CCSBT-MP/0404/04

CSIRO
MARINE RESEARCH

Comparison of the performance of tuned candidate management procedures for southern bluefin tuna based on the final trial specifications and testing procedures

Tom Polacheck
Paige Eveson
Jason Hartog
Marinelle Basson
Dale Kolody

Prepared for the CCSBT MP3WS
19-24 April 2004, Busan, Korea

Comparison of the performance of tuned candidate management procedures for southern bluefin tuna based on the final trial specifications and testing procedures.

T. Polacheck, P. Eveson, J. Hartog, M. Basson, and D. Kolody

Abstract

This document presents a comparison of the performance of a number of possible candidate management procedures (CMPs) for southern bluefin tuna based on the final trial specifications and testing procedures agreed to at the 2003 Stock Assessment Group (SAG) meeting (Anon 2003). In addition, as requested at the 2003 SAG meeting, three decision rules were selected (based on our judgement of “best” overall performance) for consideration by the April 2004 CCSBT Management Procedure Workshop in its final selection of rules to present to the Commission.

The comparison and selection of CMPs is complex because of (1) the large number of performance indicators that can be considered, (2) potential interactions between tuning levels and relative performance, (3) maximization of “average” performance versus minimization of risk (i.e. low probability of really poor performance), (4) short and long term trade-offs in performance and (5) the need to consider the robustness of a rule's performance. Results indicated that for high tuning levels, all CMPs had similar performance, while the greatest potential differences in performance were seen at lower tuning levels. Some differences in the relative performance of CMPs were found between tuning levels of 0.7 and 1.1. All rules tuned to level 0.7 resulted in a greater than 50% probability that the spawning stock would decline between 2022 and 2032 in addition to the 50% probability that the spawning stock would decline by 30% between 2002 and 2022, and thus showed little or no prospect for rebuilding. As such, in our evaluation and selection, the main focus was given to comparing performance for rules tuned to level 1.1, given the long standing objective of the Commission to rebuild the spawning stock. The main difference in performance among the three rules selected was in the trade-off they made in shorter and longer term catches. Their performance in terms of spawning biomass were relatively similar.

For the selected CMPs tuned to level 1.1, results are presented that compare performance across the three different intervals that TAC changes are permitted (1, 3 and 5 years). In addition, in the case of a 3-year TAC interval, results are presented to compare performance when the first TAC change is implemented in 2006 versus 2008.

Introduction

This paper compares the performance of a number of possible candidate management procedures (CMPs) for southern bluefin tuna based on the final trial specifications and testing procedures agreed to at the 2003 Stock Assessment Group (SAG) meeting (Anon 2003). The results from the testing of CMPs based on these final trial specifications are to form the basis upon which the Third CCSBT Management Workshop will select a reduced set of preferred CMPs to present to the Commission for their consideration.

As part of the testing procedure, it was agreed that CMPs would be “tuned” to three different levels for the median of the ratio B2022/B2002 (0.7, 1.1 and 1.5). The 2003 SAG considered that it was essential for CMPs to be tuned to a common biomass level in order to make meaningful comparisons of their relative performance in terms of their catch/biomass trade-offs. The need to compare CMPs at different tuning levels was also considered important to examine whether consistency existed among

CMPs in their relative performance at different tuning levels (e.g. does the CMP judged to provide the "best" trade-off between catch and biomass at one tuning level also provide the best performance at other tuning levels?). All results presented here are based on comparison of CMPs tuned to the three agreed tuning levels.

In order to facilitate the selection process at the Workshop, CMP developers were requested to come up with a small number of "best" CMPs for final evaluation and selection at the Workshop. However, it was recognized that strong guidance from the Commission about the relative importance of different aspects of performance was unlikely and that developers would have to use their judgement in selecting "best" CMPs based on limited general guidance from the Commission (plus feedback from the national consultation processes with stakeholders and managers).

This paper first presents some general comparisons from a larger number of CMPs that were tested and examined by authors. These provide an indication of the range and limits of performance in terms of catch/biomass trade-offs that were achieved with the CMPs that we developed. Based on these comparisons, we selected three CMPs for consideration by the Workshop in its final selection process. The three CMPs selected were chosen based on having "reasonable" general overall performance and then based on differences with respect to specific trade-offs in shorter and longer term catch and/or biomass performance.

In addition to the basic algorithm within any CMP that determines the next TAC, CMPs being tested also contain specifications which determine:

1. the frequency with which the TAC is allowed to be adjusted;
2. the first year in which a TAC change is allowed and
3. the maximum amount by which the TAC can change.

As part of the agreed testing specifications, CMPs were to be tested with TAC frequency changes of 1, 3 and 5 years. The default year for the first TAC change was 2006 in the case of annual changes and 2008 in the case of 3 and 5 year changes. In addition, the SAG agreed on defaults for the maximum amount the TAC could change (3,000t in the case of annual changes, 5,000t in the case of 3-year changes, and 8,000t in the case of 5-year changes[1]). In the present paper, all our comparisons among CMPs are based on these default maximum TAC changes.

Methods

Results for a range of decisions rules are presented. A decision rule is defined here as a basic algorithm that can be used to determine the TAC in the next year given available information (e.g. past catches, CPUE trends, etc.). All of the decision rules considered here have "control" parameters, one of which was used as the parameter that was varied to achieve a specified tuning level. We defined a CMP as a decision rule with all of its control parameters specified, except the one that was varied for tuning. While a range of control parameters was explored for the different decision rules, in most cases we only present results for one version of each decision rule (i.e.

---

[1] TACs are applied on an annual basis and thus the maximum change is defined to be the maximum amount the annual TAC could change and not the maximum change in the sum of the TAC over the period between changes. Thus, in the case of 3-year intervals between TAC changes, a 5000t change means that the maximum increase or decrease would be 5000t in the year that the TAC changed and then the TAC would be maintained at this increased or decreased level for 3 years.

all of the control parameters were fixed except the one for tuning). Altogether, comparisons are presented for nine different CMPs. However for the CMPs named BRU_10 and BRU_15 results are only presented for the 1.1 tuning level. Time did not permit tuning these two CMPs to other levels and, furthermore, these CMPs were not considered to provide the "best" performance at the 1.1 level. Detailed descriptions of all CMPs presented in this document are contained in Appendix 1.

The graphical summaries and associated software initially developed by Eveson and Ricard (2003), modified by Eveson (2003) and further modified for the final trial by Eveson (and distributed through e-mail) were used to evaluate the performance of different CMPs as the tuning level was varied and to compare results between decision rules.

General Considerations in the Selection of a Management Procedure

The final selection of a management procedure can be broken up into a number of specific decisions:

(1) Deciding on a decision rule,
(2) Deciding on the frequency of TAC changes
(3) Deciding on the first year a TAC change will be implemented,
(4) Deciding on the maximum amount the TAC can vary,
(5) Deciding on a tuning level,
(6) Deciding on meta-rules ("exceptional" circumstances).

The final management procedure will need to encompass all of the above elements and its overall performance will be dependent upon the full suite of decisions adopted. Nevertheless, it is possible to address each of these questions in a sequential approach in the order given above, with the understanding that some iterative reconsideration may be required in light of decisions made on subsequent issues. For example, within limits, we have found the general performance of a decision rule is independent of decisions about the frequency of TAC changes, the maximum change permitted (as long as maximum change and frequency are appropriately compensated), and the first year a TAC change is permitted. However, there are potential interactions between the choice of decision rule and tuning levels in terms of both the relative performance of different decision rules and the ability of a CMP to meet a specific objective (as measured by the suite of performance indicators). Thus, for example, all CMPs tuned to the same level will have equal performance in terms of the median spawning biomass in 2022 but will vary with regard to other measures of biomass performance. In this regard, it is important that a tuning level is not confounded as defining a biomass objective as it does not provide any indication of "risk" or shorter and longer term performances.

The simulation testing of CMPs provides estimates of the probability of future stock sizes and catches. In some cases, the estimates from the reference set of operating model scenarios differ substantially from those obtained from the robustness trials. This raises questions about how to consolidate the evaluation and advice across the reference and robustness trials. This is particularly problematical in cases where a "high weight or plausibility" was given to a robustness scenario. This will be particularly important to consider when presenting results to the Commission. This is because in presenting any meaningful advice that would allow the Commission to decide on a tuning level, the results presented need to be considered reasonable estimates of the actual likely performance of a management procedure in order to be able to judge the risks of adopting a particular tuning level in terms of both the stock and the fishery.

To some extent, in advising on the selection of a CMP, only the relative differences in the estimated probabilities need to be considered since the objective is to find the CMP with the “best” overall performance. Relative performance in some cases can vary depending upon the tuning level, which means that the choice of a CMP is potentially tuning level dependent. However, relative performance can even vary between the reference and robustness scenarios. In such cases, the magnitude of the difference as well as the relative rank can be important (e.g. little importance may be given to the relative rank in the reference set if the absolute differences in a performance indicator are relatively small compared to differences seen in a robustness test). In addition, even when considering trade-offs in performance within the reference set at a given tuning level, absolute differences in the magnitude of performance indicators between CMPs can be important, not just relative rankings.

In considering the results, it must be recognized that the risks associated with low stock sizes are poorly estimated. This is because at low stock sizes (1) recruitment dynamics are being extrapolated beyond the range previously observed (e.g. no allowance for the probability a recruitment collapse) and (2) it may not be commercially viable to take the full TAC. This potential underestimation is particularly of concern for CMPs and/or tuning levels that indicate a reasonable probability of spawning stock declining substantially below its current level during the 30-year time period being tested. However, it should be noted that even for the highest tuning level, the results suggest that there is a reasonable probability (e.g. >50%) that the spawning stock will decrease below its current (2002) level by a significant amount (see below) at some point during the 30-year period.

The current paper focuses on the selection of a decision rule, recognizing that the selection of tuning level is a matter for the Commission to decide based on their competing biomass/catch and catch stability objectives. Even so, the selection of a CMP is complex because

(1) there are a large number of performance indicators that can be considered,
(2) there are potential interactions between tuning levels and relative performance,
(3) improved performance in conservation objectives will have negative consequences in catch objectives,
(4) maximization of “average” performance can compete with minimization of risk (i.e. ensuring low probability of really poor performance), and
(5) there are trade-offs between short and long term performance.

Evaluating and Comparing CMPs

In the comparison and selection of CMPs, we think it is important to consider how well any performance measure does in maximizing the probability of desirable outcomes and minimizing the probability (or “risk”) of undesirable ones – i.e. an informal “mini-max” approach. In this regard, we have found that plots overlaying cumulative probabilities for various performance statistics to be useful. Figure 1 provides an example of such a plot with the cumulative probabilities for the spawning stock in 2022 relative to 2002 (B2022/B2002) for nine CMPs overlaid (results from the reference operating model). Note that all the CMPs in this figure have been tuned to 1.1. CMPs with a low probability of being below a small value of B2022/B2002 would be considered less "risky". For example, CON_01 (the constant catch rule) and BRU_15 both yield a much higher probability that the spawning stock biomass in 2022 will be below 50% of the 2002 level (i.e. B2022/B2002<0.5) than the other CMPs, indicating that all of these rules have better performance in terms of risk when measured by the B2022/B2002 performance indicator. Similarly, in terms of this example, a desirable outcome would be for a CMP to have a low probability of being below a target level of B2022/B2002 (i.e. a high probability of being above the target). Thus, if the tuning level is taken to be the "rebuilding" target, then all CMPs will have the same probability (i.e. 50%) of achieving that target; therefore CMPs considered less risky (e.g. those further to the right on the graph at low values of B2022/B2002, say values less than 1) would be judged to have "better" performance. However, if the rebuilding target is above the tuning level, then CMPs that maximize the chance of achieving the rebuilding target are often not the least risky.

Relative Performance Across Tuning Levels

Figures 1 and 2 compare cumulative probability curves for the spawning stock in 2022 relative to 2002 for three different tuning levels. Evident in these figures is that as the tuning level increases the relative performance across all CMPs becomes similar. This is because as the tuning level increases, there is little scope for a CMP to do anything but to cut catches by the maximum permissible amount to achieve the tuning level. In fact a tuning level of ~1.7 is the maximum possible that can be achieved since this is the median value for the B2022/B2002 statistic under no catches. As such, in term of CMP selection, results for higher tuning levels (e.g. 1.5) are essentially irrelevant since all CMPs have similar performance. As the tuning level decreases, the differences found among CMPs increase.

However, depending on the performance statistic and time frame, the relative performance (or rank) of CMPs can vary substantially between the 0.7 and 1.1 tuning levels. For example, we found with the CMPs we tested that in terms of median performance of B2032/B2022, the CMPs with the best performance when tuned to 1.1 had the worst performance when tuned to 0.7, and similarly for the median average 30-year catch (Figure 3). This complicates both the evaluation and selection process.

Comparison of Biomass and Catch Trends

Figure 4 (a-c) compares the median, 10th and 90th percentiles of the time trajectories for spawning biomass (relative to 2002) for a range of CMPs for each of the three tuning levels. Similarly, Figure 5 compares the median, 10th and 90th percentiles of the time trajectories for catch. In terms of the median spawning biomass trajectories, CMPs tend to have very similar performance in the short to medium term (i.e. to ~2022), which is due in part to all the trajectories having to go through the same point in 2022 combined with the relatively long time lags for recruitments to flow into the spawning stock. For the 0.7 tuning level there is somewhat more variability among CMPs in their short term median biomass performance; however, at all levels, the greatest distinguishing feature among CMPs in their median biomass performance is in their longer term performance (i.e. post 2022). As such, median biomass performance in the short to medium term did not provide a useful basis for judging the relative performance. However, long term biomass performance over the 30-year time horizon is one feature that we judged to be important in our selection of CMPs for consideration by the Workshop.

Although little differences are seen in the median biomass trajectories in the short to medium term, substantial differences are seen in the median catch trajectories among the CMPs within short to medium time frames. This is in spite of the fact that there is not a very large difference in the median value of the average catch over the 2002 to 2022 time frame among CMPs (Figure 12). Essentially, to achieve the same tuning level, CMPs can make different trade-offs between short and medium term catches. In considering how such trade-offs affect the overall longer term performance, CMPs that make smaller catch adjustments initially end up with substantially worse median catch performance in the long term (particularly for the 0.7 and 1.1. tuning levels – Figure 5). It is also worth noting that the median catch level for the 1.1 tuning was always below current (2002) catch levels (except in the case of BRU_10 for which the lower 10th percentile in biomass declines to a very low level by 2032).

