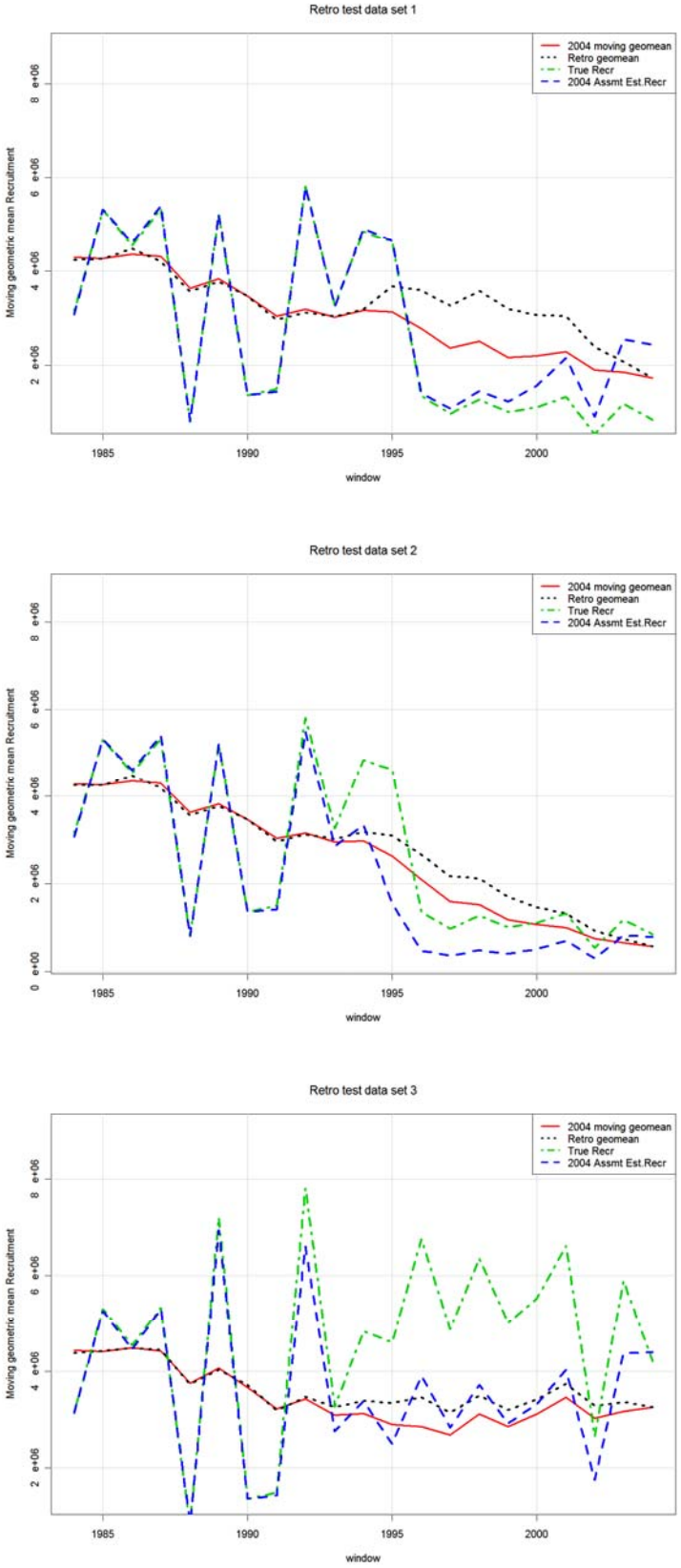

Figure 3.11. Comparison of true (green dot-dash) versus estimated recruitment (blue dash). Dashed black line is 10 year geometric mean recruitment estimated for each retrospective VPA, while solid red line is 10 year moving geometric mean from 2004 assessment.

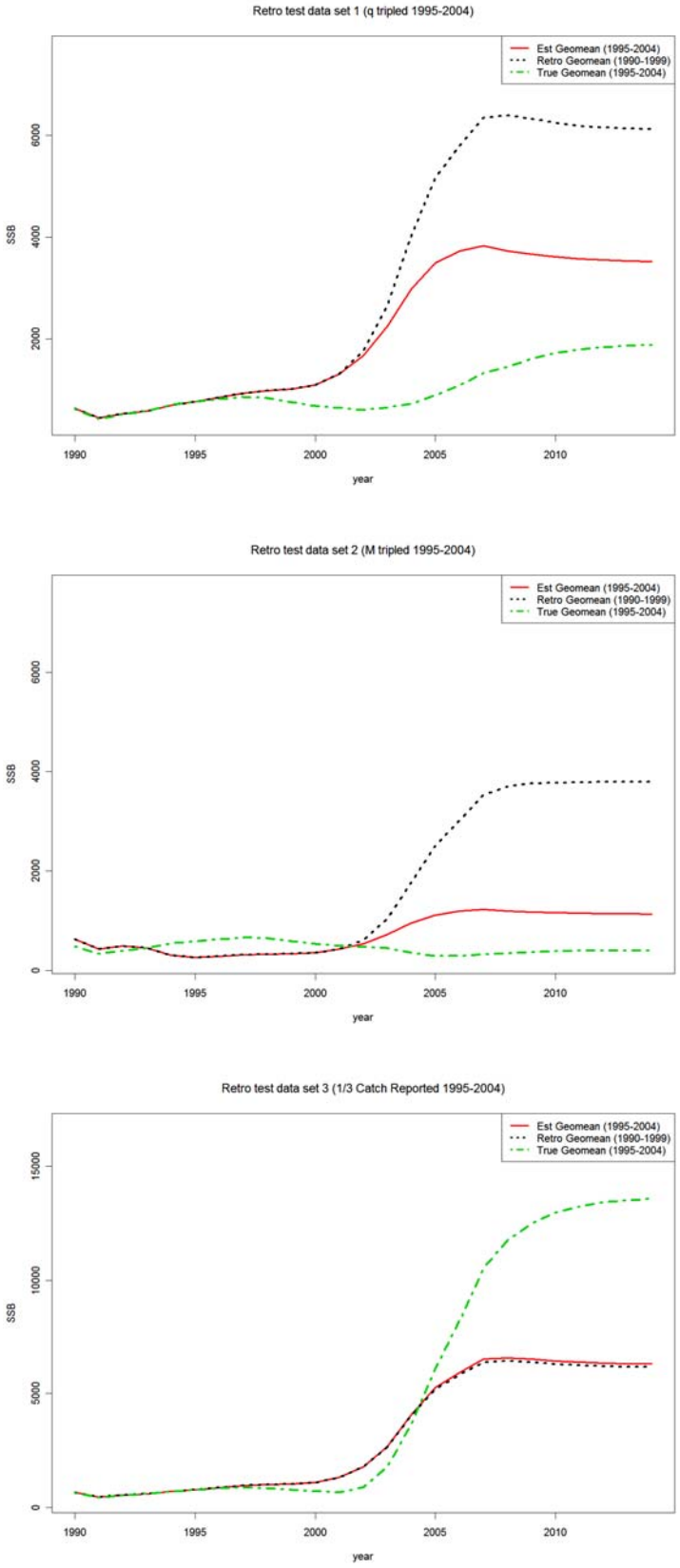

Figure 3.12. Comparison of true (green dot-dash) versus estimated SSB from projections based on 10 year geometric mean from 2004 assessment (1995–2004, solid red line) or 10 year retrospective geometric mean (1990–1999, dashed black line).

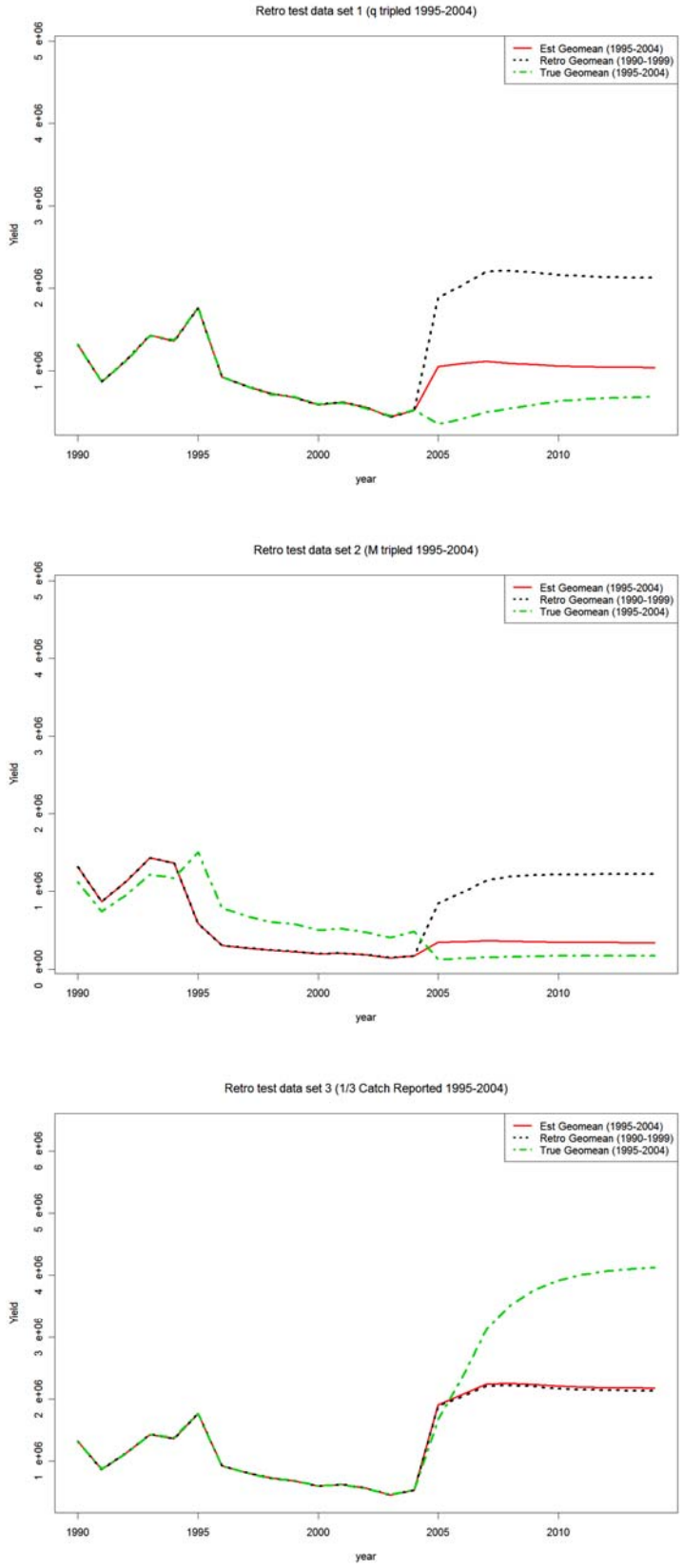

Figure 3.13. Comparison of true (green dot-dash) versus estimated yields from projections based on 10 year geometric mean from 2004 assessment (1995–2004, solid red line) or 10 year retrospective geometric mean (1990–1999, dashed black line).

Traffic-light approach

A simulated dataset was created to cover the period 1970–2005 (see Section 3.3.1). Survey catchability was increased substantially for the period after 1994. The simulated dataset provided true values that were then used as input for a modified version of an ADAPT virtual population analysis (VPA) model. The Traffic Light Approach (WP10) was used to summarise the total biomass and fishing mortalities from the true and VPA estimations. The input data for the comparative traffic light performance reports was divided into three equal parts such that green, yellow and red colours were equally represented within each report. Horizontal lines were used to divide the data into lower, middle and upper thirds.

The objective for the exercise was to check whether the Traffic Light Approach could correctly identify that changes occurred after 1994. Simulated data were used as a check against the VPA output.

ADAPT VPA and true biomass estimates are presented in Figure 3.14. VPA and true biomass estimates were similar until 1994 after which the VPA estimate increased to 3,205 t in 1995, declined to 2,177 t in 1998; and finally increased to 4,198 t in 2005. The true biomass estimates increased to 3,141 t in 1995 and then gradually decreased to 1,244 t by 2005. The performance reports were able to illustrate major changes in biomass (Figure 3.15). The early part of the true and VPA datasets were green because biomass was high in both cases; however, the two datasets did not track each other well after 1994. As a result of the upward divergence, the VPA bounds were higher than the 1/3 bounds produced for the simulated data. After 1994, VPA estimates fluctuated around the lower bound and then finally increased above the upper bound. Therefore after 1994, the VPA performance report was red intermixed with yellow and then in the final year became green. The performance report for the true biomass estimates tracked from green to yellow and finally red as the biomass gradually decreased over the years.

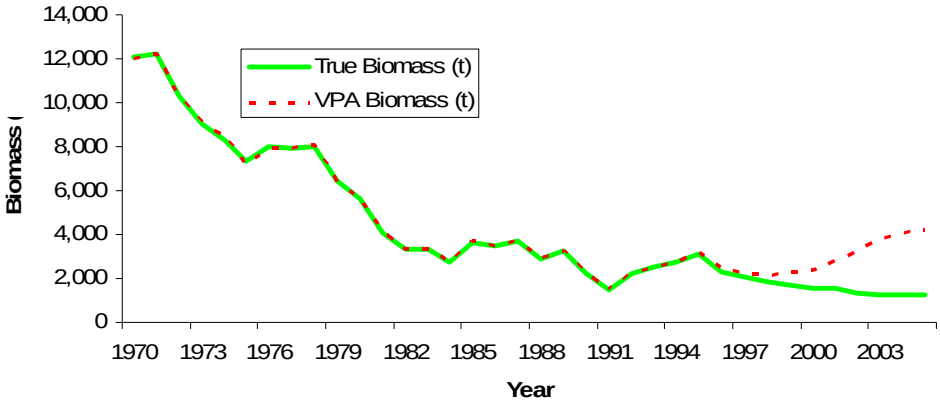

Figure 3.14. Simulated and VPA biomass estimates.

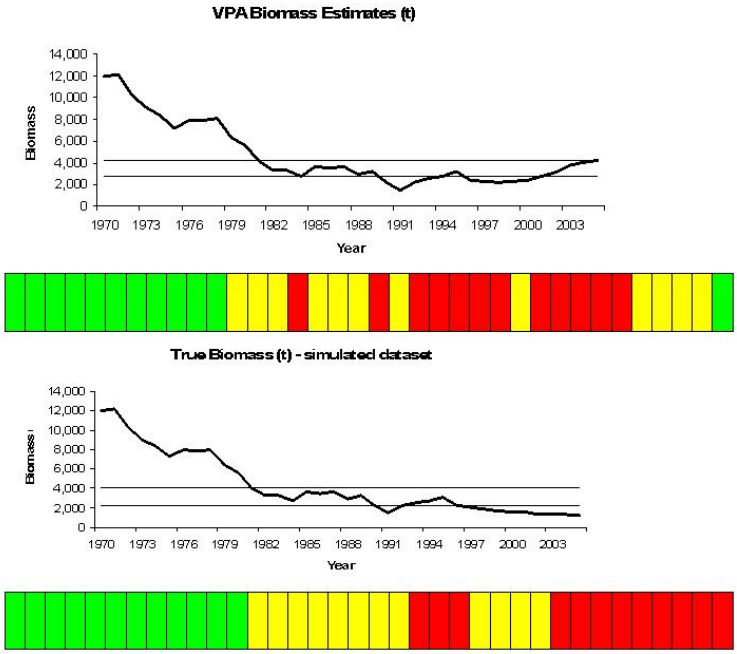

Figure 3.15. ADAPT VPA and simulated total biomass comparative performance reports. The horizontal lines divide the data into three equal sets of values.

Similarly, the VPA and simulated fishery mortalities tracked each other until 1997 when the VPA output indicated that fishing mortality decreased in relation to the simulated values (Figure 3.16). For this reason, the lower and upper bounds for the simulated performance reports were higher than they were for the VPA data performance reports. Once again this had implications for the traffic light summaries within each report (Figure 3.17). Even though the data tracked each other well until 1997, the performance report was more positive at the early portion of the simulated time series than for the VPA time series. Both sets of performance reports fluctuated between red and yellow between 1978 and 1994 after which the VPA performance report provided a more positive impression of the resource than did the simulated data report.

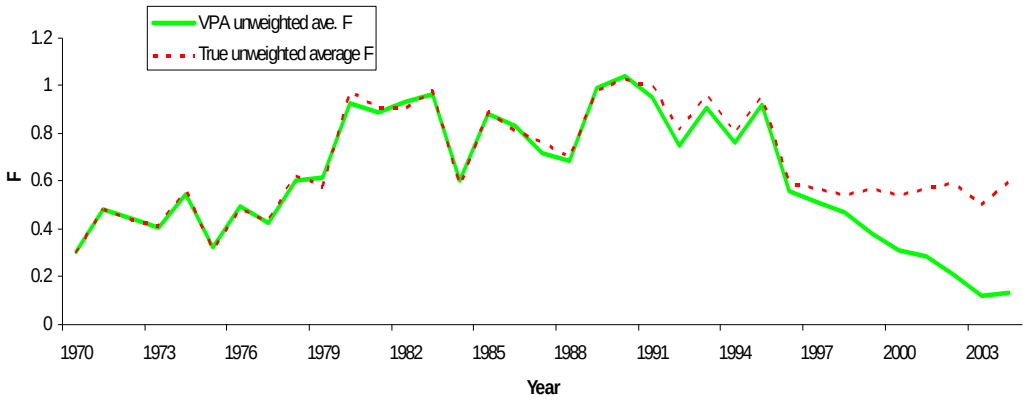

Figure 3.16. ADAPT VPA and simulated data fishery mortalities.

Figures 3.18 and 3.19 present the semi-quantitative summary performance reports. In this instance, the individual parameters are weighted equally and each colour is given a score; -1 for red, 0 for yellow and 1 for green. Through summing columns it is possible to obtain a numeric value that could feasibly, in a more fully developed performance report, be used in the development of harvest control rules. The VPA final score was positive in the final four years while the simulated data had a negative score in three out of the final four years. Therefore the VPA data would have given the managers the false belief that the resource was in good shape and that possible increases in catch could be achieved while maintaining the prescribed F value.

In well-developed traffic light reports additional parameters must be included. Care should be taken when choosing the parameters to ensure that they are clearly linked to the resource, that they do not duplicate information thereby biasing the results and are appropriately weighted. An appropriate weighting scheme is important because not all variables should have equal importance. For instance, trends in long term fishery independent biomass and recruitment indices are critical to overall status and could merit a higher score relative to trends in fishery dependent indices which are less reliable as stock indicators.

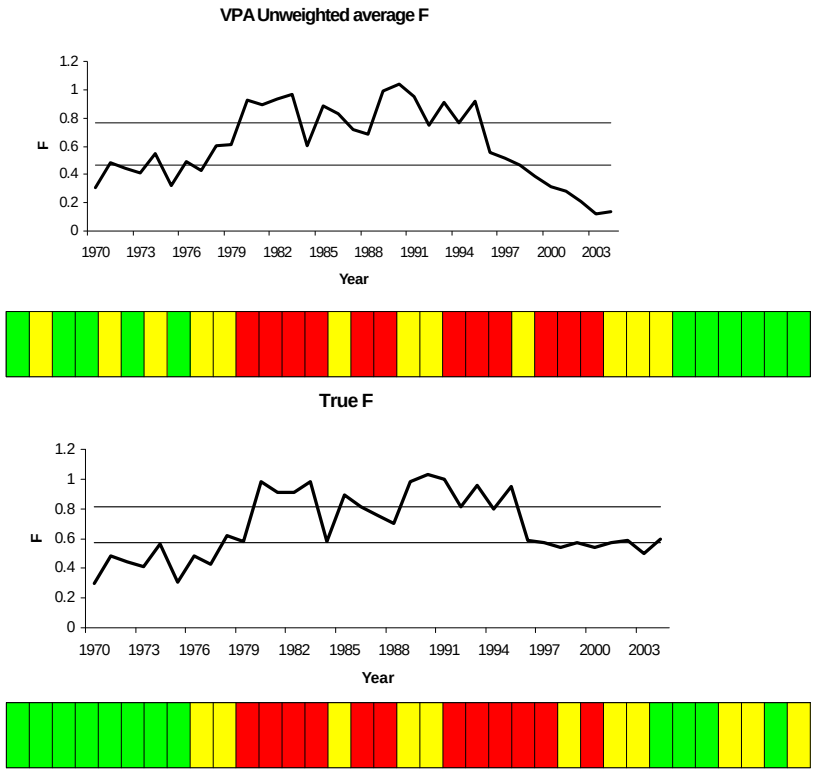

Figure 3.17. ADAPT VPA and simulated fishery mortality comparative performance reports. The horizontal lines divide the data into three equal sets of values.

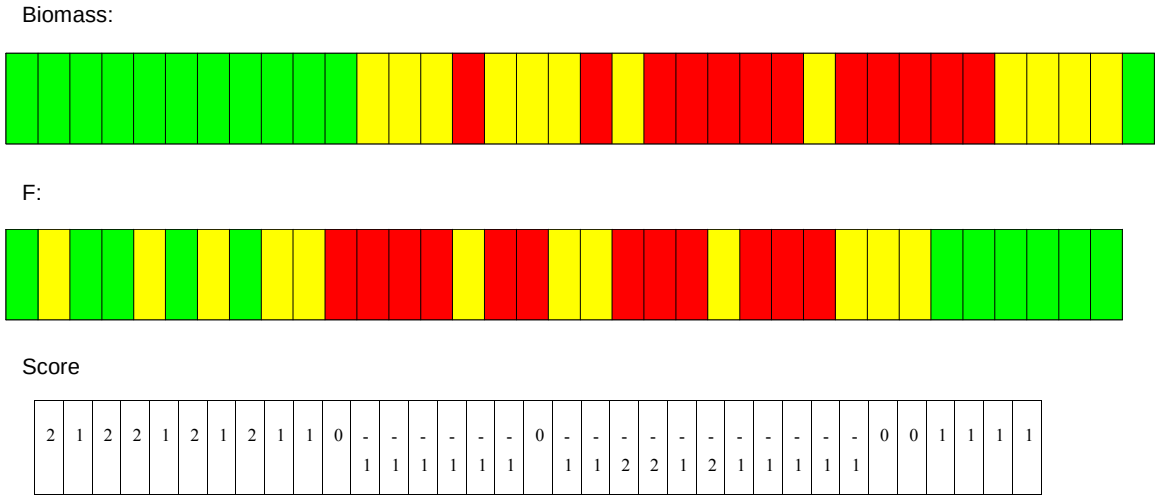

Figure 3.18. VPA summary performance reports.

Figure 3.19. Simulated data summary performance reports.

3.4 Examples from other management areas

Assessment and management approaches for three harvested resources in the Southern Ocean

A review was presented (WP20) on assessment and management approaches for three harvested resources in the Southern Ocean: Antarctic krill (*E. superba*); mackerel icefish (*C. gunnari*); and toothfish (*Dissostichus* spp.). The relevance of this review to WGMG is that while the management issues faced in the Southern Ocean are different in some ways to those encountered in the ICES, NAFO and ICCAT areas with which participants had most experience, there is a degree of cross-over and much that can be learnt by both sides.

Although the resources in the Southern Ocean have unique biological, life history, population, and ecological characteristics, their assessment and management approaches are united under the application of decision rules intended to ensure they meet the objectives and acceptable levels of risk defined under Article II of the Convention for the Conservation of Antarctic

Marine Living Resources (CCAMLR). The objectives require that all harvested resources in Southern Ocean (except seals and whales) must be managed to 1) prevent a decrease in size of harvested populations below that necessary for stable recruitment; 2) maintain ecological relationships between harvested, dependent and related species; and 3) prevent or minimize risk of changes not reversible over two or three decades. Accordingly, assessment methods must provide advice in relation to long term stock status, must be precautionary, and must consider the needs of predators and dependent species of the harvested resource.

The three objectives have been scientifically interpreted for implementation of resource management decision rules. The rules set the conditions for trigger levels in the assessment process for the purpose of generating advice on precautionary yield, and are represented as 1) a *depletion rule*, which finds the proportion (γ) of unexploited biomass (B_0) such that the probability of spawning stock dropping below 20% of its pre-exploitation median level is 10% (reduces risk to recruitment stability); 2) an *escapement rule*, which finds γ so that median spawning stock escapement is at 75%, or another ecologically appropriate level, of the pre-exploitation median level (permits a proportion of spawning biomass to escape the fishery to safeguard predators). Simulated populations are projected for at least 20 years, which leads to the third decision rule – 3) choose the lower value of yield that triggers either of the first two decision rules.

Some degree of flexibility is given within the management process in order to tailor the decision rules to the unique characteristics of the resources. In addition, assessment methods used to generate management advice are chosen that make optimal use of the data available. In all cases, advice must strictly adhere to the decision rules.

Antarctic krill are a key prey species in the Antarctic ecosystem, with a spatial distribution that can overlap with important land based predator foraging ranges. Information available for assessments include large scale acoustic surveys designed to estimate B_0, as well as localized annual surveys, low resolution fishery data, and length at age data from surveys. Assessments and management advice for this species are generated using the Generalized Yield Model (GYM; see Constable *et al*., 2000), an age-structured Monte Carlo population simulation and projection tool that allows for integration across uncertainties in population parameters. This approach (a cohort model) requires external estimation of population parameters, and draws upon probability density functions of abundance, growth, mortality, maturity and recruitment to form the basis of the population projections. In the case of krill, the estimate of precautionary γ is based on spawning biomass according to the decision rules described above, with natural variability in biomass driven by recruitment variability. The γ triggered by the decision rule is then used with the survey estimate of B_0 to form the basis of advice on total allowable catch. Further, because the fishery overlaps with several critical land-based predator forging grounds, there is an additional precautionary trigger level of catch, after which the fishery must adhere to quotas that are spatially subdivided according to small scale management units. This has been adopted as a further precaution so that krill will not be severely limited for land-based predators.

Mackerel icefish possess a different type of complexity in terms of their ecological role in several parts of the Southern Ocean. This species is short lived, relatively fast growing, a krill predator, a prey species for top level predators, and demonstrates large and short lived pulses in year-class strength, with a complex and variable natural mortality rate *M*. The data available for assessments include semi-annual (every two years) trawl surveys, fishery data, and length at age from surveys and the fishery. Assessments are based primarily on results of scientific surveys. However because of the short lived nature of the species and pulse recruitment, precautionary catch is determined in biomass rather than relative to an estimate of pre-exploitation biomass, with biomass determined from projections of actual numbers in each cohort of the population, and decision rules based on based on total biomass at the end of a two year projection period. This approach was developed as an interim step to enable higher

catches in seasons when strong year classes are present and vice versa. It requires the calculation of a fishing mortality which would result in a probability of no more than 0.05 that the spawning stock (after fishing) would be less than 75% of the level that would have occurred in the absence of any fishing. The estimation of fishing mortality is achieved by using the bootstrap one-sided lower 95% confidence bound on the trawl survey estimate as the current stock biomass estimate.

Toothfish is a higher trophic-level species that demonstrates low comparative fecundity, spawns at a relatively late age, reaches 2.2m in length and 120kg, is slow growing, and lives up to 50 years. They are also the most high valued and economically important finfish species in the Southern Ocean. There are a substantially greater number of available data sources with which to conduct an assessment, including surveys, fishery data, catch at age data, CPUE indices, high resolution spatially-structured sampling, and mark-recapture data. This species is assessed using the C++ Algorithmic Stock Assessment Laboratory (CASAL; see Bull *et al*. 2005), a fully integrated modelling framework that allows multiple data sources to be combined into a single assessment. In the example given, the data used in the assessment included catch at age, mark-recapture, CPUE, and process error. The estimation included Maximum Posterior Density and likelihood profiles, as well as Monte Carlo Markov Chain (MCMC) estimates to derive long term yield under some modifications of the decision rules. These modifications include a reduction in the median spawning stock escapement to 50% to reflect the higher order trophic-level of this species, and a 35 year forward projection period, owing to the longer life-span of the species. The decision rules are triggered (or not) by projecting forward each draw of the parameter values from the posterior distribution of the MCMC sample. Some of the challenges identified for the assessment of toothfish, and the use of the integrated modelling framework, include uncertainties in model structure, reliable estimates of M and steepness, and weighting factors of datasets when they demonstrate conflicting information.

The next logical step for in the progression of Southern Ocean resource assessments is a comprehensive evaluation of the assessment methods employed and the management advice that they generate, best carried out in a management strategy evaluation framework. This could be approached by using CASAL as the assessment model and the CCAMLR precautionary approach as the harvest rule, leading to an evaluation of the robustness of the assessment model and harvest rules to the uncertainties in the assessments. It would be useful for WGMG to follow developments in this work, as methodologies may arise that would be appropriate in the broader context.

3.5 Conclusions and implications for management

In this Section, summaries of the work done by Subgroup A on management strategy evaluation have been presented. We focussed on the possible effect of retrospective bias (generated in a number of different ways) on the subsequent ability of managers to manage the stock using a simple harvest control rule (HCR). Note that the approach would be equally valid no matter what HCR was used – the purpose of WGMG with regards to this topic is to collate and develop methods for *evaluating* HCRs, not to develop and propose the HCRs themselves. We also explored methods for determining the onset of retrospective bias, and for summarising management information.

The first conclusion must be that we are not yet in a position to answer the questions of whether and how management should proceed in the presence of retrospective bias. While it is very clear that management in the presence of retrospective bias should be more precautionary (as bias of this kind will generally lead to overestimates of stock size), it is certainly not clear how cautious such management should be nor how such caution should be achieved. To a large extent, programming solutions to management strategy evaluation problems must be created specifically for each case, and this takes more time than was available during the

meeting. Added to this is the complexity of models generated using such toolboxes as FLR and FPRESS – such complexity allows for great flexibility, but greatly reduces our ability to produce rapid solutions.

The second conclusion is that any such management-strategy evaluation toolbox must allow for assessments to be run "live" as part of the evaluation loop. This is important in any case, but especially so when we are trying to explore the effect of assessment problems on future management. Thus far, FLR is the only toolbox that allows for this, but work is under way to add this facility to the NOAA toolbox and WGMG welcomes this.

The third and final conclusion is that managers will get it wrong if they base their decisions on biased advice. This is perhaps obvious, but the simulation work presented in this Section highlights the problem with great clarity. The options for what to do in such a case are less clear – here there must be cognisance of the concurrent work carried out by Subgroup C (Section 5) in particular. In other words, if we can detect retrospective bias and determine its proximate cause, then we may attempt to do something about it and determine the likely effects of such remedial action via management evaluations. We are not yet in a position to do this, but progress has been made.

Several groups within ICES and elsewhere are involved in management strategy evaluations. The principal contribution that WGMG should make to this process is in the collation and further development of methods to enable management questions to be answered. We should not propose management plans, but we should be able to advise managers what the likely consequences would be of any approach that they choose to suggest. Evaluations in this context need to recognise that assessments are not perfect, and that advice may need to be modified accordingly. This is the reasoning behind our focus on the effect of retrospective bias, and it is hoped that this work can be continued into the future.

4 Bayesian and bootstrap-based inference for surplus production models

4.1 Introduction

Subgroup B was tasked with exploring parameter estimation uncertainty for surplus production models. This was a rather different approach to that called for in ToR e) (the most relevant ToR), which asked the WG to "provide guidance on incorporation in assessments of estimates of variance in input data."However, the expertise available in the WG was better suited to examining output uncertainty rather than input uncertainty, and the WG considered this to be a justifiable interpretation of the ToR. Given this, the goal was to compare precision of classical and Bayesian estimators of model parameters. Bayesian posterior credible intervals and bootstrap confidence intervals for biomass in each year were compared. When doing this, the difference in interpretation between bootstrap confidence intervals and Bayesian credible intervals must be kept in mind.

The bootstrap confidence interval provides an approximation to a true confidence interval, which is an interval that would contain the true unknown value (e.g. biomass) in 95% of replicates from idealized repeated sampling from the survey experimental design or the statistical model used for inferences (and assuming that the model used is correct). On the other hand, the 95% Bayesian posterior credible interval represents the degree of belief that the parameter value belongs to the interval given the series of survey data that was actually observed (while also assuming that the model used is correct). This difference in the interpretation of uncertainty intervals stems directly from the different interpretations of probability under the Frequentist (as a limiting proportion) and the Bayesian (as a degree of belief) paradigms, and it is important that the meaning intended for "probability" is clearly specified every time the word is used in an assessment context. Suggested wording for different cases will be considered by the next meeting of this WG,

In large sample settings with well identified parameters in the likelihood, both maximum likelihood and Bayesian methods produce very similar answers. However, there can be substantial differences in data poor situations, such as with small sample sizes or when the likelihood contains badly identified parameters leading to rather flat areas in the likelihood or possibly containing ridges. The likelihood contours shown in this Section indicate that there are identifiably issues with the likelihood of the surplus production model considered here. Bootstrap confidence intervals were obtained using the program ASPIC, whereas Bayesian credible intervals were obtained by simulating the posterior distribution via Markov Chain Monte Carlo (MCMC), with the algorithms coded in R (R Development Core Team 2005) and WinBUGS.

A key consideration in specifying estimation models is whether process error (or temporal stochasticity) in population biomass is present along with observation error (or measurement error) in yearly indices. Traditionally, estimation models have assumed no process error (e.g. Prager, 1994), but the trend in recent analyses has concentrated on properly incorporating or accounting for these two error components using a state-space construct (e.g. Meyer and Millar, 1999; Punt, 2003; de Valpine and Hilborn, 2005). To properly account for process error in likelihood-based inference for surplus production models, the marginal probability distribution of the (observed) indices forms the likelihood that is maximized. This generally requires integrating the joint distribution of unobserved biomass and observed indices over the unobserved biomasses, which is straightforwardly achieved by the Kalman filter when both observed and unobserved data arise from normal distributions and the data are linearly related through time. However, the relationship of (log) biomasses in a surplus production model (i.e. the Schaefer model) is nonlinear over time. The nonlinear relationship of biomasses is not an issue for Bayesian inference because the numerical methods for obtaining posterior

distributions readily deal with these types of transformations of random variables (although the fitting process itself is rarely straightforward). However, approximations of the marginal distribution via various methods including the extended Kalman filter (e.g. Durbin and Koopman, 2001), unscented Kalman filter (e.g. Wang, 2007), numerical integration (de Valpine and Hastings, 2002), Monte Carlo kernel likelihood method (de Valpine, 2004) and particle filtering (e.g. Kitagawa, 1996) are required for likelihood-based inference.

Although we feel it is desirable to account for process error in population dynamics models through a state-space framework, this was not allowed by the non-Bayesian tool at hand (ASPIC). Hence, most of the analysis presented here relates to the no process error setting. However, we can run a Bayesian implementation of the model with process error in WinBUGS and results for this are presented towards the end of this Section.

4.2 Methods

4.2.1 Data simulation

We simulated two twenty-year time series of abundance, index and catch with expected initial biomass equal to carrying capacity ($K = 1\times10^8$) with no process error, but observation error in yearly indices. Both the initial biomass, B_1, and yearly index, I_t, were log-normally distributed random variables with $CV(B_1) = 0.3$ and $CV(I_t \mid B_t) = 0.3$. In addition, $E\left[\log(I_t) \mid B_t\right] = \log(qB_t)$ where index catchability is $q = e^{-7}$.

We used the Schaefer production model,

$$B_{t+1} = B_t + B_t r\left(1 - \frac{B_t}{K}\right) - C_t,$$

to determine yearly biomass and we chose catches that would yield one data set that would provide poor information for estimation of population parameters (one-way trip; Table 4.1) and one data set that would provide good information for estimation (informative; Table 4.2).

We also explored two types of bias (and their interaction) in the information used to fit the surplus production model for the informative and one-way trip time series: a 1/3 reduction in survey index catchability after year 10 in the time series and underreporting of all yearly catches by 20%. Thus, we considered models for eight data sets (see Table 4.3) where observations either reflect correct or mis-specified estimation models.

Table 4.1. Yearly biomass, survey index and catch (all in log scale) for the uninformative scenario.

YEAR	BIOMASS	INDEX	CATCH
1	17.87632	10.60427	16.72849
2	17.55427	10.47729	16.58648
3	17.16571	10.13409	15.96232
4	16.90605	9.845474	15.66472
5	16.66907	9.911835	14.87243
6	16.5822	9.685647	15.61749
7	16.2298	9.182972	14.93765
8	16.02418	8.776982	14.93322
9	15.74315	8.505444	13.9114
10	15.67379	8.886099	14.45437
11	15.44838	8.613998	14.20212
12	15.23432	8.577164	13.87263
13	15.05954	7.544464	13.57134
14	14.92115	8.259409	13.52547
15	14.75782	7.6885	13.94243
16	14.33465	7.046218	13.21409
17	14.07617	6.919658	12.75777
18	13.89122	6.431449	12.421
19	13.75086	6.876871	12.44799
20	13.56128	6.126399	11.9415

Table 4.2. Yearly biomass, survey index and catch (all in log scale) for the informative scenario.

YEAR	BIOMASS	INDEX	CATCH
1	18.10395	11.56734	15.51704
2	18.2126	11.47647	15.86834
3	18.24757	10.85076	17.42778
4	17.8482	10.74642	17.17605
5	17.61816	10.78815	16.45653
6	17.68889	10.27104	17.44436
7	17.14449	10.32704	16.89709
8	16.82125	9.936434	16.31448
9	16.77642	9.473897	16.28462
10	16.72856	9.308249	14.80156
11	17.0831	10.59037	15.16012
12	17.39798	10.24365	14.96708
13	17.70569	10.87514	15.15182
14	17.95238	10.98196	16.00884
15	18.06442	10.51518	15.23531
16	18.20482	11.34332	16.31378
17	18.1897	11.22044	16.45615
18	18.15688	10.78851	16.67741
19	18.08919	11.10594	17.02609
20	17.92916	10.69974	17.32608

Table 4.3. Different data scenarios for each informative and one-way trip data scenarios in Tables 4.1 and 4.2.

NAME	OBSERVED CATCH	TRUE INDEX CATCHABILITY
CgIg: Catch good Index good	100% of true catch	e^{-7} years 1–20
CbIg: Catch bad Index good	80% of true catch	e^{-7} years 1–20
CgIb: Catch good Index bad	100% of true catch	e^{-7} years 1–10 and $e^{-7}/3$ year 11–20
CbIb: Catch bad index bad	80% of true catch	e^{-7} years 1–10 and $e^{-7}/3$ year 11–20

4.2.2 Likelihood function

The likelihood-based estimates of yearly biomass and biomass at MSY were obtained from ASPIC (Prager 2005) and Bayesian posterior distributions of these parameters were obtained from programs written in R (see Annex 4). ASPIC can make use of different types of data, in that the observations can be CPUE or catch, but the program assumes that the CPUE and catch data are directly related. Although the survey index in our data assumption is a CPUE, it is not directly related to catch. However, of the two types of data ASPIC may use, CPUE observation is most appropriate. The general likelihood we use to model observed indices is

$$L = \prod_{t=1}^{n} f\left[\log\left(I_t\right)\right] = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi v}} \exp\left\{-\frac{1}{2v}\left[\log\left(I_t\right) - f\left(\theta_t, C_t\right)\right]^2\right\}.$$

where n is the number of years of data and the variance parameter is v. The expected value of the log-index, $f(\theta_t, C_t)$, appears to be a complex function of parameters in the ASPIC algorithm (although it is not clear why this should be; see equations 8a and 8b and following objective functions in Prager, 1994) whereas it is simply $\log(qB_t)$ (where B_t is defined above) for the likelihood used in the Bayesian analyses.