For the 1.1 tuning level, there was a tendency for rules that took higher catch in the short term (at the expense of medium term catch) to continue to reduce catches after 2022 so that the median catch was low by 2032 (i.e.CPU_05,CPR_11, and BRU_10). Unfortunately, these same rules yielded the highest median biomass levels by 2032. Thus, it would seem that such large catch reductions were not necessary after 2022 given that these rules achieved the 1.1 rebuilding goal, particularly in light of the median catch trajectories for the other CMPs. In essence, these rules appear to overcompensate beyond 2022 for maintaining higher early catches. None of the CMPs that we investigated had relatively higher catches in the short term (e.g. before 2012) and lower catches in the medium term (between 2012 and 2022), and then maintained or increased catches in the long term. It should also be pointed out that the actual amount of trade-off between short and medium term catches that we explored was relatively modest with the median catches for all CMPs, dropping by essentially the maximum amount permissible in the first year of a TAC change. We did not find or deliberately investigate a rule that explored the catch timing and biomass trade-off on an equal footing (i.e. as would have happened if the rule was simultaneously tuned to a B2022 and B2032 biomass targets – such a simultaneous tuning would have required time dependent and/or multiple tuning parameters).

Comparison of the lower 10th percentile for the catch trajectories shows that there was a large difference among CMPs in the risk with respect to catches reaching a low level. This is particularly true for the 1.1 tuning in which the lower 10th percentile approached 0 by 2032 for some CMPs while for others it remained ~4,000t (Figure 5b).

Continuous Rebuilding

Rebuilding of the SBT spawning stock has been a long standing objective of the SBT management. This objective was adopted because of the current low level of the spawning stock and the associated unquantifiable risk of recruitment collapse. However, given the current status of the stock and estimates of productivity, the potential amount of rebuilding that can be achieved within the 30-year time frame is constrained, particularly given the limits on catch reductions used in the testing process. Ensuring that the spawning stock has a high probability of continually increasing on average will reduce this risk. As such, we considered the probability that the spawning stock increases over time to be an important measure of a CMP's performance. However, the trend in median spawning biomass was downward until about 2010 for all tuning levels and CMPs that we considered (Figure 4a), so we only looked at this measure for years beyond 2010. One way to evaluate this performance measure is to check that the cumulative probability curves for B20xx/B2002 are shifted to the right as xx increases (this statistic has been calculated for xx equal to 12, 22 and 32); in other words, check that the probability of being below any given spawning stock size relative to 2002 decreases with time. All of the CMPs tuned to level 1.1 have this feature (e.g. Figure 6), with the exception of constant catch and BRU_15 (which showed some cross-overs in their curves). However, it is important to note that these distributional descriptions do not mean that every single biomass trajectory continually increases over time (as can be seen in the worm plots, e.g. Figure 18). Conversely, none of the CMPs tuned to level 0.7 showed this behaviour of continual rebuilding, and in all cases the median biomass in 2032 was below the median in 2022 (again remembering that these are probabilistic statements, and that individual worms show much variability).

Year of Lowest Spawning Stock

The operating model predicts that there is a high probability that the stock will decline relative to 2002 before any MP is implemented and permitted to change the TAC (Figure 4). There is also a high probability that this decline will continue for some years even once an MP is implemented because of the lag in recruitment to the spawning stock due to the late age of maturity (Figure 4), Given the historical low level of the SBT spawning stock and the long standing objective of rebuilding it, we considered how soon a CMP begins to rebuild the spawning stock as a measure of its performance, with sooner more desirable. Figure 7 shows cumulative probability plots for the year in which the spawning stock is at its lowest level. In general, there was not much to distinguish among the CMPs in this regard. Most CMPs tuned to level 1.1. have around a 50% probability that the spawning stock will have reached its minimum level between 2009 and 2012, and with the exception of constant catch and BRU_15, there is nearly 100% probability that the lowest spawning stock will occur prior to 2032. For the 0.7 tuning level, the year by which there is a 50% probability that the stock will have reached its minimum ranges from 2017 to 2032, and for all CMPs there is a substantial probability (~25-55%) that the lowest spawning stock level will occur in 2032. For the 1.5 tuning level, all CMPs yield similar performance and there is almost 100% probability that the lowest spawning stock biomass level will occur prior to 2017.

Catch/Biomass Trade-off

Figure 8 displays the trade-off between median 20-year average catch (from 2002 to 2022) and median spawning biomass in 2022 relative to 2002 (B2022/B2002) for seven CMPs across the three tuning levels. Figure 9 shows the same results but for the 30-year time horizon. In the case of the 20-year time horizon, the median spawning biomasses for any tuning are all equal by virtue of the tuning. However, there is also relatively small difference among rules in their median catch performance (less than ~1000t at all tuning levels). There is also some cross over in relative performance among rules between the 0.7 and 1.1 tuning levels. For the 30-year time horizon, Figure 9 shows that having achieved the same level of performance in terms of median biomass in 2022, the rules can vary in the trade-off they make between catches and biomass performance over the following 10 years. However, the results for all rules tend to fall on the same trade-off line, which suggests no rule is a clear "winner" in terms of simultaneously achieving both significantly better biomass and catch performance (as would be indicated by points in the upper right-hand corner of the graph).

Figures 10 and 11 provide parallel results to Figures 8 and 9 except the lower 10th percentile of the spawning biomass statistic has been displayed to provide an indication of the trade-off between average catch and biomass risk. When considering biomass risk rather than median performance we see greater differences between CMPs, with constant catch standing out as having much higher biomass risk than the other rules while having similar average catch performance. On the contrary, FXR_01 and FXA_71 tend to minimize biomass risk while still having good catch performance (although the results depend somewhat on tuning level).

Figures 12 and 13 provide an expanded view of the results for tuning level 1.1 that were presented in Figures 10 and 11. In terms of average catch versus biomass risk, FXA_71, FXR_01 and KAL_01 form a cluster that outperform the other rules. For example, over the 30-year time horizon, these rules achieve higher average catches than any of the other rules that have similarly low biomass risk.

Relative Biomass Risk

Given the historical low level of the spawning stock, a major concern is the possibility of a further large decline or collapse in recruitment. While rigorous quantification of this is not possible (and is not taken account of in the operating model), further declines in spawning stock increase the risk of this occurring. CMPs show different performance in terms of the probability that the spawning stock will be at levels substantially below the tuning level in future years (e.g. P(B2022/B2002<0.8) for tuning level 1.1). We considered the relative performance of CMPs in this regard as one measure of risk to consider in the selection process. However, comparing performance of CMPs in terms of this risk can be measured in several ways. For example, the differences in probability of the spawning stock can be compared for any given depletion ratio relative to 2002 in any given year. Alternatively, for a given probability level, the differences in the depletion level can be compared. The cumulative probability plots in Figures 1 and 2 provide one approach for comparing the relative biomass risk among CMPs in a given year, while the trajectories of the lower 10th percentile of spawning biomass relative to 2022 in Figure 4b provide another.

In assessing biomass risk, it is not clear what should constitute substantial differences. For example, a difference between 4 and 8% in the probability that the spawning stock will be below 0.6 of the 2002 level in 2022 could be seen as representing a 100(8–4)/4 = 100% difference in performance. Alternatively, a difference between 0.55 and 0.6 in the 10th percentile the B2022/2002 in 2022 could be seen as representing a 100(.6–.55)/.55 = 9% difference in performance. The difficulty here is the lack of a quantitative measure of how consequences (e.g. risk of collapse) changes as a function of the level of depletion combined with the fact that one is concerned with relatively low probabilities.

Nevertheless, comparison of the biomass risk remains an important consideration in the selection of a CMP. Almost all CMPs result in reduction of risk when compared to a tuned constant catch scenario (i.e. indicating some positive performance from adopting a feedback approach). As expected given the results presented above, there is not much gain at a 1.5 tuning level. Also as would be expected, there is relatively little differences among CMPs prior to at least 2012 because spawning biomass trends at this point are largely determined by events prior to the implementation of a CMP and because of the similarity among CMPs in their initial TAC changes. For the 0.7 and 1.1 tuning levels, there are some “large” differences among CMPs, particularly by 2022 and beyond (e.g. Figure 1, 2 and 4). In particular, BRU_10, BRU_15 and CPU_05 have consistently worse performance at the 1.1 tuning level. CPU_05 also shows relatively poor performance at the 0.7 tuning level (remembering that BRU_10 and BRU_15 were not tuned to this level).

Relative Fishery Risk

While catch provides one measure of a CMP’s performance in terms of the fishery, catch rates are another important fishery performance measure, particularly if they decrease to levels which could impair the economic viability of the fishery. As such we considered the extent to which CPUE might decline under different CMPs to be an important consideration in the selection process. This is of primary relevance to the longline fisheries, as we have a poor understanding of how CPUE relates to abundance in the purse seine fishery. Figure 14 presents the cumulative probability curves for the minimum CPUE level for different CMPs for the three tuning levels. For tuning levels 1.1 and 1.5, little difference is seen among the tuning rules. For tuning level 0.7, there were some substantial differences with FXR_01, FXA_71 and KAL_01 having the best performance (i.e. lower probabilities of substantial declines).

It worth noting that for all three tuning levels, it is estimated that there is a near 100% probability that CPUE will decline at some point in the 30 year time horizon to below its current level, and ~40% probability that CPUE will decline by at least 30%. Additionally, the relative risk to the fishery in terms of lower CPUEs increases at lower tuning levels (i.e. even though total catches may be greater, abundances and thus catch rates have an increasing probability of being lower).

Robustness Tests

Appendix 2 provides a comparison of the performance among CMPs for tuning level 1.1 for the set of robustness tests that were defined at the 2003 SAG. In general, the performance of CMPs was relatively insensitive to the robustness tests and/or there did not appear to be substantial differences among CMPs in their performance in a robustness test compared to the reference set. For the no-AC trial, most of the CMPs generally increased their medium and long term catch performance, while at the same time maintaining about the same level of performance in terms of biomass. In other words, increased recruitment in these no-AC tests was used by the rules primarily to improve performance in terms of catch. For the low recruitment test, CMPs tended to reduce catch but maintain similar performance in terms of biomass. Both of these results are encouraging in terms of a CMP providing robust performance in the event that future recruitments are outside of the range consistent with historical data when fitted to the operating model.

The robustness tests which tended to yield the most sensitivity in term of overall performance were those involving the age range used for standardizing selectivities for CPUE predictions (i.e. Med_A12 and Low_A12). Reducing the age range from 4-30 to 8-12 resulted in substantially poorer performance for all rules in terms of both catch and biomass. The age range used for standardizing selectivities for CPUE predictions determines the relationship between age-specific changes in selectivity and age-specific changes in catchabilities. When the range is the entire age range caught, it means that age-specific changes in selectivity are directly compensated by age-specific changes in catchability (e.g. if selectivity for age 4 increases by 10% then there is a corresponding increase in the catchability for that age class and decreases in other age classes). When the age range is narrow, then there is little compensation in catchabilities within the age range and much greater compensation outside of it (e.g. if the age range is reduced to a single age class, the catchability for that age class would be constant no matter how much the relative selectivity for that age class changed over time). There is little empirical basis for assessing how changes in selectivity (e.g. as the result of changes in targeting for specific age classes) affects the relative catchability of other age classes. The actual relationship is likely to be highly complex and not necessarily constant across years as it will depend upon the overlap in spatial/temporal distribution of age classes, differences in vulnerability with age to gear, and changes in gear configuration and setting practices.

The results of the Med_A12 and Low1_A12 robustness trials do not appear to have substantial implications for the selection of a CMP since CMPs showed relatively, although not totally, similar performance (this is discussed further below). However, the results do have substantial implications regarding the biomass and catch results achievable at any particular tuning level and, as such, on the selection of a tuning level. At the 2003 SAG, the 4-30 and 8-12 age ranges were considered “equally plausible and therefore a high weight or plausibility was given to this robustness scenario” (Anon 2003). Thus, given the substantially poorer performance in these sensitivity trials, the overall performance of a CMP when tuned to a specific value would be expected to be substantially worse (e.g. given the equal plausibility conclusion of the 2003 SAG, it would be approximately half way between the reference set and the A12 results if A12 was run on the full set of scenarios run in the reference set and the two were combined with equal weight).

Selection of “Best” CMPs

Based on the general comparison of the rules we tested and developed, we selected decision rules FXA_71, FXR_01 and KAL_01 for consideration by the Workshop in its final selection process. Figure 15 presents the performance of these three rules in terms of the median and 80% confidence interval for all of the agreed performance indicators. Figures 16 and 17 show the median, 10th and 90th percentiles of the time trajectories for spawning biomass and catch. Figure 6 compares the relative performance of these three rules in terms of biomass risk in 2012, 2022 and 2032 (i.e. cumulative probability relative to 2002). Figure 18 provides comparative examples of worm plots for the 1.1 tuning level for these three CMPs.

We note that the results of general comparisons presented above indicate a lot of similarities among CMPs when tuned to the same level. The major differences in overall performance stem from different tuning levels rather than from choice of a

CMP. This meant that the selection of CMPs to put forward was based on the finer and more subtle aspects of performance rather than on major differences in overall catch, biomass and stability of catch objectives. The three rules put forward show a high degree of similarity in performance; without more definitive selection criteria, we were unable to select among them.

One feature that we took into consideration in the selection of the rules we put forward was the notion that a rule has the "ability to learn" – i.e. adapt the magnitude of its response based on estimated quantities related to the productivity of the stock. This seems like a desirable feature because the magnitude of the response is not fixed over time for a given magnitude of change in the input data. The three rules being put forward all have this characteristic although KAL_01 to a lesser extent.

In our evaluation and selection, the main focus was given to comparing performance for rules tuned to level 1.1, given the long standing objective of the Commission to rebuild the spawning stock. This was because all rules tuned to the 0.7 level resulted in a greater than 50% probability of the stock declining between 2022 and 2032 on top of the 50% probability of that the stock would decline by 30% between 2022 and 2002; thus, they showed little or no prospect for rebuilding, much less continuous rebuilding over the 30-year time frame used for the evaluation process. Conversely, at the 1.5 tuning level, all rules were restricted to essentially identical behaviour.

In our evaluation and selection process, we concentrated primarily on the trade-offs in performance between catch and spawning biomass (not just in terms of median performance, but also in terms of the probability, or "risk", of having low catches or low spawning biomass). We did not focus on the catch stability performance indicators as we assumed that the catch stability objectives would be largely met by decision about the frequency and maximum magnitude of TAC changes (as well as the one year lag between a TAC decision and its implementation).