4.2.2.1 Investigation of likelihood surfaces for r and K

As a preliminary exploratory data analysis, here we look at the profile likelihood surfaces of log(r) and log(K) to investigate the information available in the data for fitting the Schaefer production model. Given the above likelihood with $f(\theta_t, C_t) = \log(qB_t)$, it can be shown that the maximum likelihood estimates of log(q) and v are given by

$$\log(\hat{q}) = \frac{1}{n}\sum_{t=1}^{n}\left[\log(I_t) - \log(\hat{B}_t)\right],$$

and

$$\hat{v} = \frac{1}{n}\sum_{t=1}^{n}\left[\log(I_t) - \log(\hat{q}) - \log(\hat{B}_t)\right]^2 .$$

This allows us to compute the joint profile likelihood of log(r) and log(K). This is done over a grid of values of log(r) and log(K), assuming that we can find the maximum value in the entire space of r and K, we can compute a likelihood ratio surface where each point is approximately distributed on a χ^2 distribution with 2 degrees of freedom. We plot approximate 90, 95 and 99% confidence sets on the log likelihood surfaces for each simulated data set and refer to them as confidence contour plots understanding that they apply to log(r) and log(K).

The informative case

The confidence contour plots for the four data sets indicate a strong negative correlation between $\log(\hat{r})$ and $\log(\hat{K})$, and show that both these parameters are identifiable in the sense that the confidence contours are closed (that is, the maximum in each case is fully enclosed by the confidence contours; Figures 4.1–4.4). However, the high correlation means that the 95% confidence contour contains a wide range of plausible values for log(r) and log(K), but in particular log(r). This effect worsens in the case where the survey index, q, is reduced by one third in the second half of the series (the cases CgIb and CbIb) where $\log(\hat{K})$ and $\log(\hat{r})$ are even more correlated than the constant survey catchability cases. The shapes of the confidence contours are less sensitive to constant underreporting (cases CbIg and CbIb).

The One Way Trip case

The main features of the confidence contours are that $\log(\hat{r})$ and $\log(\hat{K})$ are negatively correlated, but not identifiable, in the sense that the confidence contours are not closed and extend to negative infinity on the log(r) axis (Figures 4.5–4.8). The main features of the confidence contours do not change between the different data sets. It appears to be possible to put upper and lower bounds on log(K), and this is perhaps the effect of fixing initial biomass to be K. It is possible to estimate an upper bound on r, the intrinsic population growth rate, with a lower bound of zero. This is not surprising as the underlying biomass follows a consistent decline which; given the catches can be explained in the extreme case by a growth rate of zero.

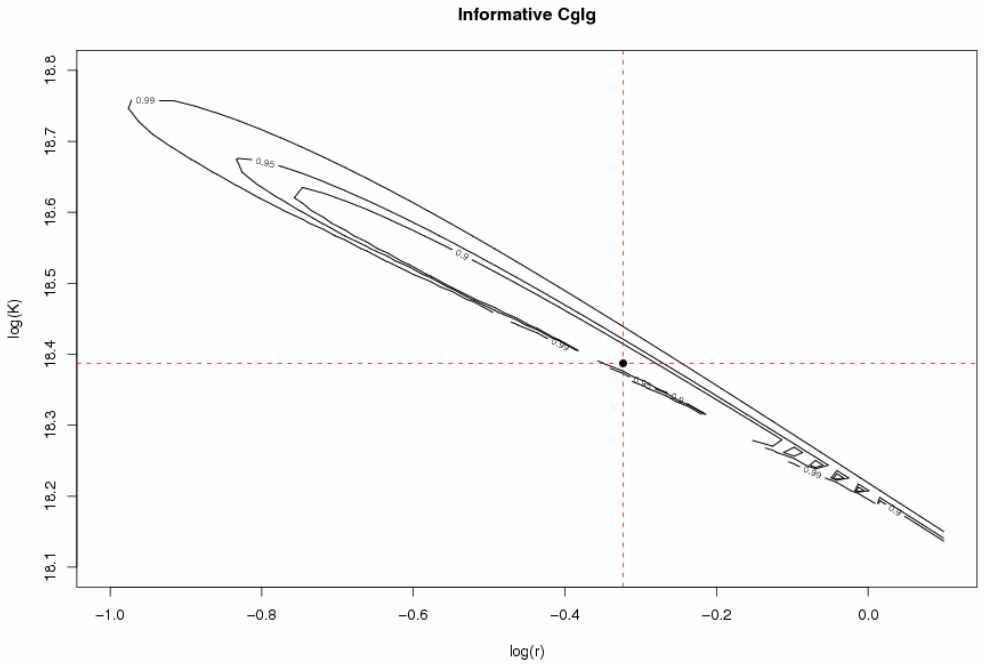

Figure 4.1. Joint-likelihood profile with 90, 95 and 99% confidence contours for log(*r*) and log(*K*) using the informative CgIg data scenario.

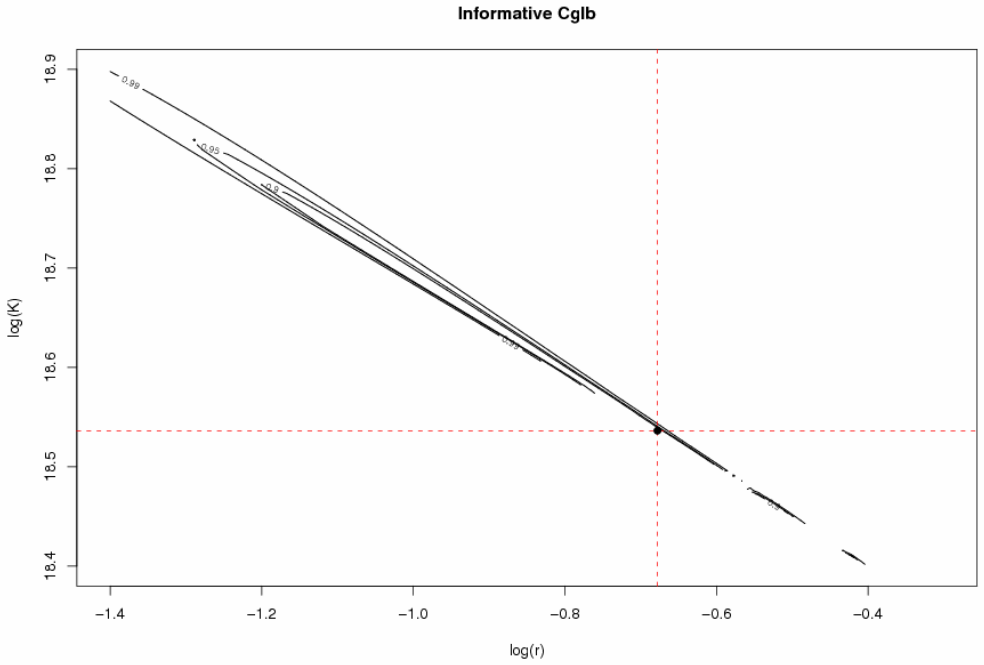

Figure 4.2. Joint-likelihood profile with 90, 95 and 99% confidence contours for log(*r*) and log(*K*) using the informative CgIb data scenario.

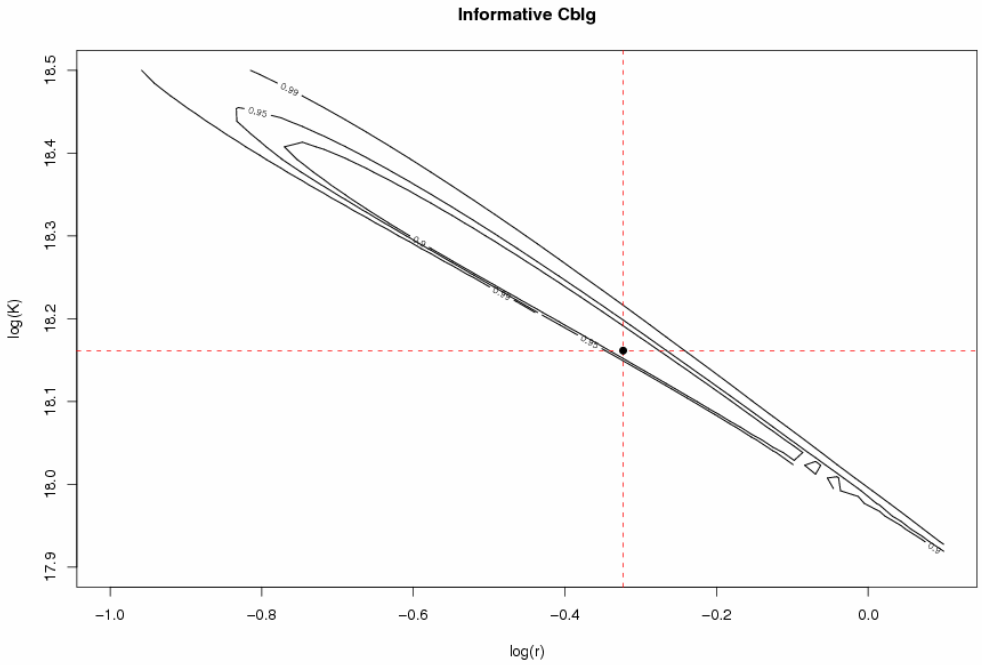

Figure 4.3. Joint-likelihood profile with 90, 95 and 99% confidence contours for log(*r*) and log(*K*) using the informative CbIg data scenario.

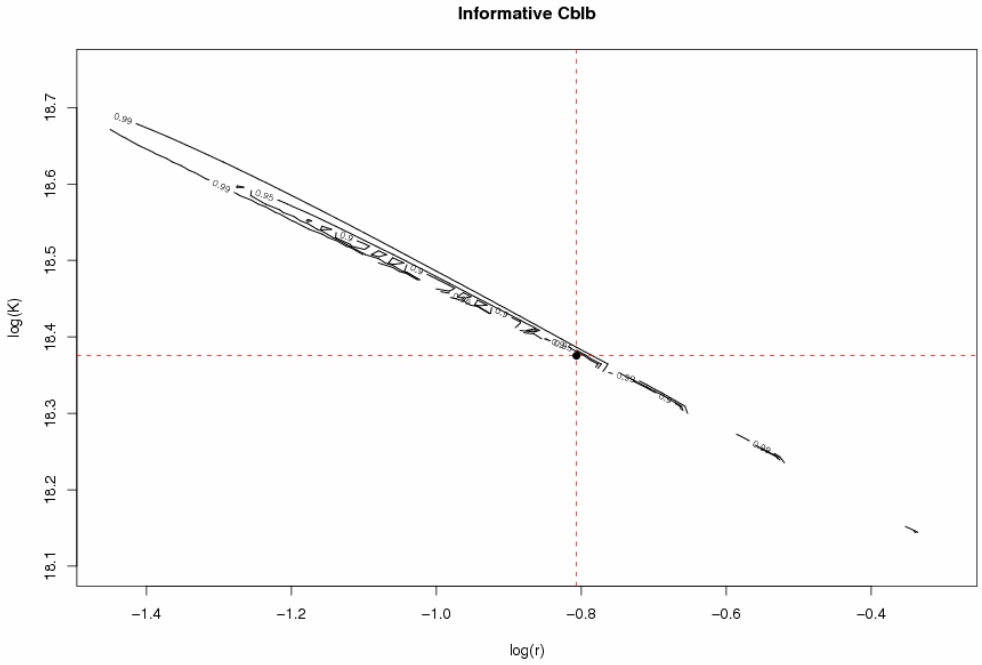

Figure 4.4. Joint-likelihood profile with 90, 95 and 99% confidence contours for log(*r*) and log(*K*) using the informative CgIb data scenario.

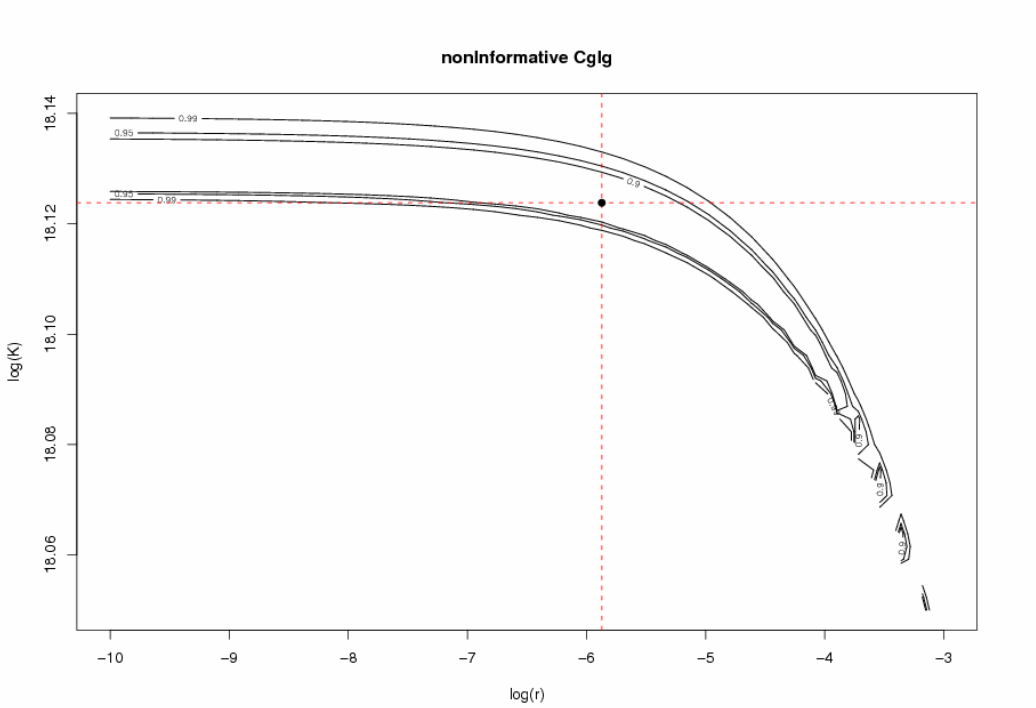

Figure 4.5. Joint-likelihood profile with 90, 95 and 99% confidence contours for log(*r*) and log(*K*) using the one-way trip CgIg data scenario.

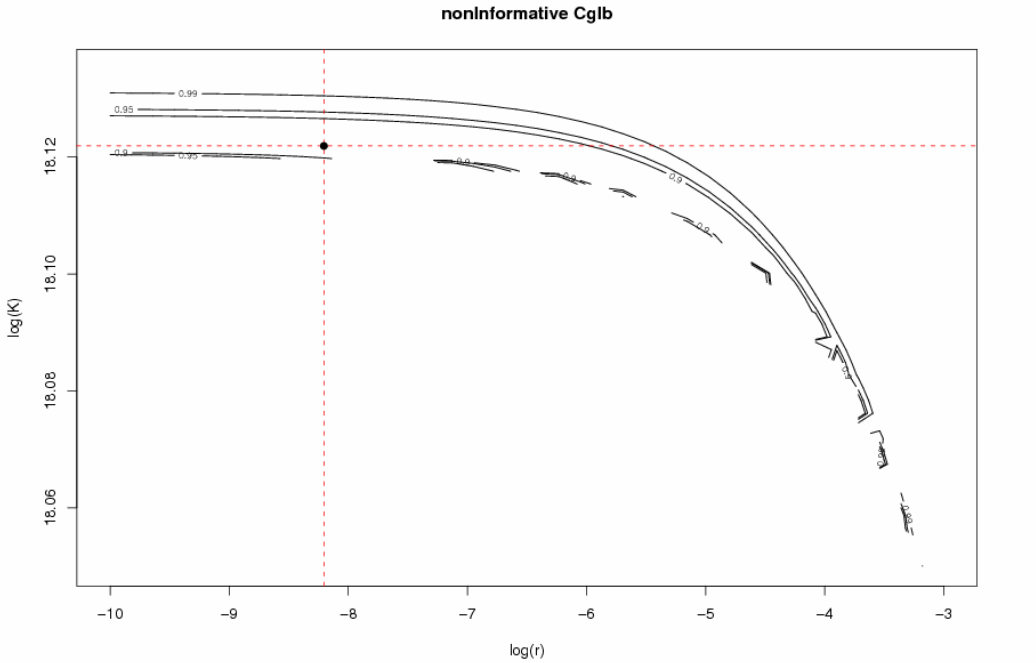

Figure 4.6. Joint-likelihood profile with 90, 95 and 99% confidence contours for log(*r*) and log(*K*) using the one-way trip CgIb data scenario.

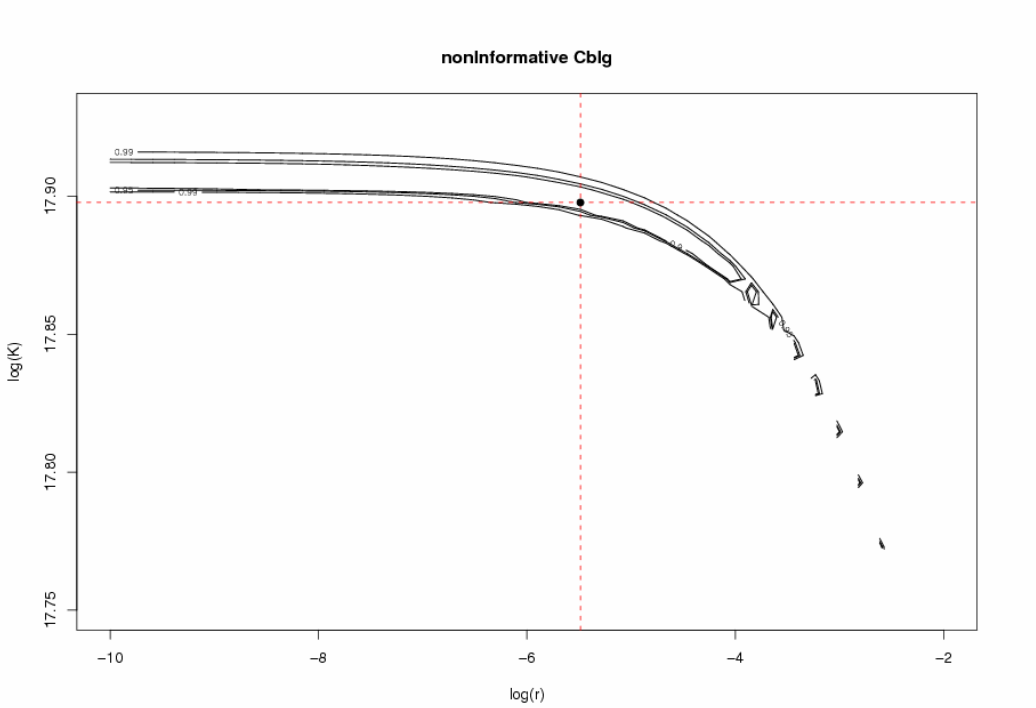

Figure 4.7. Joint-likelihood profile with 90, 95 and 99% confidence contours for log(*r*) and log(*K*) using the one-way trip CbIg data scenario.

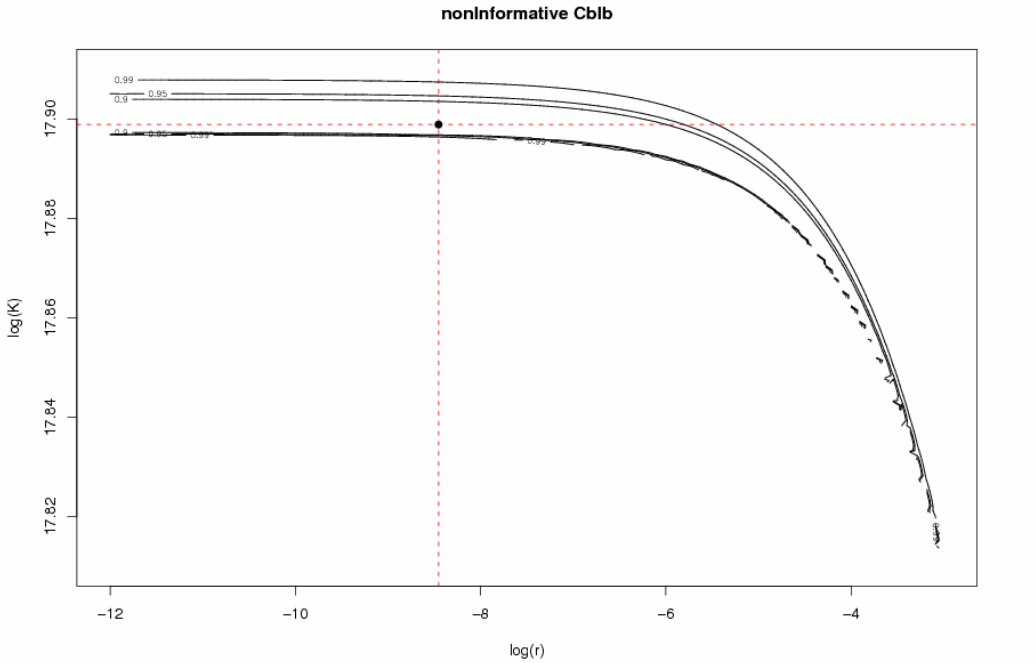

Figure 4.8. Joint-likelihood profile with 90, 95 and 99% confidence contours for log(*r*) and log(*K*) using the one-way trip CbIb data scenario.

4.2.3 Bayesian Priors and computational algorithms

For the Bayesian analyses prior distributions on *K*, *r*, *q* and *v* were considered. In a Bayesian context, a prior distribution is meant to capture the knowledge and uncertainty available before observing the data. In the absence of such knowledge, a common strategy is to use a so-

called vague (or dispersed) prior, which is in general terms a prior distribution that is fairly flat in the areas where the likelihood is high. To aid comparability of Bayesian and bootstrap results, fairly dispersed priors were chosen, in order to prevent them from having a strong impact on posterior results. The prior distributions were as follows:

$$\log(K) \sim N\left(\mu = 18.5, \sigma^2 = 3.26\right)$$
$$\log(r) \sim N\left(\mu = -0.35, \sigma^2 = 3.26\right)$$
$$\log(q) \sim N\left(\mu = -7, \sigma^2 = 3.26\right)$$

where the chosen values for the variances correspond to a coefficient of variation of 5 in the original scale before taking logarithms.

A Gamma prior distribution was chosen for the inverse variance of the survey index, as follows:

$$1/v \sim \text{Gamma}\left(\text{shape} = 0.592, \text{rate} = 0.013\right)$$

which implies a prior distribution for the coefficient of variation of the survey with median value 0.2 and a 90% central credible interval of (0.008, 3.30).

These prior distributions were used in all the cases examined.

Markov Chain Monte Carlo (MCMC) methods are the standard computational algorithms for simulating from the posterior distribution in a Bayesian context. They work by breaking down the dimensionality of the posterior distribution. The various model parameters are split into groups of one or more parameters each (often each group consists of just one parameter). The algorithm is initialised at some random value. Then instead of simulating draws for the multidimensional parameter in a single step, each individual group of parameters is simulated in turn, conditioned on the current values of the other groups of parameters in the Markov chain. In other words, the simulation is done from the conditional posterior distribution (sometimes termed full conditional distribution) of the group of parameters, which corresponds to drawing from a lower dimensional (often unidimensional) slice of the joint posterior distribution. MCMC algorithms are extremely powerful and can be generally tuned to simulate from complex high-dimensional distributions effectively. Gilks *et al*. (1996) provide a useful practical introduction to MCMC.

The parameters of the Schaefer surplus production model were estimated in a Bayesian setting using random walk MCMC for the log catchability log(q), log intrinsic growth rate log(r), and log carrying capacity log(K) and an independence sampler was used for the standard deviation of the observation process, $\sqrt{v}$. All data sets were fitted using the same proposal distribution for $\sqrt{v}$, while log(K) and log(r) were block updated using a bivariate normal random walk with a correlation of -0.5 for all data sets and log(q) was updated on its own using a normal random walk. Only the random walk variances changed when fitting a different data set. Much fine tuning was required of the random walk variances in order to achieve the presented results. With regard to this, some concerns remain as to whether the MCMC chains have converged. The convergence difficulties experienced stem from the ill conditioning of the likelihood, as the profile likelihood surfaces of r and K presented above have illustrated. For the Informative dataset, the situations that caused most trouble were the two in which survey catchability changed half way through the data set. Severe convergence issues for the one-way trip dataset made it impossible to present Bayesian results for it in this report.

4.2.4 ASPIC initialization

The ASPIC algorithm requires the user to specify initial guesses for K, q, and *MSY* and bounds that the algorithm will search for K and *MSY*. The initial guesses for *MSY* and q for all 8 data

scenarios were functions of the observed catches and indices, but the exact values that were ultimately used were determined by trial and error because the algorithm frequently failed to run. The initial guesses for *MSY* were chosen as a scalar of either the maximum or mean of observed catches whereas the initial guesses for *q* were chosen as a scalar of the mean ratio of observed indices and catches. The initial guess for *K* was set at the true value of *K* for the data set (1×10^8, not the realized initial biomass). Ranges of *MSY* were also chosen as scalars of either the mean or maximum of observed catches and the upper bound of *K* was required to be larger than the upper bound on *MSY*.

To assess uncertainty, 95% confidence intervals were calculated for yearly biomasses by fitting 1000 bootstrapped data sets using procedures also available within ASPIC.

4.3 Results

4.3.1 Results for Informative Dataset:

For the 4 data scenarios considered, namely Catch good Index good (CgIg), Catch good Index bad (CgIb), Catch bad Index good (CbIg) and Catch bad Index bad (CbIb), Figures 4.9 and 4.11 – 4.13 plot the prior density (green solid line), posterior histogram and true values used to simulate the data (red vertical lines). The posterior distribution is much more concentrated than the prior, which appears in the figures as an almost horizontal line over the range where the posterior mass is.

For the CgIg case, Figure 4.10 presents a pairwise scatterplot of the posterior distribution, highlighting very strong negative correlation between *K* and each of *r* and *q*, and positive correlation between the latter two parameters. The coefficient of variation of the survey index displays some posterior correlation with the model parameters, although this does not appear to be very severe. Similar pairwise scatterplots (not presented) were found for the other 3 data scenarios examined. As the priors used on parameters were fairly flat and independent, the severe correlations seen in the posterior distribution reflect the behaviour of the likelihood function. The strong negative association between *K* (carrying capacity and also assumed biomass in the initial year) and *r* (intrinsic growth rate) is not surprising, as large *K* values and small *r* values can lead to similar looking biomass trends to those obtained from small values of *K* coupled with large *r* values. As *K* is the assumed biomass in the initial year and the survey provides only a relative biomass index, it is not surprising either to see negative posterior correlation between *K* and *q* (survey catchability). These severe correlations were the cause of the computational difficulties alluded to earlier.

Figures 4.14 to 4.17 summarize Bayesian and bootstrap inference on yearly log(biomass) (with the vertical axis given in the original non-logged scale). The true log(biomass) series is plotted in red. For each time point, there are 2 vertical lines: the left one (in black) is the 95% posterior credible interval, with the posterior median indicated by an open circle; the right one (in grey) is the 95% bootstrap interval obtained from the ASPIC software, with the point estimate indicated by an open circle.

In the CgIg case, both Bayesian and bootstrap methods perform quite well, with the Bayesian method leading to somewhat wider intervals.

In the CgIb case, the survey index values were divided by 3 in the last 10 years, corresponding to a three-fold decrease in catchability, whereas the fitted model assumes constant catchability throughout the entire series. Hence, there is model misspecification. Figure 4.15 shows that the Bayesian method overestimates biomass at the beginning of the time series and underestimates it at the end. This is consistent with the posterior distributions depicted in Figure 4.11 which show an overestimation of *K* (assumed to be the initial biomass value), underestimation of *r* and an estimate of *q* which is in between the two true values in the series. The bootstrap estimates are always below the true value, being quite good at the beginning of

the series, but breaking down when the change in catchability occurs. The difference between the Bayesian and bootstrap estimates is striking. However, given that only a rather limited simulation study has been conducted and the convergence difficulties experienced with the computational methods, it does not seem sensible to draw any firm conclusions from this.

In the CbIg case, the survey catchability remains constant throughout the time series, but nominal catches (the ones used in the model fitted) are 80% of the value of the true catches (the ones used to generate the true biomass series). One might expect this to lead to underestimation of K (since the catches only enter the model via C_t / K) and, hence, also underestimation of the initial biomass (assumed equal to K in the model). In turn, from the posterior correlations observed in the pairwise scatterplot corresponding to the CgIg case, one might expect both r and q to be overestimated. Figure 4.12 confirms that the posterior distribution of K has shifted towards lower values with respect to the one obtained for the CgIg case, whereas the opposite holds for the posterior distributions of r and q. However, the net effect of these changes on the estimated biomass is not large (see Figure 4.16), causing just a shift towards somewhat lower values than those found in the CgIg case. A similar comment can be made for the results corresponding to the CbIb situation with respect to those for the CgIb case.

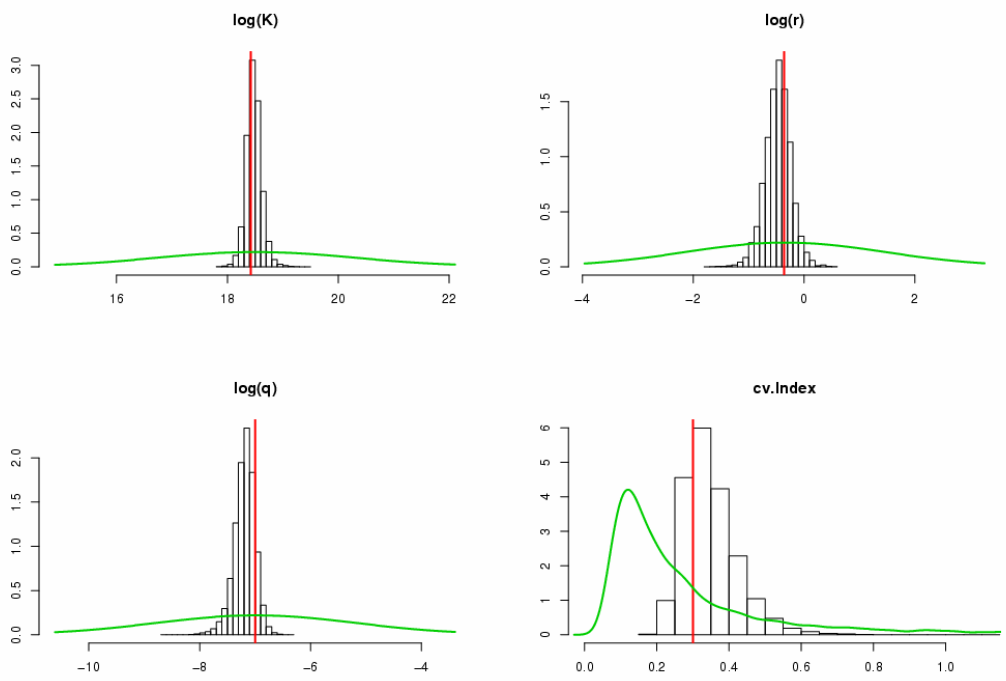

Figure 4.9. Prior (green) and posterior (histogram) distributions for parameters for the informative CgIg data scenario. The true parameter value assumed for data generation is denoted by the intersection of the red vertical line on the horizontal axis.

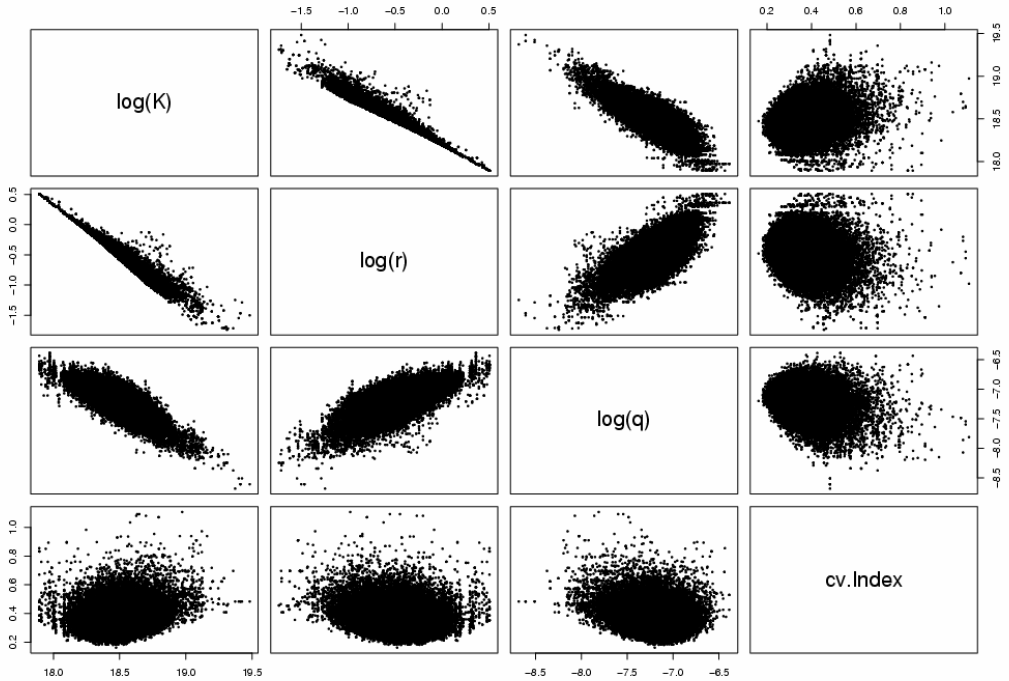

Figure 4.10. Pairwise joint posterior distributions for the informative CgIg data scenario.

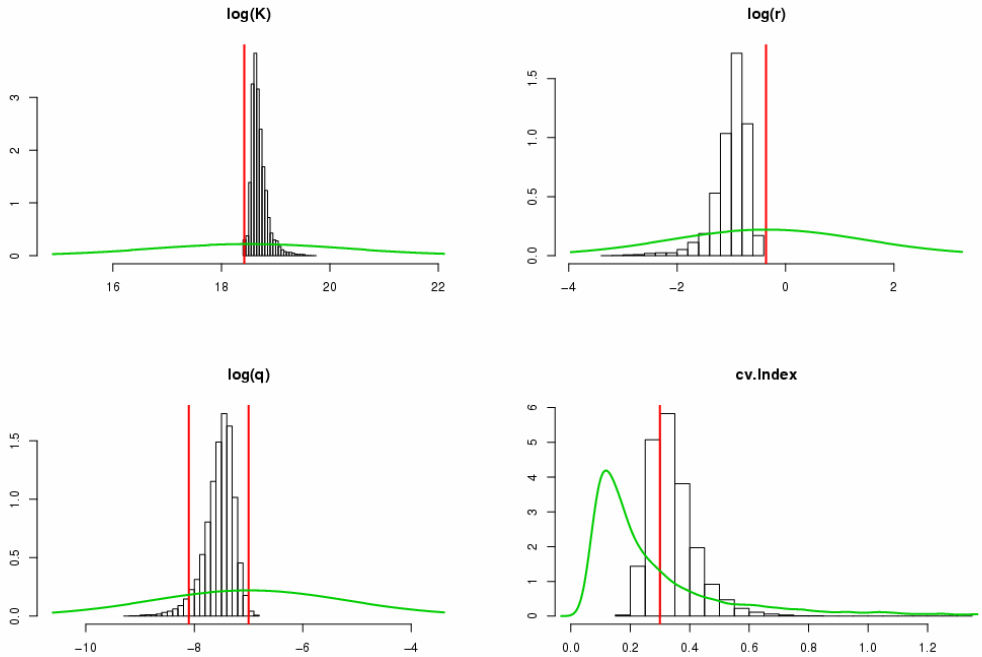

Figure 4.11. Prior (green) and posterior (histogram) distributions for parameters for the informative CgIb data scenario. The true parameter value assumed for data generation is denoted by the intersection of the red vertical line on the horizontal axis. The log(q) plot has two red lines because there are two true values of survey catchability, one for the first part of the time series and one for the second.

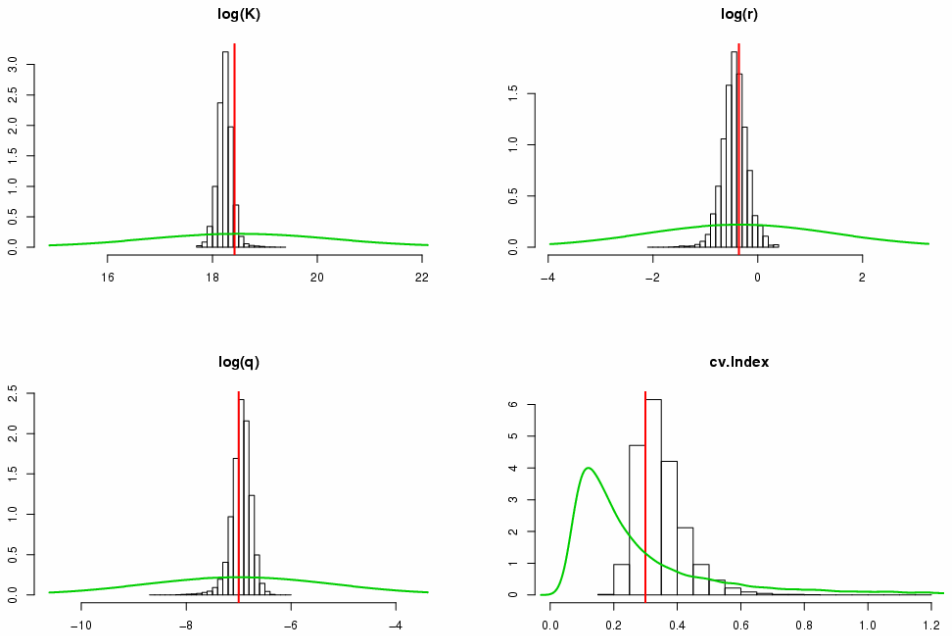

Figure 4.12. Prior (green) and posterior (histogram) distributions for parameters for the informative CbIg data scenario. The true parameter value assumed for data generation is denoted by the intersection of the red vertical line on the horizontal axis.

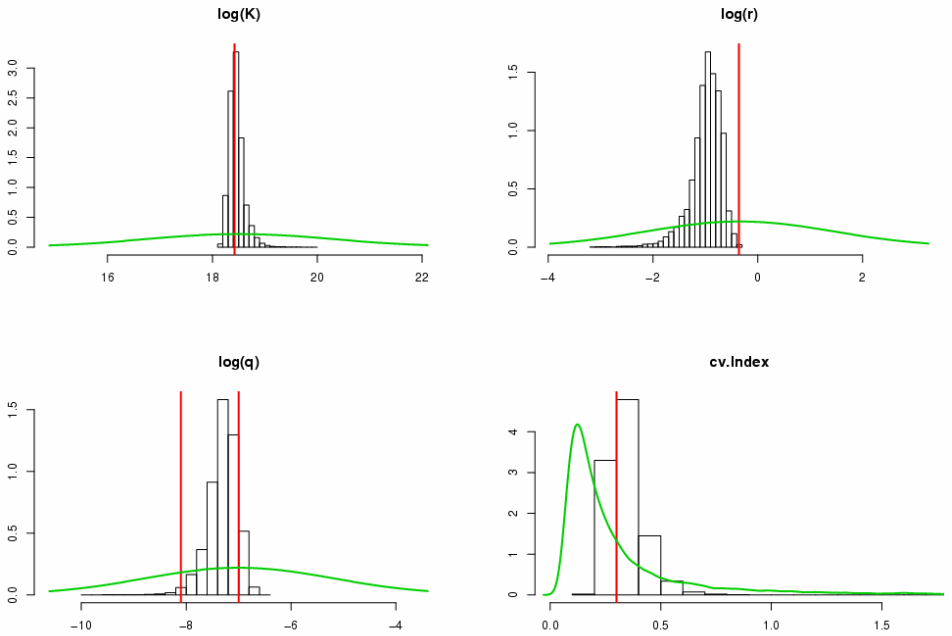

Figure 4.13. Prior (green) and posterior (histogram) distributions for parameters for the informative CbIb data scenario. The true parameter value assumed for data generation is denoted by the intersection of the red vertical line on the horizontal axis. The log(q) plot has two red lines because there are two true values of survey catchability, one for the first part of the time series and one for the second.

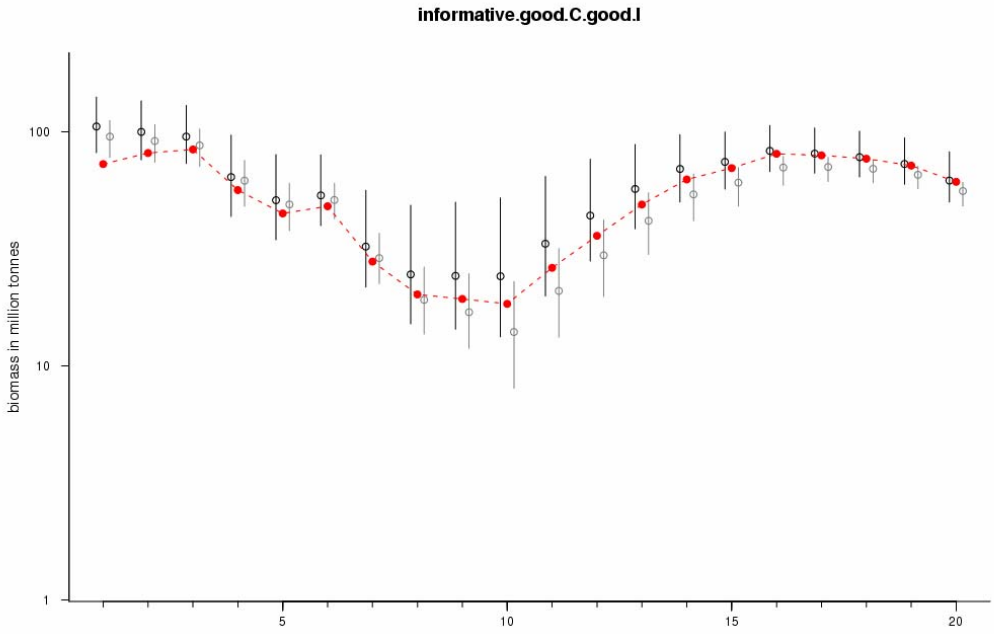

Figure 4.14. Bayesian 95% credible intervals (black), bootstrap 95% confidence intervals (grey) and true biomasses (red) for the informative CgIg data scenario.

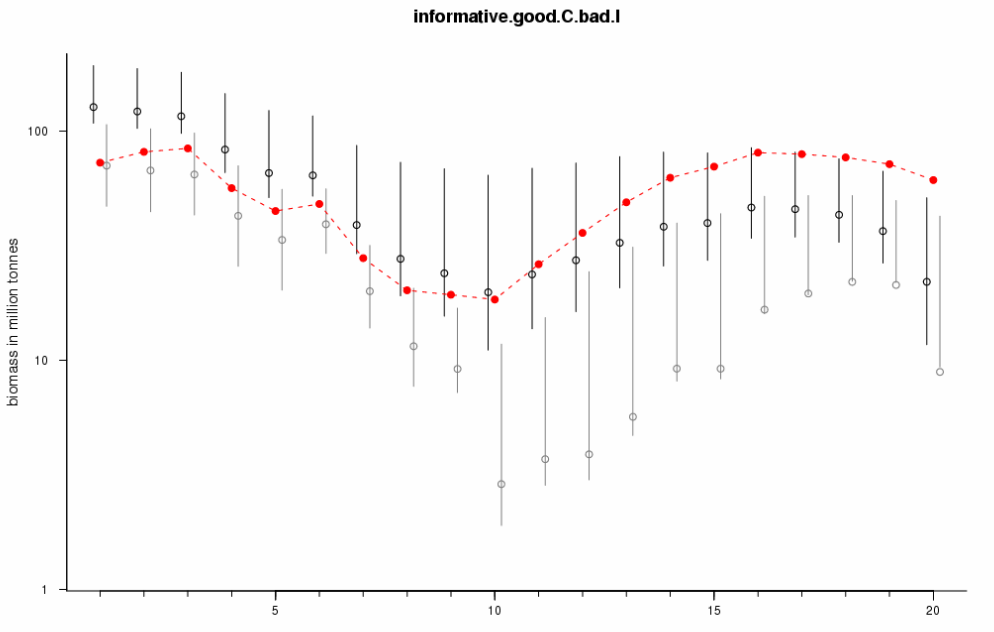

Figure 4.15. Bayesian 95% credible intervals (black), bootstrap 95% confidence intervals (grey) and true biomasses (red) for the informative CgIb data scenario.

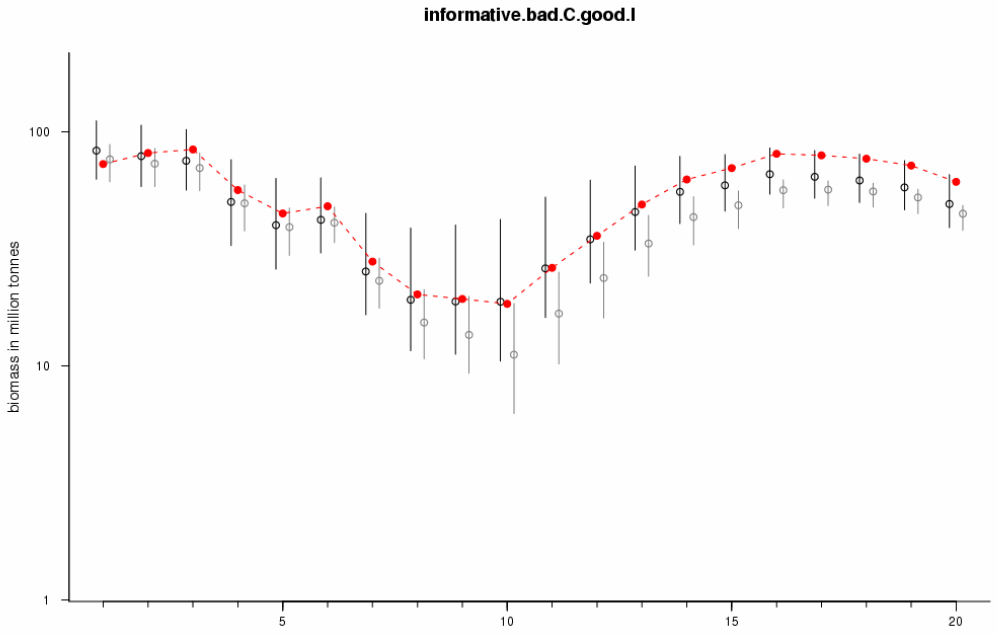

Figure 4.16. Bayesian 95% credible intervals (black), bootstrap 95% confidence intervals (grey) and true biomasses (red) for the informative CbIg data scenario.

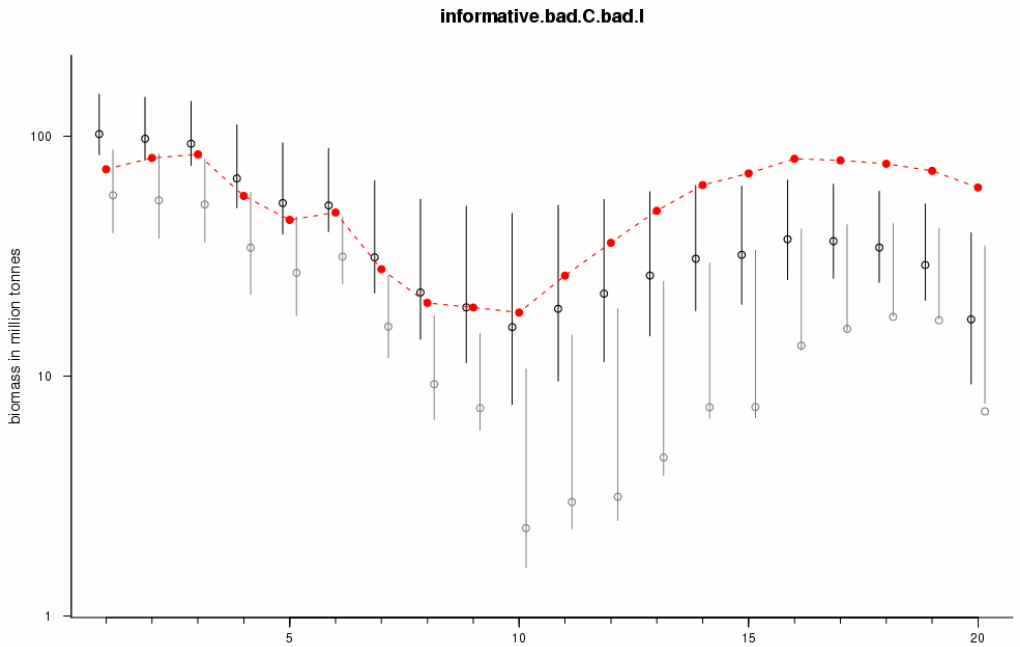

Figure 4.17. Bayesian 95% credible intervals (black), bootstrap 95% confidence intervals (grey) and true biomasses (red) for the informative CbIb data scenario.

4.3.2 Results for One Way Trip dataset:

The second dataset corresponds to a true series of biomasses which is always sloping downwards, which is typically thought of as a harder situation in terms of parameter estimation. The same four data scenarios were considered as in Section 4.3.1.

As indicated above, when trying to fit the Bayesian model with an MCMC program in R, convergence problems were so severe that the results were considered too unreliable for presentation. Hence, only bootstrap estimates, obtained via the software ASPIC, could be presented. These are displayed in Figures 4.18–4.21, with true log(biomass) in red and bootstrap 95% intervals in grey, with point estimates shown with an open circle. The bootstrap estimates for the CgIg case are quite reasonable, while they show some deterioration in the CgIb case. The CbIg case is not too different from that obtained in the CgIg situation. What is striking is the complete break down of the estimate in the CbIb situation. However, the computational procedure leading to this estimate should be explored more carefully before any firm conclusion could be drawn. It was noticed that the bootstrap confidence intervals for K and r were very far from the true values and close to the boundaries initially set for them in ASPIC. So our conjecture is that the algorithm has not found the likelihood maximum.

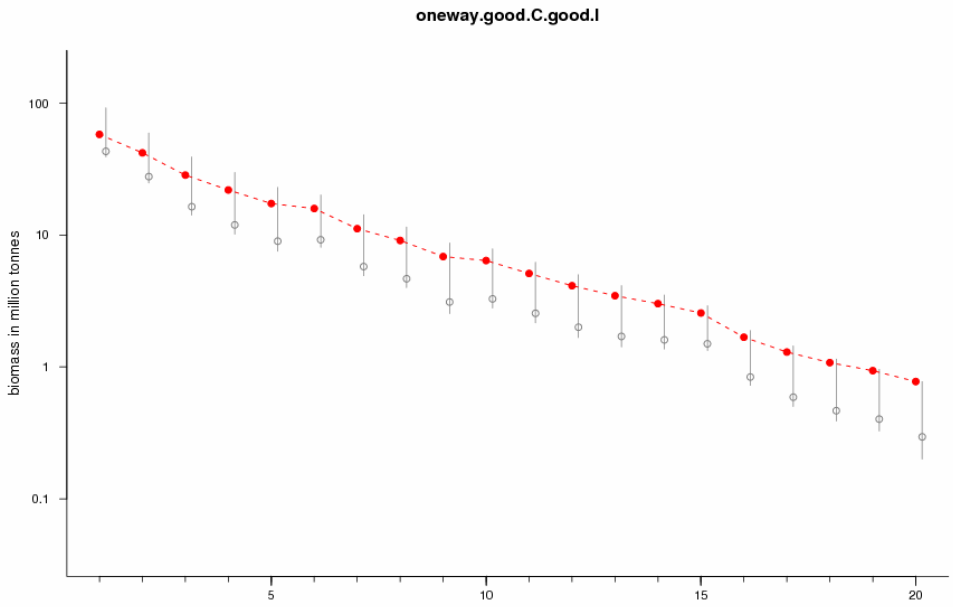

Figure 4.18. Bayesian 95% credible intervals (black), bootstrap 95% confidence intervals (grey) and true biomasses (red) for the one-way trip CgIg data scenario.

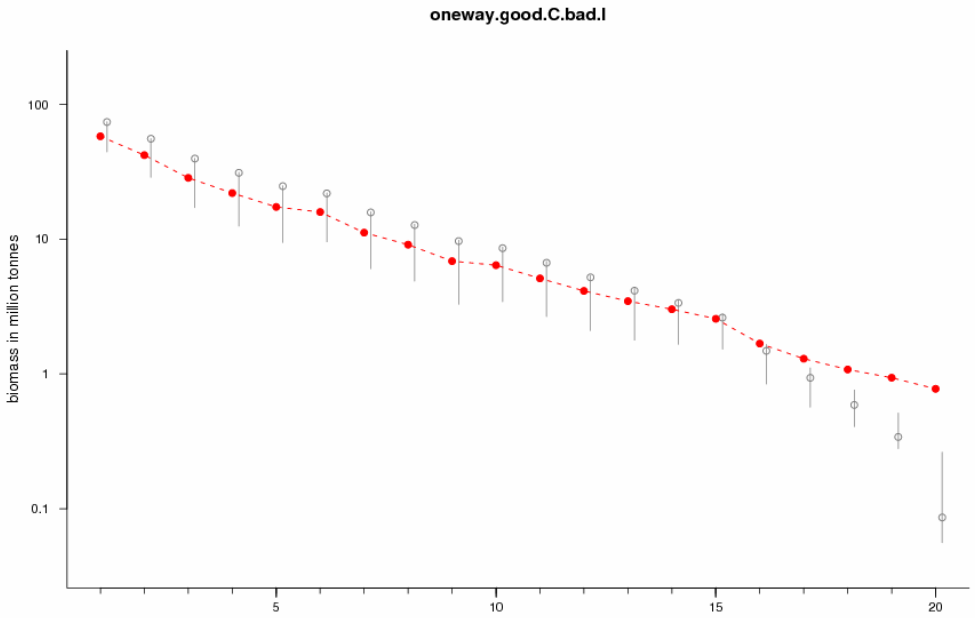

Figure 4.19. Bayesian 95% credible intervals (black), bootstrap 95% confidence intervals (grey) and true biomasses (red) for the one-way trip CgIb data scenario.

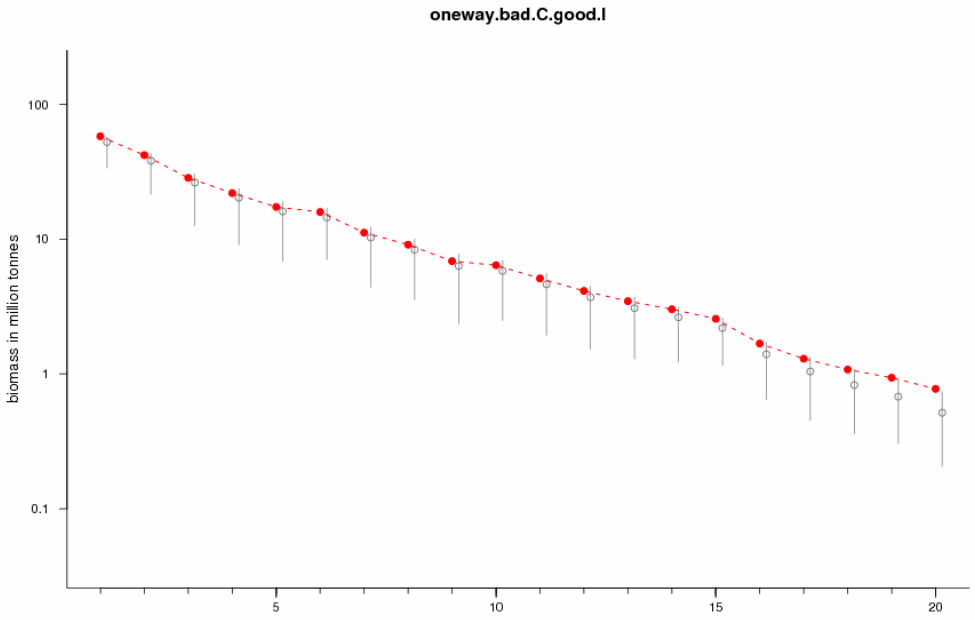

Figure 4.20. Bayesian 95% credible intervals (black), bootstrap 95% confidence intervals (grey) and true biomasses (red) for the one-way trip CbIg data scenario.

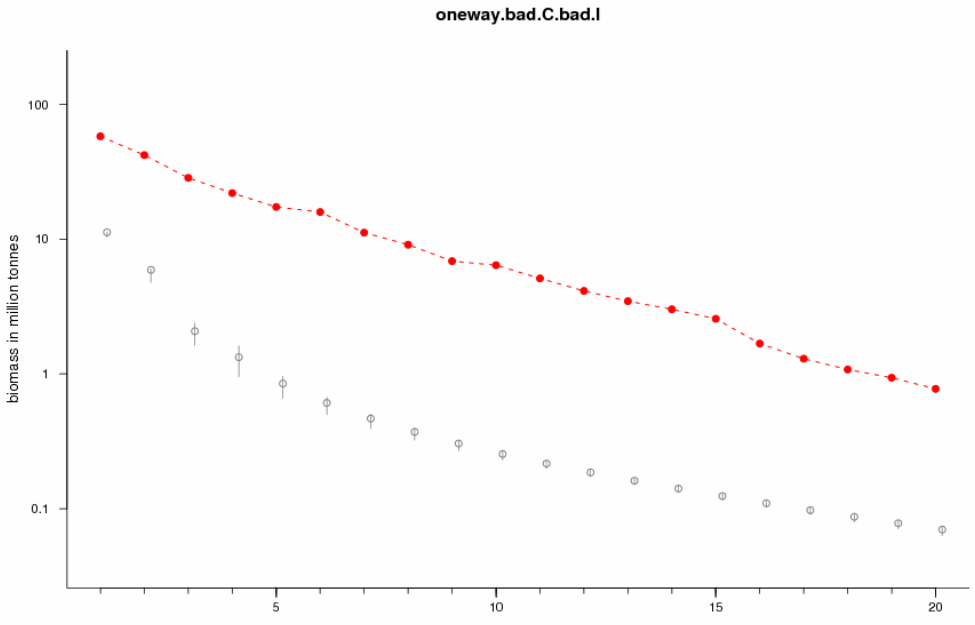

Figure 4.21. Bayesian 95% credible intervals (black), bootstrap 95% confidence intervals (grey) and true biomasses (red) for the one-way trip CbIb data scenario.

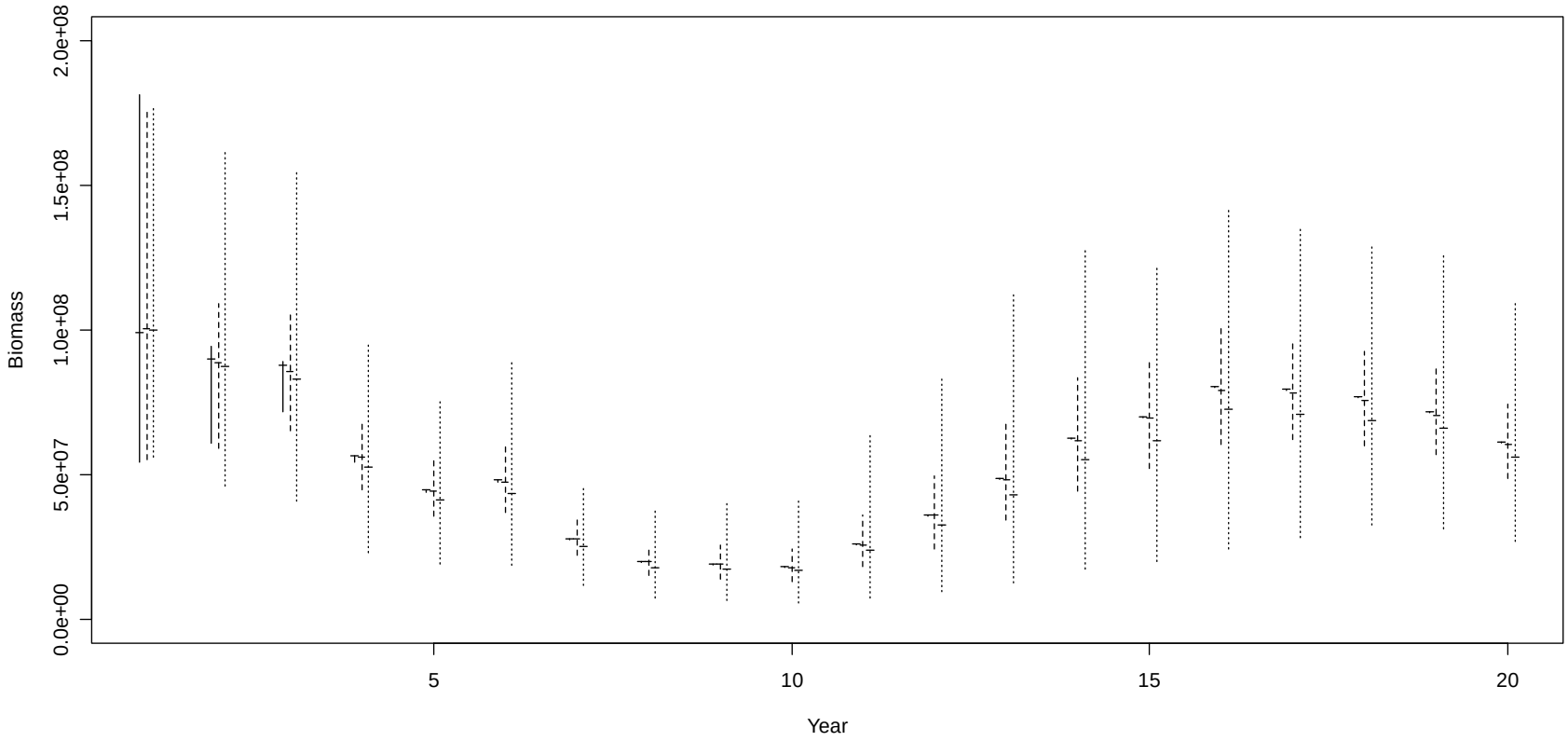

Figure 4.22. Variability in yearly biomass of a hypothetical population assumed with a Schaefer production model when the yearly process errors have lognormal distributions with CV=0 (solid line), CV = 0.1 (dashed line) and CV = 0.3 (dotted line) as reflected by 95% probability intervals for 1000 simulations of each scenario. For all scenarios the initial biomass (and carrying capacity) is a lognormal random variable with CV=0.3 and yearly F is the ratio of catch to biomass in Table 4.2.

4.4 Surplus production models with process error

The results presented so far correspond to a surplus production model without process error. In other words,

$$B_{t+1} = B_t + B_t r\left(1 - \frac{B_t}{K}\right) - C_t$$

is assumed to hold exactly, so that B_{t+1} is entirely determined by the values of B_t, C_t, K and r. This assumption does not seem realistic, given the simplicity of the equation. A way to relax the assumption is by including process error. For example, including log-Normal process error means that

$$B_{t+1} = \left[B_t + B_t r\left(1 - \frac{B_t}{K}\right) - C_t\right] e^{\varepsilon_t},$$

where $\varepsilon_t \sim N\left(0, \sigma^2\right)$. Now B_{t+1} is modelled as random, log-normally distributed, with median value given by deterministic surplus production equation and with some variance σ^2 in the log-scale (hence, the coefficient of variation of B_{t+1} is $\sqrt{e^{\sigma^2} - 1}$. This casts the model as what is now termed a state-space model, including both process and observation errors. For fixed values of *K*, *r* and an example series of catches (a selection only of the possible values is used here), Figure 4.22 gives some indication of the much wider range of biomass trajectories that may be obtained by allowing for process error.

At the meeting, we did not have any software available to compute bootstrap estimates for the surplus production normal with log-normal process error. However, the model can be fitted in the Bayesian framework, using WinBUGS code (Annex 4) which is a slight modification of the one provided by R. Millar (http://www.stat.auckland.ac.nz/~millar/Bayesian/BayesIndex.html). The prior distributions for *K*, *r*, *q* and the variance of the survey log-index, *v*, are the same as chosen in the no-process-error case. The prior distribution for process error was taken to be

$$1/\sigma^2 \sim Gamma\left(s = 1, r = 0.039\right)$$

leading to a prior distribution on the coefficient of variation of the process $\sqrt{e^{\sigma^2} - 1}$ with median value of 0.24 and (0.11,1.07) as 90% central credible interval. Hence, it is a fairly flexible prior distribution on the process error, although it does not accommodate low values of process error (such as values lower than 10% for the coefficient of variation). The reason for making this choice is also partly driven by the difficulties experienced with the MCMC computational algorithm implemented by WinBUGS when the process error is small. In the model with process error, yearly biomasses are no longer determined by *K*, *r* and the catches, and have to be treated as additional unknown variables. Hence, they need to be also sampled in the computational algorithm for the posterior distribution. This increases very substantially the dimension of the posterior distribution, which goes from 4 (corresponding to *K*, *r*, *q*, *v*) to $4 + n - 1$, where *n* is the number of years in the time series, since $B_2, \ldots, B_n$ need to be sampled as well (note that the model assumes $B_1 = K$). The MCMC algorithms used by WinBUGS work site by site: in other words, the variables (model parameters and yearly biomasses) are sampled from the full conditional posterior distribution one at a time. This means that each variable is sampled conditioned on the values of all other variables. When process error is very small, it means that each B_t is almost deterministically given by B_{t-1} and B_{t+1}, leading to very badly convergent results. Hence, in the WinBUGS program we could not allow for very small process error.

The same datasets (informative, one-way trip) with 4 different scenarios each (CgIg, CgIb, CbIg, CbIb) as for the no process error model were analysed, so that results were directly comparable.

Starting with the informative dataset, Figures 4.23 and 4.25–4.27 display the prior density (green solid line), posterior histogram and true value (red vertical line) of the model parameters for each of the 4 scenarios. For the CgIg case, Figure 4.24 gives the pairwise scatterplot of the posterior distribution. Again, the unsurprising strong negative correlation between K and each of r and q appears. There also seems to be some negative association between the index and process errors, as somewhat similar looking survey data series might have arisen from low process error and high index error or the other way around. It is generally acknowledged that separating (i.e. making inference on both) process and observation errors is a difficult problem and often requires making judicious choices with respect to their relative magnitudes. The marginal posterior distributions in Figure 4.23 (with process error) and those in Figure 4.9 (with no process error) are generally in agreement, although those under process error are wider (in other words, the reduction in uncertainty from prior to posterior is smaller in the model with more unknowns). Figure 4.28 displays the true log(biomass) series (in red) and the Bayesian posterior median and 95% posterior credible interval. The results are good, although the intervals are very wide, meaning that there is very large uncertainty *a posteriori*. The results also demonstrate that unless we are very certain that the surplus production model is correct for the population, which will usually not be the case, then Bayesian credible intervals and bootstrap confidence intervals based on assuming the model is exact with no process error may be much too narrow and also mis-leading in that the true population values, if process error exists, may lie outside these intervals with a much higher probability than the nominal values associated with the intervals.

In the CgIb (three-fold reduction in catchability of survey half way through the time series) case, the posterior distribution of the single catchability fitted (Figure 4.25) has shifted towards smaller values with respect to that in the CgIg case. This implies some shift of the posterior distribution of K towards larger values and an opposite shift in the posterior distribution of r. As a consequence, biomasses are overestimated at the beginning of the time series although this is corrected later in the series. Results for the CbIg only differ slightly from those in CgIg, and the same can be said about those for CbIb with respect to those for CgIb.

The 4 different scenarios were also examined for the one-way trip data. Prior densities (in green), posterior histograms and true values (red vertical lines) are displayed in Figures 32–35. For the CbIb case, Figure 4.36 presents a pairwise scatterplot, with the same correlations as found in the other cases. Figures 37–40 gives Bayesian 95% posterior credible intervals for yearly log(biomass) and the true values (in red). Results are surprisingly good and the break down encountered with bootstrap in the CbIb case (without process error) is not present here (with process error).

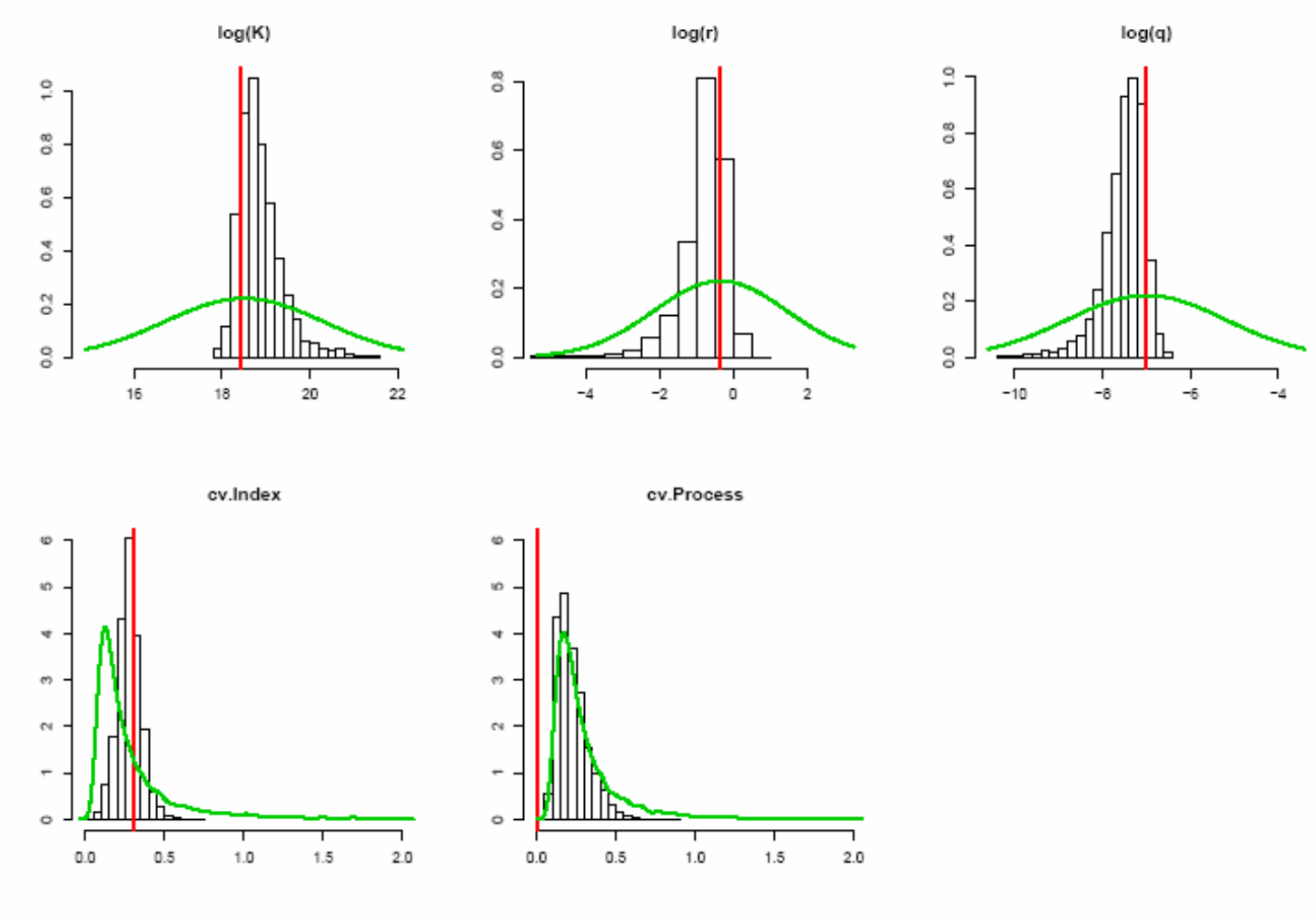

Figure 4.23. Prior (green) and posterior (histogram) distributions for parameters for the informative CgIg data scenario when process error is specified. The true parameter value assumed for data generation is denoted by the intersection of the red vertical line on the horizontal axis.

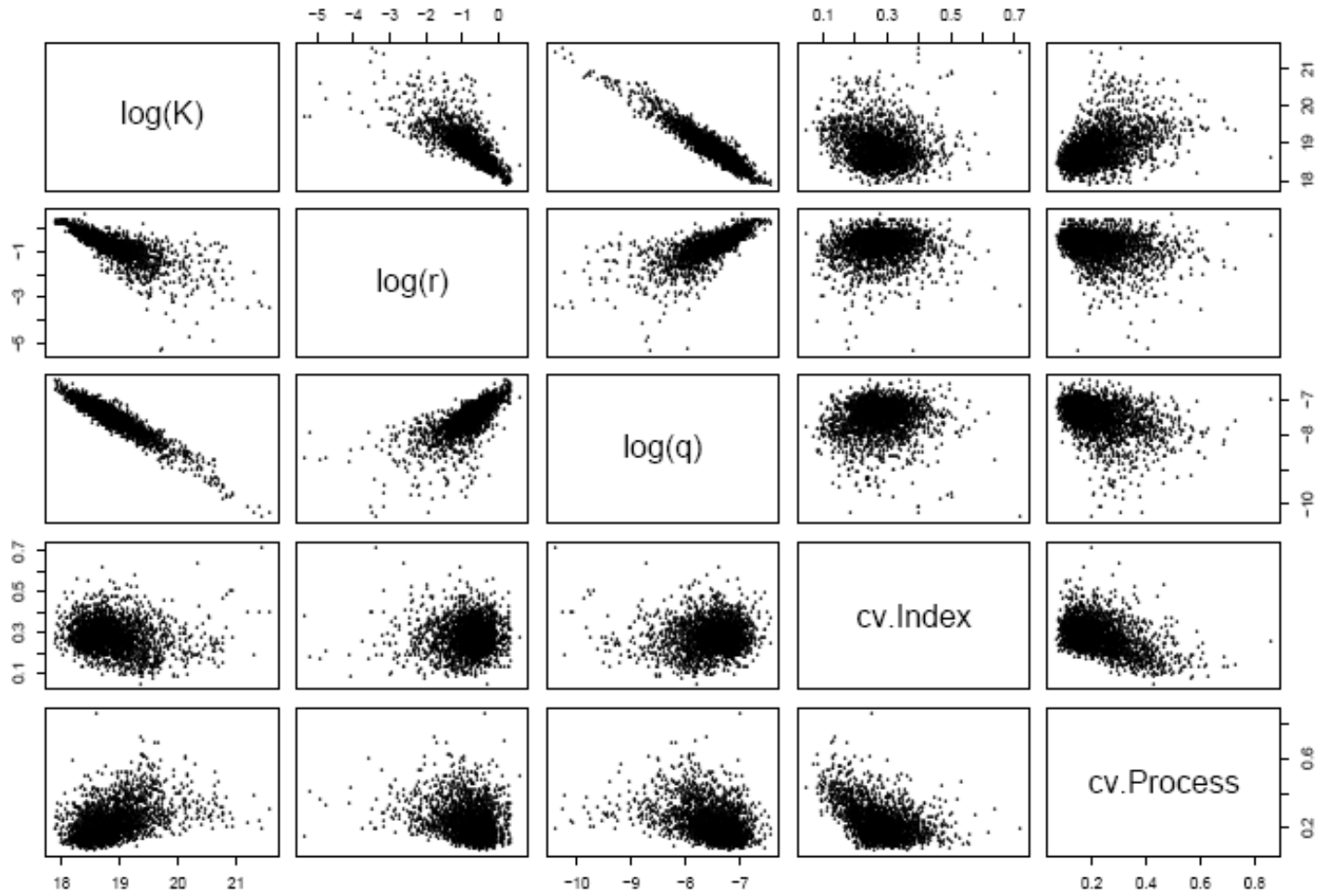

Figure 4.24. Pairwise joint posterior distributions for the informative CgIg data scenario when process error is specified.

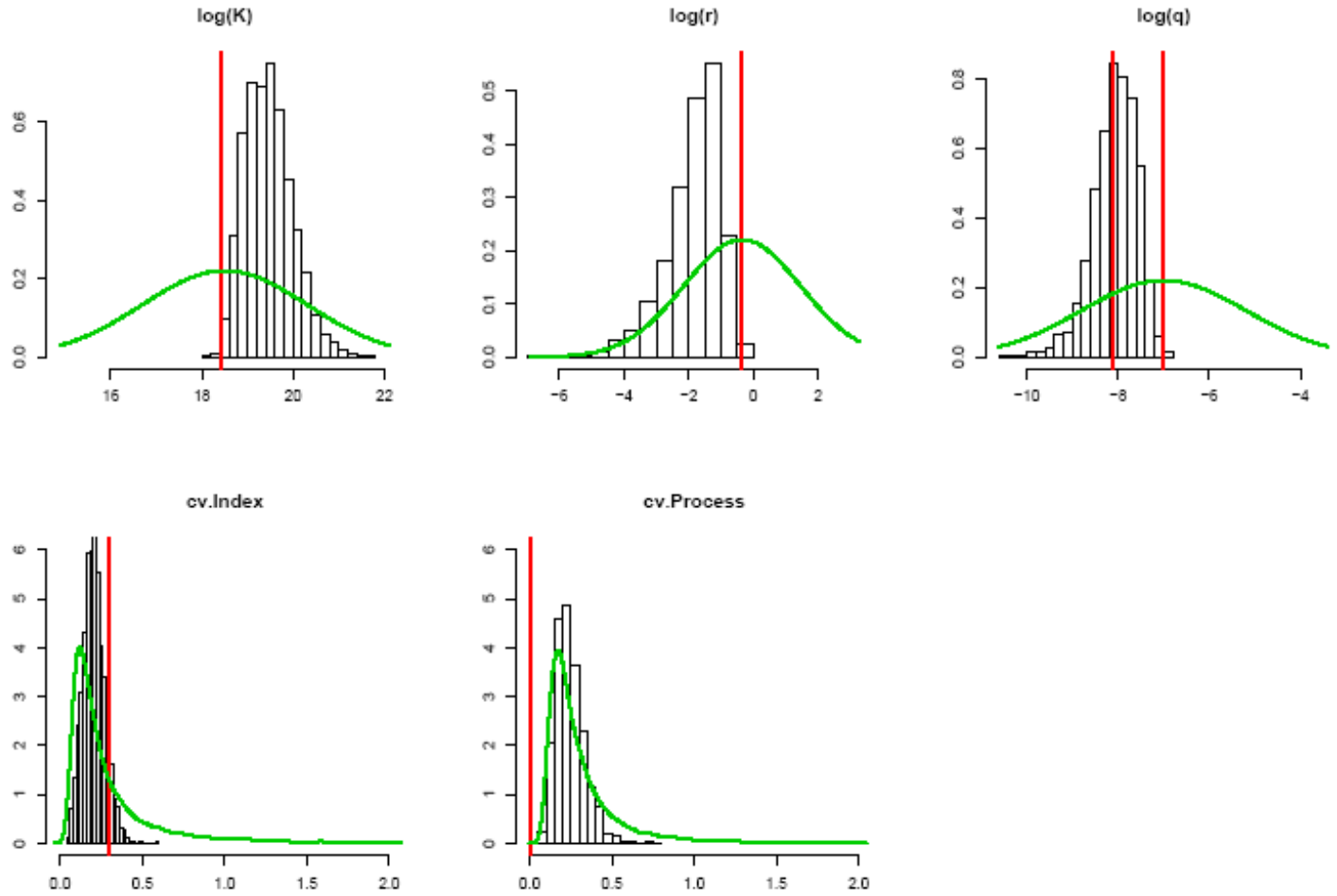

Figure 4.25. Prior (green) and posterior (histogram) distributions for parameters for the informative CgIb data scenario when process error is specified. The true parameter value assumed for data generation is denoted by the intersection of the red vertical line on the horizontal axis.