The three rules selected were chosen based on the fact that they all showed increasing probabilities of the spawning biomass increasing with time (i.e. relatively continuous rebuilding of the spawning stock; Figure 6), had good relative performance in terms of biomass (especially in terms of biomass risk), and also tended to do the best in maximizing average catches over the medium to long term. Overall, these three rules had very similar performance in terms of their spawning biomass trajectories, although KAL_01 has somewhat better longer term performance in the median and 90th percentiles, and FXR_01 has slightly poorer performance (Figure 16). Not surprisingly, FXR_01 has somewhat better average catch performance and KAL_01 somewhat worse (Figure 17). However, the biggest difference among them is in their trade-off between short and long term catch performance, with FXR_01 showing the larger probability of cutting catches in the short term, but compensated for by better performance in the long term (Figure 17). In particular, for KAL_01, there is a relatively high risk (e.g. 10% for 1.1 tuning level) that the TAC would be reduced to close to zero by 2014. In contrast, for FXR_01, there is less than a 10% probability that catches would ever be reduced much below 5,000t; however, this is achieved by a much higher probability that catches will be reduced by the maximum permitted when the first TAC change is allowed. FXA_71 has intermediate performance in this regard.

The degree of stability and consistency of TAC changes is one aspect of a CMP that industry emphasized as important in consultations held with them (Anon. 2003). In this regard KAL_01 appears to have somewhat poorer performance than the other two rules based on its AAV and B-TAC inconsistency statistics (Figure 15). This is also somewhat evident in the comparison of worm plots (Figure 18).

These three rules also had differences in performance in the A12 robustness trials. In particular, KAL_01 had substantially lower risk in terms of spawning biomass and FXR_01 had substantially higher risk. Thus, in the Med1_A12 trial, FXR_01 had about an 80% probability that the spawning stock in 2002 would be below the 2002 level compared to around 65% for KAL_01 (this compares to ~55 and 60% respectively for the reference Med1 trials). Similar increased relative risk is seen for levels of depletion below 1 and in the Low1_A12 trials (Figures 19 and 20). Given the high weight that the 2003 SAG gave to the A12 robustness test, the better performance of KAL_01 in these tests would perhaps be the single most important factor favouring it over the other two CMPs.

Figure 21 compares the cumulative probability curves for a) spawning biomass in 2032 relative to 2002 and b) 30 year average catch relative to 2002 for the no-AC robustness test in relationship to the reference set for FXR_01, FXA_15 and KAL_01 tuned to level 1.1. The three rules had relatively similar but slightly improved performance in terms of spawning biomass relative to the reference set. However, the biggest differences are in the catch performance, with KAL_01 showing somewhat more improvement. For all three decision rules, the potential trade-off between catch and biomass as the result of the increased recruitment in the no-AC trial is put into catch rather than rebuilding.

Frequency of TAC change

One component in the selection of a management procedure is the frequency that TAC changes are permitted. The performance of the three selected decision rules (i.e. FXR_01, FXA_71 and KAL_01) was explored for frequency changes of 1, 3 and 5 years for a 1.1 tuning. Figures 22 and 23 compare the median, 10th and 90th percentiles of the time trajectories for spawning biomass and catch using a 1, 3 and 5 year TAC frequency for each of the three selected decision rules.

For all three rules, there is no substantial difference in the median spawning stock trajectories for a 1, 3 or 5 year TAC change, except to some extent in the case of KAL_01, for which the median biomass at the end of the projection period decreases when the period between TAC changes increases (Figure 22). For all three rules, the lower 10th percentile in spawning biomass is lowest with a 5 year interval. For FXR_01 and FXA_71, an annual TAC change has noticeably better performance in terms of the lower 10th percentile than either a 3 or 5 year interval. In particular, the difference in the lowest level of the 10th percentile in biomass compared to 2002 was around 0.05, which could be considered a substantial reduction in risk. In contrast, for an annual TAC change, the upper 90th percentile for an annual TAC was smaller for the biomass trajectories starting in around 2015, suggesting that that annual TAC changes allowed for earlier and larger overall increases in the more optimistic scenarios.

Comparisons of the 90th percentiles of the catch time trajectories suggest that in the scenarios where catches are increased, the increases will occur earlier, and for KAL_01 they will increase to higher levels. Conversely, based on the lower 10th percentile of catches, in situations where catches are reduced, the reduction in catch tends to occur earlier, which is the source of the better performance in the lower 10th percentile trajectories of spawning biomass.

Comparison of worm plots for 1, 3 and 5 year TAC changes suggest that there was considerably more variability in the magnitude of TAC changes with an annual TAC change than with either a 3 or 5 year change (examples of this can be seen in comparing Figures 24 and 18). This appears to be particularly notable in the results for KAL_01.

Initial Year TAC applied

The above comparisons of performance for an annual TAC change versus a 3 or 5 year change are confounded by the fact that in the annual case the first TAC change was permitted in 2006 compared to 2008 for the 3 and 5 year cases. In order to get an indication of whether the differences seen above were the result of the different start year or the TAC interval, we re-tuned FXR_01 to the 1.1 level using a 3-year TAC interval but allowing a 2006 start year (referred to as FXR_99). Figure 25 compares the median, 10th and 90th percentiles of the catch and spawning biomass trajectories. There results when compared to Figure 22 suggest that the frequency of TAC change was the primary source of the differences seen in the spawning biomass trajectories and not the two year difference in start year.

Concluding Remarks

Based on the comparison of the CMPs tested, very large differences in performance among CMPs were not found when CMPs were tuned to the same median spawning biomass level in the first 20 years. This was particularly true for performance with respect to biomass. The main differences between CMP performances were in their trade-offs between short and medium term catches (up to 2022), and in their trade-offs between longer term catches versus levels of rebuilding after 2022. However, the major differences in overall biomass and catch performances stem from different tuning levels rather than from differences among CMPs. It appears, at least based on our experience, that the selection of CMPs to put forward for consideration by the Commission will be based on the finer and more subtle aspects of performance rather than on major differences with respect to overall catch, biomass and stability of catch objectives.

Literature Cited

Anon. 2003. Report of the Fourth Stock Assessment Group Meeting. 25-29 August, 2004. Christchurch, New Zealand. CCSBT.

Eveson. P. 2003. An update of the graphics used for evaluating the performance of candidate management procedures for southern bluefin tuna. CCSBT-ESC/0309/23.

Eveson, P. and Ricard, D. 2003. An overview of potential graphics for evaluating the performance of candidate management procedures for southern bluefin tuna. CCSBT-MP/0304/05.

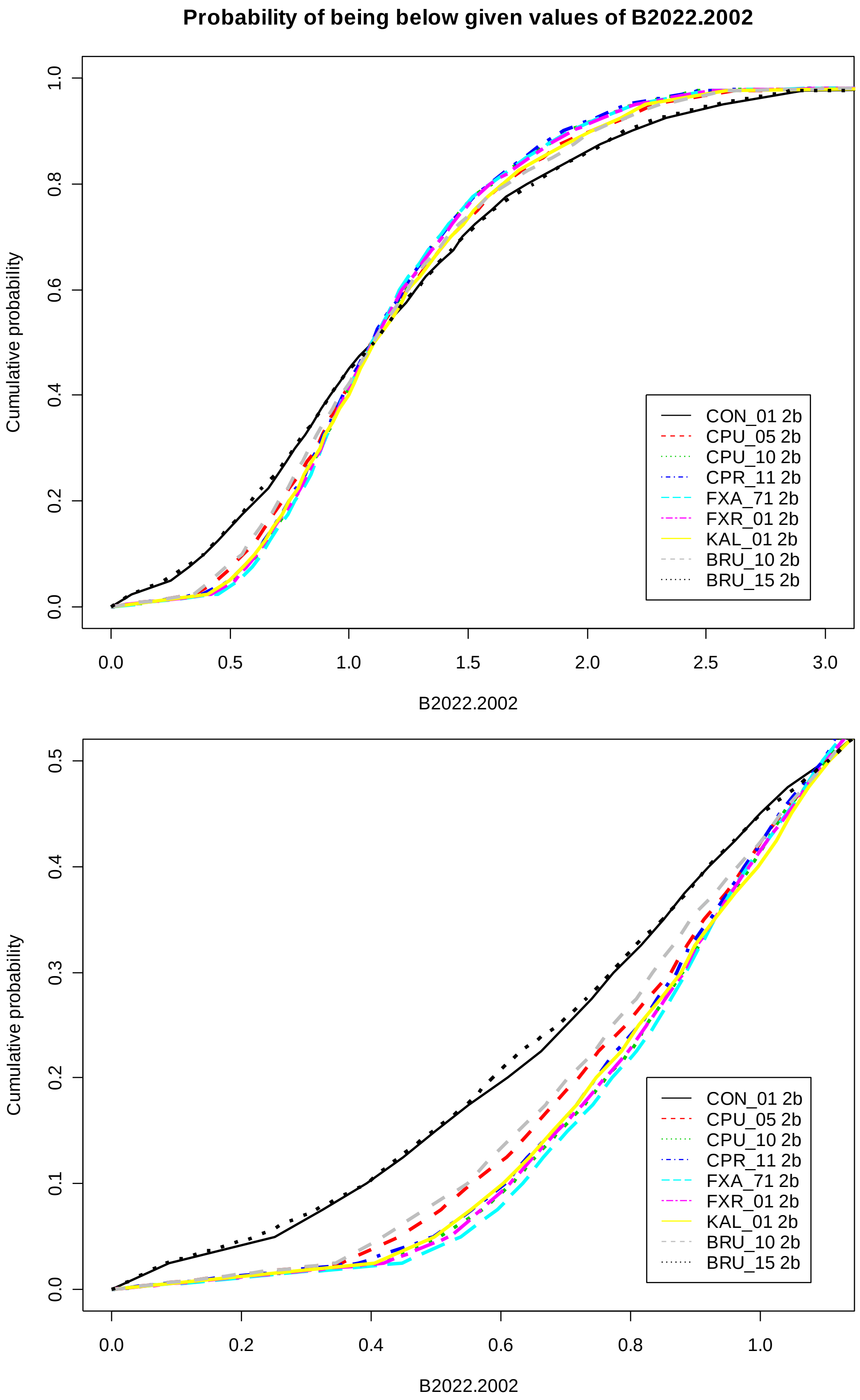


Figure 1: Cumulative probability of the spawning stock in 2022 relative to 2002 (B2022/B2002) being below a given value for CMPs tuned to 1.1. The upper panel gives results over the entire range of values for B2022/B2002; the lower panel is a close-up for values below 1.1.

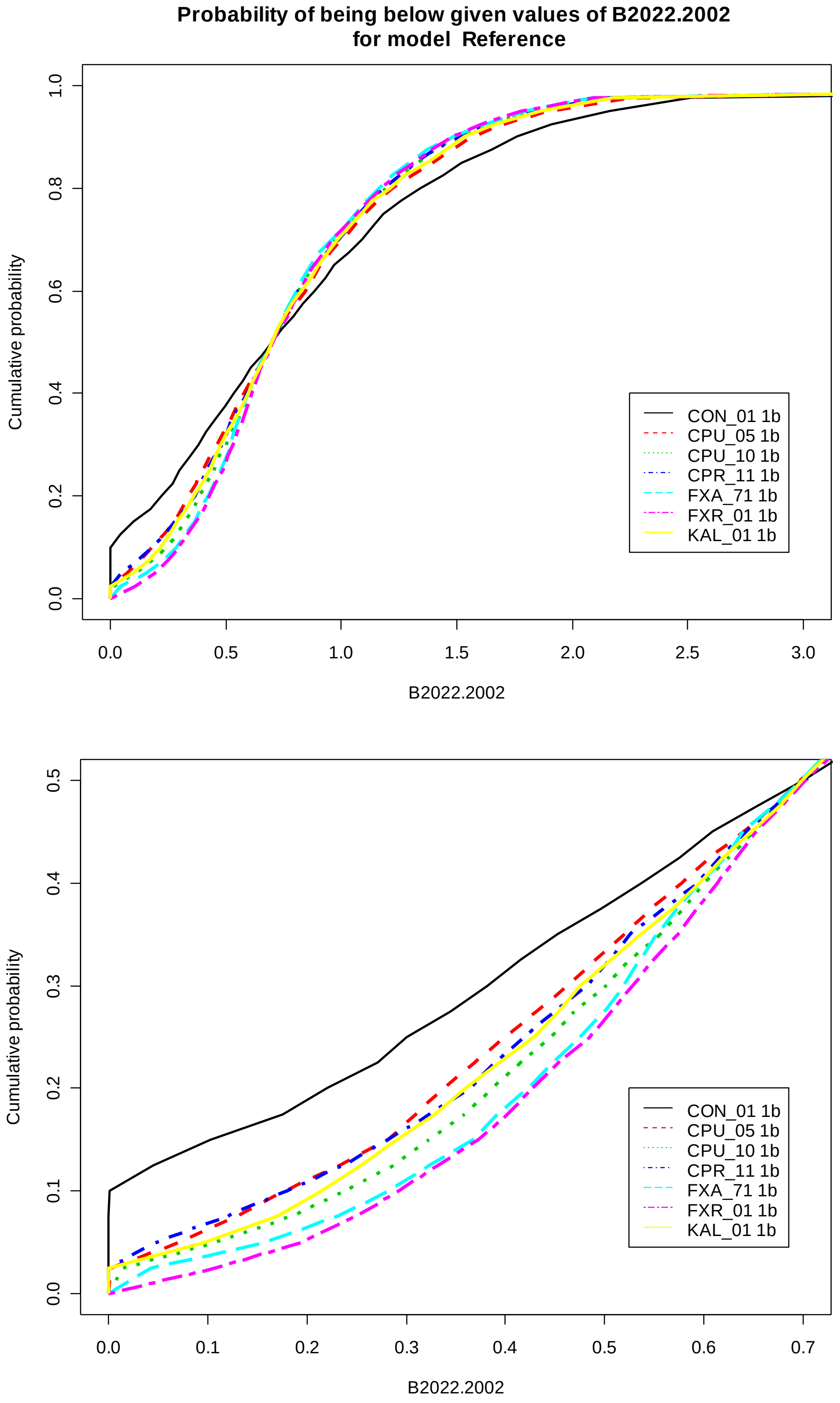


Figure 2a: Cumulative probability of the spawning stock in 2022 relative to 2002 (B2022/B2002) being below a given value for CMPs tuned to 0.7. The upper panel gives results over the entire range of values for B2022/B2002; the lower panel is a close-up for values below 0.7.

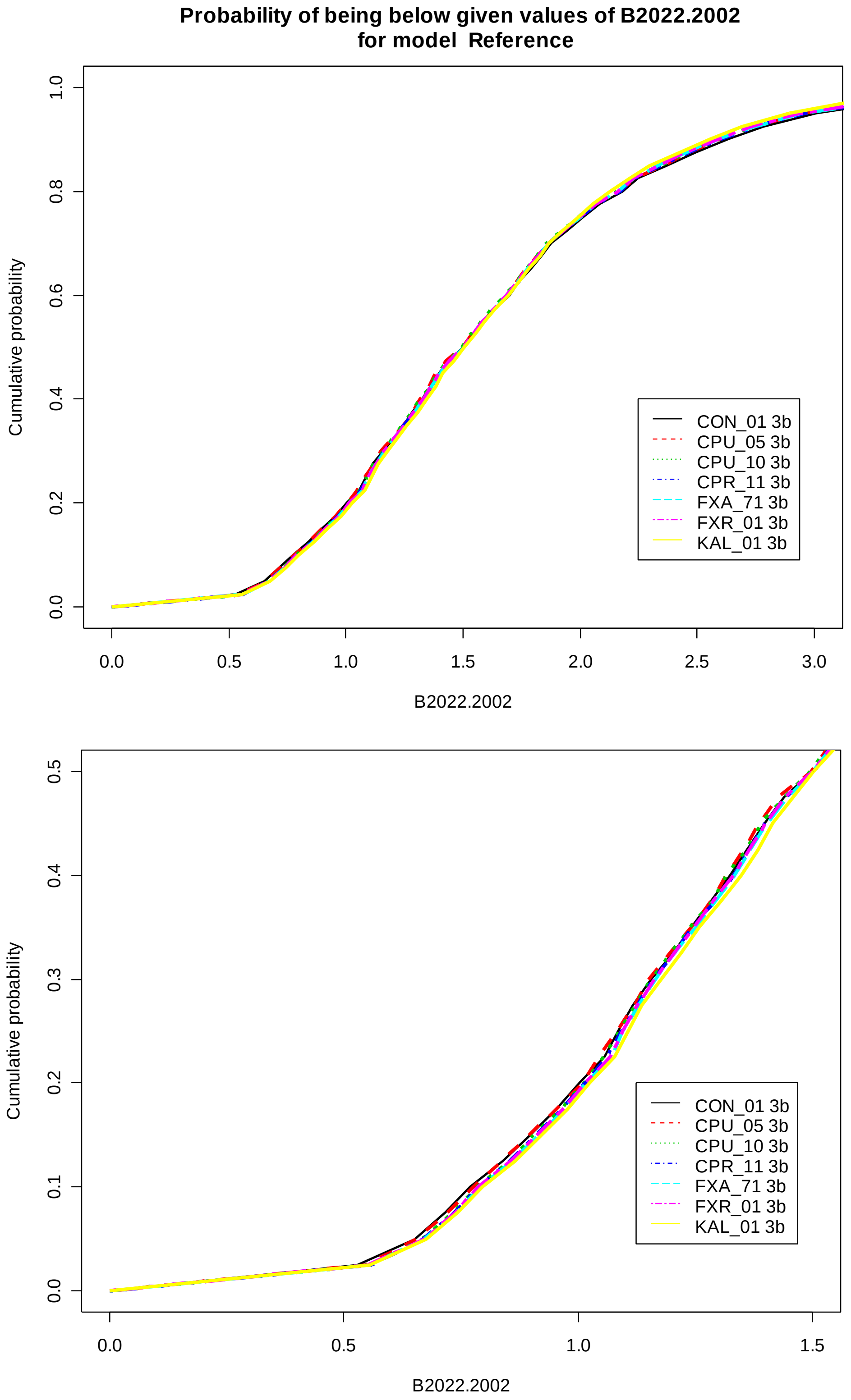
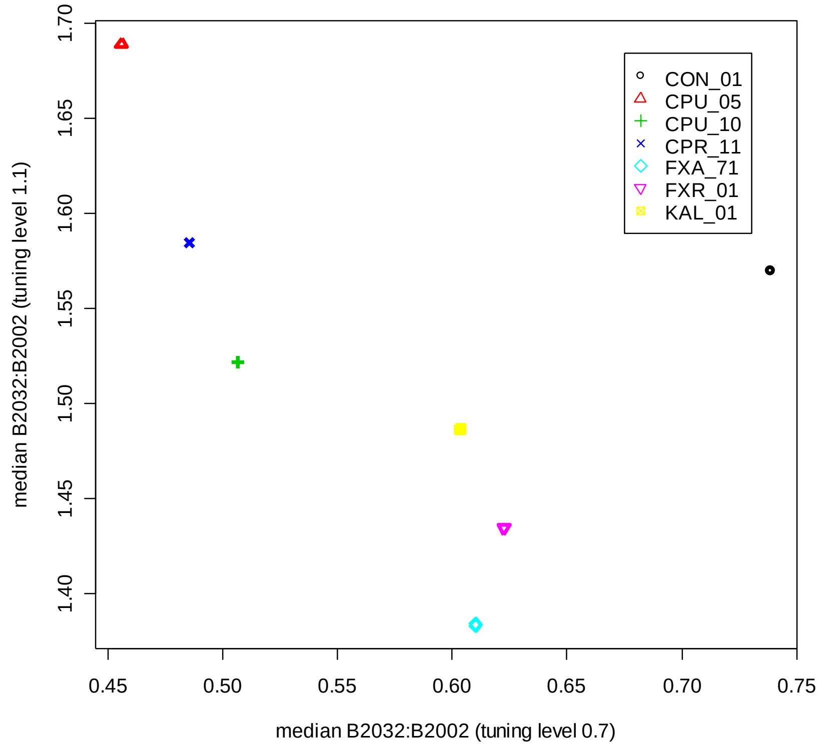


Figure 2b: Cumulative probability of the spawning stock in 2022 relative to 2002 (B2022/B2002) being below a given value for CMPs tuned to 1.5. The upper panel gives results over the entire range of values for B2022/B2002; the lower panel is a close-up for values below 1.5.

.



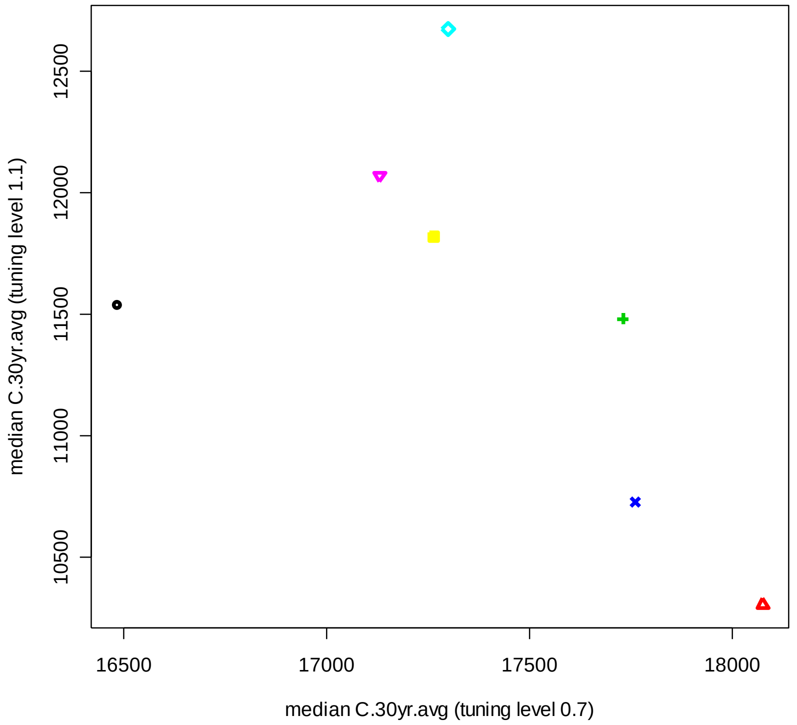


Figure 3: Comparison of CMPs' performance at tuning level 1.1 versus 0.7 in terms of: (upper panel) the median 2032 spawning biomass level relative to 2002 (B2032/B2002); (lower panel) the median 30-year average catch.

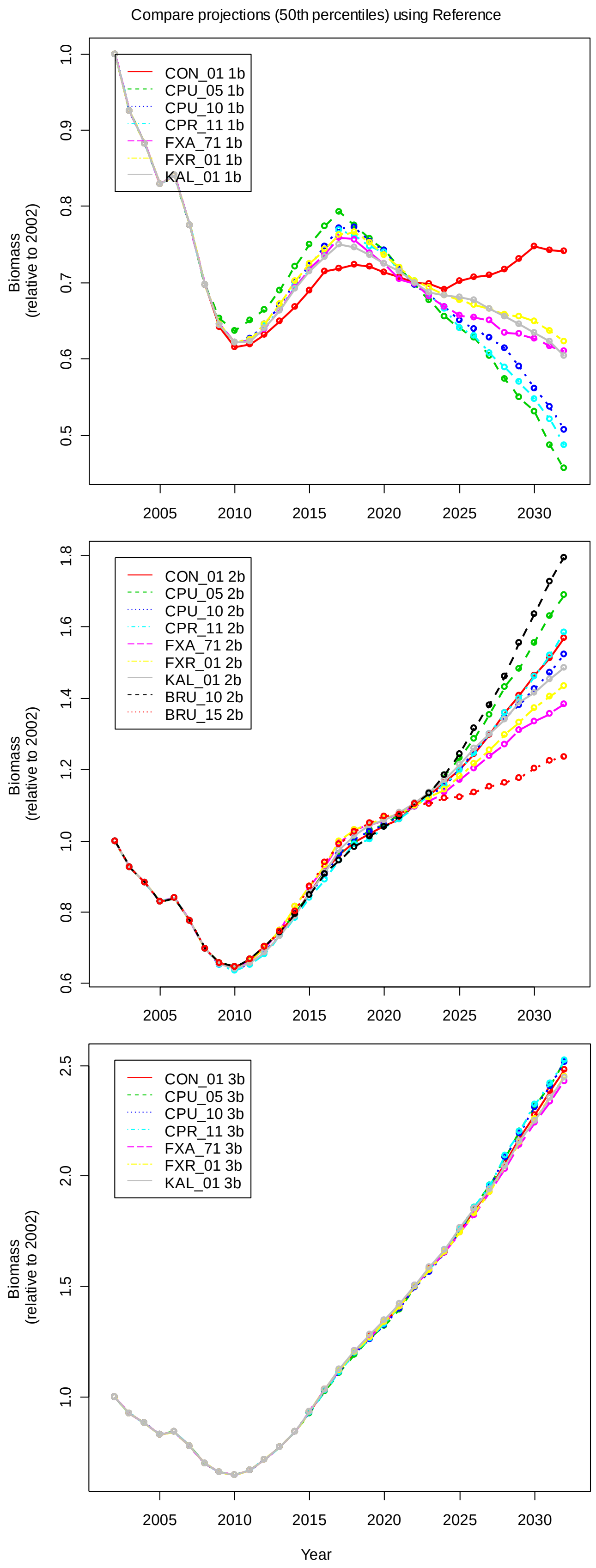


Figure 4a: Comparison of the time trajectories of the **median** spawning biomass relative to 2002 for a range of CMPs tuned to three different levels (upper panel 0.7, middle 1.1 and lower 1.5).

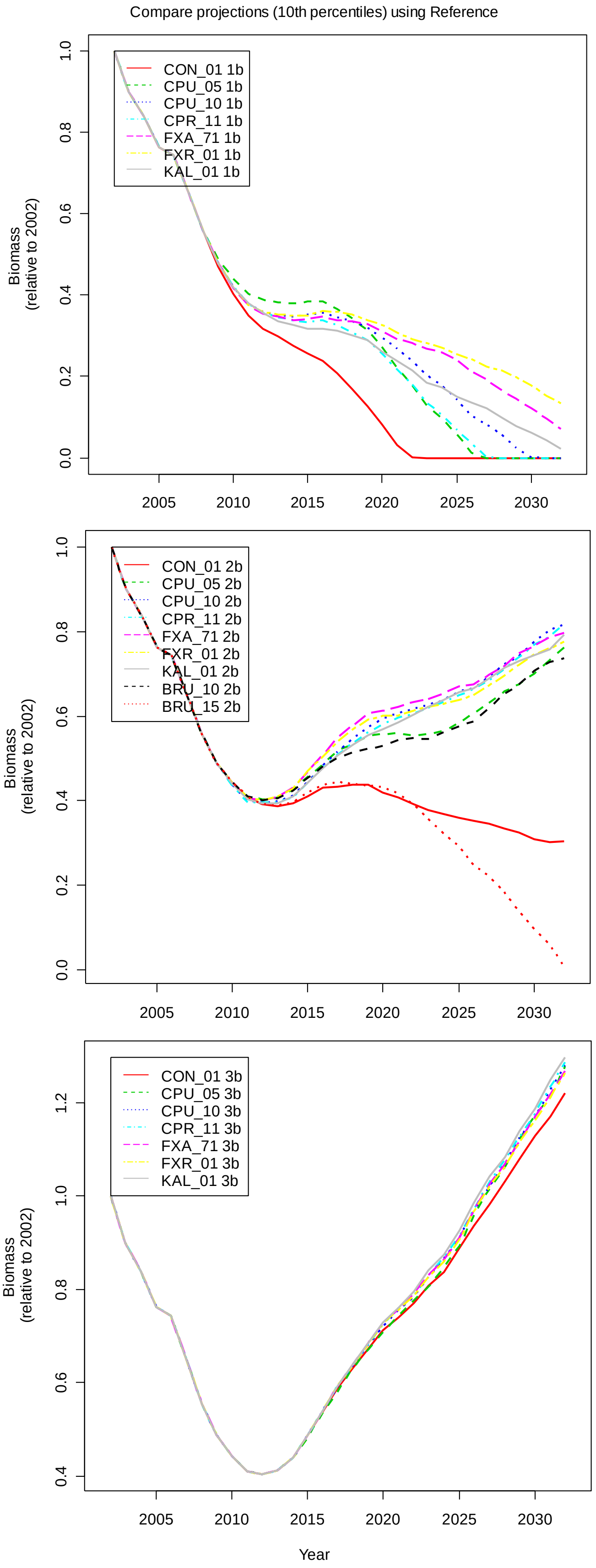


Figure 4b: Comparison of the time trajectories of the **10th percentiles** of spawning biomass relative to 2002 for a range of CMPs tuned to three different levels (upper panel 0.7, middle 1.1 and lower 1.5).