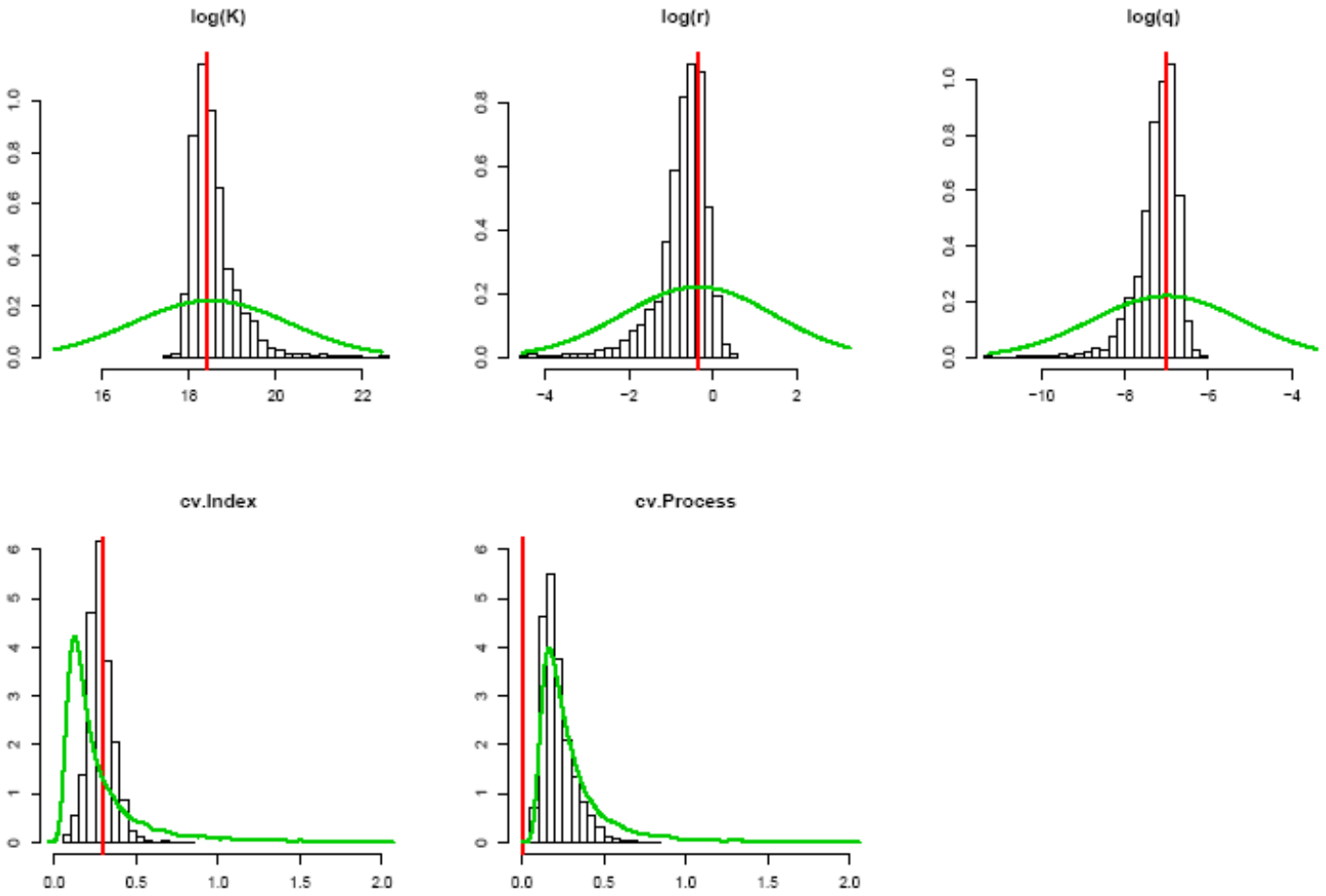

Figure 4.26. Prior (green) and posterior (histogram) distributions for parameters for the informative CbIg data scenario when process error is specified. The true parameter value assumed for data generation is denoted by the intersection of the red vertical line on the horizontal axis.

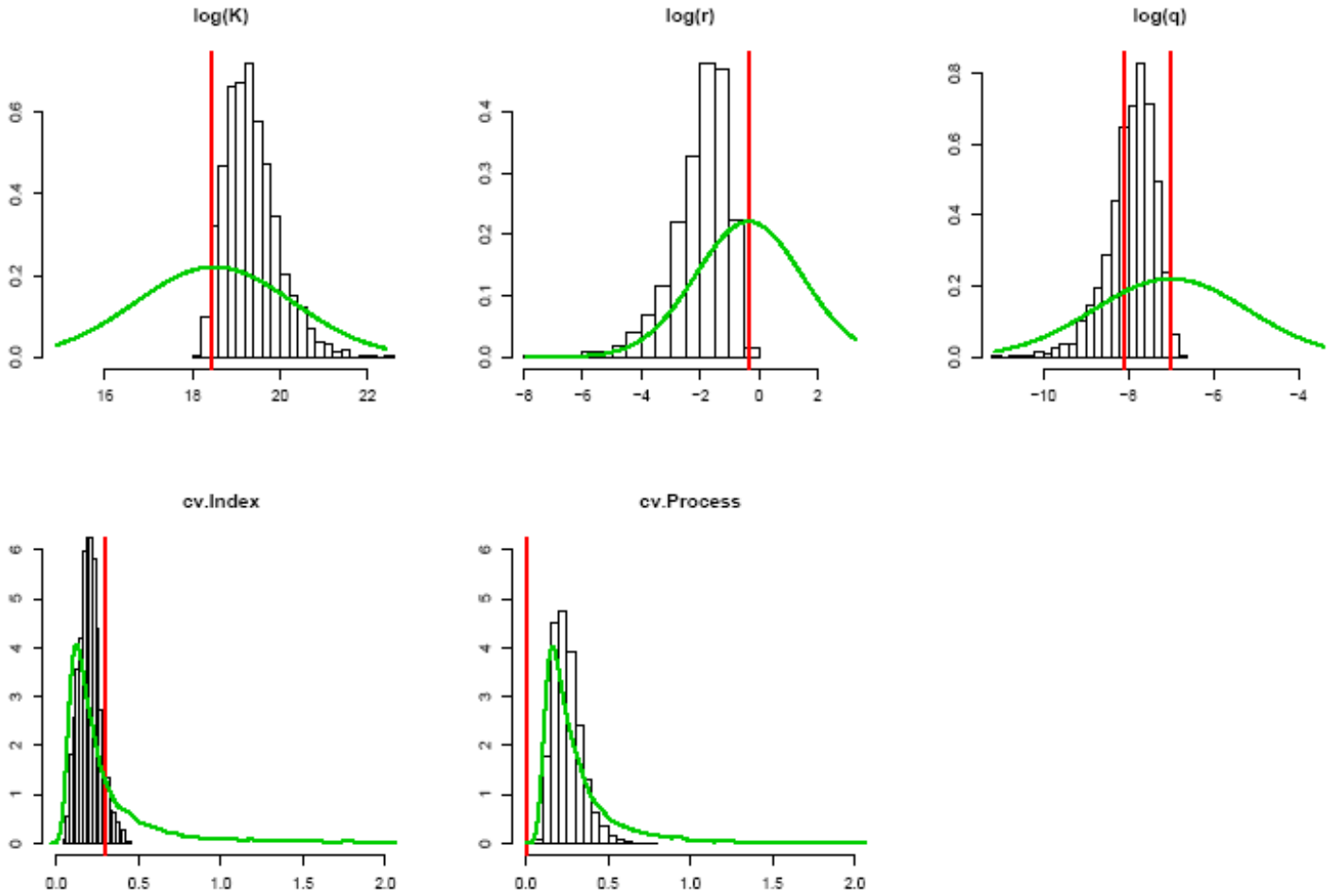

Figure 4.27. Prior (green) and posterior (histogram) distributions for parameters for the informative CbIb data scenario when process error is specified. The true parameter value assumed for data generation is denoted by the intersection of the red vertical line on the horizontal axis.

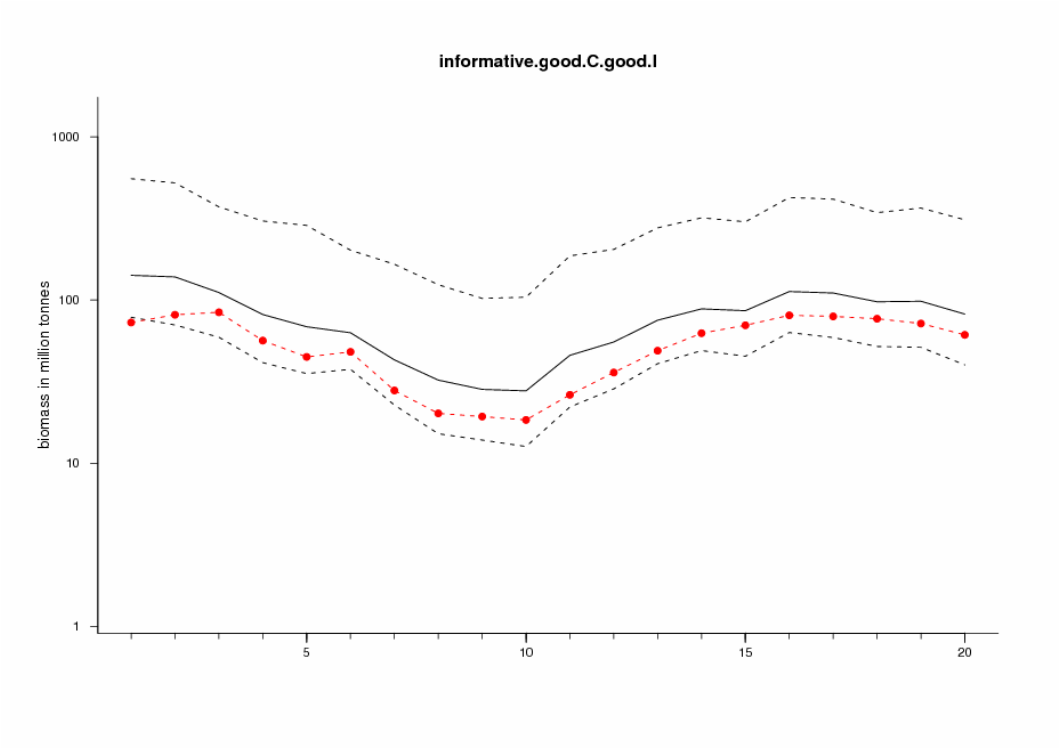

Figure 4.28. Bayesian posterior median (solid black), 95% credible intervals (dashed black) and true biomasses (red) for the informative CgIg data scenario when process error is specified.

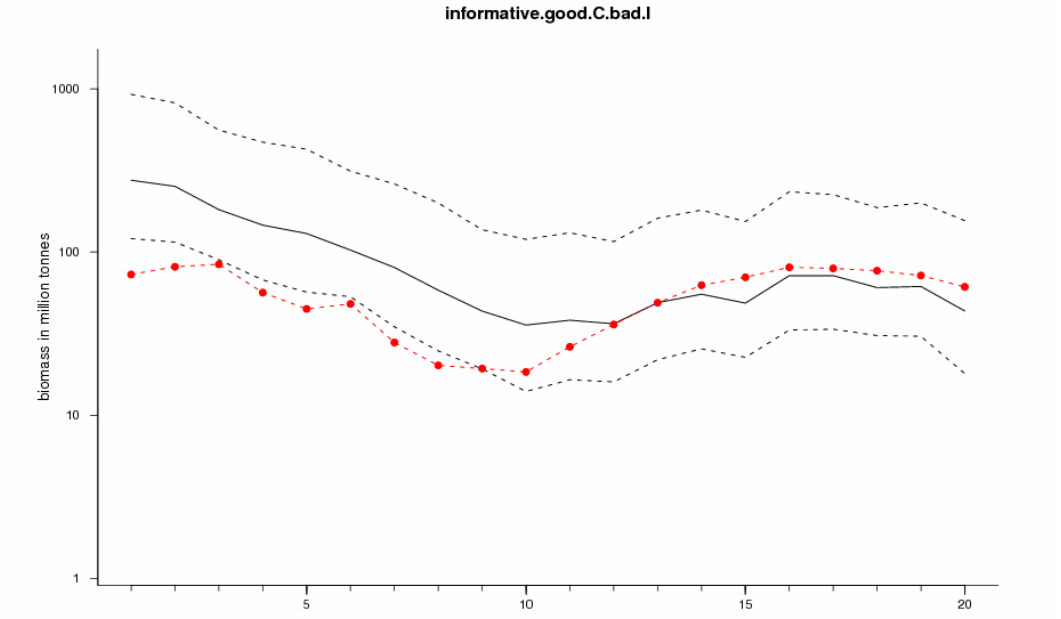

Figure 4.29. Bayesian posterior median (solid black), 95% credible intervals (dashed black) and true biomasses (red) for the informative CgIb data scenario when process error is specified.

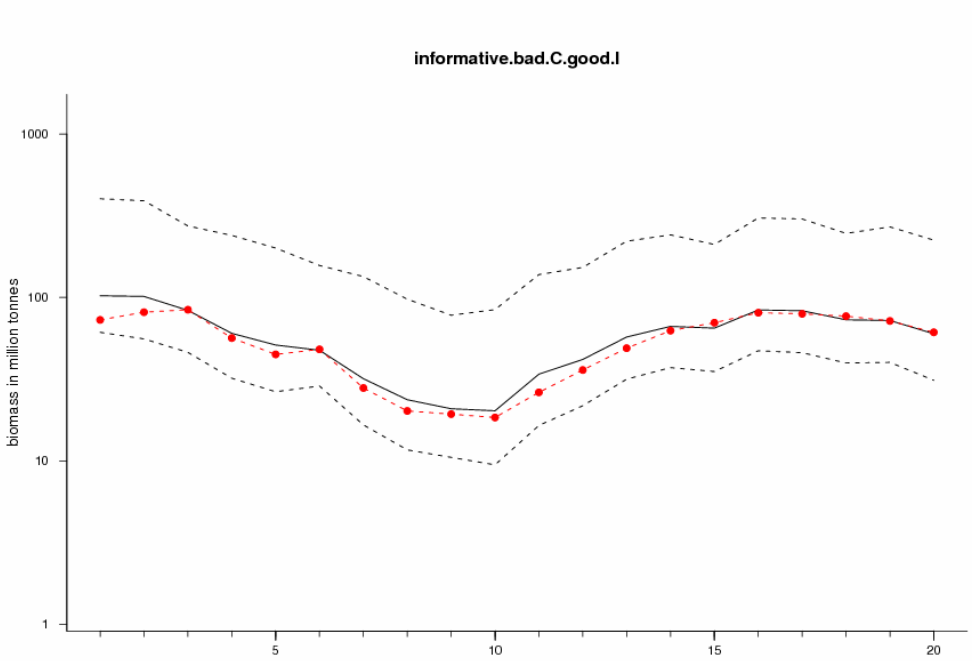

Figure 4.30. Bayesian posterior median (solid black), 95% credible intervals (dashed black) and true biomasses (red) for the informative CbIg data scenario when process error is specified.

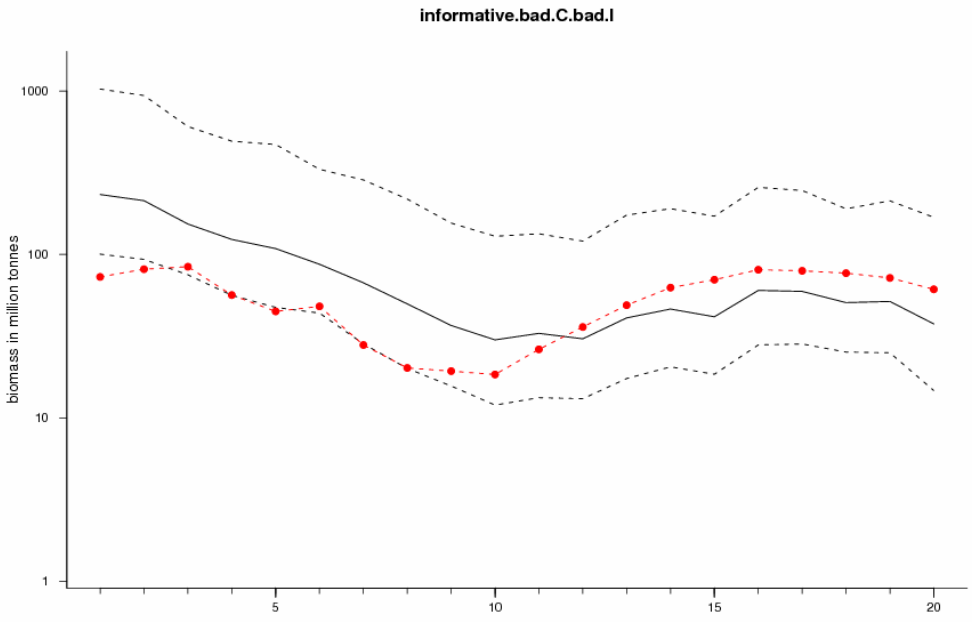

Figure 4.31. Bayesian posterior median (solid black), 95% credible intervals (dashed black) and true biomasses (red) for the informative CbIb data scenario when process error is specified.

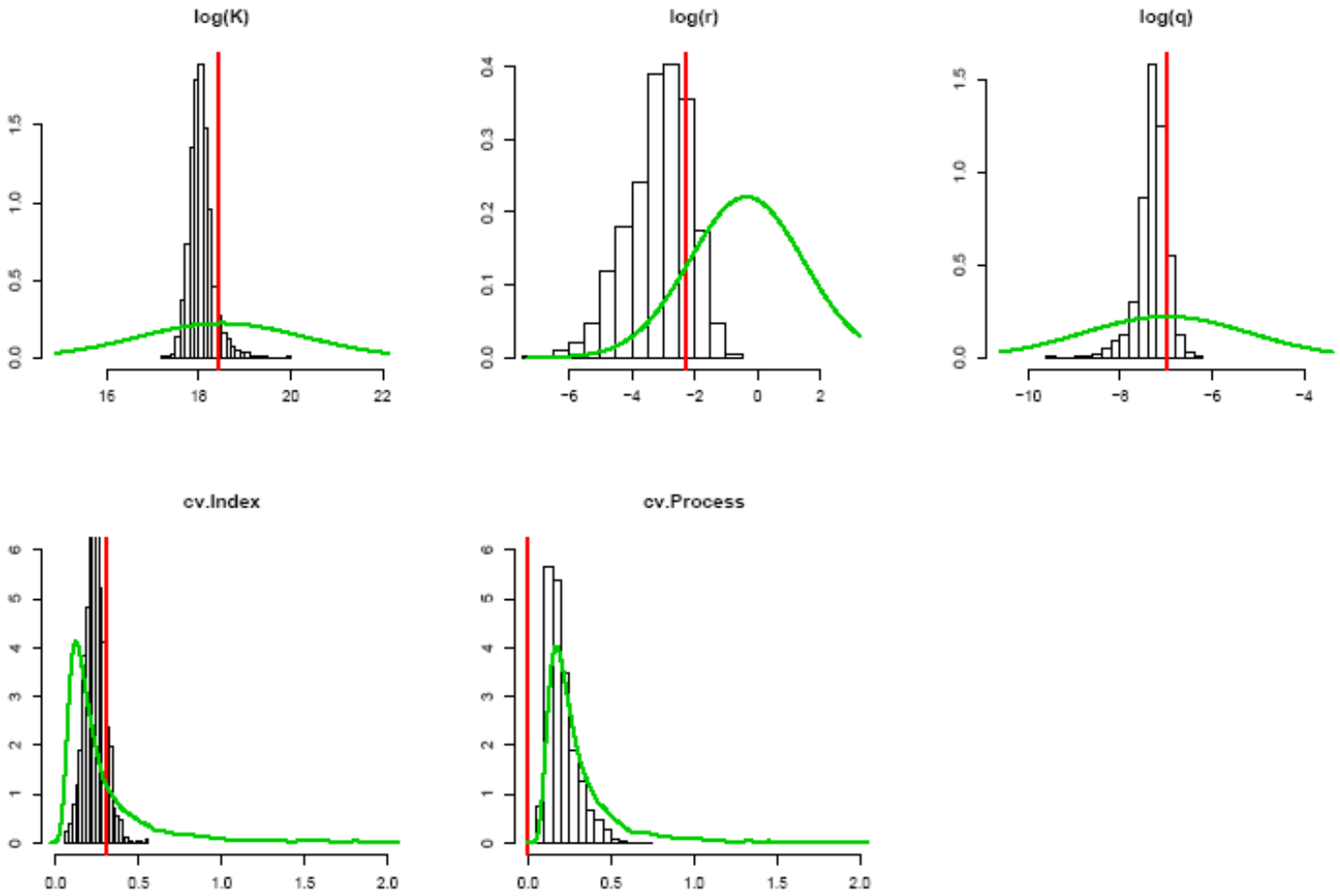

Figure 4.32. Prior (green) and posterior (histogram) distributions for parameters for the one-way trip CgIg data scenario when process error is specified. The true parameter value assumed for data generation is denoted by the intersection of the red vertical line on the horizontal axis.

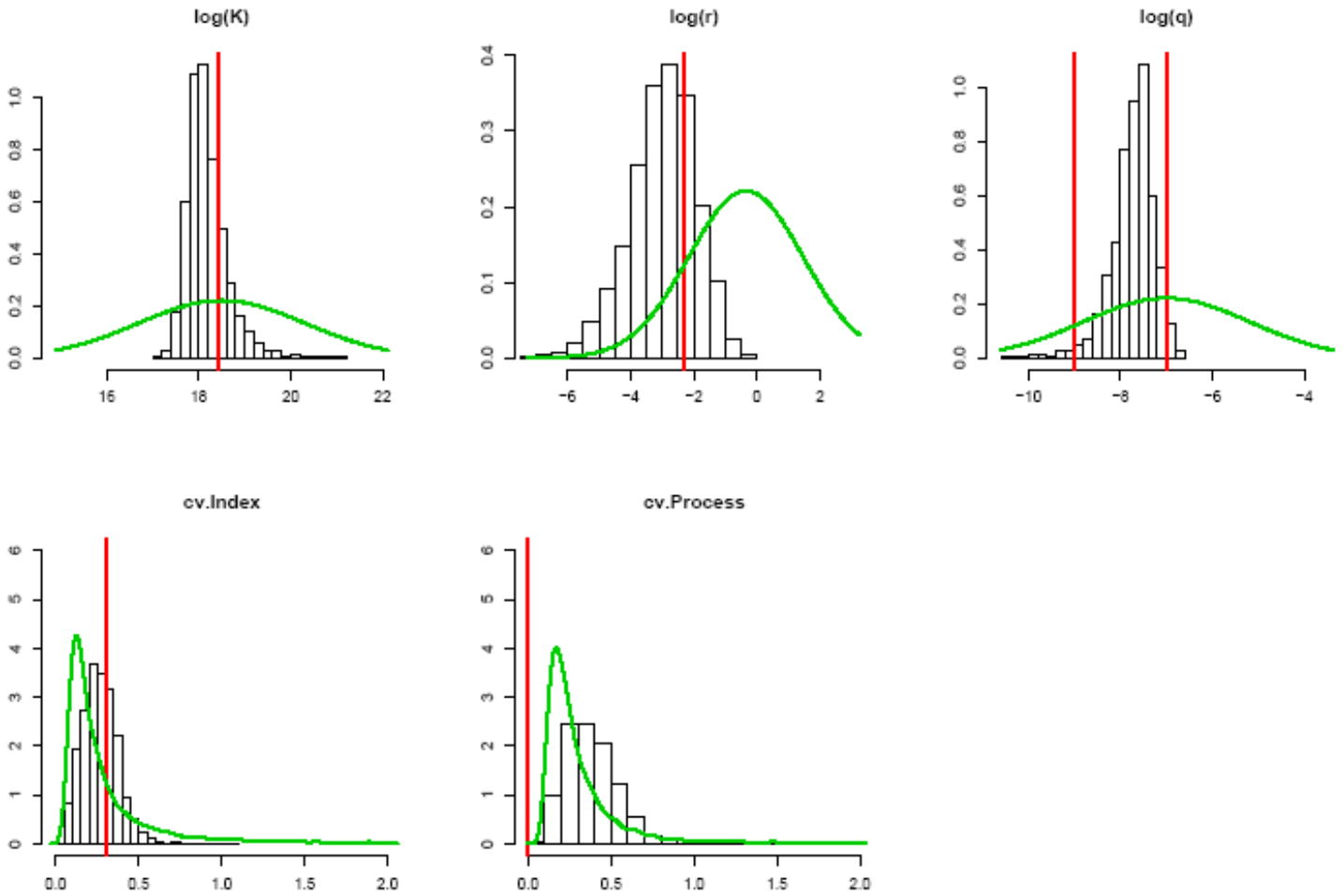

Figure 4.33. Prior (green) and posterior (histogram) distributions for parameters for the one-way trip CgIb data scenario when process error is specified. The true parameter value assumed for data generation is denoted by the intersection of the red vertical line on the horizontal axis.

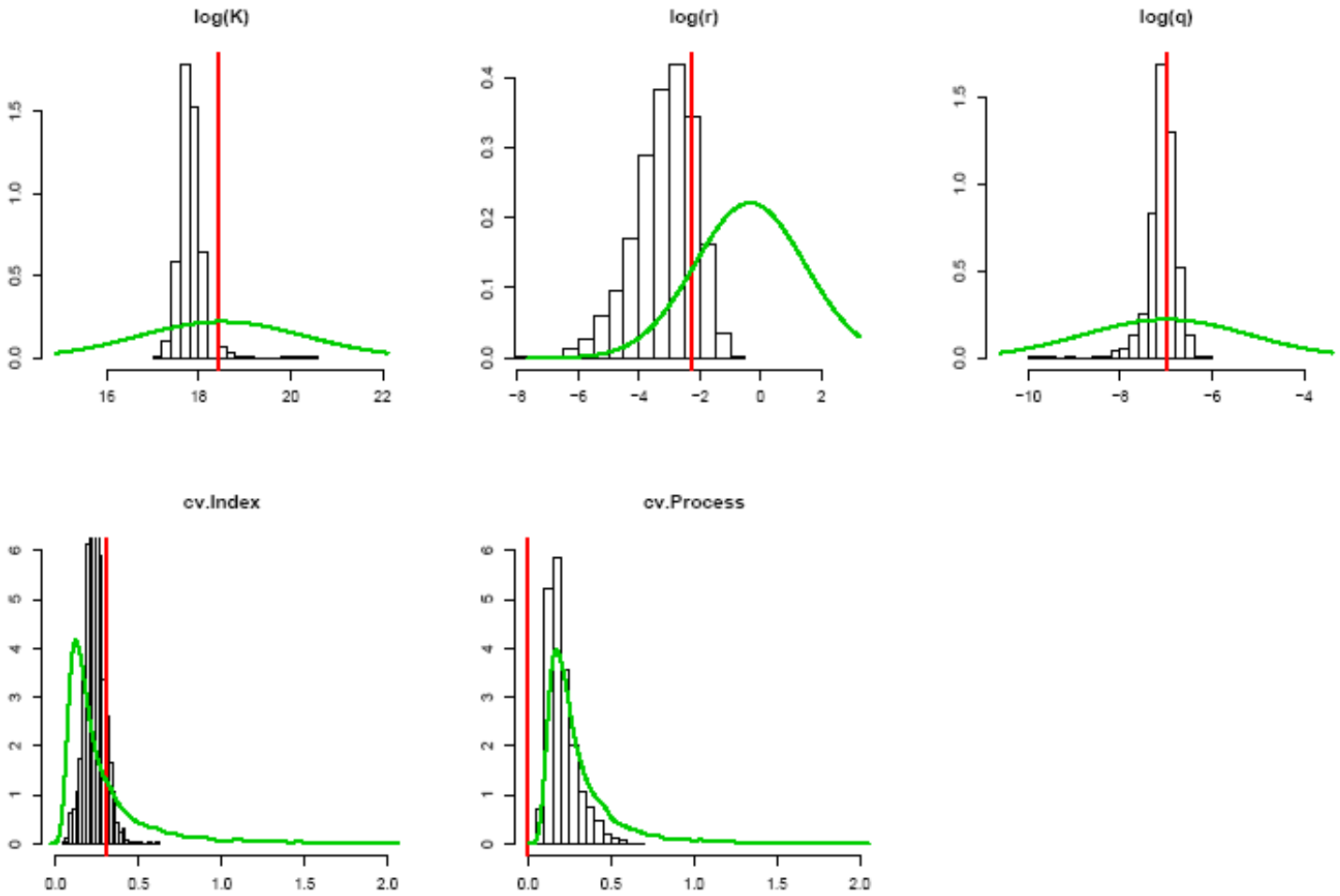

Figure 4.34. Prior (green) and posterior (histogram) distributions for parameters for the one-way trip CbIg data scenario when process error is specified. The true parameter value assumed for data generation is denoted by the intersection of the red vertical line on the horizontal axis.

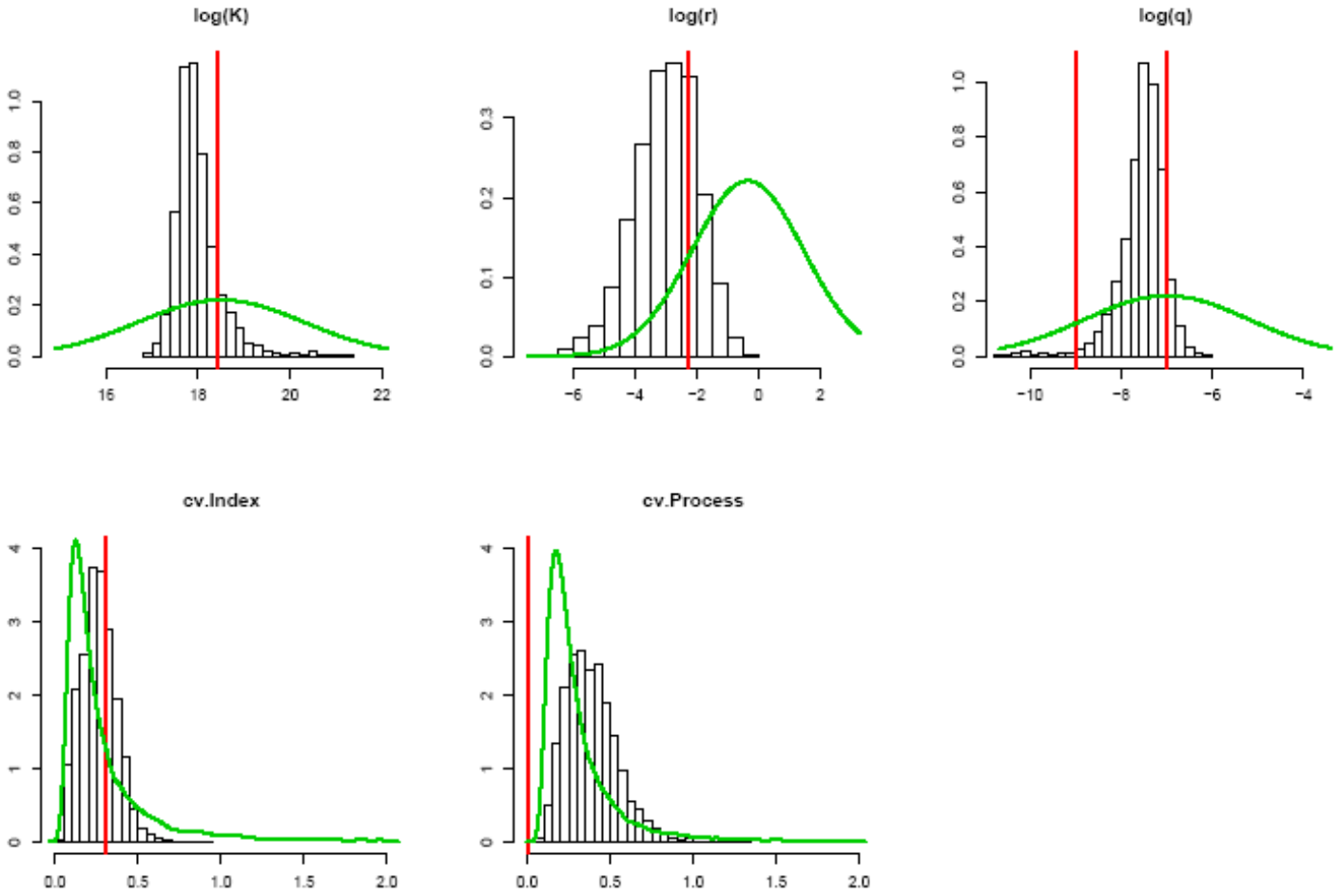

Figure 4.35. Prior (green) and posterior (histogram) distributions for parameters for the one-way trip CbIb data scenario when process error is specified. The true parameter value assumed for data generation is denoted by the intersection of the red vertical line on the horizontal axis.

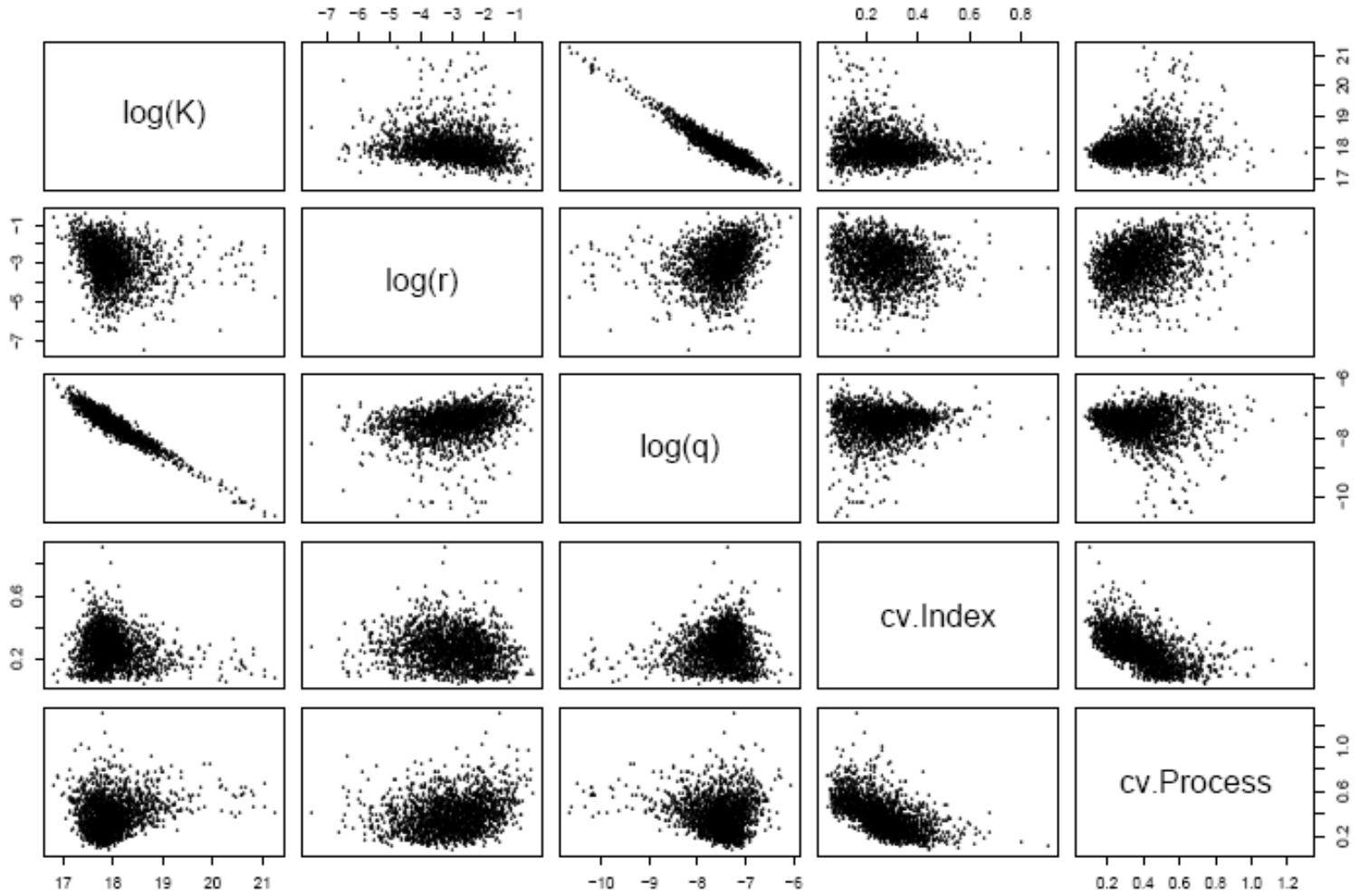

Figure 4.36. Pairwise joint posterior distributions for the one-way trip CbIb data scenario when process error is specified.

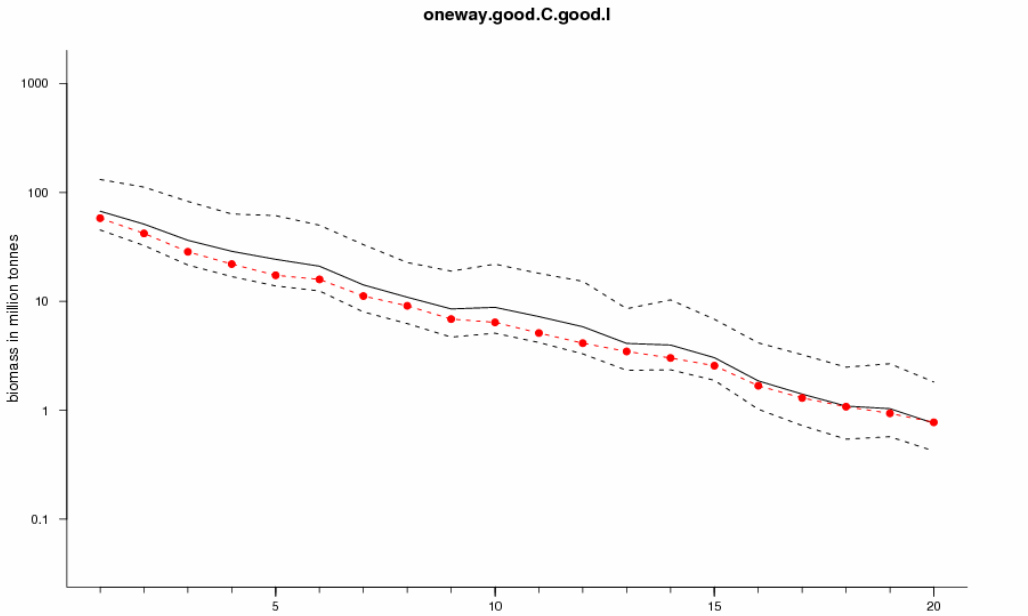

Figure 4.37. Bayesian posterior median (solid black), 95% credible intervals (dashed black) and true biomasses (red) for the one-way trip CgIg data scenario when process error is specified.