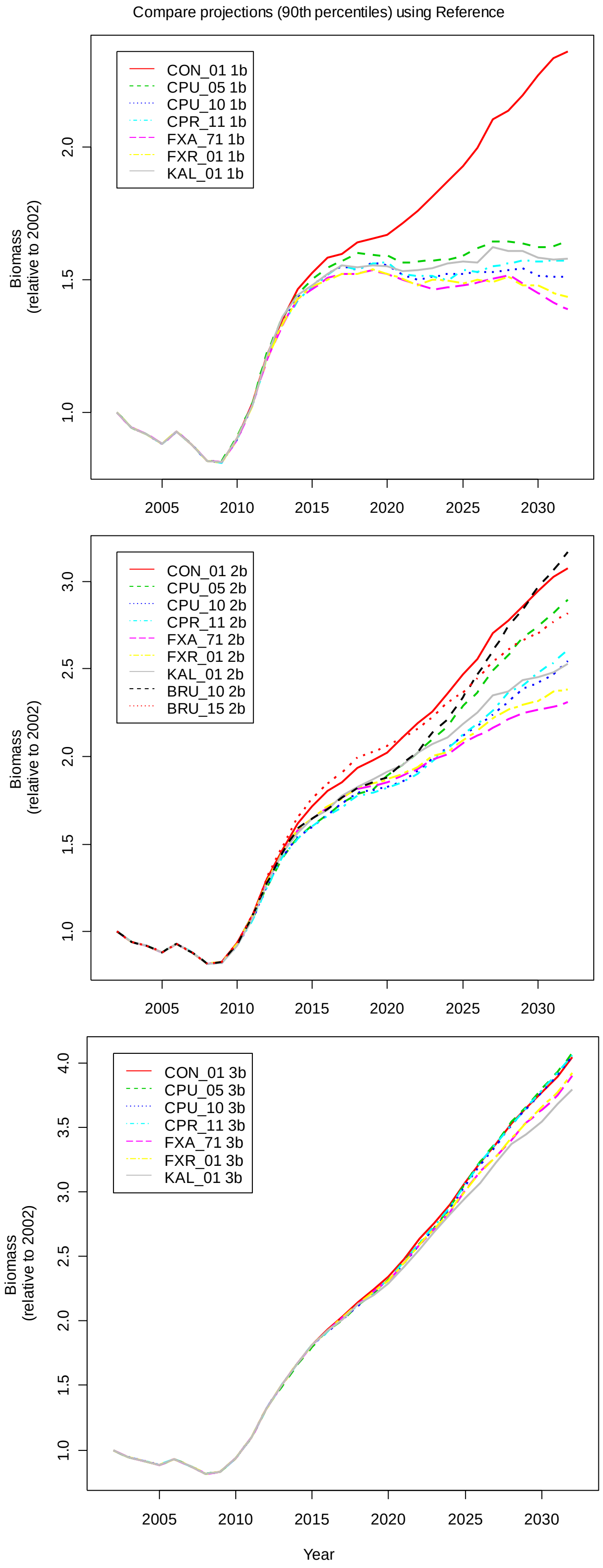


Figure 4c: Comparison of the time trajectories of the **90th percentiles** of spawning biomass relative to 2002 for a range of CMPs tuned to three different levels (upper panel 0.7, middle 1.1 and lower 1.5).

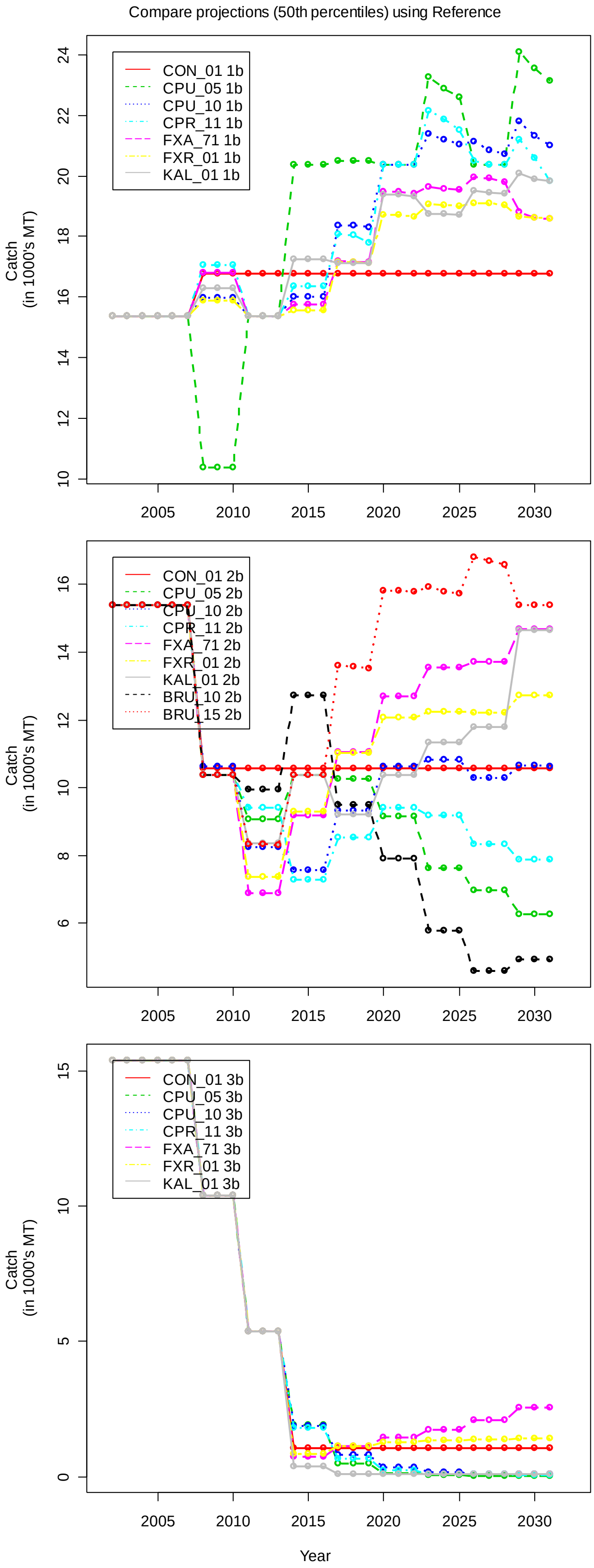


Figure 5a: Comparison of the time trajectories of the **median** catch for a range of CMPs tuned to three different levels (upper panel 0.7, middle 1.1 and lower 1.5).

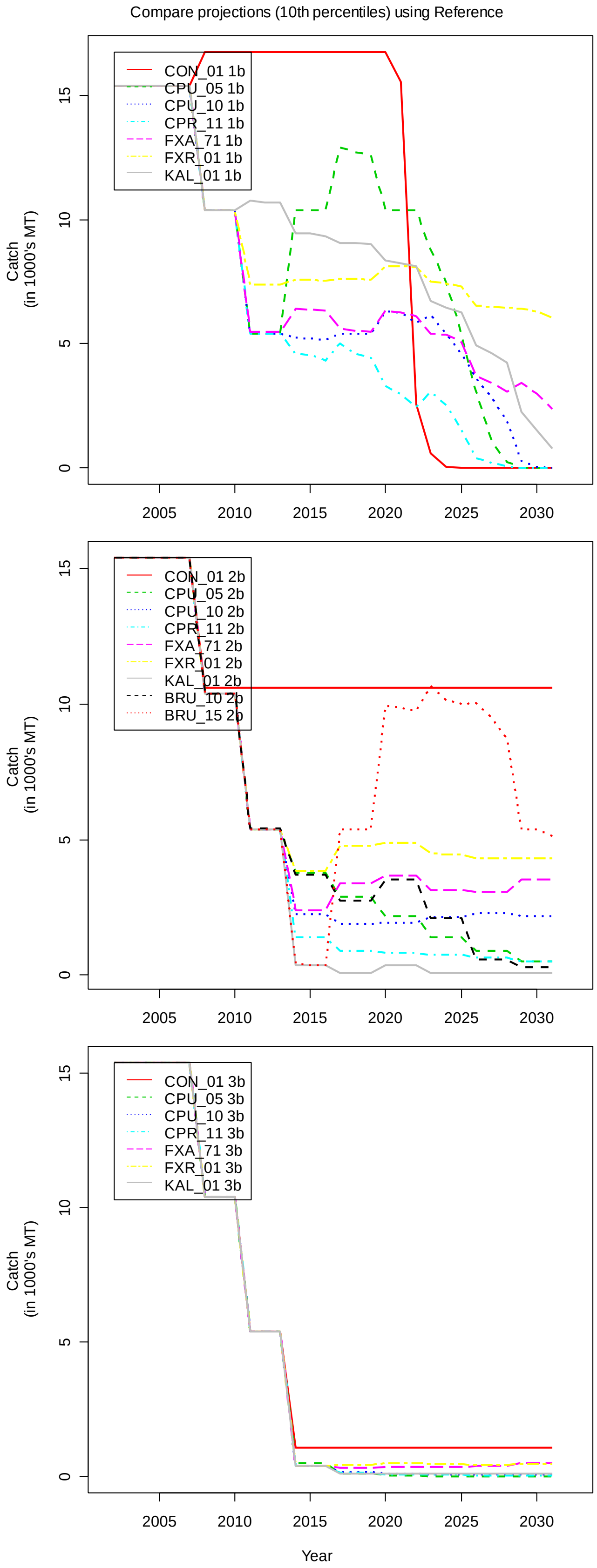


Figure 5b: Comparison of the time trajectories of the **10th percentiles** of catch for a range of CMPs tuned to three different levels (upper panel 0.7, middle 1.1 and lower 1.5).

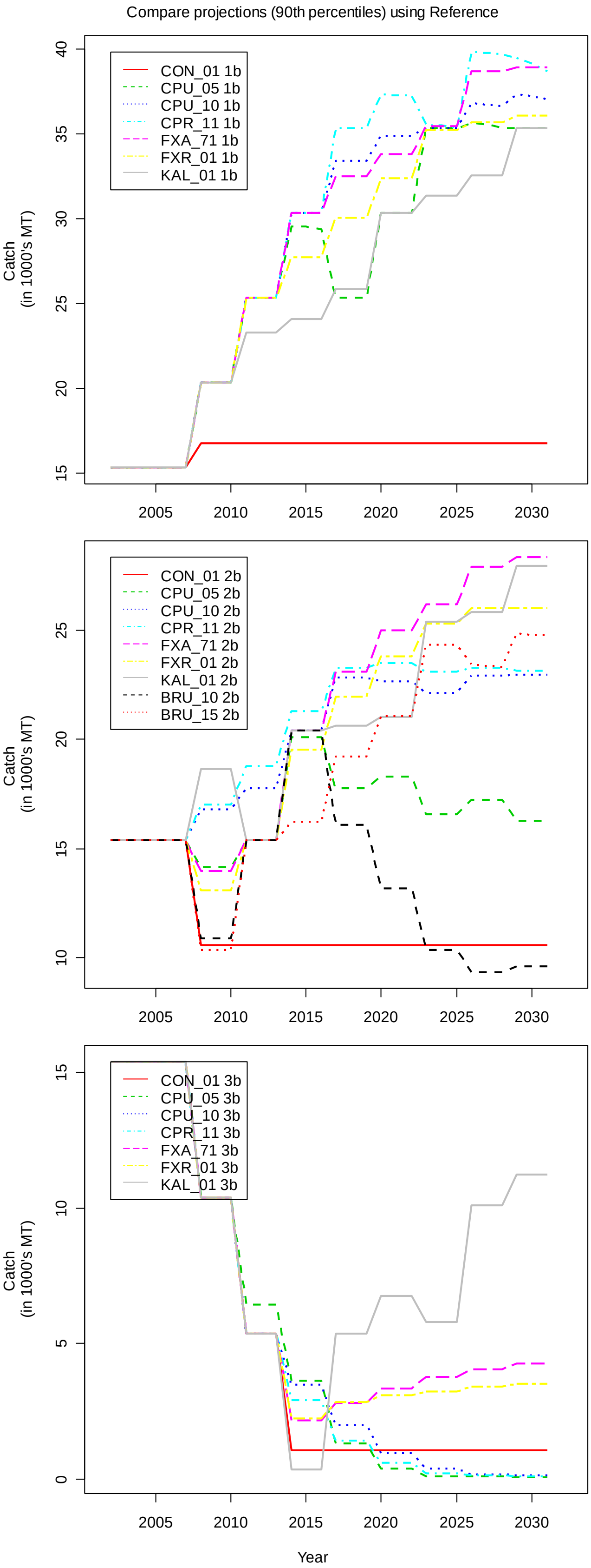


Figure 5c: Comparison of the time trajectories of the **90th percentiles** of catch for a range of CMPs tuned to three different levels (upper panel 0.7, middle 1.1 and lower 1.5).

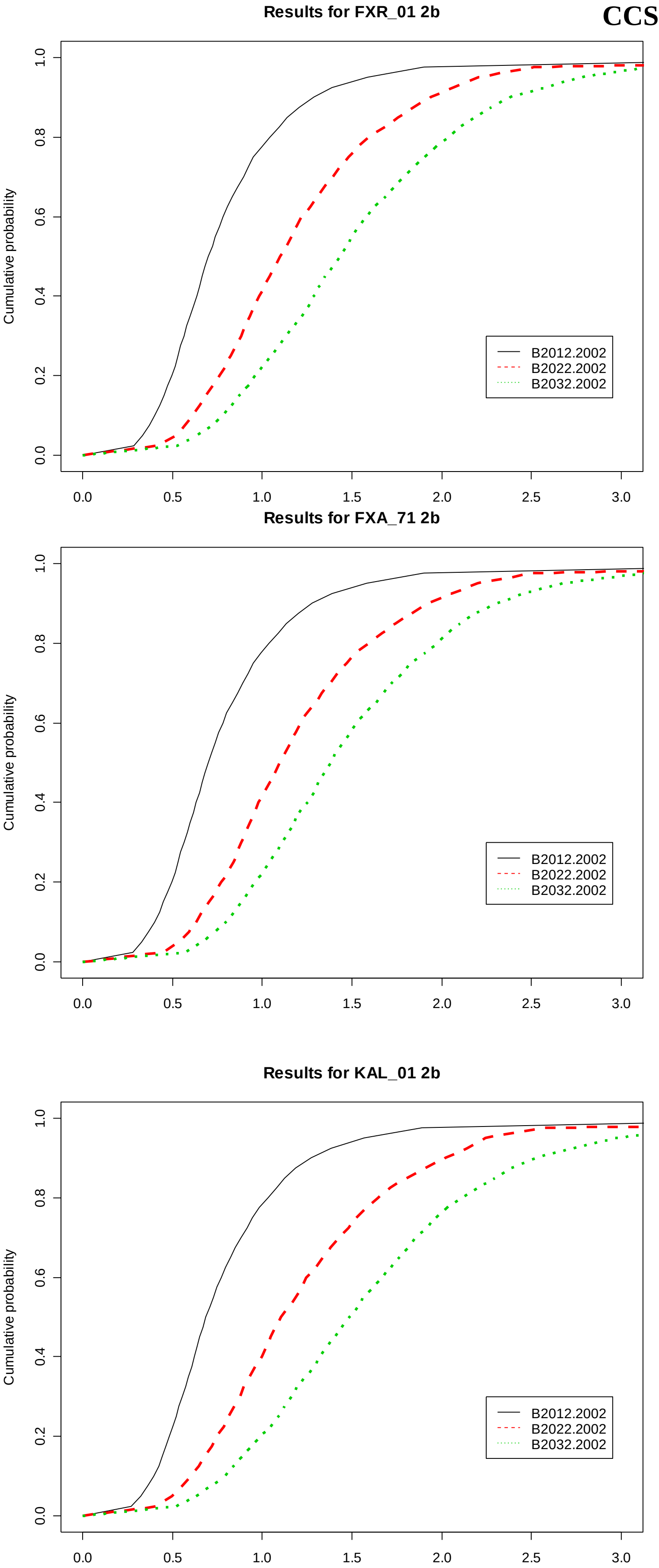


Figure 6: Comparison of the cumulative probability curves for the spawning stock in 2012, 2022 and 2032 relative to 2002 for the 1.1 tuning level. Upper panel is for CMP FXR_01, middle panel is for FXA_71 and lower panel is for KAL_01.

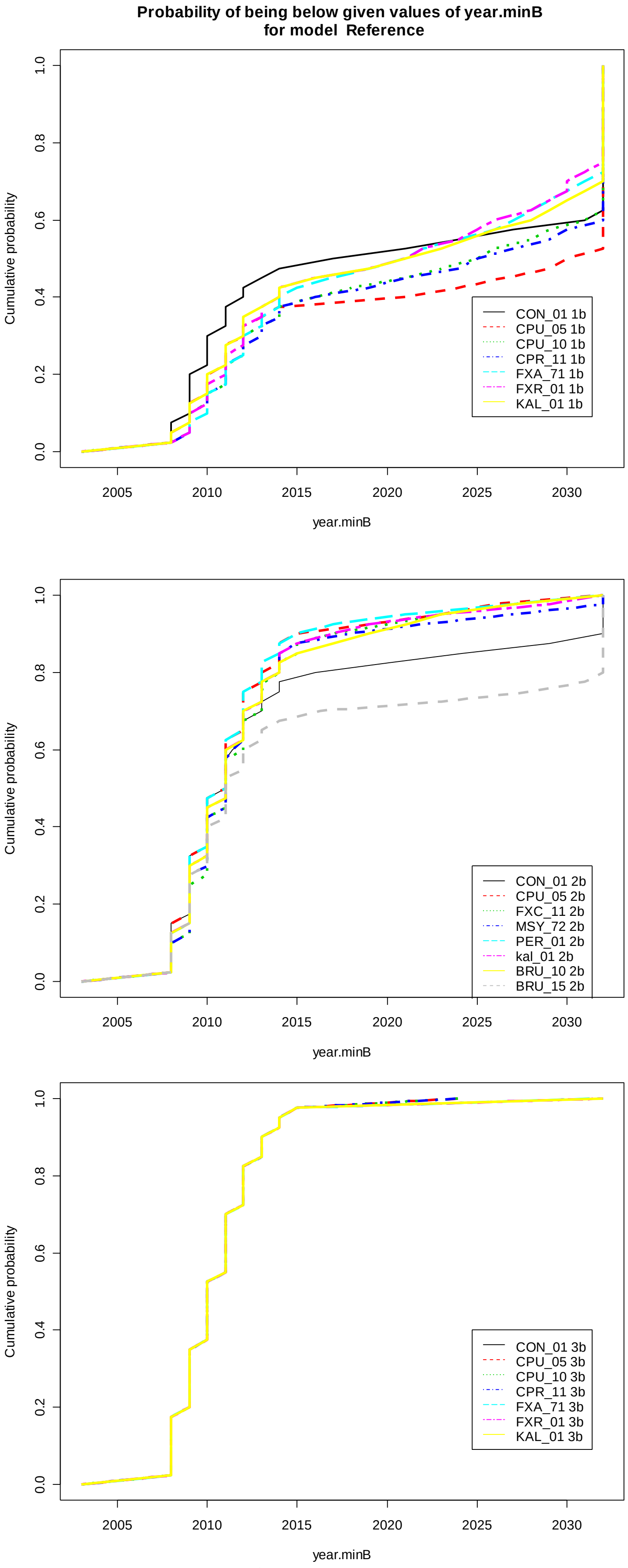


Figure 7: Cumulative probability for the year in which the spawning would be at its lowest level for three different tuning levels for a range of CMPs (upper panel 0.7, middle 1.1 and lower 1.5) .

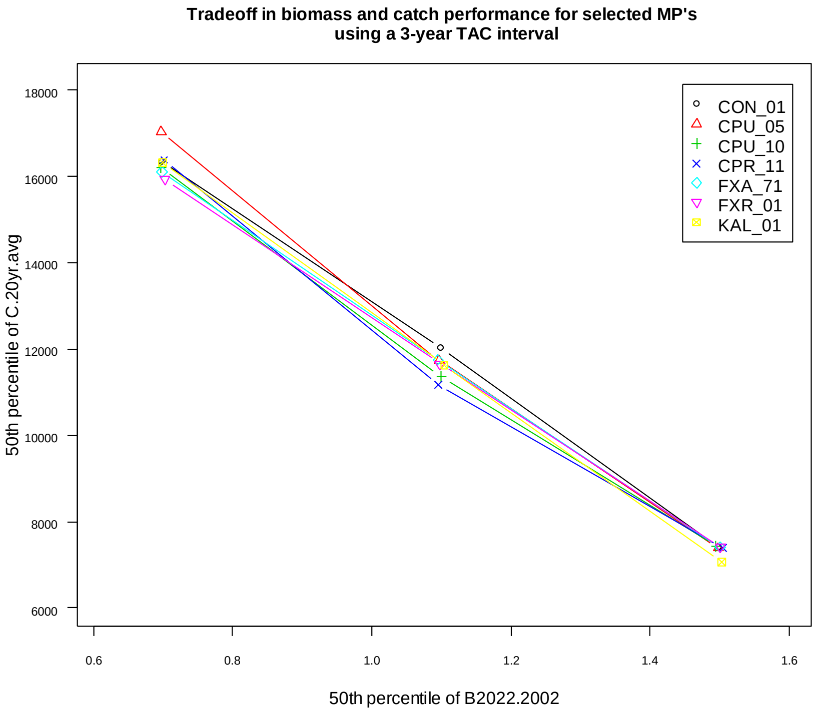


Figure 8: Comparison of the trade-off between median 20-year average catch (from 2002 to 2022) and median spawning biomass in 2022 relative to 2002 (B2022/B2002) for the seven decision rules across the three tuning levels. (Note results are not shown for BRU_10 and BRU_15 because these two CMPs were only tuned to 1.1).

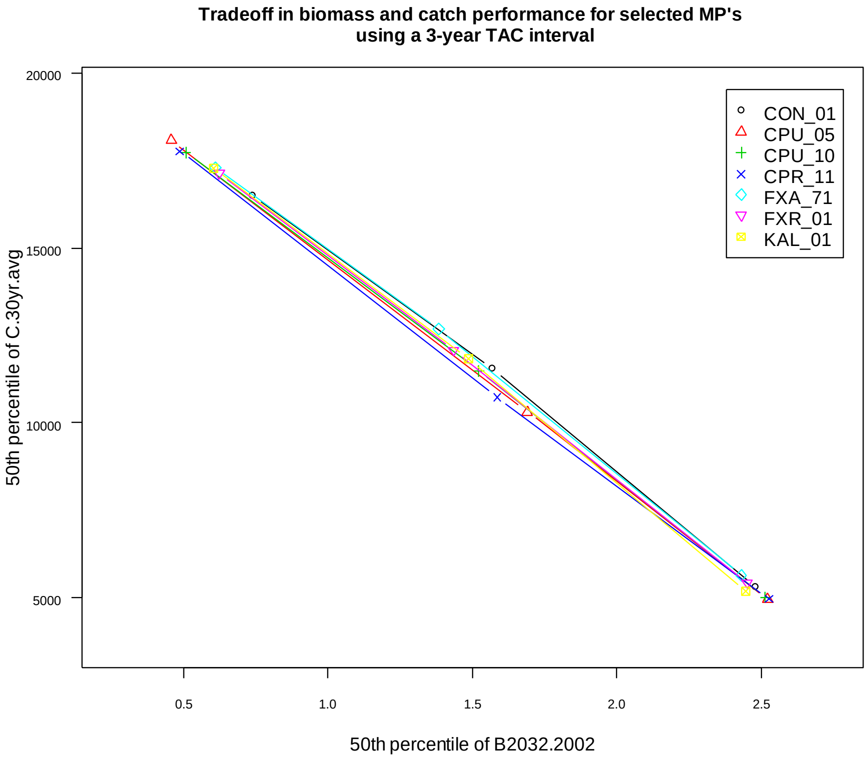


Figure 9: Comparison of the trade-off between median 30-year average catch (from 2002 to 2032) and median spawning biomass in 2032 relative to 2002 (B2032/B2002) for the seven decision rules across the three tuning levels. (Note results are not shown for BRU_10 and BRU_15 because these two CMPs were only tuned to 1.1).

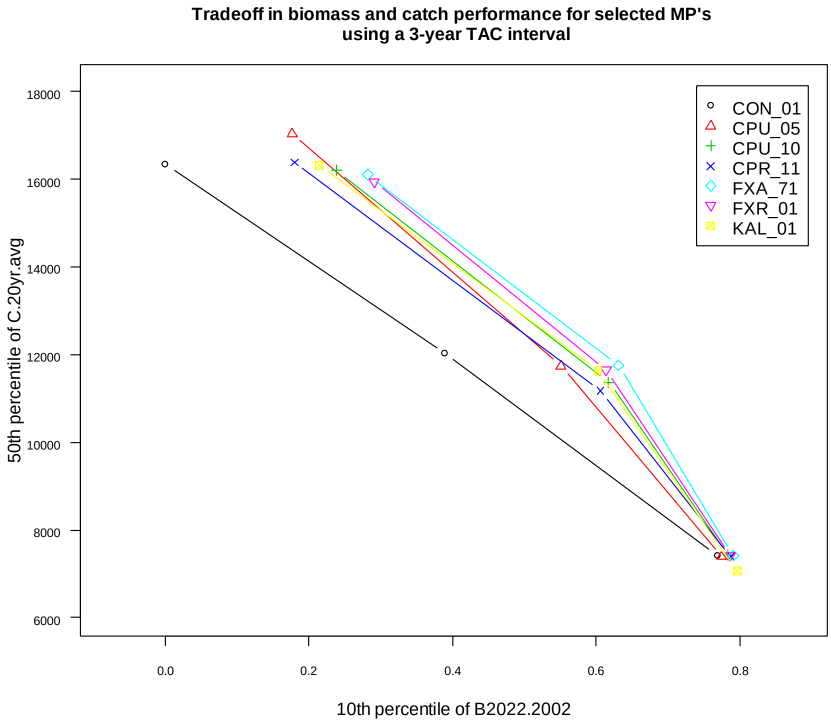


Figure 10: Comparison of the trade-off between median average catch between 2002 and 2022 and lower 10th percentile of spawning biomass in 2022 relative to 2002 (B2022/B2002) for the seven decision rules across the three tuning levels (Note results are not shown for BRU_10 and BRU_15 because these two CMPs were only tuned to 1.1).

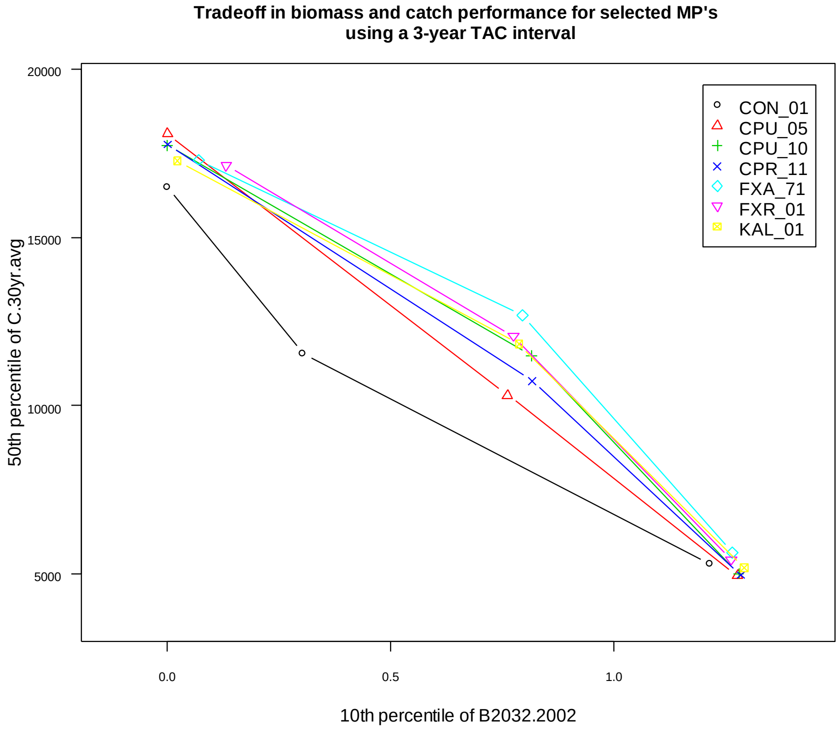


Figure 11: Comparison of the trade-off between median average catch between 2002 and 2032 and lower 10th percentile of spawning biomass in 2032 relative to 2002 (B2032/B2002) for the seven decision rules across the three tuning levels (Note results are not shown for BRU_10 and BRU_15 because these two CMPs were only tuned to 1.1).