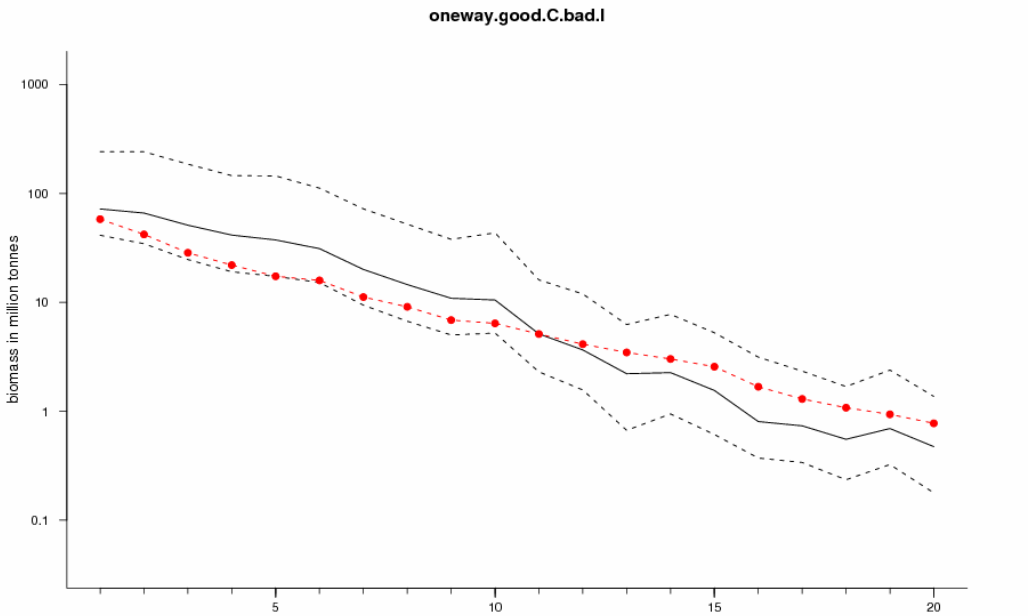

Figure 4.38. Bayesian posterior median (solid black), 95% credible intervals (dashed black) and true biomasses (red) for the one-way trip CgIb data scenario when process error is specified.

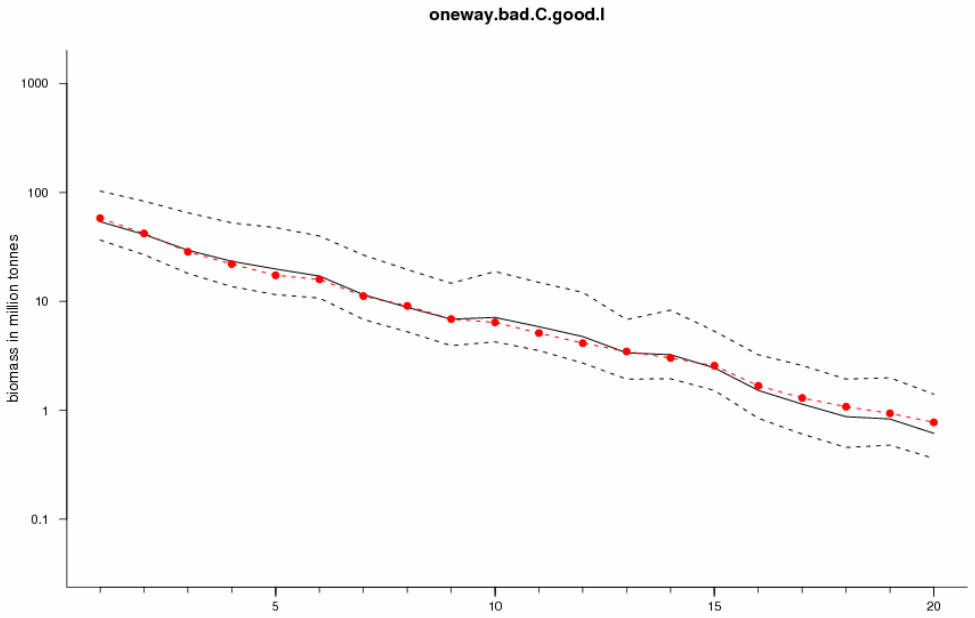

Figure 4.39. Bayesian posterior median (solid black), 95% credible intervals (dashed black) and true biomasses (red) for the one-way trip CbIg data scenario when process error is specified.

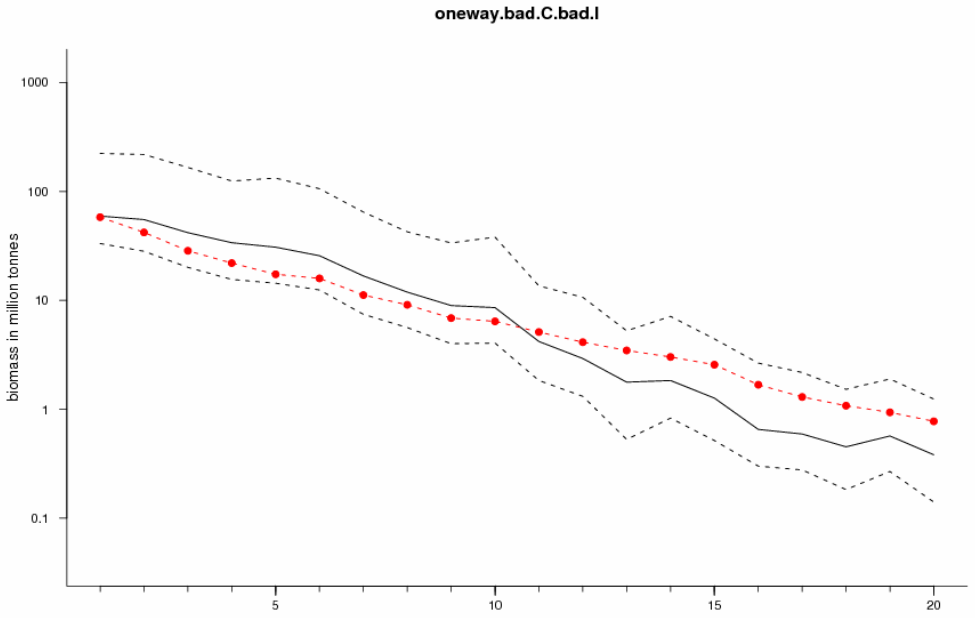

Figure 4.40. Bayesian posterior median (solid black), 95% credible intervals (dashed black) and true biomasses (red) for the one-way trip CbIb data scenario when process error is specified.

4.5 Conclusions

The difficult shape of the likelihood function, illustrated earlier in this chapter, caused severe difficulties for the computational procedures. In the Bayesian setting without process error, this meant that only the Informative dataset could be fitted. For the bootstrap setting with

ASPIC, the One Way Trip dataset under the CbIb (Catch bad Index bad) scenario produced entirely wrong answers, which is suspected to be due to a failure in finding the maximum of the likelihood function. These difficulties with the computational procedures made the comparisons between Bayesian and bootstrap results less direct. On the whole, we can say that some differences were found, but it is difficult at this stage to assess whether they are really genuine or caused by the computational procedures not working as well as would be desirable. As already mentioned, it is expected that Bayesian and bootstrap estimates will be more likely to differ in data poor situations or when the likelihood has a difficult shape (as is the case here).

An interesting aspect of the ASPIC bootstrapping procedure is that the initial guesses and search bounds for the various parameters for the fitting of the bootstrap data sets are determined in an automated but unknown fashion. It would be beneficial to know the method of automation and a useful feature in a future version would be to apply this automation to the true data set. However, reporting the initial guesses that are automatically determined would be important. It would also be desirable to have a better developed MCMC algorithm to handle the surplus production model, with better convergence properties than the one considered here. This would facilitate comparison of Bayesian and bootstrap results.

It would be of interest to compare Bayesian and likelihood- or bootstrap-based inference methods when process error exists. Substantial variation in yearly biomass can be induced with increased levels of process error (Figure 4.22). de Valpine (2002) notes that models that account for the temporally stochastic nature of populations hold some promise, but they will provide results that are more uncertain than models that do not account for temporal stochasticity. However, this is an attribute not a detriment because stochasticity is a realistic attribute of populations and not accounting for it will give us a biased view of the true uncertainty in our estimates. Results of much recent work in this field have shown that models that do not properly account for process error (i.e., temporal stochasticity in biomass) when it exists will provide biased estimation of various model parameters (Punt, 2003; de Valpine and Hilborn, 2005). Wang (2007) has also explored the behaviours of estimators using different state-space approaches for a generalized logistic population growth model and found the unscented Kalman filter to provide generally lower root-mean squared errors than the extended Kalman filter. However, reliable methods that account for process and measurement errors simultaneously in stock assessment models are not yet available, and this is an important area for future research.

5 Model mis-specification retrospective bias and noise

5.1 Introduction

Model misspecification refers to any assumption made in a model that is incorrect: one example might be the assumption that natural mortality is time-invariant, which it seldom is in reality. One possible consequence of such misspecification is retrospective bias and noise, especially if there is a trend in the misspecification over time. The retrospective problem involves systematic differences in assessment model estimates of stock size or some other quantity in a reference year. The differences occur as successively more data are incorporated in the procedure for estimation. The differences appear to be structural biases that result from a misspecification of the assessment model (i.e. discrepancies between the assumptions built into the model, and the properties of the data). In some cases the retrospective problem can be so severe that the assessment model is considered to be too unreliable to serve as a basis for advice.

This retrospective problem has been recognised as widespread and serious, and has been a topic of concern in the ICES community for many years, as reflected in several ICES Methods Working Group (WGMG) reports (e.g. ICES, 2002, 2003, 2004b, 2006b). The source of the problem is difficult to identify and not always understood. Possible sources of the problem include trends or shifts in natural mortality, discards and misreporting, and misspecification of selection- and catchability-at-age (or any combination of these).

In this Section we explore the problem of model misspecification using four different diagnostic tools. One of these (local influence diagnostics) is applied directly to the problem of retrospective bias: the other three look at misspecification in more general terms, but could also be applied to the retrospective problem if required.

5.1.1 Previous work

The WGMG meeting in 2001 (ICES, 2002) considered three topic areas, each of which was thought to result in retrospective patterns in the output from sequential stock assessments and forecasts: data quality, model misspecification, and short- and medium-term prognoses. WGMG explored each of these areas in turn.

Although several data-related problems were thought to cause retrospective patterns, WGMG highlighted the failure of tuning data to comply with constant catchability assumptions, and the incomplete accounting of catch data as important causes of retrospective discrepancies related to data quality. It was suggested that assessment Working Groups should favour fewer data of good quality (as evaluated independently of the assessment model) instead of large quantities of data of unknown properties. Although retrospective patterns indicate problematic assessments, they should not be taken as the only diagnostic for detecting problems, because a lack of retrospective bias does not imply the assessment model is unbiased or well-specified. Therefore, consideration should be given to the full suite of assessment diagnostics.

When tackling the retrospective problem from the methodological point of view, WGMG recognised the need to:

- understand the mechanisms which create inconsistencies in the perception of the development of the stock;
- investigate the sensitivity of different model formulations to various causes of the retrospective bias, and the extent to which such causes can be accounted for in model structures; and
- develop diagnostics (e.g. retrospective plots of log catchability residuals, and local influence diagnostics), and understand the extent to which problems can be revealed by the diagnostic tools.

In order to achieve these aims, extensive studies were suggested on simulated data sets with known properties where both the true state of the stock and the flaws in the data were fully known. WGMG recognised at the time that, in general, the retrospective problem was caused by assessment models failing to interpret assessment data in the appropriate way, because the data represent something that is different from the *a priori* assumptions in the assessment models.

The concern with regard to short- and medium-term prognoses was that any bias in the results of the stock assessment model transfers into short- and medium-term forecasts. WGMG therefore proposed that any bias correction that is thought to be necessary to provide unbiased population estimates (e.g. starting numbers-at-age) for the forecasts should be undertaken within the assessment process itself rather than the forecast process, as retrospective bias is an assessment problem and is best dealt with in that context.

The WGMG meeting in 2003 (ICES, 2003) investigated further the utility of local influence diagnostics (LIDs) to help increase understanding of the possible causes of retrospective bias, so as to help correct the problem. In particular, the LID method could be used to find small perturbations to a variety of assessment inputs and assumptions that remove or reduce the retrospective problem. By exploring the plausibility of these perturbations, it was thought the method might be useful for identifying a smaller subset of the inputs that are more likely causes of the retrospective problem.

At the 2003 meeting, WGMG experienced some difficulty in generating simulated data sets with significant retrospective patterns. Therefore, a specific case study with a severe retrospective pattern, namely the Sequential Population Analysis (SPA) assessment of Eastern Scotian Shelf cod, was used in order to demonstrate the use of the LID method to diagnose the cause of the retrospective pattern. The analyses suggested that reasonable changes in assumptions about model errors were not a likely source of the retrospective pattern. However, additional information seemed necessary to discriminate between the possibilities that the source of the retrospective pattern was catches, natural mortality, or assumptions about survey catchability.

The WGMG meeting in 2004 did not focus on the retrospective problem *per se*, but considered it in the wider context of diagnosing model misspecification for some of the assessment models investigated using simulated data (ICES, 2004b), while the meeting in 2006 considered the problem explicitly in one of its ToRs: "investigate and test the sensitivities of catch-at-age stock assessment methods to known data problems with particular reference to the retrospective problem" (ICES, 2006b). The latter meeting provided a summary of previous work on the retrospective problem, but also considered methods to deal with catch data that do not reflect real levels of catches, a potential source of retrospective bias for assessments that rely on unbiased catch data.

5.1.2 Work plan arising from presentations

The approach adopted for the current meeting was to develop strategies to help diagnose and correct model misspecification such as could lead to retrospective bias. Four strategies are considered as follows:

1) Pre-screening of data inputs to assessment models
2) Local influence diagnostics
3) The ADAPT approach with year effects in q (SPA)
4) The ADAPT approach with year effects in a catch multiplier (B-ADAPT)

Each of these strategies was tested on six simulated data sets. The simulated data sets were generated using the population simulator POPSIM (version 3.5.2), which forms part of the NOAA Fisheries Toolbox (see WP6 for a description). Table 5.1 describes these data sets.

Note that these simulated datasets are *not* the same as those used in management strategy evaluations in Section 3, or uncertainty estimation in Section 4. Each of the strategies is described and results discussed in subsequent sections.

Table 5.1. Simulation case study description.

DATA SET	PROBLEM DESCRIPTION
Data 1	Survey catchability (Q) tripled since 1995
Data 2	Natural mortality (M) tripled since 1995
Data 3	One-third catch reported since 1995
Data 4	Q tripled in 2000 and 2002
Data 5	50% increase in M since 1998
Data 6	Q tripled in 2000 and 2002 and a 50% increase in M since 1998

5.2 Pre-screening of data inputs to assessment models

Basic pre-screening of the survey and catch data using simple models can be used to identify outliers and shifts in the data time series. Example methods were applied to the test data sets simulated to induce retrospective patterns.

The methods were able to identify a step change in the survey catch per unit effort (CPUE) induced in data set 1 (Figure 5.2.1). The increase in survey catchability was apparent in the shift at 1995 in the survey log survey CPUE and the log CPUE cohort plots, but a concomitant change is not seen in the time series of catch data (figure not shown), and Z has not declined. Therefore the most likely cause is a change in survey catchability.

The methods also highlighted the survey CPUE increases in 2000 and 2002 as outliers within all ages of the time series of CPUE data set 4 (Figures 5.2.2 and 5.2.3). In Figure 5.2.2 the log survey CPUE mean standardised plots by year and the log survey CPUE cohort plots indicate strong increases in 2000 and 2002 that are not apparent in other years. Within survey comparisons (Figure 5.2.3) indicate that the 2000 and 2002 data at each age lie outside the robust regression prediction confidence intervals (except the 1997 cohort at age 3). This identification of the year effects would suggest their removal from the time series or fitting a model with year effects in catchability for those years. The same analysis technique identified the outlier year effects within data set 6 (results not shown).

The pre-screening methods were not able to distinguish changes to mortality rates, such as the changes in natural mortality and under reporting of catch (data sets 2, 3 and 5), which affect catch and survey data series with equal impact. They were also unable to identify the cause of the problem, only the years in which the deviations occur. Obviously the power of the methods is conditional on the observation noise occurring in the time series.

Figure 5.2.1. Data set 1 (survey catchability increased 3 fold after 1995):
a) Log survey CPUE, mean standardised, plotted by year
b) Log survey CPUE, mean standardised, plotted by cohort
c) Log survey CPUE cohort curves
d) The slope of the decline in log numbers between ages 4 and 5 (the reciprocal of an estimate of total mortality).

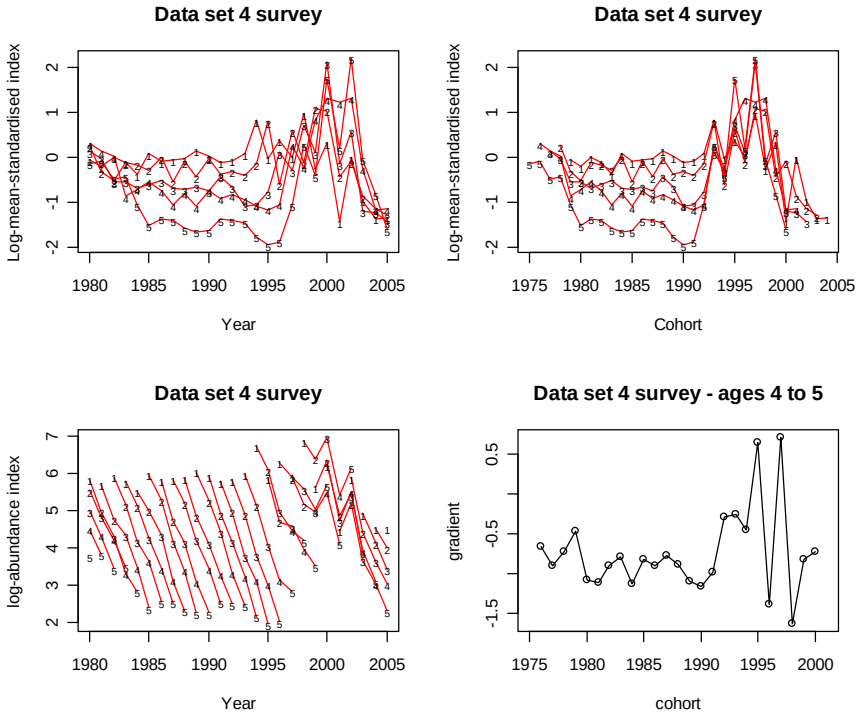

Figure 5.2.2. Data set 4 (catchability increased 3 fold in 2000 and 2002). Figure description is given in Figure 5.2.1.

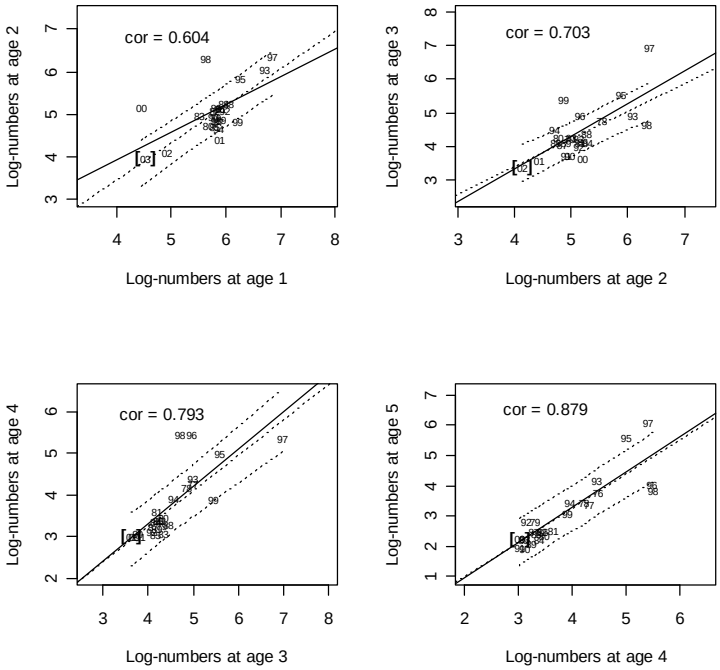

Figure 5.2.3. Data set 4 (survey catchability increased 3 fold in 2000 and 2002) Log CPUE at age plotted against the log of the CPUE at the previous age in the previous year, dates indicate cohort. Square brackets indicate the most recent year.

5.3 Local influence diagnostics

Local influence diagnostics (LIDs) are metrics that describe the effect of small perturbations of model components on important model results. These perturbations can be used to generate an influence surface for the model result. LIDs are based on the geometry of this surface, using the slope and possibly the curvature of the surface near the origin. They are a computationally feasible way to examine high-dimensional perturbations. An important feature of the local influence surface is the direction of maximum slope, d_{max}. LIDs can be used to find changes in model inputs that have large effects on outputs, especially those changes in inputs indicated by d_{max}. More specifically, LIDs can be used to find changes in VPA inputs or assumptions that remove retrospective patterns (Cadigan and Farrell, 2004). It has been suggested that the model input component that requires the least change to remove the retrospective pattern could be considered to be the most likely source of the problem. However, as pointed out in Cadigan and Farrell (2004), this requires the assumption that the perturbations to the various VPA inputs should be comparable; for example, a 20% change in survey catchability is the same size "error" as a 20% change in M.

Recent applications of the method to simulated data revealed a problem (e.g. WP 7). The LIDs were based on a metric of the size of the retrospective pattern (Mohn's ρ; see Cadigan and Farrell, 2004); however, the LIDs were sensitive to the number of retrospective years used to compute the metric. The LIDs often suggested changes to VPA inputs that removed retrospective patterns, but these changes were substantially different to the ones used to generate the patterns. That is, retrospective patterns could be removed for the wrong reasons. This could lead to biased assessment models that were apparently right because they did not have retrospective patterns. Also, reducing the metric to zero did not always result in VPAs without retrospective patterns. It was possible to have $\rho=0$ and still have pronounced retrospective patterns. A better metric is required for the latter problem.

We postulate that retrospective patterns are another manifestation of residual patterns, and we propose that a good statistic to examine in diagnosing the source of retrospective patterns is a measure of the size of the residual problem. Retrospective patterns are almost always associated with time trends in residuals, so we examine the mean square average annual residual (MSAAE),

$$MSAAE = \sum_y \bar{e}_y^2$$

as a measure of the size of the residual problem. In a well specified model the residuals should be centered about zero each year, so that $\bar{e}_y$ should be close to zero, and the sum of squares of these terms should also be close to zero. However, in a mis-specified model with time trends in residuals, $\bar{e}_y$ will be different from zero for many years, so that the sum of squares will be substantially greater than zero.

We can easily apply the local influence methods presented in Cadigan and Farrell (2002) using MSAAE to find perturbations to VPA inputs to greatly reduce MSAAE and thereby also reduce or remove retrospective patterns. This approach is much easier than the one described in Cadigan and Farrell (2004) because it does not require retrospective evaluations of the VPA in the perturbation analysis. However, a drawback of the approach is nonlinearity in the perturbation surface. The first-order local influence approach of Cadigan and Farrell (2002), which is based only on local slopes, works best when the influence surface is linear; however, MSAAE is non-negative and clearly its influence surface cannot be linear for all perturbations. Nonetheless, the approach can still be applied to find perturbations to reduce MSAAE, but we may not be able to make MSAAE=0 and there may be smaller perturbations to reduce MSAAE to a given level than the one's found using the local influence approach.

We examined perturbations to VPA catches, natural mortalities (M), survey catchabilities (Q), and estimation weights to find changes in these inputs that removed residual patterns and also retrospective patterns. We applied the approach to the six simulated case studies (Table 5.1) with misspecifications that resulted in retrospective patterns. We restricted our presentation to the Q and M perturbation analyses, although the M results were broadly similar to the catch results. The perturbations to M were of the form

$$m_{w,a,y} = (1+w_{a,y})\times m,$$

and the perturbations to Q were of the form

$$q_{w,a,y} = (1+w_{a,y})\times q_a$$

The Q perturbations can be interpreted as year-effects in survey catchability.

The results are shown in the left-hand panels of Figures 5.3.1–5.3.6. Each figure corresponds to the data sets in Table 5.1, in the same order.

For Data 1 (Figure 5.3.1), the LIDs suggested that an increase in Q in 1995, followed by a decline, was the easiest way to reduce MSAAE. The LIDs for M suggested that a decrease prior to 1995, followed by an increase and then a decline in M after 2000 was the easiest way to reduce MSAAE. However, the M perturbation maximum slope was roughly half the Q value which suggested that larger changes in M were required to remove the retrospective pattern than changes in Q.

In a separate analysis, the Qs and Ms were modified by applying scaled perturbations using the d_{max}'s to reduce MSAAE to close to zero. The scaling was chosen by trial and error. The scaled perturbations are shown in the middle panels of Figure 5.3.1. The scaling values and the values of MSAAE are shown in the figure legend. Note that we could reduce MSAAE closer to zero using Q perturbations (MSAAE = 0.003) than using M perturbations (MSAAE = 0.02). These results give a clearer indication of the size of perturbations required to reduce

MSAAE to near zero. Larger relative changes to the nominal M value of 0.2 were required than changes in Q. Even with these larger changes, the residual pattern was not completely removed (see right panels in Figure 5.3.1), and the pattern of the residuals from the M perturbation analysis suggests that this was not the source of the misspecification. These can be compared with the original residual presented later in Figure 5.4.1.

The LIDs for Data 1 pointed to the correct source of the residual problem, including the timing of the abrupt change in Q, but they did not accurately detect the form of the misspecification (given in Table 5.1). However, this correct detection appears to be an artefact rather than a success. The results in Figures 5.3.2–5.3.6 demonstrate that smaller changes to Q could always reduce MSAAE to near zero, independent of the actual type of VPA misspecification. It seems that the magnitude of the perturbations were not comparable. To assess this, we computed the bootstrap mean and standard deviation for the maximum slope. The results in Table 5.3.1 show that the null expectation of the maximum slope for all datasets is much greater for Q perturbations than for M perturbations. This suggests that these two perturbation schemes are not directly comparable. However, we were unable to utilize the bootstrap results to obtain a better indicator of the source of the retrospective problem.

Table 5.3.1. Maximum local slopes with bootstrap means and standard deviations (S.D.). Results are in percent of MSAAE.

	SURVEY CATCHABILITY (Q)			NATURAL MORTALITY (M)		
Data	Max Slope	Mean	S.D.	Max Slope	Mean	S.D.
1	38.8	15.72	2.21	17.3	3.37	0.70
2	67.5	28.11	3.99	21.2	5.25	0.97
3	41.5	16.88	2.42	23.7	4.21	0.99
4	61.5	25.72	4.15	13.3	5.61	1.41
5	165.0	93.91	14.27	75.3	19.54	4.43
6	52.0	21.79	3.53	15.5	5.91	17.57

The LIDs correctly determined the major timing of change in Data 2 and Data 3 (Figures 5.3.2 and 5.3.3), but they did not get the form right. For example, in Data 2 the LIDs (Figure 5.3.2) from the M perturbation analysis suggested a decrease in M prior to 1995 which was not correct. We were not able to remove residual patterns in Data 3 as successfully as in Data 2 (compare Figure 5.3.3 with Figure 5.3.2). In Data 3, catches were under-reported since 1995, and the fact that we could not completely account for the residual pattern using changes in M suggests there is some potential to differentiate between catch misreporting and changes in M. However, this requires further investigation.

The LID results for Data 4 and Data 5 were more promising. These were examples with smaller misspecifications for which we expect the LID's to perform better. For Data 4 (Figure 5.2.4), the LIDs correctly detected the major increases in Q in 2000 and 2002, although the LIDs suggested a 2-fold increase was required to remove residual patterns, and smaller decreases in some other years, when in fact the problem was a three-fold increase in Q only in 2000 and 2002. For Data 5 (Figure 5.2.5), the LIDs detected the basic pattern of an increase in M, although the decrease in 2003-04 was not correct.

As expected, the results for Data 6 (Figure 5.3.6) showed that the LIDs could not detect misspecifications of more than one VPA input, although separately they did, to some extent, detect the problems with the individual components.

To assess our speculation that reducing MSAAE would also result in reduced retrospective patterns we performed retrospective analysis for the Q and M corrected VPAs. The LID-based corrections were treated as fixed effects and not re-estimated retrospectively. The results are shown in Figures 5.3.7–5.3.12 for Data 1-Data 6, respectively. The corrected VPAs have small

or no retrospective patterns; however, the corrections do not lead to a reduction in misspecification bias. For example, in Data 1 the Q-corrected estimates of SSB (Figure 5.3.7) are more biased than the last year estimates from the original VPA. In general, **we conclude that LIDs cannot be used to correct a misspecified VPA**. The exception is Data 5 (Figure 5.3.11). The M-corrected VPA had little bias and was improved compared to the original VPA. This is because the M LIDs accurately detected the M misspecification in this case (Figure 5.3.5). Hence, if misspecifications are not large then LID corrected VPAs may be less biased. However, this is contingent upon selecting the correct source of the misspecification, which the LIDs cannot seem to do, at least directly. We re-visit this point in Section 5.4.

There are several ways that the LIDs may be improved. The dimensionality of the perturbations schemes may be too high and provide too much flexibility resulting in reduced efficacy. Constraints or smoothing of perturbation schemes may improve the reliability of the diagnostics; however, this will depend on the appropriateness of the constraints or smoothing. A major problem with the approach is that the diagnostics are based on slopes at the origin, and the origin is taken to be the unperturbed and biased VPA parameter estimates. We suspect that this is an important reason why the real VPA problems are masked when the problems are large. This may be similar to masking of outliers in linear regression. The LIDs can also be computed at other points on the influence graph, and more reliable diagnostics may be obtained from a different point; for example, one where SSB is at a more "typical value". This may be a useful avenue for future research.

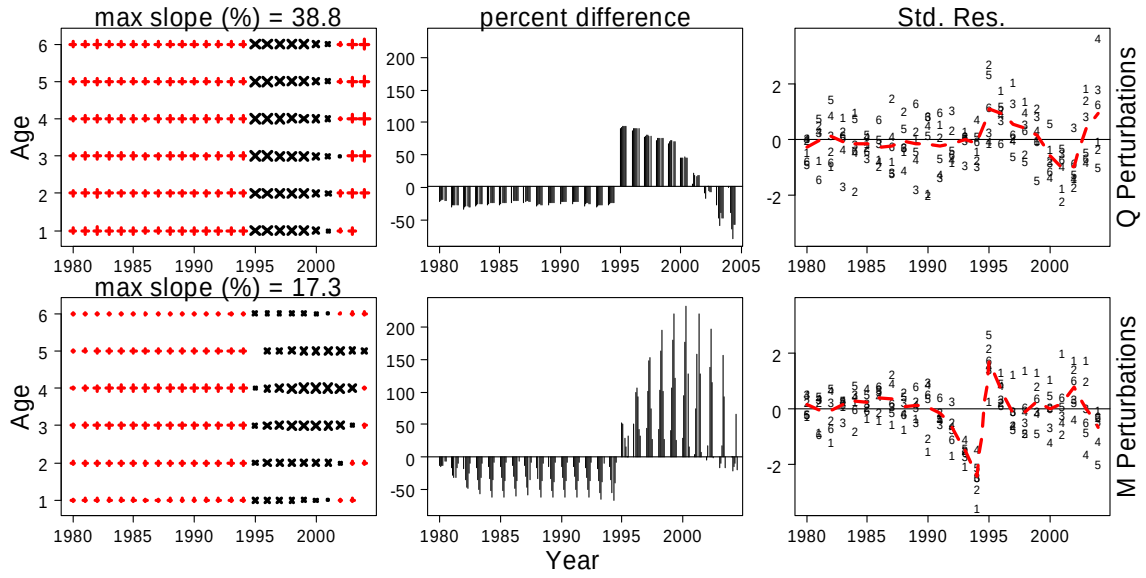

Figure 5.3.1. Local influence results for Data 1. Survey catchability (Q) perturbation results are shown in the top panels, and natural mortality (M) results are shown in the bottom panels. Left Column: Elements of the direction of maximum slope, dmax, for MSAAE. The size and type of the plotting symbols are proportional to the absolute value and sign of the elements, respectively. Negative is denoted by an ×. The slope of dmax is shown at the top of each panel, in percent of MSAAE=0.17 from the unperturbed standard VPA. Middle panels: VPA perturbations to reduce MSAAE. Each vertical line shows the percent perturbations; they are grouped by year and shown sequentially for ages 1-6. The values are obtained as –h×dmax, with h=5.5 for Q perturbations resulting in MSAAE=0.003, and h=8.6 for M perturbations resulting in MSAAE=0.02. Right panels: Time series of residuals after pertubations. Red dashed lines connect the annual averages. The plotting symbols are ages.

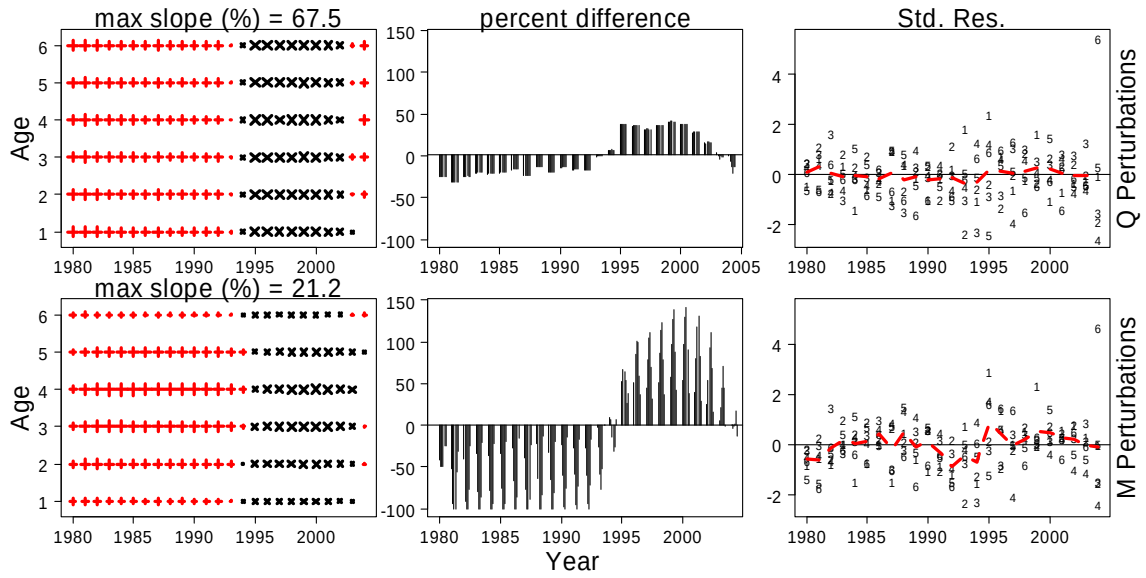

Figure 5.3.2. Local influence results for Data 2. Figure description is given in Figure 5.3.1. MSAAE=0.06 from the unperturbed standard VPA. H=3.0 for Q perturbations resulting in MSAAE=0.0003, and h=9.5 for M perturbations resulting in MSAAE=0.003.

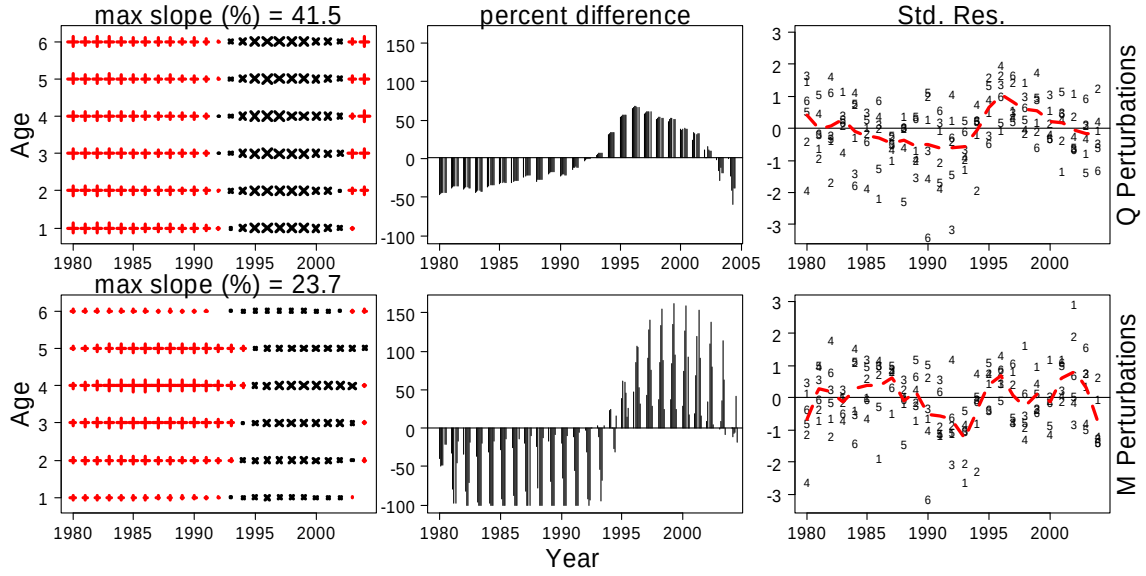

Figure 5.3.3. Local influence results for Data 3. Figure description is given in Figure 5.3.1. MSAAE=0.15 from the unperturbed standard VPA. h=4.5 for Q perturbations resulting in MSAAE=0.003, and h=12.0 for M perturbations resulting in MSAAE=0.005.

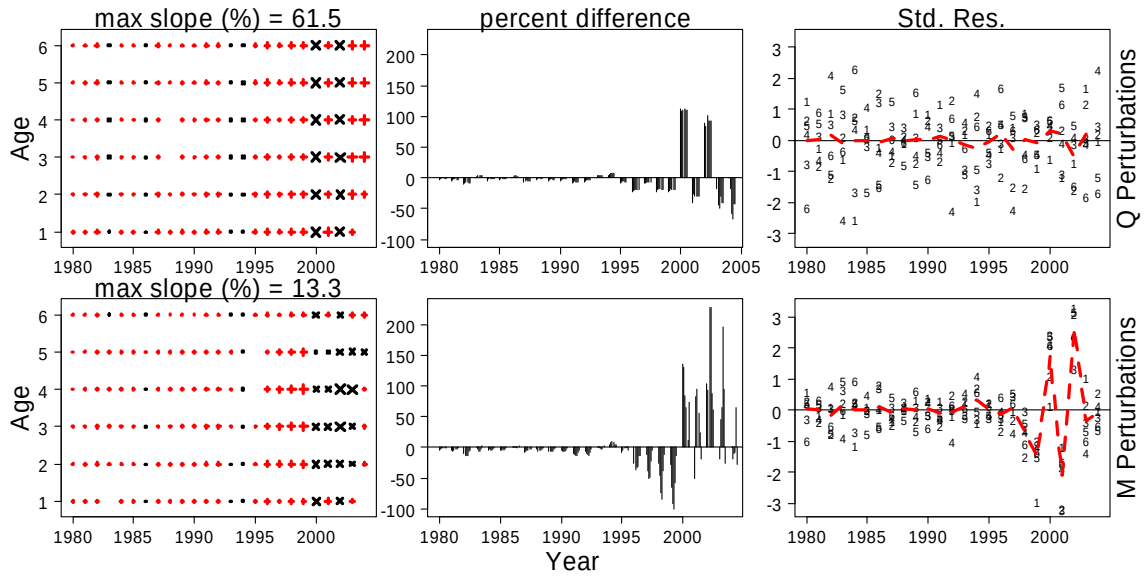

Figure 5.3.4. Local influence results for Data 4. Figure description is given in Figure 5.3.1. MSAAE=0.07 from the unperturbed standard VPA. h=4.0 for Q perturbations resulting in MSAAE=0.0003, and h=5.8 for M perturbations resulting in MSAAE=0.04.