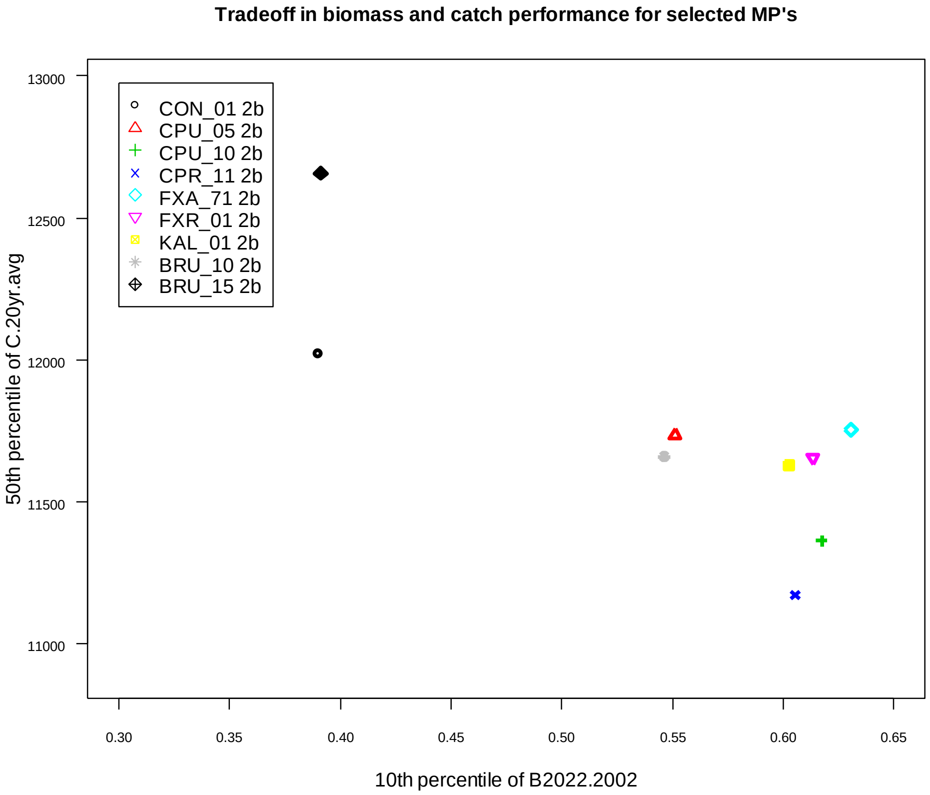


Figure 12: Comparison of the trade-off between median average catch between 2002 and 2022 and lower 10th percentile of spawning biomass in 2022 relative to 2002 (B2022/B2002) for nine decision rules for the 1.1 tuning level.

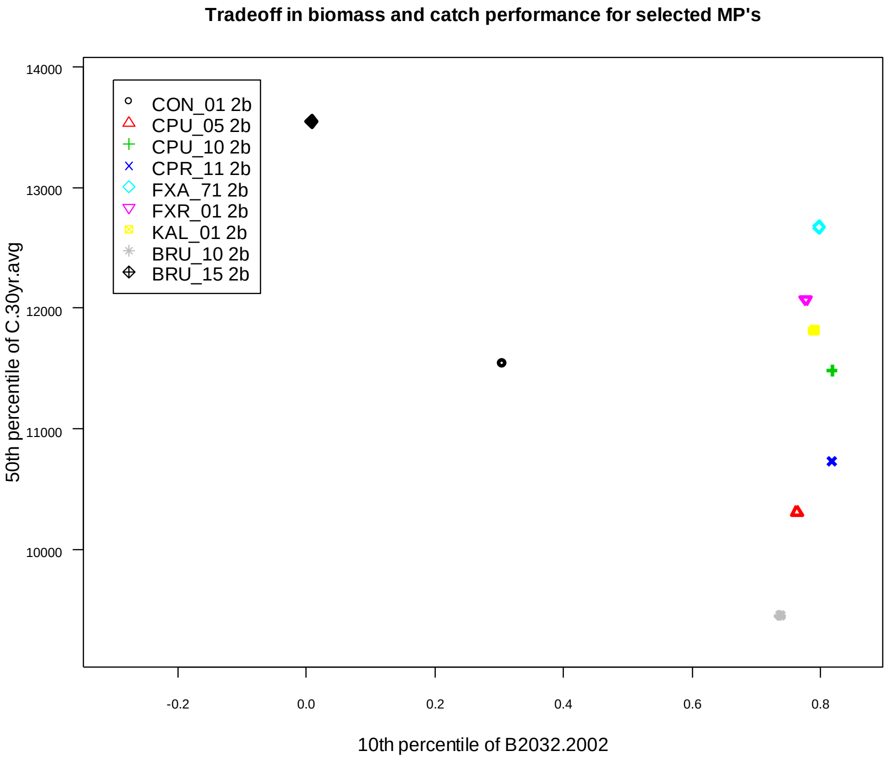


Figure 13: Comparison of the trade-off between median average catch between 2002 and 2032 and lower 10th percentile of spawning biomass in 2032 relative to 2002 (B2032/B2002) for nine decision rules the 1.1 tuning level.

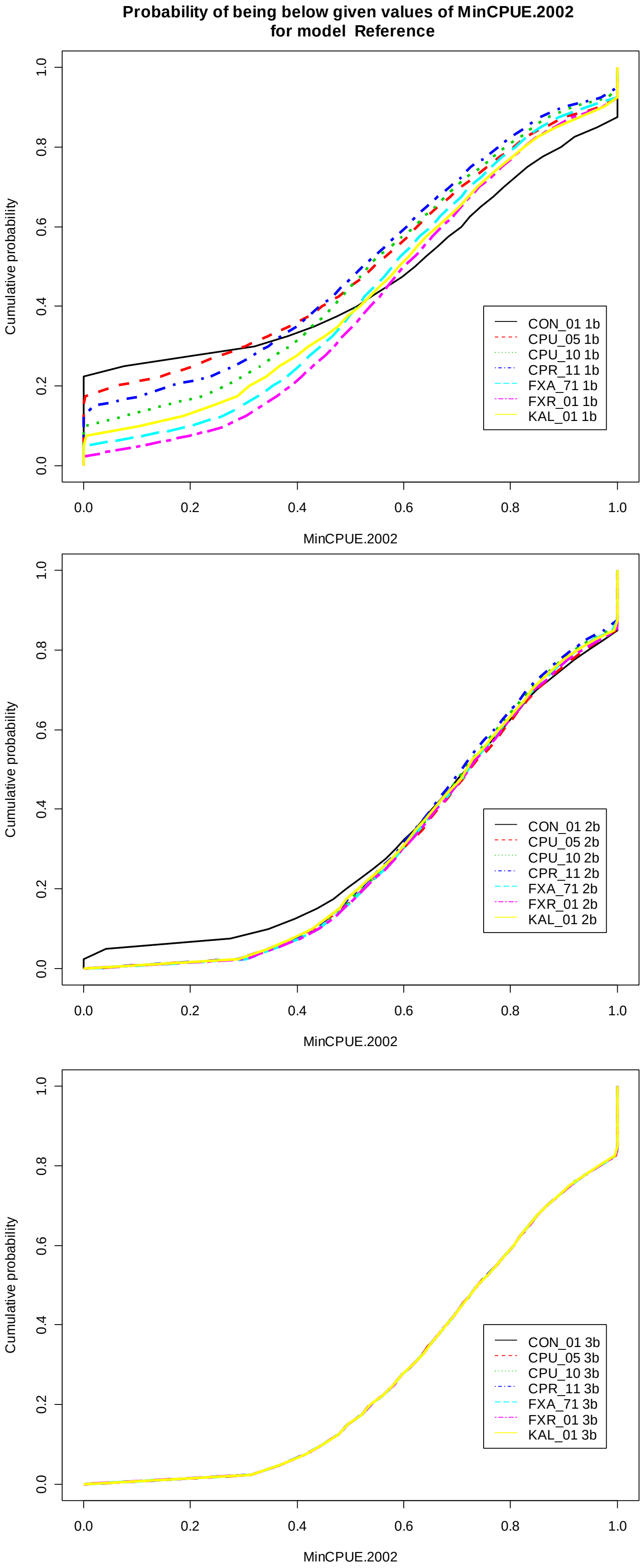


Figure 14: Cumulative probability for the minimum CPUE over the 30-year projection period relative to 2002 for three different tuning levels for a range of CMPs (upper panel 0.7, middle 1.1 and lower 1.5) .

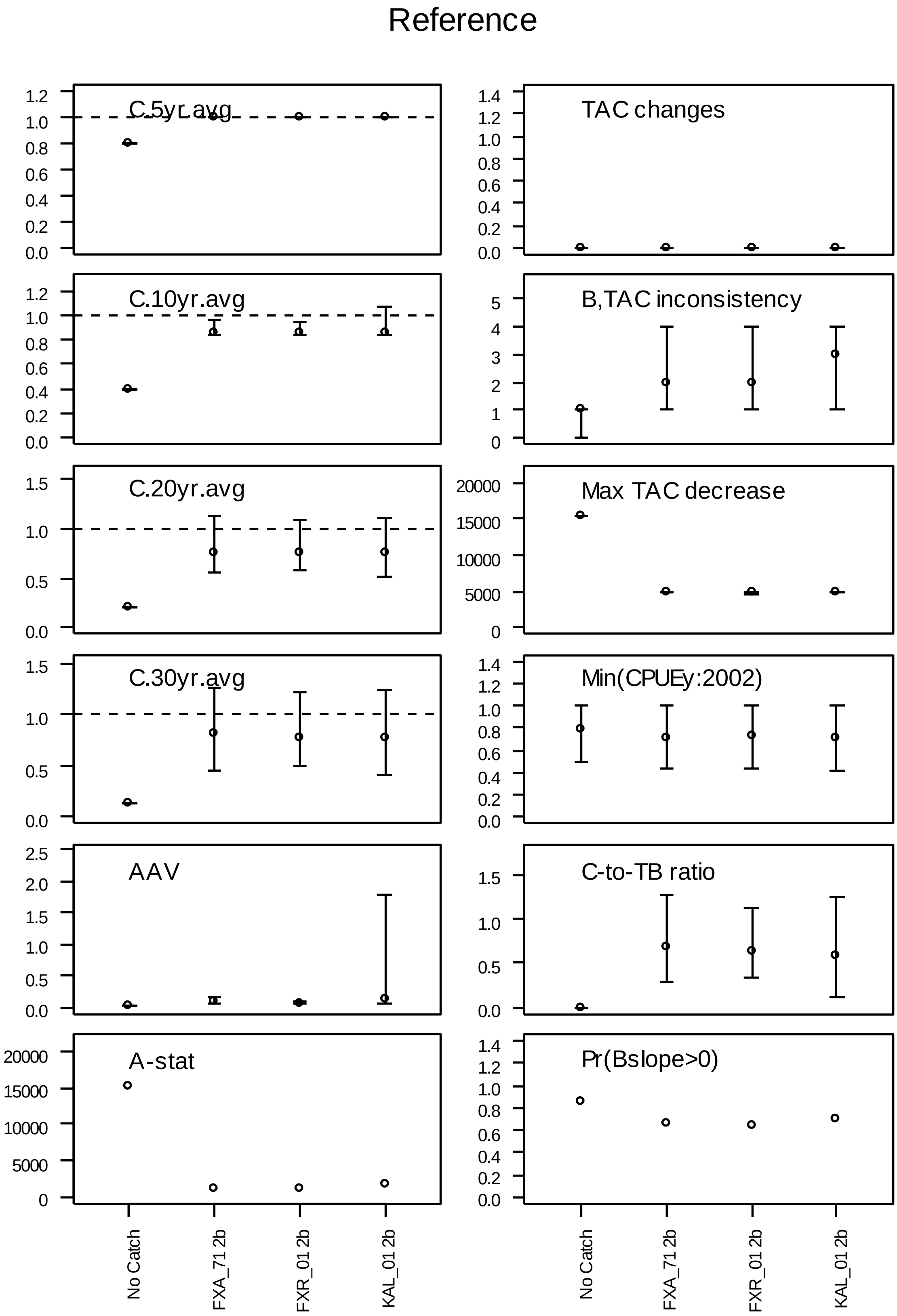


Figure 15: Comparison of the median and lower 10th percentiles for the set of agreed performance indicators for the FXA_71, FXR_01 and KAL_01 for the 1.1 tuning level.

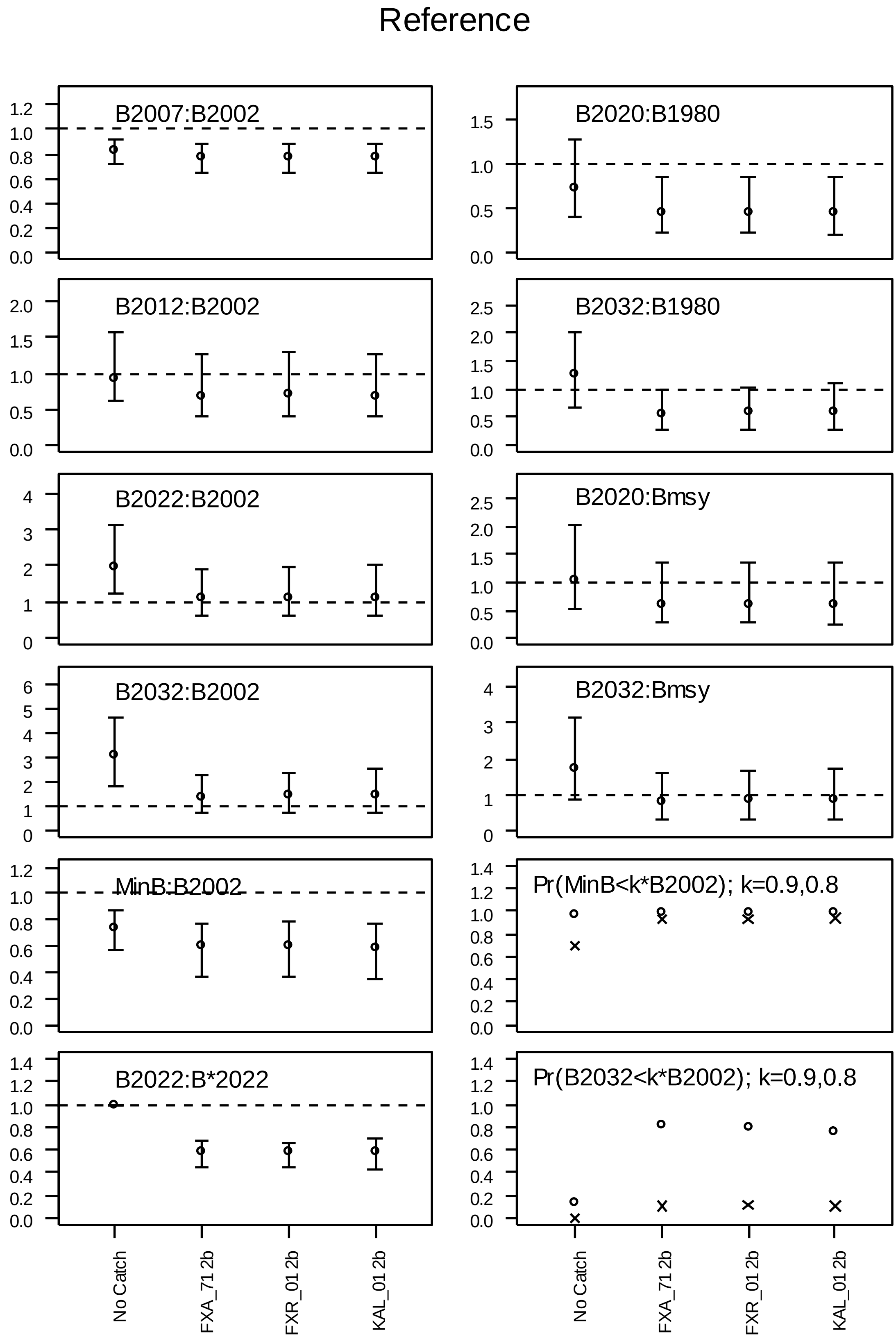


Figure 15(continued): Comparison of the median and lower 10th percentiles for the set of agreed performance indicators for the FXA_71, FXR_01 and KAL_01 for the 1.1 tuning level.