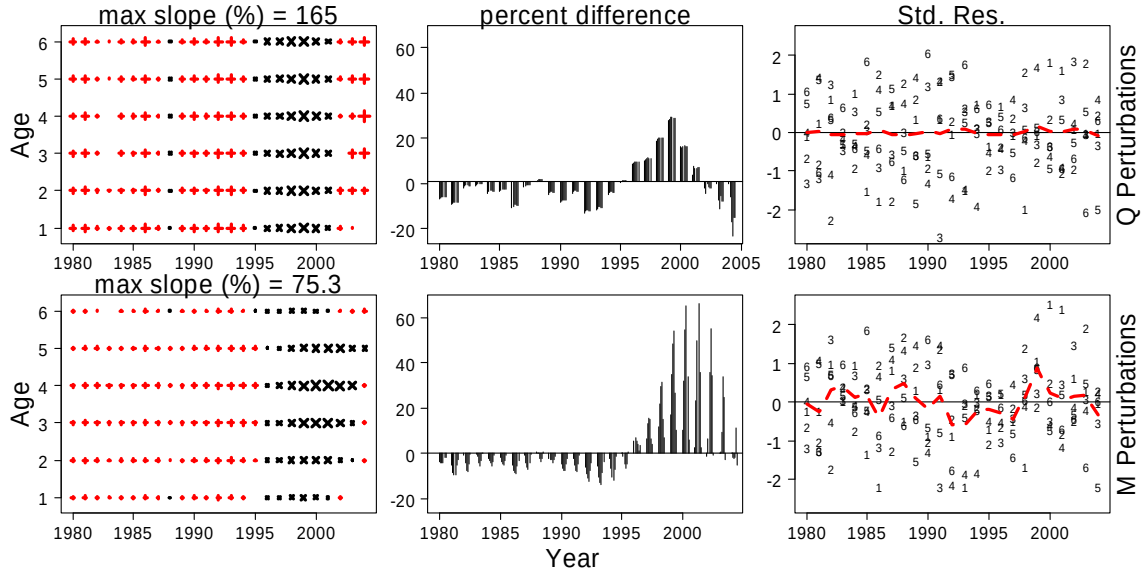

Figure 5.3.5. Local influence results for Data 5. Figure description is given in Figure 5.3.1. MSAAE=0.01 from the unperturbed standard VPA. h=1.3 for Q perturbations resulting in MSAAE<0.0001, and h=2.0 for M perturbations resulting in MSAAE=0.002.

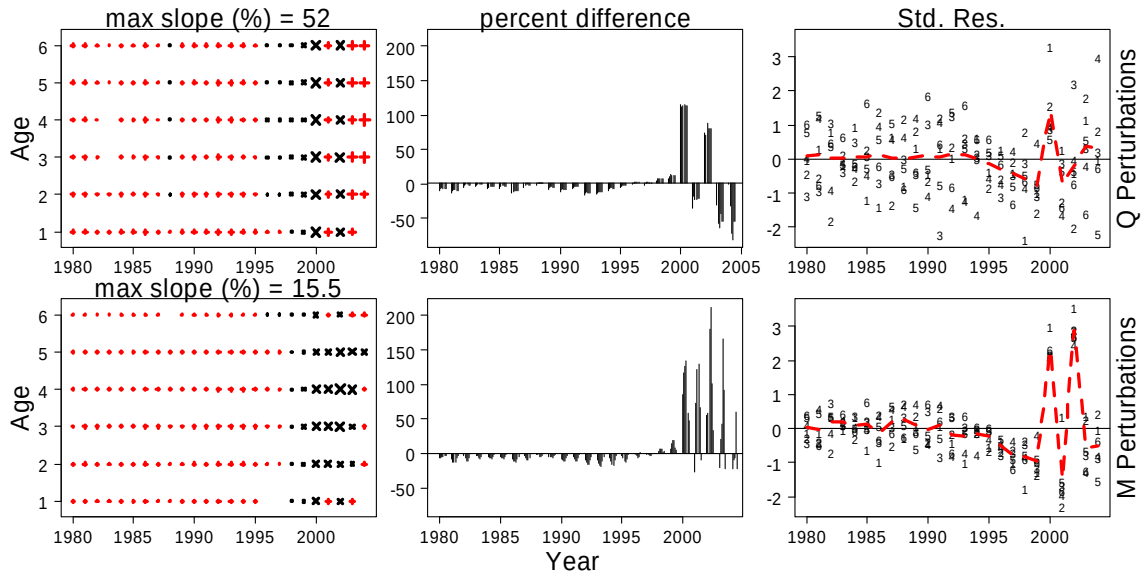

Figure 5.3.6. Local influence results for Data 6. Figure description is given in Figure 5.3.1. MSAAE=0.09 from the unperturbed standard VPA. h=4.0 for Q perturbations resulting in MSAAE=0.003, and h=5.0 for M perturbations resulting in MSAAE=0.06.

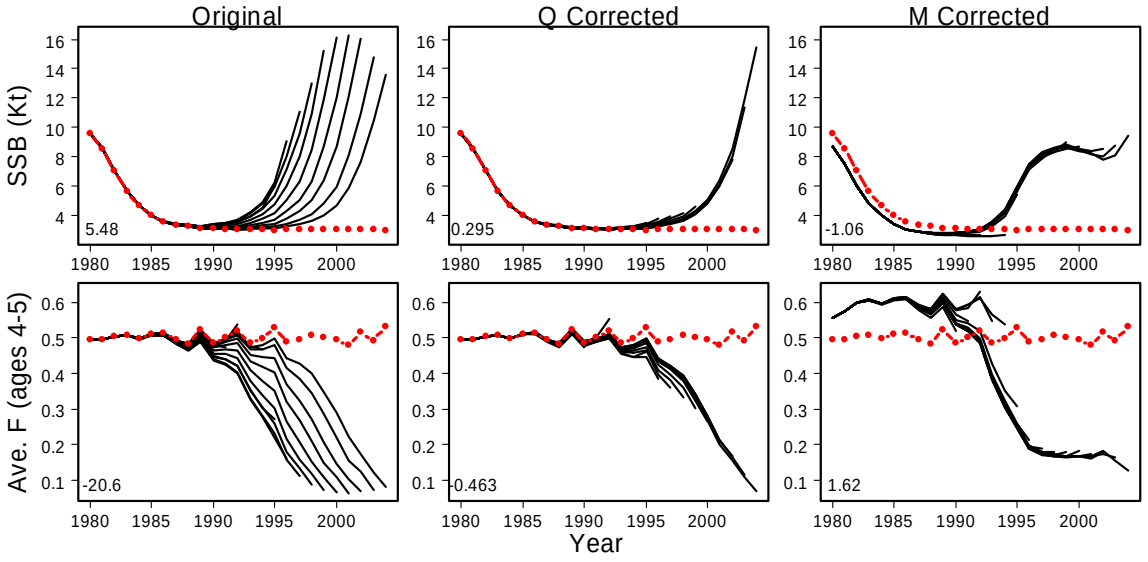

Figure 5.3.7. Retrospective patterns in SSB and F for Data 1. In all panels the red points and lines show the results from the VPA in which the correct mis-specification has been applied when fitting. Mohn's ρ statistic is shown in the bottom left-hand corner of each panel. The left panels show the results from the standard (but mis-specified) VPA. The middle panels show the results for the survey catchability (Q) corrected VPA based on the local influence direction of maximum slope; d_{max} (left panels in Figure 5.3.1). The right panels show the results for the natural mortality (M) corrected VPA based on d_{max} (right panels in Figure 5.3.1).

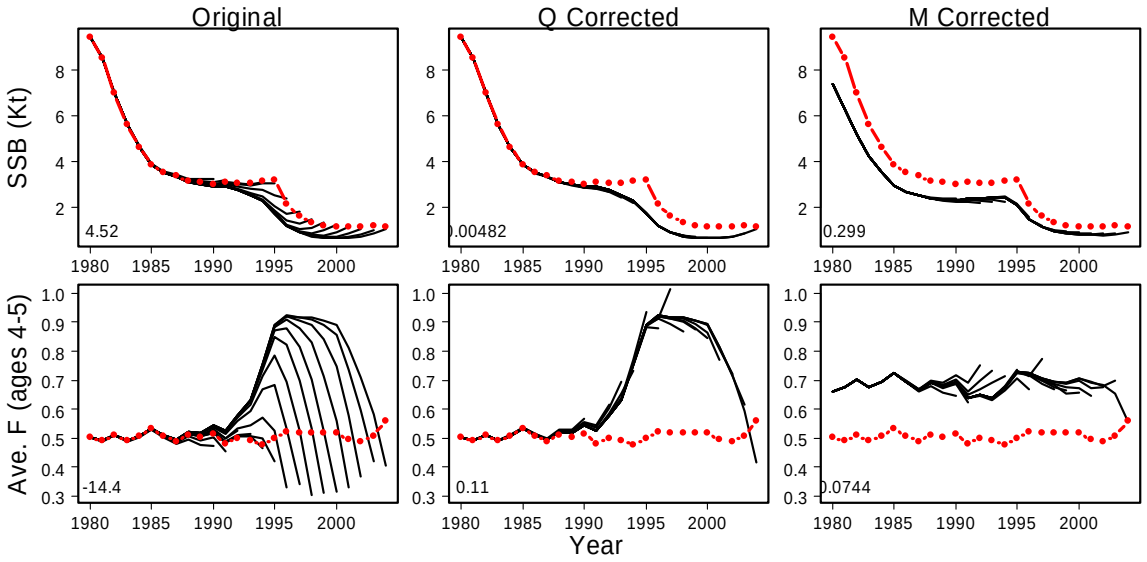

Figure 5.3.8. Retrospective patterns in SSB and F for Data 2. Figure description is given in Figure 5.3.7. Perturbations are shown in Figure 5.3.2.

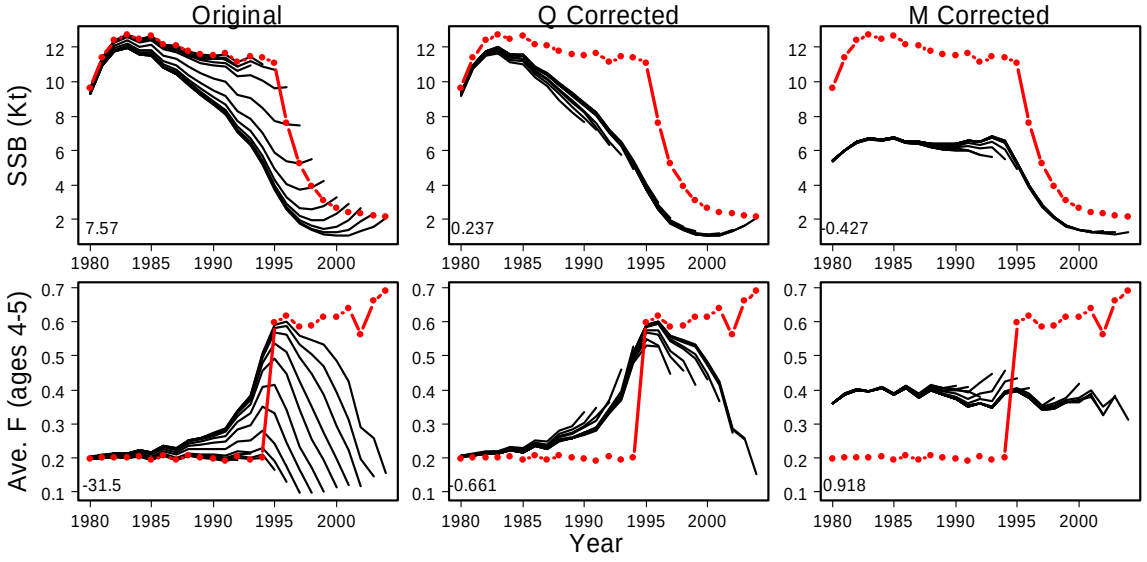

Figure 5.3.9. Retrospective patterns in SSB and F for Data 3. Figure description is given in Figure 5.3.7. Perturbations are shown in Figure 5.3.3.

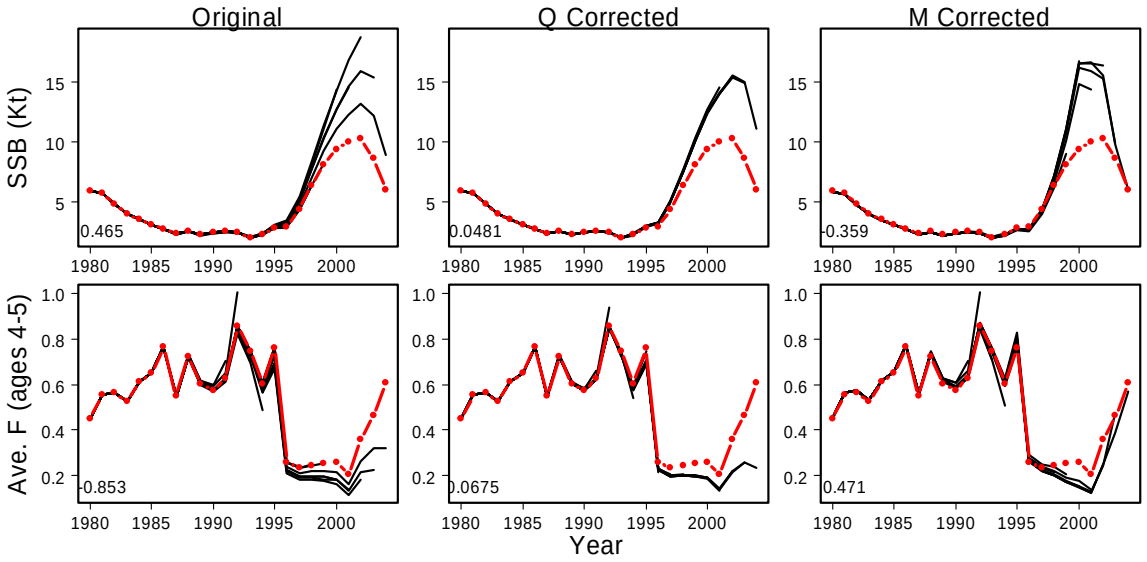

Figure 5.3.10. Retrospective patterns in SSB and F for Data 4. Figure description is given in Figure 5.3.7. Perturbations are shown in Figure 5.3.4.

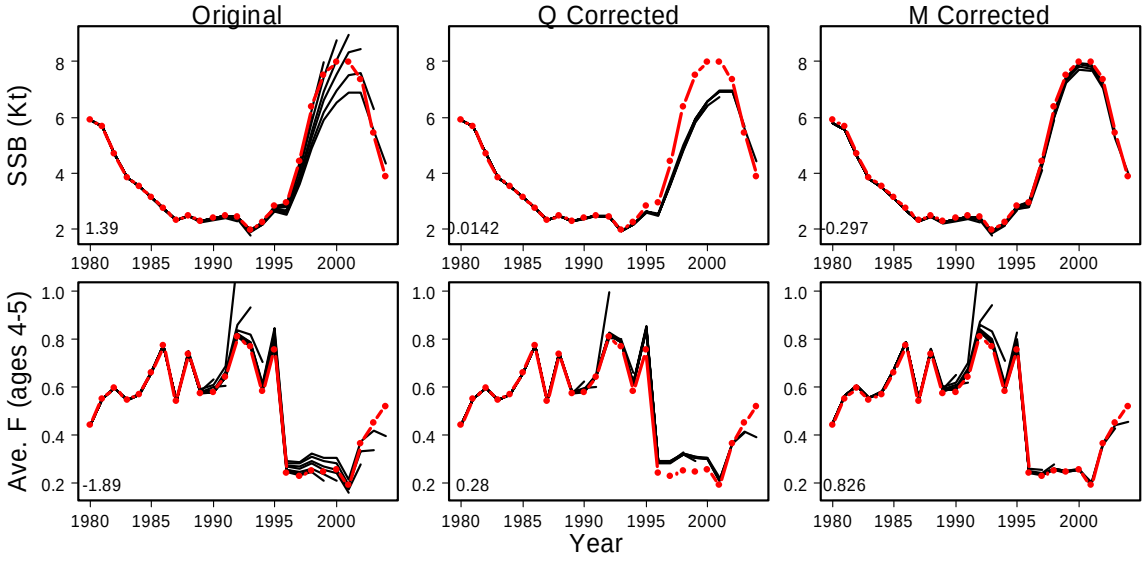

Figure 5.3.11. Retrospective patterns in SSB and F for Data 5. Figure description is given in Figure 5.3.7. Perturbations are shown in Figure 5.3.5.

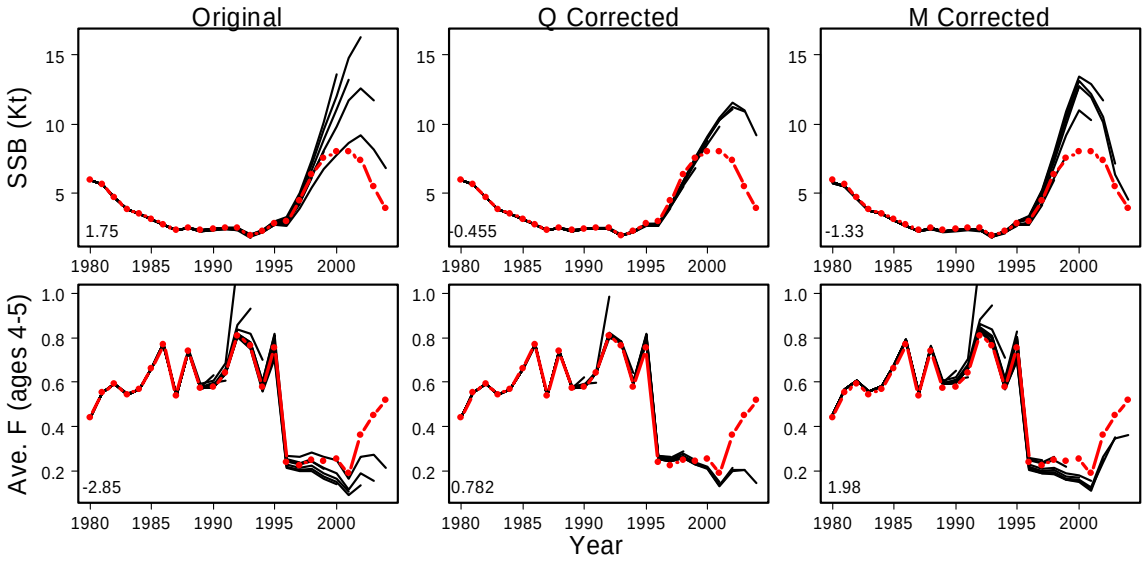

Figure 5.3.12. Retrospective patterns in SSB and F for Data 6. Figure description is given in Figure 5.3.7. Perturbations are shown in Figure 5.3.6.

5.4 The ADAPT approach with year effects in q (SPA)

Another approach to detect model misspecification is to imbed the basic model (here SPA) in a more complex framework, and estimate the various correction factors required to remove the misspecification or improve the residual problem (e.g. Carota *et al*., 1996). This type of approach can provide an advantage over the local influence approach because it is not based only on the unperturbed and biased VPA parameter estimates. However, we are limited in terms of the complexity of models we can estimate, so this restricts the types of misspecifications we can investigate. Nonetheless, the approach has been useful in other applications.

The strategy we investigate in this section is to estimate year effects in survey catchabilities (YEs), which would usually be assumed to be constant through time. The catchability model can be decomposed into age and year effects,

$$q_{ay} = \tau_y \times q_a$$

with some restriction on τ's to ensure identifiability. We constrain τ to be one for the first five years in the time series. In addition, if we suspect the problem is YE's then we propose that the estimated τ's be shrunk to the long term average, and if we suspect the problem is a trend in τ's then we propose that the drift in τ's be penalized. The year effects penalty we used was the sum of squares of $\log(\tau)$, and the drift penalty we used was the sum of squares of $\tau_{y+1} - \tau_y$. Time series of residuals from SPA applied to the six case studies listed in Table 5.1 are shown in Figure 5.4.1. Based on the residuals, we used a drift penalty for all cases except Data 4 and Data 6 for which we used the year-effects penalty.

The results for Data 1 are shown in Figure 5.4.2. The estimated YE's were very close to the true YE's and the residual pattern was removed satisfactorily. The population estimates were also much closer to the true values compared to the original VPA estimates.

The penalty weight we selected to fit the YE model was small. It was selected to give as much smoothing in estimated YE's as possible while not reducing goodness-of-fit very much. A more specific and algorithmic approach needs to be developed before this approach could be proposed for routine use.

The trend in residuals in Data 2 could also be removed using the YE model (Figure 5.4.3), concomitant with a large change in F during 1990-2004. The estimates of SSB and F were similar to the original VPA estimates, and the F's were seriously biased. Similarly the residual trend could be removed with YE's for Data 3 (Figure 5.4.4). The estimates of F were improved; however, they increased rapidly in the last few years.

The YEs estimated for Data 4 were reasonably close to the true values. Their confidence intervals covered the true values. The residual problems were removed. However, the population estimates were very similar to the original VPA estimates. These YE diagnostics were very similar to the LIDs in Figure 5.3.4. They were slightly more accurate in that the LIDs suggested a doubling of Q's in 2000 and 2002 whereas the YE estimates suggest slightly more than a doubling of Q's. The real change was three-fold. However, a reasonable conclusion from the YE diagnostics is that Q changed in 2000 and 2002, and not in other years. This is the correct model formulation whose results are also shown in Figure 5.4.5.

The YE model results for Data 5 are shown in Figure 5.4.6. In this case the YE model estimates are more biased, although the residual pattern was removed. However, the rapid increase in F in the most recent period may indicate that this is not the right correction if there is no supporting evidence for this increase.

For Data 6, the YE estimates were very close to the true values (Figure 5.4.7) and the residual patterns were removed. Stock estimates were closer to the exact model estimates, but there was still bias because of the mis-specified M component. This example demonstrates the difficulty that will occur when multiple VPA inputs are mis-specified.

The YE model diagnostics were similar to the local influence diagnostics for Data 4 and Data 6 (compare Figures 5.3.4 and 5.3.6 with Figures 5.4.5 and 5.4.7). These were two simulated data sets in which smaller misspecifications were used to produce retrospective patterns. Data 5 was also based on a relatively small misspecification, but the YE model diagnostics and the LIDs were more different. This may be due to the drift penalty used in the YE model. No such type of smoothing was used to generate the LIDs.

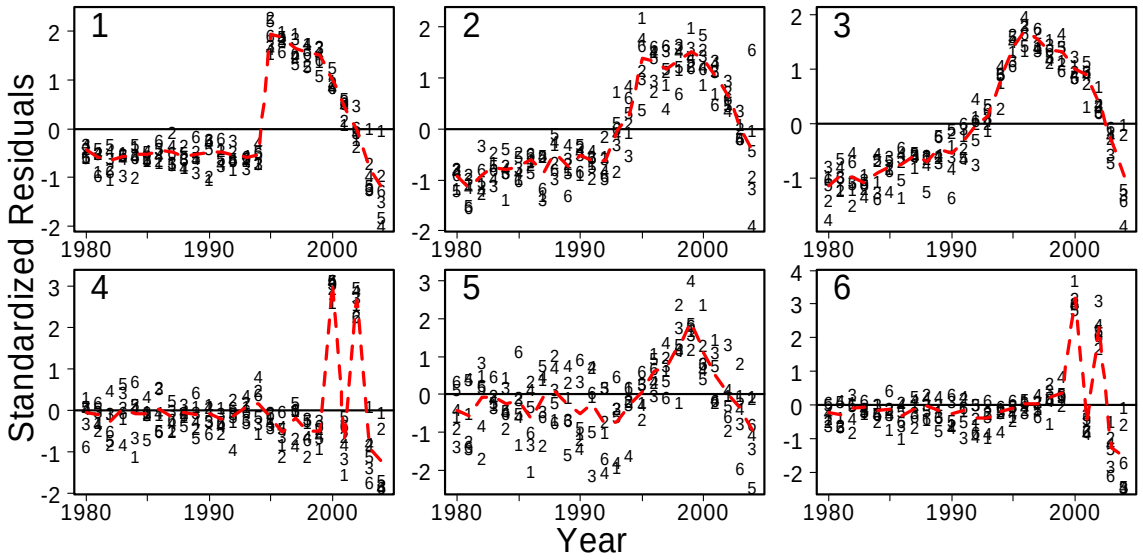

Figure 5.4.1. Time series of residuals for the six simulation case studies in Table 5.1. Case numbers are listed in the top right-hand corner of each panel. The red dotted lines connect the annual averages. The plotting symbols are ages.

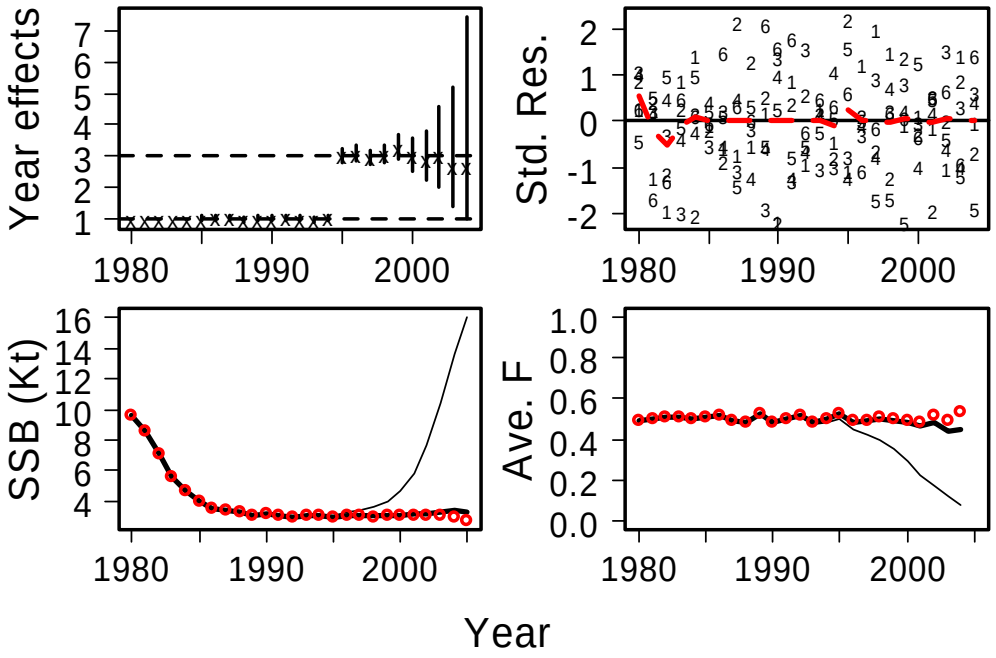

Figure 5.4.2. Data 1. Top left: Year effect (YE) model estimates of survey catchability, with approximate 95% confidence intervals. Top right: Time series of residuals from the YE model. Red dotted lines connect the annual averages. The plotting symbols are ages. Bottom: SSB (left) and average fishing mortality at ages 4–5 (right) from the YE model (heavy solid line), the exactly specified model (red circles), and the original VPA (thick solid line).

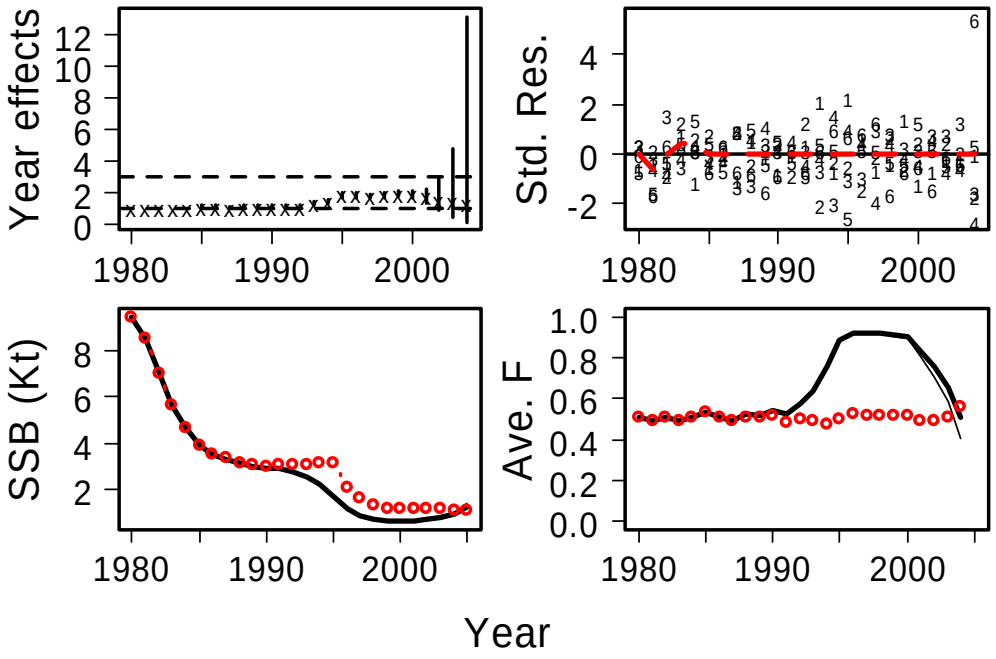

Figure 5.4.3. Data 2. Figure description is given in Figure 5.4.2.

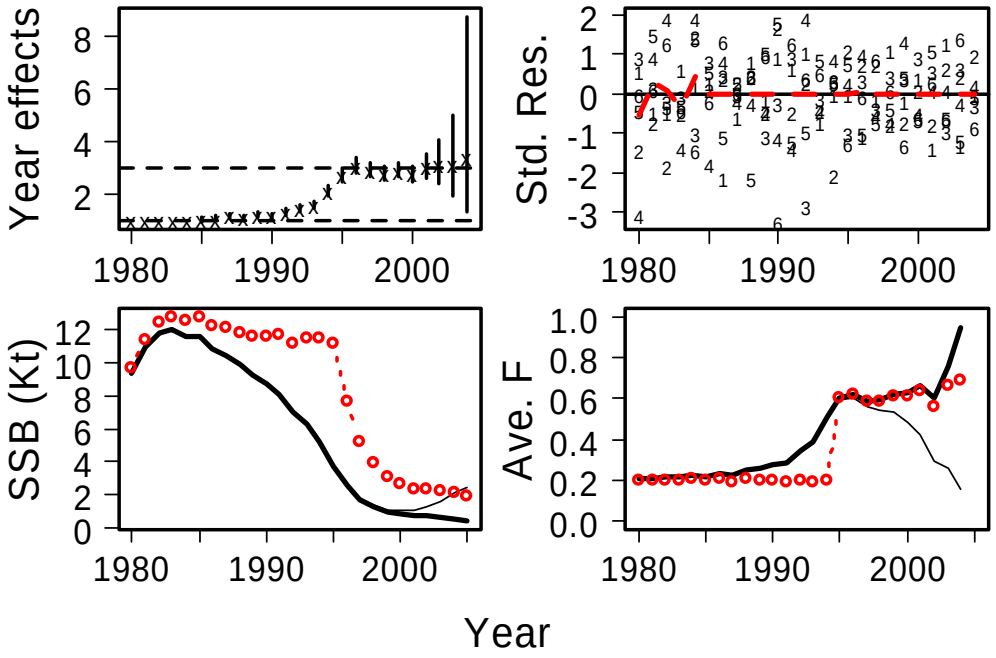

Figure 5.4.4. Data 3. Figure description is given in Figure 5.4.2.

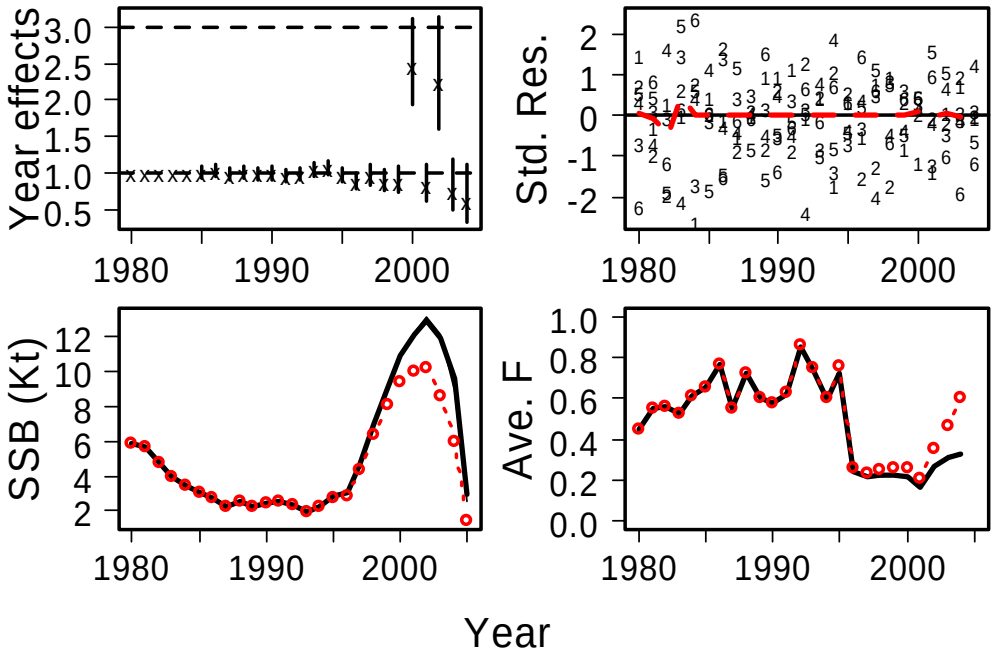

Figure 5.4.5. Data 4. Figure description is given in Figure 5.4.2.

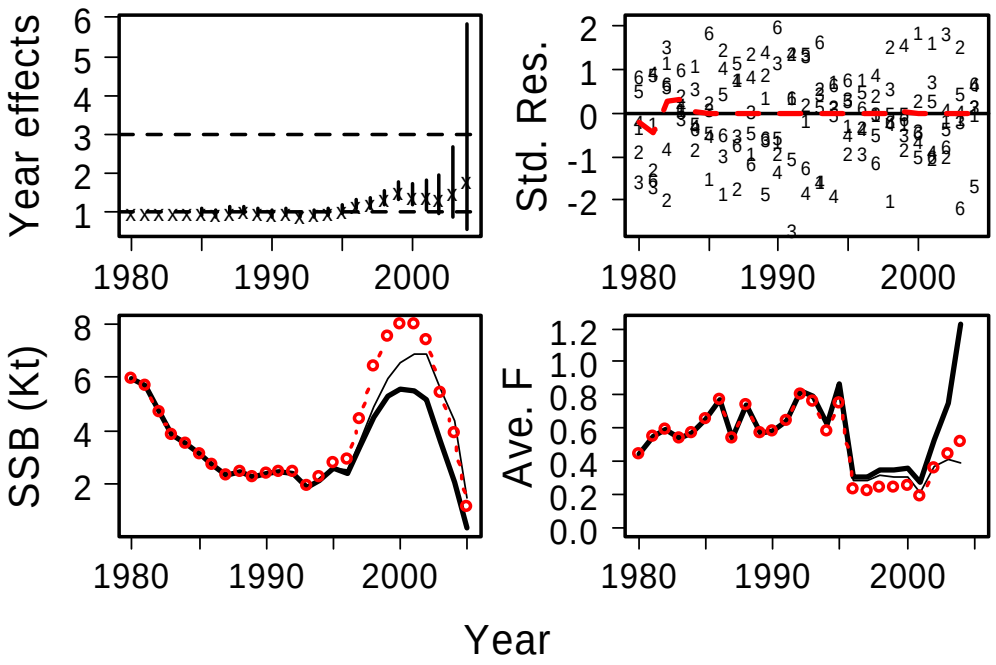

Figure 5.4.6. Data 5. Figure description is given in Figure 5.4.2.

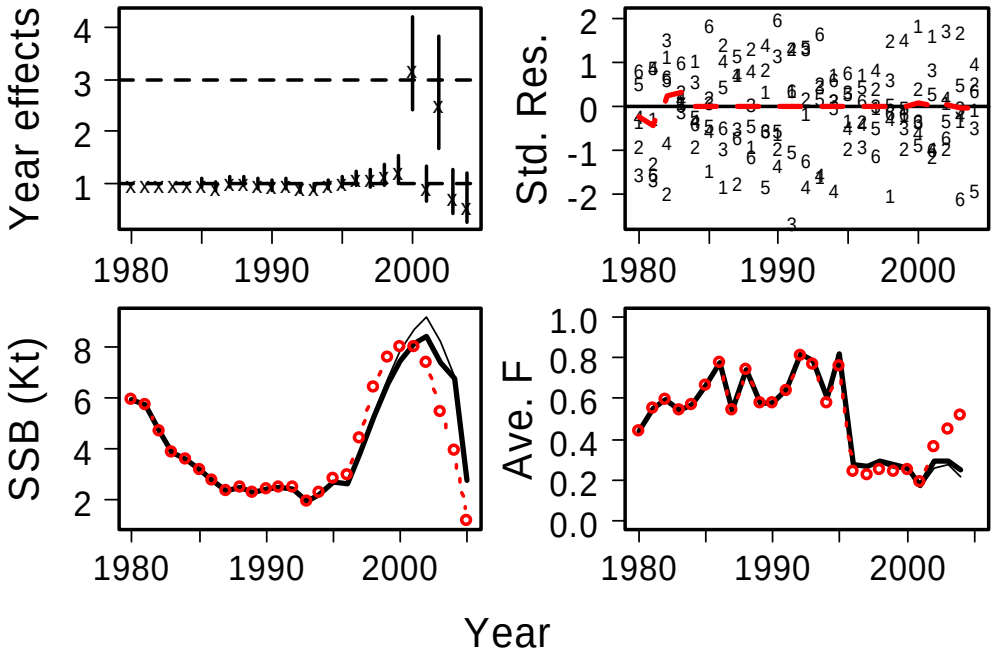

Figure 5.4.7. Data 6. Figure description is given in Figure 5.4.2.

5.5 The ADAPT approach with year effects in a catch multiplier (B-ADAPT)

B-ADAPT (Darby 2004, 2005, WP8) is a method for fitting year effects in catch within the ADAPT model structure. The model was applied to the test data sets generated to induce retrospective patterns in order to examine the effect of estimating catch multipliers when catch, survey catchability and natural mortality components of the model are mis-specified.

Data set 1 – catchability increased threefold in 1995 and subsequent years

Figure 5.5.1 presents the results of the B-ADAPT model fits to the data set in which survey catchability is increased threefold in 1995 and subsequent years. A penalty weight of 0.1 was applied to year to year changes in fishing mortality in order to stabilise the year effect estimates (effectively, year-effect estimates are smoothed by penalising large inter-annual variations).

To reduce the residual deviances in the most recent years of the time series (Figure 5.4.1 shows residuals for each data set), the model fits catch multiplier effects with the same pattern and scale as the induced changes in catchability used to generate the simulated data. The B-ADAPT residuals are low in all years apart from the period either side of the catchability change point. The recent trend in spawning stock biomass is poorly estimated by the model by a factor of three, fishing mortality is appropriately estimated at the start and end of the series but is estimated to have a strong decrease around the time of the catchability change point.

If the increase in survey catch rates were compared with data from other surveys (if available) or a change in survey catchability explored based on the results of the pre-screening (Section 5.2) then the retrospective pattern could be modelled and removed by fitting a change in survey catchability, as discussed in Section 5.4. In the absence of such external information, however, this would certainly be the wrong thing to do.

Data set 3 – 1/3 of the true catch reported in 1995 and subsequent years

This is the scenario that the B-ADAPT model was designed to analyse. Figure 5.5.2 presents the results of the B-ADAPT model fits to the data set in which 1/3 of the true catch reported in 1995 and subsequent years. Following exploratory analyses, it was decided that no penalty weight was needed to stabilise the model estimates.

The model fits catch multiplier effects with the same pattern and scale as the induced changes in catch. The year effect fitted model residuals are reduced to low values in all years. The model estimates the trends and scale of spawning stock biomass and fishing mortality correctly.

Data sets 2 and 5 – natural mortality increased threefold in 1995 and subsequent years and natural mortality increased by 50% in 1998 and subsequent years

Figures 5.5.3 and 5.5.5 present the results of the B-ADAPT model fits to the data sets in which natural mortality is increased. No penalty weight was needed to stabilise the model estimates.

The model fits catch multiplier effects with the same pattern and scale as the induced changes in natural mortality. The year effect fitted model residuals are reduced to low values in all years. The model estimates the trends and scale of spawning stock biomass correctly, fishing mortality is estimated correctly at the start of the series but is over-estimated by the same amount as the change in natural mortality in the years in which natural mortality is increased, consequently total mortality Z is estimated correctly.

Data set 4 – catchability increased threefold in 2000 and 2002

Figure 5.5.4 presents the results of the B-ADAPT model fits to the data set in which survey catchability is increased threefold in 2002 and 2004. A penalty weight of 0.1 was applied to year to year changes in fishing mortality used to stabilise the year effect estimates.

The model fits increases in catches in an attempt to remove the residual pattern. However, the final log catchability residual pattern still has strong outliers in 2000 and 2002. Negative residuals are induced in adjacent years. The model over-estimates the biomass levels in 1999–2003 but recent values are only slightly over estimated. Recent fishing mortality estimates are also over-estimated.

If the two years of survey CPUE data are removed (or down weighted) in the model fit as suggested by their identification as outliers by the pre-screening (Section 5.2) the model fits the data sets appropriately and the stock estimates are improved (results not shown).

Data set 6 – catchability increased threefold in 2000 and 2002 and natural mortality increased by 50% in 1998 and subsequent years

Figure 5.5.6 presents the results of the B-ADAPT model fits to the data set in which survey catchability is increased threefold in 2002 and 2004 and natural mortality increased by 50% in 1998 and subsequent years. A penalty weight of 0.1 was applied to year to year changes in fishing mortality used to stabilise the year effect estimates.

The model fits increases in catch in an attempt to remove the residual pattern. However, the final log catchability residual pattern still has strong outliers in 2000, 2001 and 2002. Negative residuals are induced in adjacent years. The model over-estimates the biomass levels in 1999–2003 but recent values are only slightly over estimated. Recent fishing mortality estimates are over-estimated as the increase in mortality is attributed to fishing.

If the two years of survey CPUE data are removed (or down weighted) in the model fit as suggested by their identification as outliers by the pre-screening (Section 5.2) the data sets and model fits revert to the data set 5 scenario (results not shown).

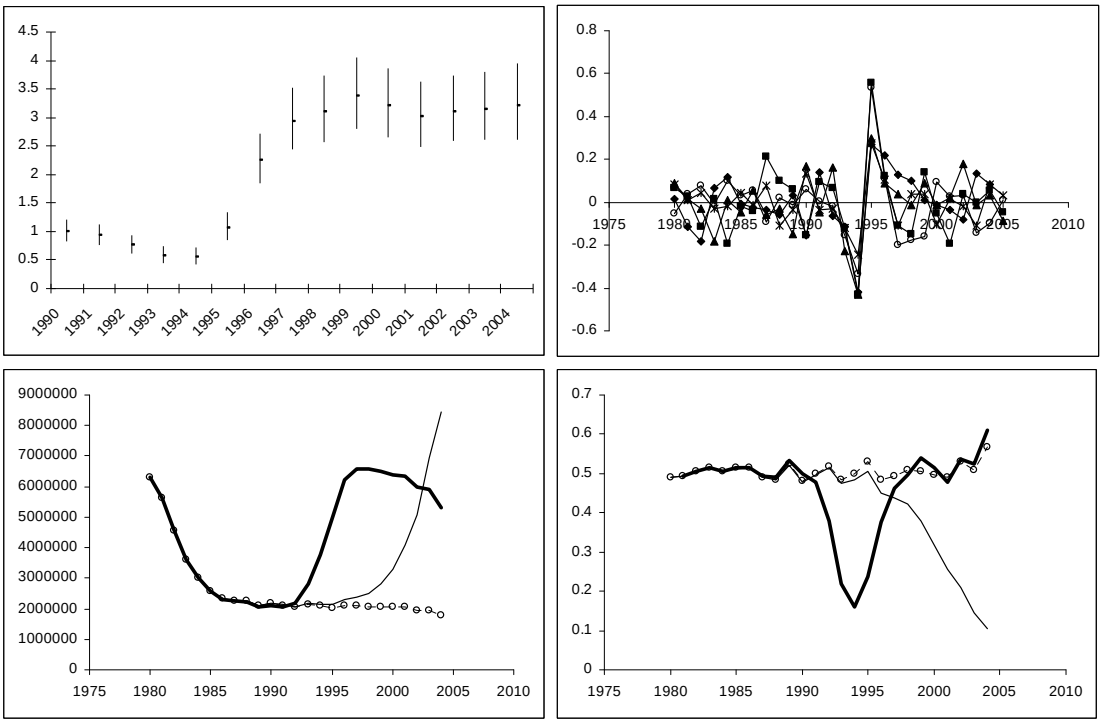

Figure 5.5.1. Data set 1 catchability three fold increase after 1994.
a) (top left) The estimated catch multiplier year effects required to reduce the log catchability residual patterns resulting from the change in catchability.
b) (top right) The resulting time series of log catchability residuals.
c) (bottom left) Spawning stock biomass estimated without fitting the catch multiplier year effects (thin line), with estimation of year effects (thick line) and the values estimated with the correct specification of catchability (circles).
d) (bottom right) Fishing mortality estimated without fitting the catch multiplier year effects (thin line), with estimation of year effects (thick line) and the values estimated with the correct specification of catchability (circles).

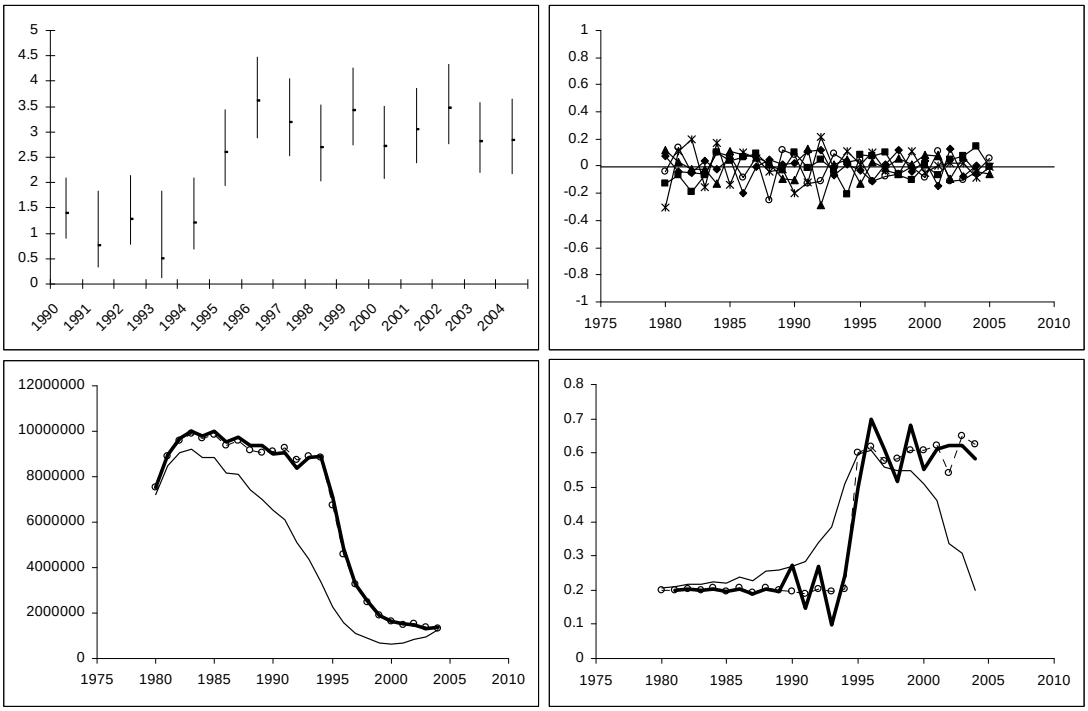

Figure 5.5.2. Data set 2 – 1/3 of the true catch reported in 1995 and subsequent years. Figure description is given in Figure 5.5.1.

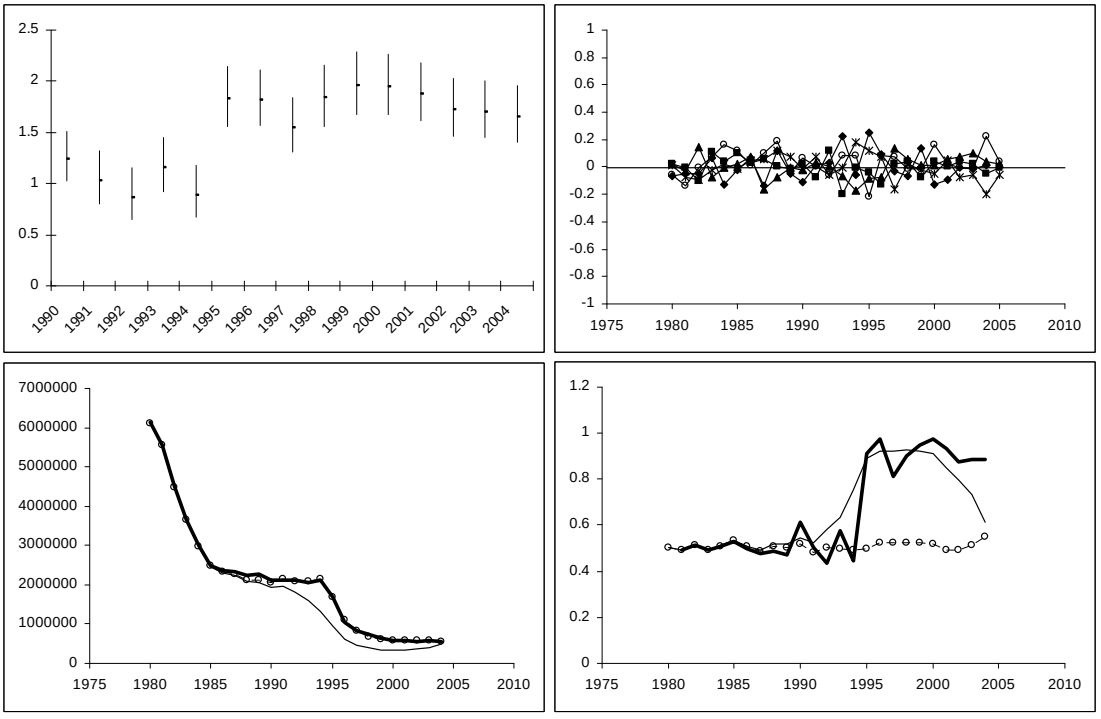

Figure 5.5.3. Data set 3 - three fold increase in M increase after 1994. Figure description is given in Figure 5.5.1.

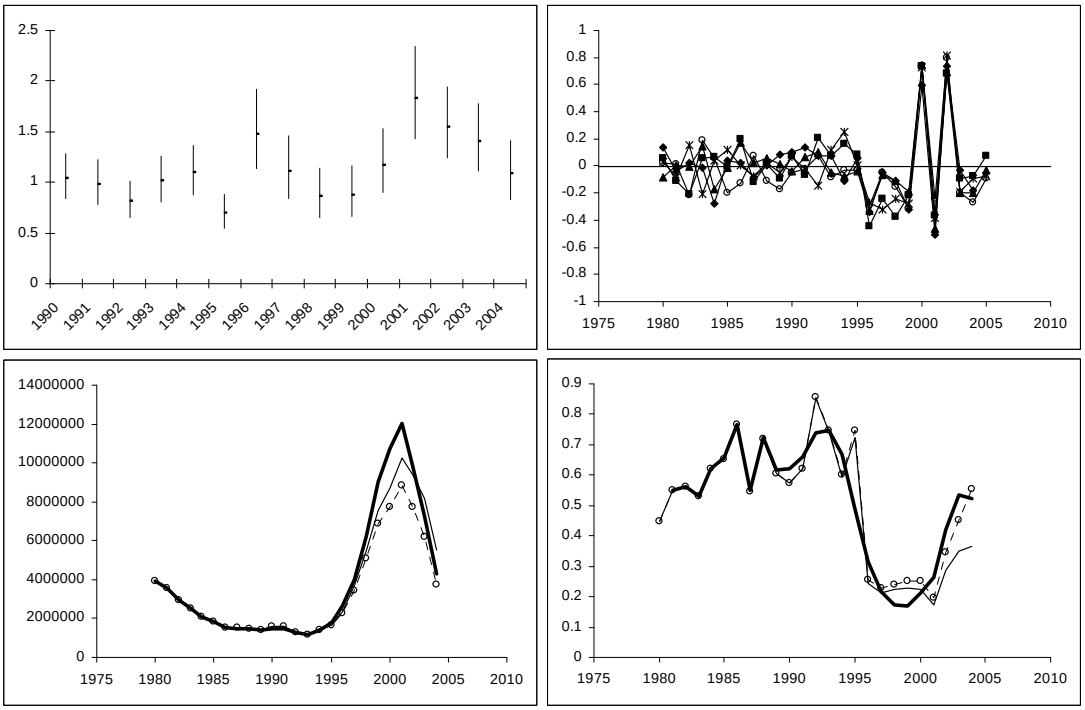

Figure 5.5.4. Data set 4 – three fold increase in catchability in 2000 and 2002. Figure description is given in Figure 5.5.1.

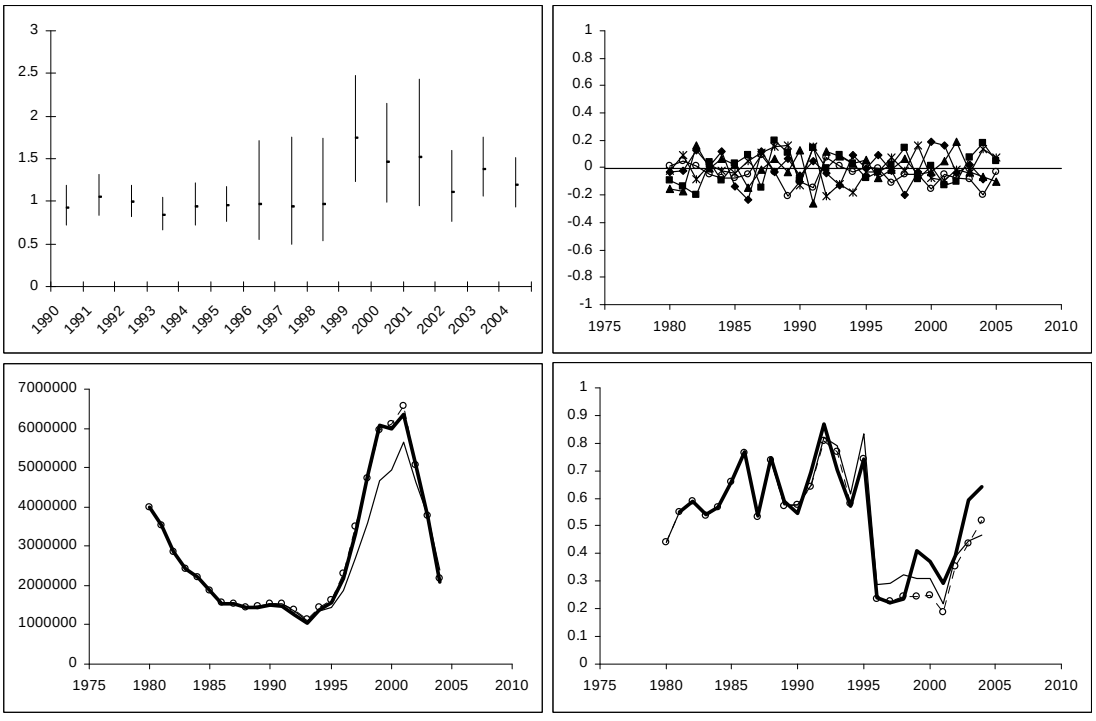

Figure 5.5.5. Data set 5 - natural mortality increased by 50% in 1998 and subsequent years. Figure description is given in Figure 5.5.1.

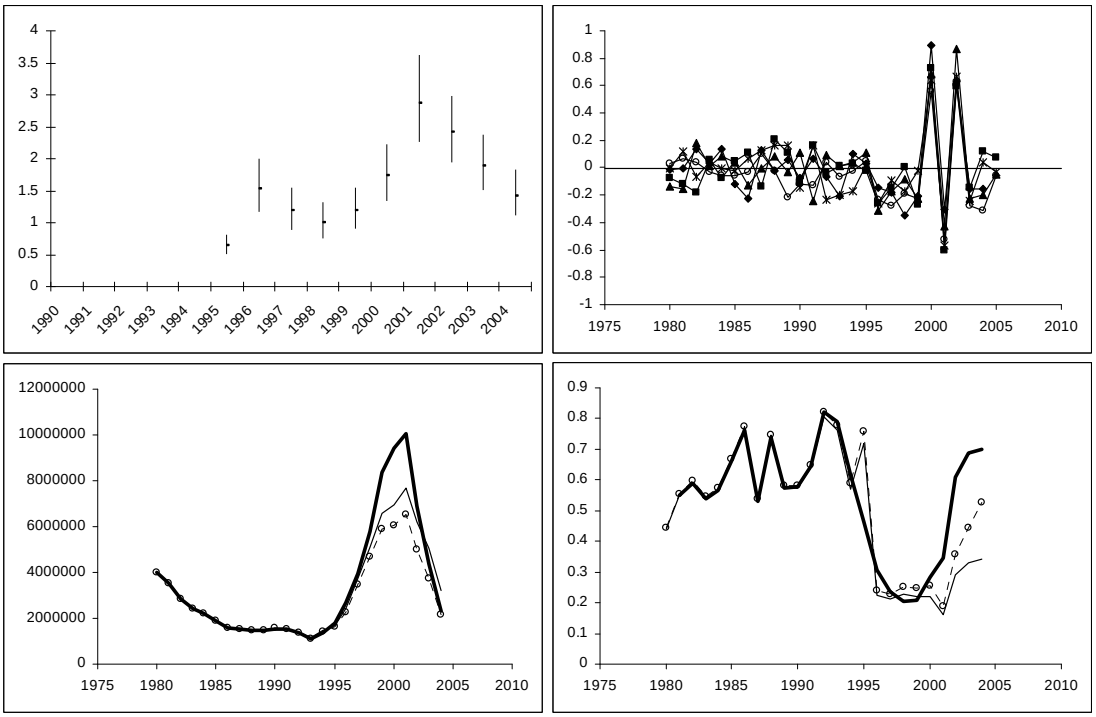

Figure 5.5.6. Data set 6 - catchability increased threefold in 2000 and 2002 and natural mortality increased by 50% in 1998 and subsequent years. Figure description is given in Figure 5.5.1.

5.6 Conclusions

Pre-screening techniques applied to assessment input data are able to identify large changes in survey catchability (both step changes over time and individual year effects), because these changes will show up in the survey data, but not the catch data. Used in isolation, they are not able to identify changes in natural mortality or misreported catch, as these impact both survey and catch data, and simple pre-screening techniques are not sufficient to detect or correctly identify simultaneous changes in these data sets.

In all six simulated case studies the LIDs suggested that smaller changes in catchabilities could reduce the residual patterns than could changes in the other components. This was independent of the real source of the problem and suggests that the approach cannot be used to diagnose the source of the problem. In addition, bootstrap results suggested that the four perturbation schemes were not directly comparable. For example, a multiplicative perturbation to catchability appeared to be larger than a multiplicative perturbation to natural mortality. However, a more positive result was that the diagnostics could more reliably detect the timing and direction of the problem.

We do not recommend correcting VPA retrospective patterns directly using the LIDs. In almost all cases this does not lead to less biased estimates. Correcting for retrospective patterns the wrong way can lead to a more biased assessment.

If the misspecification in the VPA is not large then the local influence approach may be reasonable for diagnosing the form of the misspecification (see Data 5). Also, the corrected VPA based on the LIDs may be less biased than the original VPA. This is not a trivial point, because in practise large retrospective patterns can occur in situations when the inputs do not appear to be seriously mis-specified. Although in simulations we have not been able to generate substantial retrospective patterns from small misspecifications, this may still happen in practise, and the LIDs seem more useful for this situation. However, their successful use still requires that the correct input is selected (e.g. M, catch, etc). Although on their own the LIDs were not useful for selecting the source of the misspecification, additional analyses of

corrected VPA diagnostics may provide additional information for this purpose, as demonstrated in Section 5.4.

The SPA catchability year effect (YE) model approach can reliably estimate year effects in survey catchability. It can also lead to improved estimates of SSB and F when survey year effects are the correct source of misspecification, but otherwise the stock size estimates may be worse.
It is important to choose the correct penalty function; otherwise, the diagnostics could be misleading even for the correct source of misspecification. More objective methods for determine penalty weights would be useful (e.g. L curve; Hansen and O'Leary, 1993).

The YE model is not proposed for assessment purposes, unless the underlying cause of the retrospective bias has been investigated or is known from other information. More research is required before its usage could be recommended, because estimates of standard errors and confidence intervals may not be reliable due to the use of the penalty function and weighting. Furthermore, it is not straight-forward to determine the number of parameters in this type of model, but it could be high. Both of these factors may lead to statistical inference difficulties.

The B-ADAPT model is able to reliably estimate year effects in catch, and reproduces estimates of SSB and F accurately, in situations where the catch is the correct source of the misspecification. It also provides accurate estimates of SSB in situations where M is the source of the misspecification, but then it allocates mortality to F instead of M, leading to biased estimates of F. It performs poorly in situations where the misspecification is in q.

Both ADAPT approaches use additional flexibility (estimating year effects in q, M or catches) to deal with model misspecifications that typically lead to retrospective bias, but even when they successfully remove residual patterns they do not always solve the correct problem. However, in these cases, the plausibility of corresponding population estimates (e.g. trends in F, etc) may also be used to indicate the correct source of the mis-specificartion. Clearly, additional information is needed to assess this plausibility, and this information may be available in some cases. For example, a corrected model that estimates a sudden increase in F may not be plausible if fishing effort is known to have not changed much. This approach requires further development, including simulation testing and application to real case studies.

6 Other analyses

6.1 Methods to estimate mean *F*

Estimates of F at age from virtual population analysis can be highly variable for "fully selected" ages due to the assumption that catch at age is known without error (or other model misspecifications). Managers desire a single value of *F* for each year to compare against reference points. The annual values for a specified range of ages could be calculated either as an arithmetic average or as a weighted average with the weights supplied by population abundance (*N*), biomass (*B*), or catch (*C*). Two simulation approaches were used to address the question of which of these four methods produced the least biased and smallest confidence intervals. The first simulation approach used a top-down approach of applying a gamma error to each *F*-at-age and deriving *N*, *B*, and *C* from an initial population structure. The second simulation approach used the NFT PopSim program (see WP6) to create datasets with noise in catch at age and survey indices which were supplied to a VPA to estimate the *F*-at-age.

The first simulation approach was conducted in Excel using the add-in @Risk. A time series of fully selected fishing mortality rates was chosen for years 1980 to 2006 that contained both high and low values. Selectivity at age was set to one for ages 5 to 10 and was approximately logistic for younger ages. Expected fishing mortality at age by year was simply the product of the year specific full F and the selectivity at age. A gamma distribution with CV=0.1, 0.2, or 0.4 was used to add noise to the fishing mortality rates at age and year independently for each value in the matrix. This F matrix was then used to project forward an initial population in equilibrium and a given stream of recruitment to generate matrices of population numbers and catch at age and year. A constant weight at age vector was applied to the population numbers matrix to compute biomass at age and year. Various recruitment time series for the population were considered including both low noise about a mean and pulses of large recruitment. Average F values were computed for ages 5–7 or 5–10 as unweighted or weighted by population numbers, population biomass, or catch. There were no observation errors for the population numbers, biomass, or catch. These simulations were made 10,000 times. This approach can be thought of as an examination of process error in F, or a top-down approach, with a fully consistent set of N, B, F, and C matrices.

Results for the various combinations of gamma distribution CV, recruitment streams, and ages in the average F calculation were completely consistent. The catch weighted F estimates were always biased, while the other three were not (Figure 6.1). Unweighted F always produced the smallest confidence intervals, sometimes as low as half the size of the intervals from the three weighted F estimates (e.g. Figure 6.2). Based on these results, the unweighted F was clearly the preferred metric.

The second simulation was a bottom-up approach whereby the F matrix was held fixed and datasets for input to VPA were generated with noise. Each dataset was fit by VPA and the four methods to calculate average F were computed from the estimated F matrix. An additional method, suggest during the WGMG meeting, of estimating F from the survivors of the cohorts was also examined. These simulations used the PopSim software program from the NOAA Fisheries Toolbox to generate the datasets with relatively large levels of noise in the catch at age and survey indices (40%–70% coefficients of variation). There were no changes in parameters during the time series, so the VPAs did not exhibit retrospective patterns. Estimates of F from the five methods were computed for 1,000 simulated datasets and compared to the underlying true values.

Results from this second approach did not show a clearly superior method of estimating average F. Each average F could perform best in any particular year of one particular simulation (Figure 6.3). When the bias was averaged over the entire time series for a specific

simulation, the best method varied depending on the particulars of the simulation in a non-predictable manner (Table 6.1). These results demonstrate that there is not one method for computing average F that clearly performs best in bottom-up simulations. All five methods performed well on average, as demonstrated by the low bias in almost all cases, but could be highly biased for a particular year in a specific simulation.

Based on the results from both simulations, care must be taken when summarizing F matrices as one method cannot be expected to always perform best. Comparison of the multiple methods to calculate average F should be undertaken and when they differ a specific reason should be given for why one method was selected over the others.

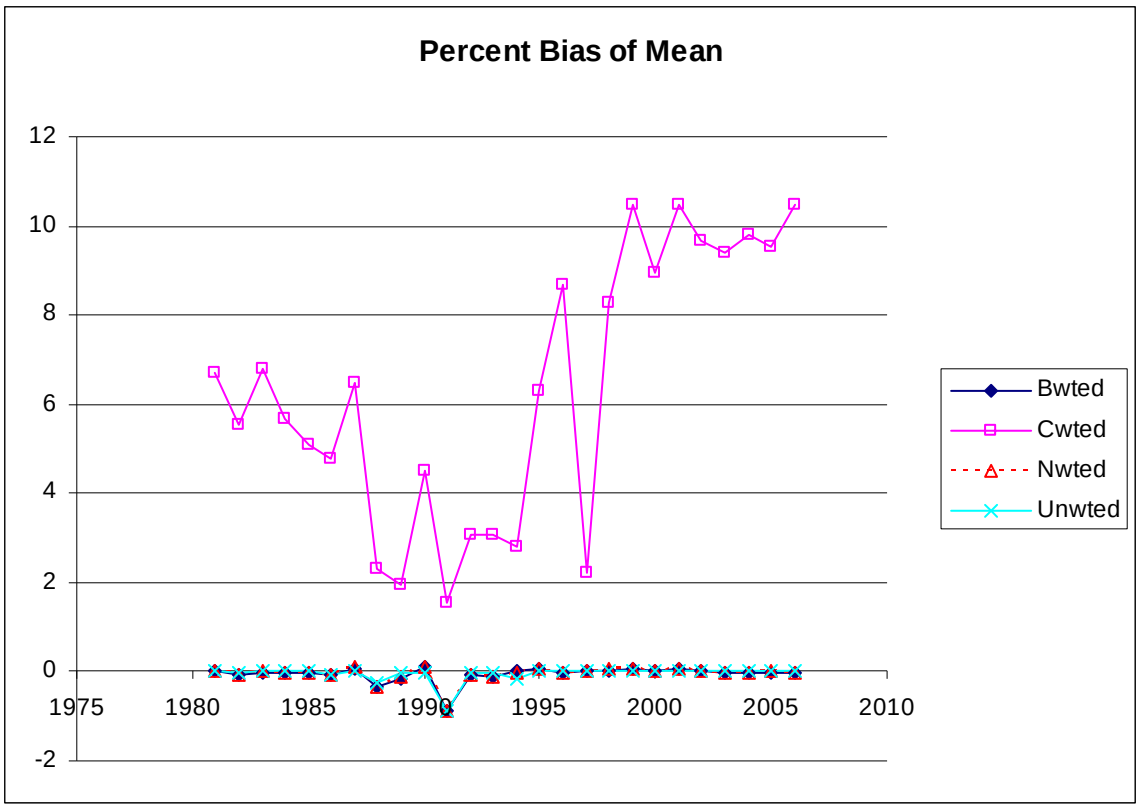

Figure 6.1. Percent bias from four estimates of average F using simulation approach 1 with a gamma CV of 0.2 applied to each F at age and year, relatively constant recruitment, and ages 5–7 used in the calculation of average F.

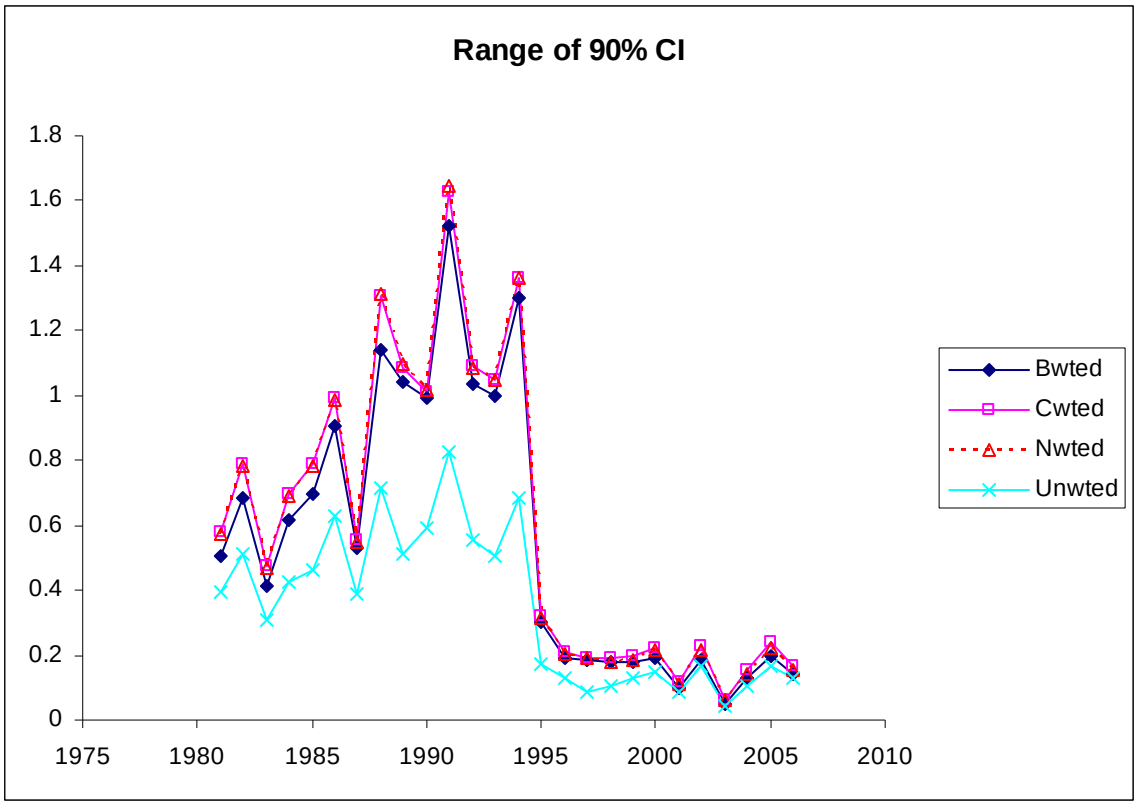

Figure 6.2. Comparison of the width of the 95% confidence intervals from four methods of estimating average F using simulation approach 1 with a gamma CV of 0.1 applied to each F at age and year, relatively constant recruitment, and ages 5–10 used in the calculation of average F.

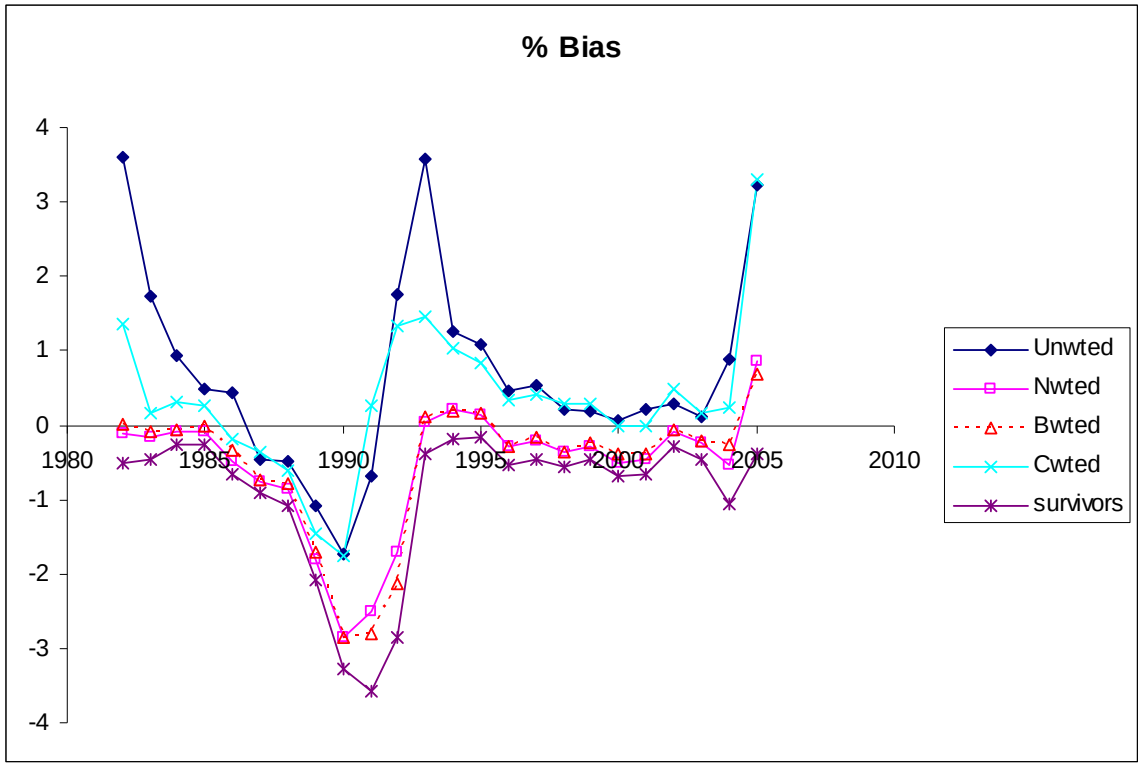

Figure 6.3. Percent bias in estimates of annual F for ages 4–6 from five methods in simulation approach 2.

Table 6.1 Average percent bias over the entire time series for five methods of estimating average F from six simulations using approach 2. Highlighted cells show the method with lowest (or nearly lowest) absolute bias for each simulation.

Simulation	Unwted F	N wted F	B wted F	C wted F	survivors F
1	10.02	1.66	1.83	6.03	0.65
2	2.66	0.80	0.88	1.99	0.44
3	0.81	-0.59	-0.58	0.45	-1.05
4	7.74	4.59	4.84	6.09	4.48
5	7.43	4.52	4.76	5.79	4.43
6	0.69	-0.55	-0.53	0.34	-0.93

7 Conclusions

7.1 Conclusions

WGMG worked this year via three subgroups. Subgroup A looked at methods for running management strategy evaluations (MSEs), and began the process of setting up simulations that would enable an answer to the question of how management advice might be affected by errors in assessments (in particular, retrospective bias). Subgroup B investigated ways in which the uncertainty in outputs from assessment models could be estimated – this was done by comparing Bayesian and bootstrap estimates of uncertainty arising from a comparatively simple surplus production model. Subgroup C looked further into the problem of retrospective assessment bias; that is, where each successive annual assessment substantially alters the perception of historical stock in a systematic way (either consistently increasing or decreasing it).

Subgroup A reached three main conclusions. Firstly, WGMG is not yet in a position to answer the questions of whether and how management should proceed in the presence of retrospective bias. The presence of such bias should lead to more cautious management, but how to implement this and how cautious such management should be is less clear. This is due principally to the complexity of programming management-strategy evaluations, but answers to these questions are certainly feasible using current approaches. Secondly, any management-strategy evaluation toolbox must allow for assessments to be run "live" as part of the evaluation loop. And thirdly, managers will get management decisions wrong if these are based on biased advice. This last point may seem obvious, but the analyses presented by Subgroup A highlight the issue with great clarity.

Assessments will always have problems of one sort or another, and it is important that MSEs are able to accommodate this fact. The function of WGMG in this regard is then to provide methods to do this. This endeavour therefore links the work of all three Subgroups.

Subgroup B provided important advances in the implementation of MCMC algorithms for model fitting, and went a considerable distance in generating comparisons of uncertainty estimates from bootstrap and Bayesian methods, with observation and/or process error, using a number of different datasets with different problems (one-way trips, under-reporting and changes in survey catchability). They were able to explore the varying reactions of models to these situations, but firm conclusions remained elusive due to considerable problems in software coding. The Section should be viewed as a strong advance in a work-in-progress. Nonetheless, it seems that not accounting for process errors can lead to a biased view of the true uncertainty in stock estimates based on approximate populations models. Reliable methods that account for process and measurement errors simultaneously in stock assessment models are not yet available.

Subgroup C used four different techniques to try and detect model mis-specification in six simulated datasets. The techniques were:

1) Pre-screening of data inputs to assessment models.
2) Local influence diagnostics (LIDs).
3) The ADAPT approach with year effects in survey catchability (SPA YE).
4) The ADAPT approach with year effects in a catch multiplier (BADAPT).

In the case of LIDs, the method was used to try and ascertain the cause of retrospective bias *directly*; the use of the other methods was restricted to an evaluation of which model mis-specification had been applied (and when), without a concomitant analysis of the effect on retrospective bias (although this would be the next step). Pre-screening techniques can only be used to identify large changes in survey catchability. Similarly, the SPA YE model can only

improve assessments when mis-specification of survey catchability is known to be the problem; and the BADAPT model performs best when errors in catch are the true source of mis-specification. LIDs suggested that survey catchability changes were responsible for retrospective bias in *all* simulations, even those in which the true cause was under-reporting and/or changes in natural mortality. In addition, correcting assessments using LIDs often removed retrospective bias but resulted in an incorrect assessment. These LIDs cannot therefore be considered reliable indicators for such problems, although they may still have utility when the VPA mis-specification is known to be small. However, a more positive result was that the diagnostics could more reliably detect the timing and direction of the problem when the source was known (e.g. M or survey catchability), especially in the more converged part of the VPA, Such models and diagnostics will perform best when used in combination with a) each other, and b) (more importantly) external information about the likely source of mis-specification.

Finally, analyses of different approaches to calculating a representative average F estimate for a given year were not able to determine any particular method that consistently performed well. Sensitivity of management advice to the method used needs to be evaluated on a case-by-case basis.

The main recommendations from the 2007 meeting of WGMG are summarised above. Of most direct relevance to this year's assessment Working Groups are the conclusions from Subgroup C, regarding testing for and correcting retrospective bias. The work of the other two Subgroups is at an earlier stage, but strong foundations for further work have been laid and plans are in train to continue collaborations. In addition, it was agreed that WGMG was an appropriate forum within which to carry forward certain aspects of size-based analyses; specifically, an exploration of the biases inherent in assuming size-based processes are age-based.

7.2 Recommendations

The recommendations from the 2007 WGMG meeting are summarised in the previous section.

7.3 Terms of Reference for next meeting

It is the opinion of WGMG that the existing ToRs for the group are too vague to encourage focussed work. The work in this year's meeting as been quite clearly directed towards a list of key objectives, but this has been achieved despite the ToRs rather than because of them. WGMG has also become a repository for an expanding list of special requests from other Expert Groups within ICES, which have in general been impossible to address because of lack of time, expertise or data.

WGMG therefore proposes that the ToRs for the next meeting be aimed quite specifically at addressing a shorter list (no more than four or five) of key methods-related problems. In the suggested ToRs below, these problems have been drawn largely from the experience of this year's subgroups (ToRs a, b, c) or from presentations and plenary discussions within this year's meeting (ToR d). WGMG believes that this more focussed approach will encourage participants to address specific problems, rather than simply attending the meeting to "work on methods."

The issue of special requests is yet more problematic. While the primary function of WGMG is to provide methodological advice to assessment Working Groups (via ACFM and RMC), it cannot do so by attempting to address a long list of requests from all and sundry. If such requests do not fit in with the ToRs and with the expertise of participants, they are very unlikely to be addressed. It is therefore important that the ToRs and likely membership of WGMG be considered closely by those making requests of the Group, and kept to a minimum if possible.

The text of the draft ToRs is given below, while the justification is given in Annex 2.

The **Working Group on Methods of Fish Stock Assessments** [WGMG] (Chair: Coby Needle, UK) will meet in Nantes, France, from 15–24 April 2008 (TBD) to:

a) Develop methods for evaluating harvest control rules appropriate for biased stock assessments;

b) Develop methods for detecting and correcting retrospective bias and noise using simulations that are typical of real data;

c) Develop methods for the appropriate incorporation and estimation of uncertainty in stock assessment methods;

d) Investigate the biases inherent in assuming length-based processes are age-based, and develop approaches to reduce such biases;

e) Examine the ability of distributions that accept zero observations for use in relating observed and predicted indices in stock assessment models.

WGMG will report by 13 May 2007 for the attention of the Resource Management Committee and ACFM.

8 References

Beare, D, J., Needle, C. L., Burns, F. and Reid, D. G. 2005. Using survey data independently from commercial data in stock assessment: An example using haddock in ICES Division VIa, ICES Journal of Marine Science, 62: 996–1005.

Bull, B., Francis, R.I.C.C., Dunn, A., McKenzie, A., Gilbert, D.J., Smith, M.H. 2005. CASAL (C++ algorithmic stock assessment laboratory): CASAL user manual v2.07-2005/08/21. NIWA Technical Report 127. 272 p.

Cadigan, N. G., and Farrell, P. J. 2002. Generalized local influence with applications to fish stock cohort analysis. Appl. Statist. 51: 1–15.

Cadigan, N. G., and Farrell, P. J. 2004. Local Influence Diagnostics for the Retrospective Problem in Sequential Population Analysis. ICES Journal of Marine Science, 62: 256–265.

Carota, C., Parmigiani, G., and Polson, N. G. 1996. Diagnostic measures for model criticism. J. Amer. Stat. Assoc. 91: 753–762.

Constable, A.J., W.K. de la Mare, W.K., Agnew, D.J., Everson, I., and Miller, D. 2000. Managing fisheries to conserve the Antarctic marine ecosystem: practical implementation of the Convention on the Conservation of Antarctic Marine Living Resources (CCAMLR). ICES Journal of Marine Science, 57 (3): 778–791.

Cook, R. M. 1997. Stock trends in six North Sea stocks as revealed by an analysis of research vessel surveys. ICES Journal of Marine Science, 54: 924–933.

Cook, R. M. 2004. Estimation of the age-specific rate of natural mortality for Shetland sandeels. ICES Journal of Marine Science, 61: 159–164.

Darby C.D. 2004. Estimating systematic bias in the North Sea cod landings data. Working document to the ICES Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak, 7–16 September 2004

Darby, C.D. 2005. Estimating unallocated removals in the Irish Sea cod fishery. Report to 2005 RGNSDS meeting, August 2005.

de Valpine, P. 2002. Review of methods for fitting time-series models with process and observation error and likelihood calculations for nonlinear, non-Gaussian state-space models. Bull. Mar. Sci. 70: 455–471.

de Valpine, P. 2004. Monte Carlo state-space likelihoods by weighted posterior kernel density estimation. J.Am. Stat.Assoc. 99: 523–536.

de Valpine, P., and Hastings, A. 2002. Fitting population models with process noise and observation error. Ecol. Monogr. 72: 51–76.

de Valpine, P., and Hilborn. R. 2005. State-space likelihoods for nonlinear fisheries time-series. Can. J. Fish. Aquat. Sci. 62: 1937–1952.

Durbin, J., and Koopman, S. J. 2001. *Time Series Analysis by State Space Models.* Oxford University Press.

Efron, B. 1987. Better bootstrap confidence intervals. J. of the Amer. Stat. Assoc. 82: 171–200.

FLR Team. 2006. FLR: Fisheries Modelling in R. Version 1.2.1. Initial design by L. T. Kell and P. Grosjean. http://www.flr-project.org/doku.php.

Gavaris, S., and van Eeckhaute, L. 1998. Diagnosing systematic errors in reported fishery catch. *In* Proceedings of the International Symposium on Fishery Stock Assessment Models for the 21st Century, October 8–11 1997. Alaska Sea Grant College Program AK-SG-98-01.

Gavaris, S. 1999. Comparison of confidence statements for a fisheries assessment problem from two bootstrap methods. ICES ComFIE WP 1. 11 p.

Gedamke T., and Hoenig, J.M. 2006. Estimating mortality from mean length data in nonequilibrium situations, with application to the assessment of goosefish. Trans. Am. Fish. Soc. 135:476–487.

Gilks, W.R., Richardson, S., and Spiegehalter, D.J. 1996. Markov Chain Monte Carlo in Practice. Chapman and Hall: London.

Goodyear, C.P. 2003. FSIM version 3.0 User's Guide. Niceville, FL. USA.

Hansen, P. C. and O'Leary, D. P. 1993. The use of the L-curve in the regularization of discrete ill-posed problems. SIAM J. Sci. Comput. 14: 1487–1503.

Hirst, D., Aanes, S., Storvik, G., Huseby, R.B., and Tvete, I.F. 2004. Estimating catch at age from market sampling data by using a Bayesian hierarchical model. Appl. Statist., 53: 1–14.

ICES. 2002. Report of the Working Group on Methods on Fish Stock Assessments, ICES Headquarters, Copenhagen, Denmark, 3–7 December 2001. ICES CM 2002/D:01.

ICES. 2003. Report of the Working Group on Methods on Fish Stock Assessments, ICES Headquarters, Copenhagen, Denmark, 29 January – 5 February 2003. ICES CM 2003/D:03.

ICES. 2004a. Report of the Arctic Fisheries Working Group, Copenhagen 4–13 May 2004. ICES C.M. 2004/ACFM:28, 475 pp.

ICES. 2004b. Report of the Working Group on Methods on Fish Stock Assessments, Lisbon, Portugal, 11–18 February 2004. ICES CM 2004/D:03.

ICES, 2004c. Report of the ICES Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak. ICES CM 2005/ACFM:07.

ICES. 2005a. Report of the Arctic Fisheries Working Group, Murmansk 19–28 April 2005. ICES C.M. 2005/ACFM:20, 564 pp.

ICES. 2005b. Report of the *ad hoc* Group on Long-Term Advice. ICES CM 2005/ACFM:25.

ICES.2005c. Report of the ICES Review Group for the Assessment of Northern Shelf Demersal Stocks. Unpublished manuscript.

ICES.2005d. Report of the ICES Study Group on Management Strategies. ICES CM 2005/ACFM:04.

ICES. 2005e. Report of the ICES Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak. ICES CM 2006/ACFM:09.

ICES. 2006a. Report of the Arctic Fisheries Working Group, Copenhagen 19–28 April 2005. ICES C.M. 2006/ACFM:25, 594 pp.

ICES. 2006b. Report of the Working Group on Methods of Fish Stock Assessments (WGMG), 21–26 June 2006, Galway, Ireland. ICES CM 2006/RMC:07. 83 pp.

ICES. 2006c. Report of the ICES Working Group on the Assessment of Northern Shelf Demersal Stocks. ICES CM 2006/ACFM:30.

ICES. 2007. Report of the ICES Study Group on Management Strategies. ICES CM 2007/ACFM:04.

Kitagawa, G. 1996. Monte Carlo Filter and Smoother for Nonlinear Non-Gaussian State-Space Models. Journal of Computational and Graphical Statistics, 5: 1–25.

Lewy, P. *et al*. 2004. Survey gear calibration independent of spatial fish distribution. CJFAS 61: 636–647.

Methot, R.D. 2000. Technical description of the stock synthesis assessment program. NOAA Tech. Memo. NMFS-NWFSC-43: 1–46.

Meyer, R., and Millar, R. B. 1999. Bayesian stock assessment using a state-space implementation of the delay difference model. Can. J. Fish. Aquat. Sci. 56: 37–52.

Mohn, R. 1999. Simple tests of bias correction in SPAs using simulated assessment data or to BC or not to BC… ICES ComFIE WP 2. 15 p.

Murua, H., Cerviño, S., and Vázquez, A. 2006. A survey-based assessment of cod in Division 3M. NAFO SCR Doc. 06/32.

Needle, C. L. 2003. Survey-based assessments with SURBA. Working Document to the ICES Working Group on Methods of Fish Stock Assessment, Copenhagen, 29 January – 5 February 2003.

Needle, C. L. 2004a. Absolute abundance estimates and other developments in SURBA. Working Document to the ICES Working Group on Methods of Fish Stock Assessment, IPIMAR, Lisbon 10–18 Feb 2004.

Needle, C. L. 2004b. Data simulation and testing of XSA, SURBA and TSA. Working Paper to the ICES Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak, Bergen, September 2004.

Needle, C. L. 2005. SURBA 3.0. Working Paper to the EU-FISBOAT WP3 Workshop, Rhodes, Greece. 7–11 Nov 2005.

Needle, C. L. 2006a. Evaluating harvest control rules for North Sea haddock using FLR. Working Paper for the ICES Working Group on Methods of Stock Assessment, Galway, Ireland, 21–26 June 2006.

Needle, C. L. 2006b. Revised FLR-based evaluation of candidate harvest control rules for North Sea haddock. Working paper for the ICES Advisory Committee for Fisheries Management, Copenhagen, October 2006.

Pelletier, D. 1998. Intercalibration of research survey vessels in fisheries: a review and an application. Canadian Journal of Fisheries and Aquatic Sciences, 55: 2672–2690.

Porch, C.E. VPA-2BOX v3.01 User's guide. National Marine Fisheries Service, Miami, FL. Sustainable Fisheries Division Contribution SFD/2003-0004.

Porch, C.E. PRO-2BOX v2.01 User's guide. National Marine Fisheries Service, Miami, FL. Sustainable Fisheries Division Contribution SFD-02/03-182.

Porch, C. E., Eklund, A.-M., and G.P. Scott 2006. A catch-free stock assessment model with application to goliath grouper (*Epinephelus itajara*) off southern Florida. Fish. Bull. 104: 89–101.

Prager, M. H. 1994. A suite of extensions to a nonequilibrium surplus-production model. Fish. Bull. 92: 374–389.

Prager, M. H. 2005. User's Manual for ASPIC: A Stock-Production Model Incorporating Covariates (ver. 5) And Auxiliary Programs. National Marine Fisheries Service, Beaufort Laboratory Document BL–2004–01.

Punt, A. E. 2003. Extending production models to include process error in the population dynamics. Can. J. Fish. Aquat. Sci. 60: 1217–1228.

R Development Core Team (2005). R: A language and environment for statistical computing. http://www.R-project.org.

Restrepo, V.R., Patterson, K.R., Darby, C.D., Gavaris, S., Kell, L.T., Lewy, P., Mesnil, B., Punt, A.E., Cook, R.M., O´Brien, C.M., Skagen, D.W., and Stefánsson, G. 2000. Do different methods provide accurate probability statements in the short term? ICES CM 2000/V:08. 19 p.

Shepherd, J.G. 1983. Two measures of overall fishing mortality. J. Cons. Int. Explor. Mer, 41:76–80.

Sullivan, P. J., Lai, H.-L., and Gallucci, V. F. 1990. A catch-at-length analysis that incorporates a stochastic model of growth. Can. J. Fish. Aquat. Sci. 47:184–189.

Wang, G. 2007. On the latent state estimation of nonlinear population dynamics using Bayesian and non-Bayesian state-space models Ecol. Model. 200: 521–528.

Annex 1: List of Participants

NAME	ADDRESS	PHONE/FAX	EMAIL
Andrew Campbell *(part-time)*	Fisheries Science Services Marine Institute Rinville, Oranmore Co. Galway Ireland		andrew.campbell@marine.ie
Carmen Fernández	Instituto Español de Oceanografía Cabo Estai – Canido Apdo. 1552 36200 Vigo Spain	Tel: +34 986 492 111 Fax: +34 986 498 626	carmen.fernandez@vi.ieo.es
Chris Darby	Cefas Lowestoft Laboratory Pakefield Road Lowesoft Suffolk NR33 0HT UK	Tel: +44 1502 524 329 Fax: +44 1502 545111	chris.darby@cefas.co.uk
Chris Jones	NMFS-SWFSC 8604 La Jolla Shores Drive La Jolla CA 92037 USA	Tel: +1 858 546 5605 Fax: +1 858 546 5608	chris.d.jones@noaa.gov
Chris Legault	NMFS-NEFSC 166 Water Street Woods Hole MA 02543 USA	Tel: +1 508 495 2025 Fax: +1 508 495 2393	chris.legault@noaa.gov
Coby Needle *(Chair)*	FRS Marine Laboratory PO Box 101 375 Victoria Road Aberdeen AB11 9DB Scotland	Tel: +44 1224 295456 Fax: +44 1224 295511	needlec@marlab.ac.uk
Colin Millar	FRS Marine Laboratory PO Box 101 375 Victoria Road Aberdeen AB11 9DB Scotland	Tel: +44 1224 295575 Fax: +44 1224 295511	millarc@marlab.ac.uk
David Orr	NWAFC Dept. of Fisheries and Oceans Canada 80 White Hills Road East AIC 5X1 St John's NL Canada	Tel: +1 709 772 7343 Fax: +1 709 772 4105	orrd@dfo-mpo.gc.ca
Liz Brooks	NMFS-SEFSC 75 Virginia Beach Drive Miami FL 33149 USA	Tel: +1 305 361 4243 Fax: +1 305 361 4295	liz.brooks@noaa.gov
José De Oliveira	Cefas Lowestoft Laboratory Pakefield Road Lowesoft Suffolk NR33 0HT UK	Tel: +44 1502 527 727 Fax: +44 1502 545111	jose.deoliveira@cefas.co.uk
Noel Cadigan	NWAFC Dept. of Fisheries and Oceans Canada 80 White Hills Road East AIC 5X1 St John's NL Canada	Tel: +1 709 772 5028 Fax: +1 709 772 4105	cadigann@dfo-mpo.gc.ca
Sam Subbey	Institute of Marine Research PO Box 1870 N-5817 Bergen Norway	Tel: +47 55235383 Fax: +47 55238687	samuels@imr.no
Tim Miller	NMFS-NEFSC 166 Water Street Woods Hole MA 02543 USA	Tel: +1 508 495 2365 Fax: +1 508 495 2393	timothy.j.miller@noaa.gov

NAME	ADDRESS	PHONE/FAX	EMAIL
Yuri Kovalev	Polar Research Institute of Marine Fisheries and Oceanography (PINRO) 6 Knipovich Street 183763 Murmansk Russia	Tel: +7 8152 472469 Fax: +7 8152 473331	kovalev@pinro.ru

Several other NFMS (USA) colleagues were also involved in the meeting. Richard Methot, Alan Seaver and Dvora Hart gave presentations (respectively WP 17, WP 6 and WP 11), while Gary Shepherd, Paul Rago, Loretta O'Brien, Laurel Col, and Anne Richards attended parts of the meeting as observers.

Annex 2: Terms of Reference for next meeting

The **Working Group on Methods of Fish Stock Assessments** [WGMG] will be renamed as the **Methods Working Group (WGM)** (Chair: Coby Needle, UK) and will meet in Nantes, France, from 15–24 April 2008 (TBD) to:

a) Develop methods for evaluating harvest control rules appropriate for biased stock assessments;

b) Develop methods for detecting and correcting retrospective bias and noise using simulations that are typical of real data;

c) Develop methods for the appropriate incorporation and estimation of uncertainty in stock assessment methods;

d) Investigate the biases inherent in assuming length-based processes are age-based, and develop approaches to reduce such biases;

e) Examine the ability of distributions that accept zero observations for use in relating observed and predicted indices in stock assessment models.

WGM will report by 13 May 2008 for the attention of the Resource Management Committee and ACFM.

Supporting Information

PRIORITY:	The work of this group is essential for ICES to progress in the development of methods for fish stock assessment and for the evaluation of management strategies.
SCIENTIFIC JUSTIFICATION AND RELATION TO ACTION PLAN:	ToR a): The relevance of this work to WGM in particular lies in the development and testing of methods and tools to allow for the appropriate evaluation of management plans and strategies. It is not the function of scientists to propose or advocate management plans or targets – that is for managers themselves, along with stakeholders and the wider society. It is, however, incumbent on scientists to advise managers on the likely consequences of different management actions, and to assist managers in developing plans that have the best possible likelihood of achieving whatever it is the managers want to achieve – this, again, is not for scientists to decide. WGM should focus on the methodological aspects of this process. Specifically, WGM should neither propose management plans nor conclude whether this or that plan is more likely to succeed – rather, WGMG should test and develop methods that enable scientists to answer the questions asked of them by managers. One example of such a question would be: how would this or that management plan function in a situation where the underlying stock assessment was biased (retrospectively or otherwise)? This ToR is intended to address this question. ToR b): Considerable progress was made during the 2007 meeting in collating a suite of analysis tools that could assist in the detection (and possible correction) of causes of retrospective bias. However, testing of these tools during the meeting was carried out on simulated datasets with effects that were both simple and substantial – for example, a tripling of survey catchability combined with no other sources of bias. In reality, a number of sources of bias are likely to exist. The next stage of this analysis is therefore to generate simulated data which contains features that would be more typical of data encountered in reality, and explore the ability of

	the tools to detect and correct bias in those situations. ToR c): This ToR is intentionally rather wider than the others, because there are two main issues to be considered. Firstly, much assessment data is now available with estimates of variance, and it is important that WGM advises on the best way(s) to incorporate this variance. Secondly, the group should give advice on how to estimate and display variance in output parameters such as *F* and spawning-stock biomass (this was the area that the 2007 meeting focussed on). Both of these are important and need to be progressed. ToR d): The point has been raised within ICES about the development of methods to deal with data-poor situations. One example of such a situation is where age data is either unavailable or dubious to a greater or lesser degree. The purpose of length- or size-based assessment methods is to allow appropriate management decisions in this situation. However, length-based methods are not straightforward to implement and use, so the first question to be answered must be: is the effort involved in setting up a length-based method worthwhile? In other words, what are the biases that are introduced by assuming a process is age- rather than length-based? This is an important point which WGM should be in a position to address, and which other length-based groups have skipped over. ToR e): It is the view of WGM that the meeting should not become a repository for all methodological issues raised by assessment WGs, as this leads to a proliferation of special requests that WGM will often not be in a position to address. The issue of zero observations in surveys is somewhat different as a participant has already agreed to do the work to address it, so WGM felt the inclusion of this ToR would be an acceptable exception to the general rule. The longer meeting held in 2007 allowed for considerably more collaborative work than had been the case in 2006, and this length should be retained for the 2008 meeting.
RESOURCE REQUIREMENTS:	None.
PARTICIPANTS:	The Group is well-manned by the correct people but would benefit from new members, particularly in respect of ToR d).
SECRETARIAT FACILITIES:	Meeting facilities, production of report.
FINANCIAL:	None.
LINKAGES TO ADVISORY COMMITTEES:	ACFM has strongly supported the work of this group and has worked actively in formulating the ToRs for recent meetings. WGM will report to ACFM at its autumn meeting in 2008.
LINKAGES TO OTHER COMMITTEES OR GROUPS:	WGMG will report to the Resource Management Committee at the ICES ASC in 2008. There will also be links to the Fisheries Technology Committee.
LINKAGES TO OTHER ORGANIZATIONS:	There is similar work going on within ICCAT and NAFO. Coordination should be assured.

Annex 3: Recommendations

The main recommendations from the 2007 meeting of WGMG are in the Conclusions (Section 7).

The following table summarises the recommendations arising from the 2007 meeting of WGMG. The intended target(s) of the recommendations are listed, along with actions that need to be taken by WGMG (renamed as WGM) itself.

RECOMMENDATION	ACTION
1. Any management-strategy evaluation toolbox must allow for assessments to be run "live" as part of the evaluation loop.	**Recommendation to:** all EGs and Advisory Committees performing management-strategy evaluations. **WGM action:** Ensure that such tools are appropriately tested and made available to the ICES community.
2. Not accounting for process errors can lead to a biased view of the true uncertainty in stock estimates based on approximate populations models. If process errors are not accounted for, then uncertainty estimates are likely to be too small and should be presented with this caveat.	**Recommendation to:** assessment WGs and ACFM. **WGM action:** Work towards the provision of methods that account for process error.
3. Mis-specification diagnostics (local influence diagnostics, catch-estimation methods, pre-screening tests) can fairly reliably detect the timing and direction of assessment model problems when the source is known (e.g. M or survey catchability), and may be used in such circumstances. They should be treated far more cautiously when the source of error is not known.	**Recommendation to:** assessment WGs and ACFM. **WGM action:** Continue work on mis-specification diagnostics to improve power of tests.
4. Sensitivity of management advice to the method used to calculate average fishing mortality needs to be evaluated on a case-by-case basis, as there is no general rule that can be determined.	**Recommendation to:** assessment WGs and ACFM.

Annex 4: Program code and scripts

R code for data simulation

```
SP.pop.sim.fn <- function(r=1.3, K=1e8, E.B0=0.5e8, Fy=rep(0.4,20), M=0.2, CV.B0=0.3,
CV.By=0, CV.Fy =0.05, CV.Cy = 0, Rel.Bias.Cy = 0,
  qI.1 = exp(-7), qI.2 = exp(-7), CV.qI = 0.3){

  nfishyr <- length(Fy)

  #variance of log(By) and log(Cy)
  v.log.B0 <- log(1 + CV.B0^2)
  v.log.By <- log(1 + CV.By^2)
  v.log.Cy <- log(1 + CV.Cy^2)

  log.By <- rnorm(1, log(E.B0), sqrt(v.log.B0))
  E.Cy <- exp(log.By) * Fy[1] * (1-exp(-(Fy[1] + M)))/(Fy[1] + M)
  log.Cy <- rnorm(1, log(E.Cy), sqrt(v.log.Cy))
  for(i in 2:nfishyr){
    E.By <- exp(log.By[i-1]) + r * exp(log.By[i-1]) * (1 - exp(log.By[i-1])/K) -
exp(log.Cy[i-1])
    log.By[i] <- rnorm(1, log(E.By), sqrt(v.log.By))
    E.Cy <- exp(log.By[i]) * Fy[i] * (1-exp(-(Fy[i] + M)))/(Fy[i] + M)
    log.Cy[i] <- rnorm(1, log(E.Cy), sqrt(v.log.Cy))
  }
  log.Cy.g <- log.Cy
  log.Cy.b <- log.Cy.g + log(Rel.Bias.Cy + 1)
  v.log.qI <- log(1 + CV.qI^2)
  Index.g <- rnorm(length(log.By), log(qI.1)  + log.By, sqrt(v.log.qI))
  len.I1 <- round(0.5 * length(log.By))
  len.I2 <- length(log.By) - len.I1

  return(list(log.Cy.g= log.Cy.g, log.Cy.b = log.Cy.b, Index.g=Index.g, log.By=log.By,
Fy=Fy))
}
informative <- SP.pop.sim.fn(r=0.1, K=1e8, E.B0=1e8, Fy=informative.Fy, M=0.2,
CV.B0=0.3, CV.By=0, CV.Cy = 0,
  Rel.Bias.Cy = -0.2, qI.1 = exp(-7), qI.2 = exp(-9), CV.qI = 0.3)

oneway <- SP.pop.sim.fn(r=0.7, K=1e8, E.B0=1e8, Fy=oneway.Fy, M=0.2, CV.B0=0.3,
CV.By=0, CV.Cy = 0,
  Rel.Bias.Cy = -0.2, qI.1 = exp(-7), qI.2 = exp(-9), CV.qI = 0.3)
```

R code for Bayesian model without process error

```
##############################################################################
#surplus production model for WMG 2007 with lognormal errors
#No process error - code by C.P.Millar, FRS Marine Laboratory. 19 March 2007
##############################################################################

library(MASS)

setwd("C:\\Work\\ICES\\WGMG07\\BRugs\\Surplus_production_MCMC")
```

```
############################
## the likelihood and prior - the model
############################

# the biomass function
Bn.rec <- function(n, N, r, K, C, Bn=NULL) {
  if (n==0) return(Bn)
  if (is.null(Bn)) {
    if (N!=n) stop("N must be equal to n")
    Bn <- K
  } else {
    Bn <- Bn + r*Bn*(1-Bn/K) - C[N-n]
  }
  Bn <- Bn.rec(n-1,N,r,K,C,Bn)
  return(Bn)
}

# the likelihoods
f1 <- function(logI1, logq, logK, tau.I) dnorm(logI1, logq + logK, tau.I)
fn <- function(n,logIn, logq, logK, logr, tau.I,C) {
  Bn <- Bn.rec(n,n,exp(logr),exp(logK),C)
  if (Bn>0) {
    logBn <- log(Bn)
    out <- dnorm(logIn, logq + logBn, tau.I)
  } else {
    out <- 0
  }
  return(out)
}
Lik <- function(logI, logq, logK, logr, tau.I, C) {
  out <- f1(logI[1], logq, logK, tau.I)
  for (i in 2:length(logI)) out <- out*fn(i,logI[i], logq, logK, logr, tau.I, C)
  return(out)
}

## set up priors
mu.logq <- -7.5; cv.q <- 5
mu.logK <- 18.5; cv.K <- 5
mu.logr <- -0.35; cv.r <- 5
cv.I <- 0.15; cv.tau.I <- 1.3

p.logq <- function(logq, mu=mu.logq   , cv=cv.q) dnorm(logq, mu, sqrt(log(1+cv*cv)))
p.logK <- function(logK, mu=mu.logK , cv=cv.K) dnorm(logK, mu, sqrt(log(1+cv*cv)))
p.logr <- function(logr, mu=mu.logr, cv=cv.r) dnorm(logr, mu, sqrt(log(1+cv*cv)))
p.tau.I <- function(tau.I, cv=cv.I, cv.var=cv.tau.I) {
  s1.I <- 1/(cv.tau.I*cv.tau.I)
  s2.I <- s1.I*log(1 + cv.I*cv.I)
  dgamma(1/(tau.I^2), s1.I, s2.I)
}
prior <- function(logq, logK, logr, tau.I)
p.logq(logq)*p.logK(logK)*p.logr(logr)*p.tau.I(tau.I)
```

```
# set up startin value generation functions
gen.logq <- function(mu=mu.logq, cv=cv.q) rnorm(1,mu, sqrt(log(1+cv*cv)))
gen.logK <- function(mu=mu.logK, cv=cv.K) rnorm(1,mu, sqrt(log(1+cv*cv)))
gen.logr <- function(mu=mu.logr, cv=cv.r) rnorm(1,mu, sqrt(log(1+cv*cv)))
gen.tau.I <- function(cv=cv.I, cv.var=cv.tau.I) {
  s1.I <- 1/(cv.var*cv.var)
  s2.I <- s1.I*log(1 + cv*cv)
  1/sqrt(rgamma(1, s1.I, s2.I))
}

###################
## get some data
###################

# choose data set to use
typ <- c("Informative","nonInformative")[1]
set <- 4 # choice here is 1:4

#data
if (typ=="Informative") {
  source("C:\\Work\\ICES\\WGMG07\\BRugs\\Surplus_production_MCMC\\informative")
} else {
  source("C:\\Work\\ICES\\WGMG07\\BRugs\\Surplus_production_MCMC\\oneway"); y <- x
}
if (set==1) wk <- list(C = exp(y$log.Cy.g), logI = y$Index.g, true.B = exp(y$log.By))
if (set==2) wk <- list(C = exp(y$log.Cy.g), logI = y$Index.g-
c(rep(0,10),rep(log(3),10)), true.B = exp(y$log.By))
if (set==3) wk <- list(C = exp(y$log.Cy.b), logI = y$Index.g, true.B = exp(y$log.By))
if (set==4) wk <- list(C = exp(y$log.Cy.b), logI = y$Index.g-
c(rep(0,10),rep(log(3),10)), true.B = exp(y$log.By))

attach(wk)

######################
## MCMC prep and loop
######################

# set number of MCMC iterations
MCMC.n <- 10000

# get starting viable values
logq <- logK <- logr <- tau.I <- rep(NA,MCMC.n)
while (T) {
  logq[1] <- gen.logq()
  logK[1] <- gen.logK()
  logr[1] <- gen.logr()
  tau.I[1] <- gen.tau.I()

  if (Lik(wk$logI, logq[1], logK[1], logr[1], tau.I[1],wk$C)*
      prior(logq[1], logK[1], logr[1], tau.I[1])>0)
    break
}
```

```
# MCMC tuning parameters – mixing could still be improved to reduce iteration number
if (set %in% c(1,3)) {
  rw.logq <- 0.2
  rw.logK <- 0.1
  rw.logr <- 0.1
  rho.logK.logr <- -0.6 # correlation in logK, logr random walk (rw)
  mu.tau.I <- .4; sd.tau.I <- .5
} else if (set %in% c(2,4)) {
  rw.logq <- 0.2
  rw.logK <- 0.08
  rw.logr <- 0.09
  rho.logK.logr <- -0.7 # correlation in logK, logr random walk (rw)
  mu.tau.I <- .4; sd.tau.I <- .5
}

#set up acceptance probabilities
U <- matrix(runif(MCMC.n*2),ncol=2, nrow=MCMC.n)
accpt <- rep(0,3) # acceptance probilities
Nburn <- 1000     # burn in length

cat(paste("\nbegining mcmc for
no.process.error",typ,c("CgIg","CgIb","CbIg","CbIb")[set],sep="."),"\n\n")

for (i in 2:MCMC.n) {
if (round(mcmc/100,0)*100==i) { # update screen output
  flush.console()
  cat("\rmcmc at iteration number : ",i)
}

## update logq - a normal random walk
  logq.star <- rnorm(1, logq[i-1], rw.logq)
  ln.frac <- log(Lik(logI, logq.star, logK[i-1], logr[i-1], tau.I[i-1], C)) -
               log(Lik(logI, logq[i-1], logK[i-1], logr[i-1], tau.I[i-1], C)) +
               log(prior(logq.star,logK[i-1],logr[i-1],tau.I[i-1])) -
               log(prior(logq[i-1],logK[i-1],logr[i-1],tau.I[i-1]))
  if (is.nan(ln.frac)) ln.frac <- 1
  frac <- if (ln.frac<1) exp(ln.frac) else 2.7
  MH <- min(frac, 1)
  if (U[i-1,1] < MH) {
    if(i>(Nburn+1)) accpt[1] <- accpt[1] + 1#
        logq[i] <- logq.star
  } else {
    logq[i] <- logq[i-1]
  }
## update logK and logr - a bivariate normal random walk
  sigma <- matrix(c(rw.logK^2                    , rho.logK.logr*rw.logK*rw.logr,
                    rho.logK.logr*rw.logK*rw.logr, rw.logr^2),nrow=2)
  logK.logr <- mvrnorm(1, c(logK[i-1],logr[i-1]),sigma)
  logK.star <- logK.logr[1]
  logr.star <- logK.logr[2]
  ln.frac <- log(Lik(logI, logq[i], logK.star, logr.star, tau.I[i-1],C)) -
               log(Lik(logI, logq[i], logK[i-1], logr[i-1], tau.I[i-1],C)) +
               log(prior(logq[i],logK.star,logr.star,tau.I[i-1])) -
```

```
              log(prior(logq[i],logK[i-1],logr[i-1],tau.I[i-1]))
  if (is.nan(ln.frac)) ln.frac <- 1
  frac <- if (ln.frac<1) exp(ln.frac) else 2.7
  MH <- min(frac, 1)
  if (U[i-1,1] < MH) {
    if(i>(Nburn+1)) accpt[2] <- accpt[2] + 1
    logK[i] <- logK.star
    logr[i] <- logr.star
  } else {
    logK[i] <- logK[i-1]
    logr[i] <- logr[i-1]
  }
## update tau.I - an independance sampler
  tau.I.star <- rlnorm(1,log(mu.tau.I),sd.tau.I)
  ln.frac <- log(Lik(logI, logq[i], logK[i], logr[i], tau.I.star,C)) -
              log(Lik(logI, logq[i], logK[i], logr[i], tau.I[i-1],C)) +
              log(prior(logq[i],logK[i],logr[i],tau.I.star)) -
              log(prior(logq[i],logK[i],logr[i],tau.I[i-1])) +
              log(dlnorm(tau.I[i-1],log(mu.tau.I),sd.tau.I)) -
              log(dlnorm(tau.I.star,log(mu.tau.I),sd.tau.I))
  if (is.nan(ln.frac)) ln.frac <- 1
  frac <- if (ln.frac<1) exp(ln.frac) else 2.7
  MH <- min(frac, 1)
  if (U[i-1,2] < MH) {
    if(i>(Nburn+1)) accpt[3] <- accpt[3] + 1
    tau.I[i] <- tau.I.star
  } else {
    tau.I[i] <- tau.I[i-1]
  }
} # end mcmc loop

detach("wk")

#########################
## diagnostics and plots
#########################

# print acceptance probabilities
cat("\nmcmc finished with acceptance probabilities: ",accpt/(MCMC.n-Nburn),"\n)

# values to remove
burn <- 1:Nburn

## check independence sampler
windows()
par(mfrow=c(1,1))
plot(density(tau.I[-nburn]), main="proposal(red) and posterior(black) for tau")
lines(x<-1:500/100, dlnorm(x,log(mu.tau.I),sd.tau.I), col="red")


# plot mcmc chains
windows()
par(mfrow=c(2,2))
```

```
plot(as.ts(logq[-burn]),main="log q = -7")
plot(as.ts(logK[-burn]),main="log K = 18.4")
plot(as.ts(logr[-burn]),main="log r = 0.3?")
plot(as.ts(tau.I[-burn]),main="tau.I = ...")

# plot bivariate posterior marginal distributions
windows()
par(mfrow=c(1,1))
pairs(cbind(logq,logK,logr,tau.I)[-burn,])

## calculate Biomasses
B <- matrix(NA,nrow=MCMC.n-Nburn,ncol=length(wk$logI))
for (i in (Nburn+1):MCMC.n) {
  for (j in 1:length(wk$logI)) B[i-Nburn,j] <- Bn.rec(j, j, exp(logr[i]),exp(logK[i]),
wk$C)
}

## calculate posterior summaries
m <- apply(B,2,quantile,0.5)
l <- apply(B,2,quantile,0.025)
u <- apply(B,2,quantile,0.975)

#windows()
par(mfrow=c(1,1))
plot(0,0,xlim=c(1,20),ylim=c(0,max(m,l,u,wk$true.B)))
lines(1:20,m)
lines(1:20,u,lty=2)
lines(1:20,l,lty=2)
lines(1:20, wk$true.B, col="red", lwd=2)
```

WinBUGS code for Bayesian model with process error

```
#############################################################################
#State-space surplus production model with lognormal process errors
#############################################################################
model
{
######OBSERVATION EQUATION FOR INDEX#####

for (i in 1:N) { mu.logI[i] <- logq + logK + log(P[i])
                      logI[i] ~ dnorm(mu.logI[i], tau.I) }

s1.I <- 1/(cv.tau.I*cv.tau.I)
s2.I <- s1.I*log(1 + cv.I*cv.I)
tau.I ~ dgamma(s1.I, s2.I)

## PRIORS ON q and K:

tau.q <- 1/log(1 + cv.q*cv.q)
logq ~ dnorm(mu.logq, tau.q)
q <- exp(logq)

tau.K <- 1/log(1+(cv.K*cv.K))
```

```
logK ~ dnorm(mu.logK, tau.K)
K <- exp(logK)

## PRIOR ON P[1]=B1/K ##

P[1] <- 1

## BIOMASS DYNAMICS on P[i]=B[i]/K ##

s1.Pproc <- 1/(cv.tau.Pproc*cv.tau.Pproc)
s2.Pproc <- s1.Pproc*log(1 + cv.Pproc*cv.Pproc)
tau.Pproc ~ dgamma(s1.Pproc, s2.Pproc)

for(i in 2:N) {
    mu.logP[i-1] <- log( max(P[i-1] + r*P[i-1]*(1-P[i-1]) - C[i-1]/K, 0.01) )
    logP[i] ~ dnorm(mu.logP[i-1], tau.Pproc)
    P[i] <- exp(logP[i])
    }

## PRIOR ON r ##

tau.r <- 1/log(1 + cv.r*cv.r)
logr ~ dnorm(mu.logr, tau.r)
r <- exp(logr)

#Parameters to monitor
Pars[1]<-K; Pars[2]<-r; Pars[3]<-q; Pars[4]<-sqrt(exp(1/tau.I)-1); Pars[5]<-
sqrt(exp(1/tau.Pproc)-1)
}


# DATASET “InformativeCgIg” (Informative biomass trend, Catch good,
# Index good)
list(N=20,
C=c(5482339,7789912,37049952,28804728,14027489,37669632,21792852,
12169920,11811891,2680615,
3836698,3163140,3804961,8965007,4136276,12161488,14022205,17494667,24793466,33467161),
logI=c(11.56734,11.47647,10.85076,10.74642,10.78815,10.27104,10.32704,
9.936434,9.473897,9.30825,
10.59037,10.24365,10.87514,10.98196,10.51518,11.34332,11.22044,
10.78851,11.10594,10.69974),
cv.I=0.15, cv.tau.I=1.3,
mu.logq=-7, cv.q=5,
mu.logK=18.5, cv.K=5,
cv.Pproc=0.2, cv.tau.Pproc=1,
mu.logr=-0.35, cv.r=5
)
```