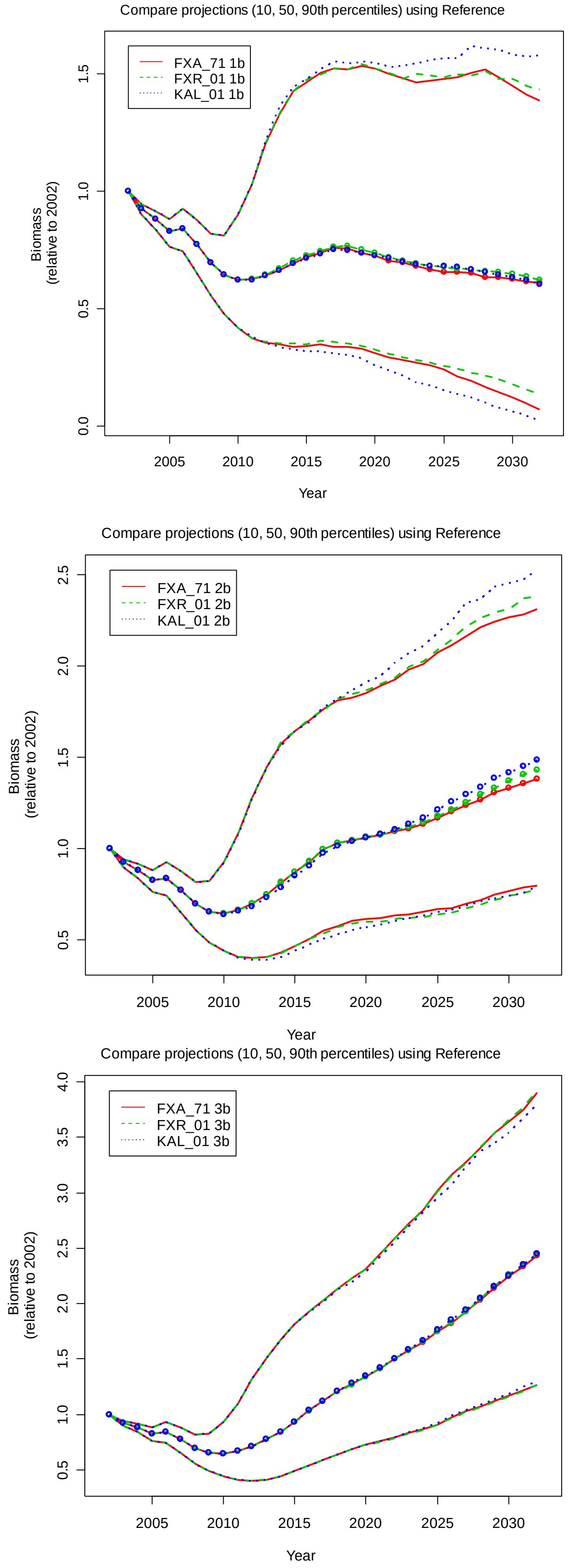


Figure 16: Comparison of the time trajectories for the median, 10th and 90th percentiles for spawning biomass relative to 2002 for the FXA_71, FXR_01 and KAL_01 for three different tuning levels (upper panel for 0.7, middle for 1.1 and lower for 1.5).

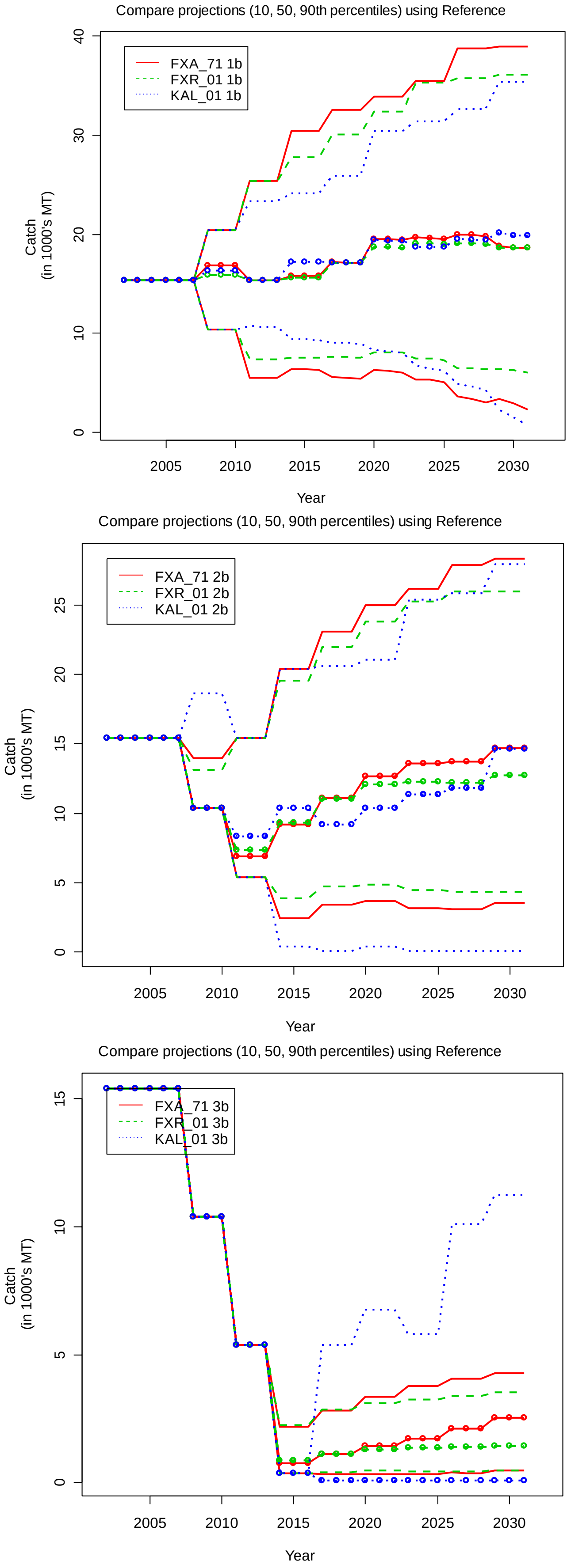
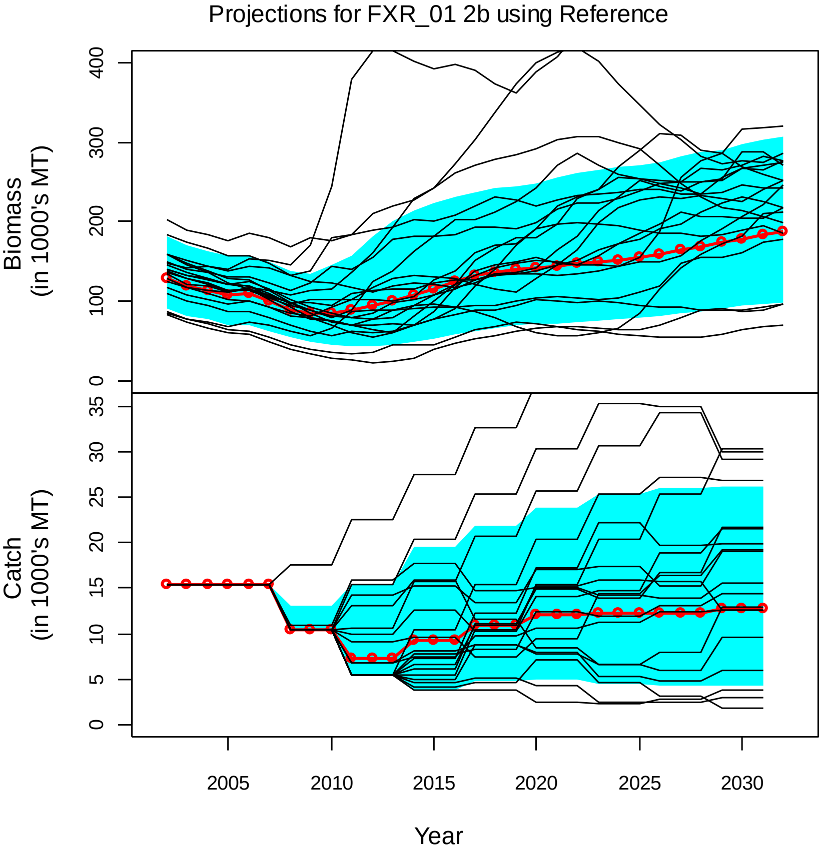
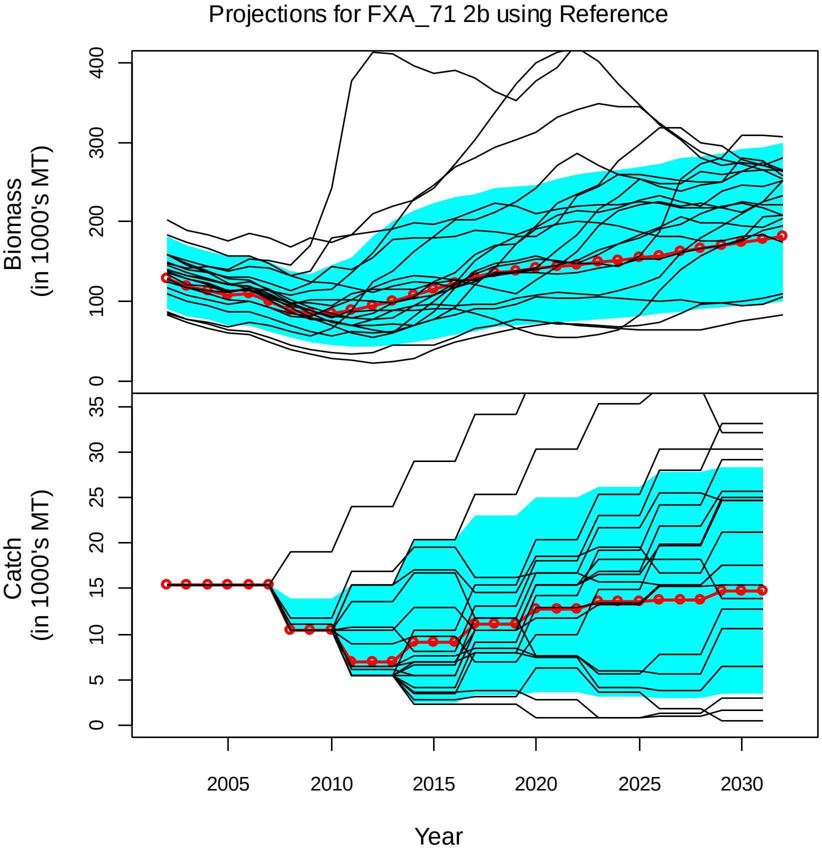


Figure 17: Comparison of the time trajectories for the median, 10th and 90th percentiles for catch for the FXA_71, FXR_01 and KAL_01 for three different tuning levels (upper panel for 0.7, middle for 1.1 and lower for 1.5).





1

Figure 18: Comparison of worm plots for the FXR_01, FXA_71 and KAL_01 decision rules for the 1.1 tuning level with a 3-year interval between TAC changes.

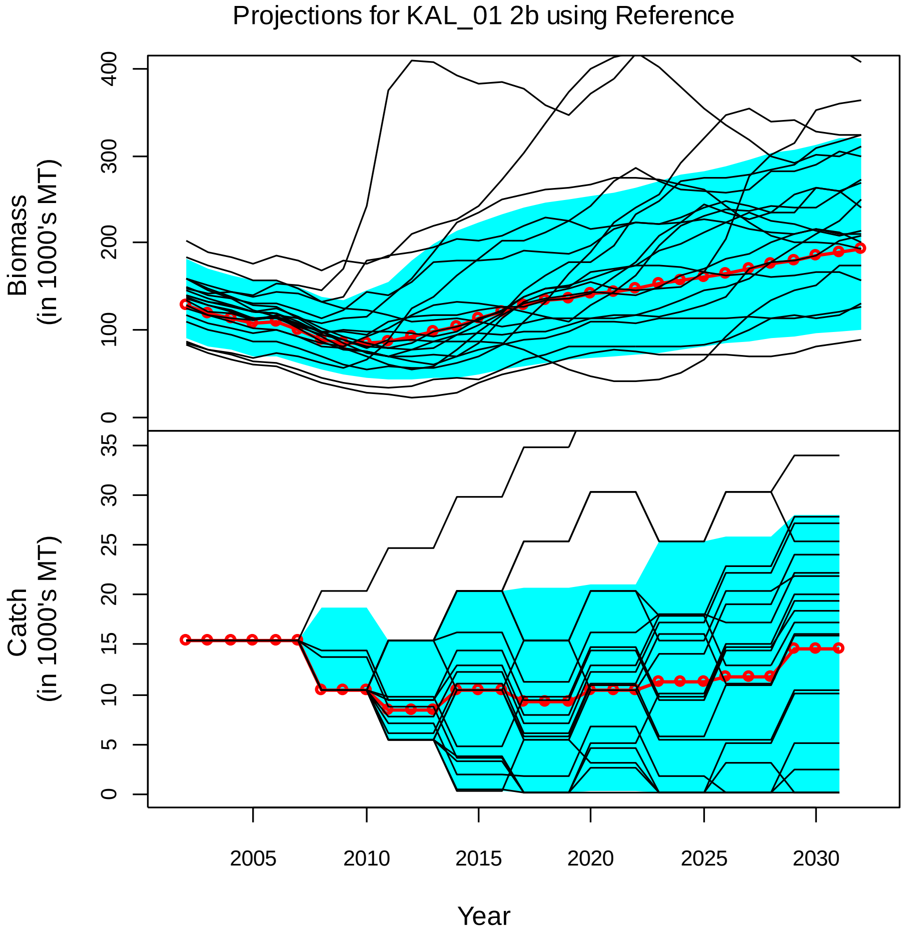


Figure 18(continued): Comparison of worm plots for the FXR_01, FXA_71 and KAL_01 decision rules for the 1.1 tuning level with a 3-year interval between TAC changes.

Figure 19: Comparison of the cumulative probability for the spawning biomass in 2032 relative to 2002 (B2032/B2002) for the Med1_A12 robustness test with the corresponding trial for the reference set for the FXA_71, FXR_01 and KAL_01 decision rule.

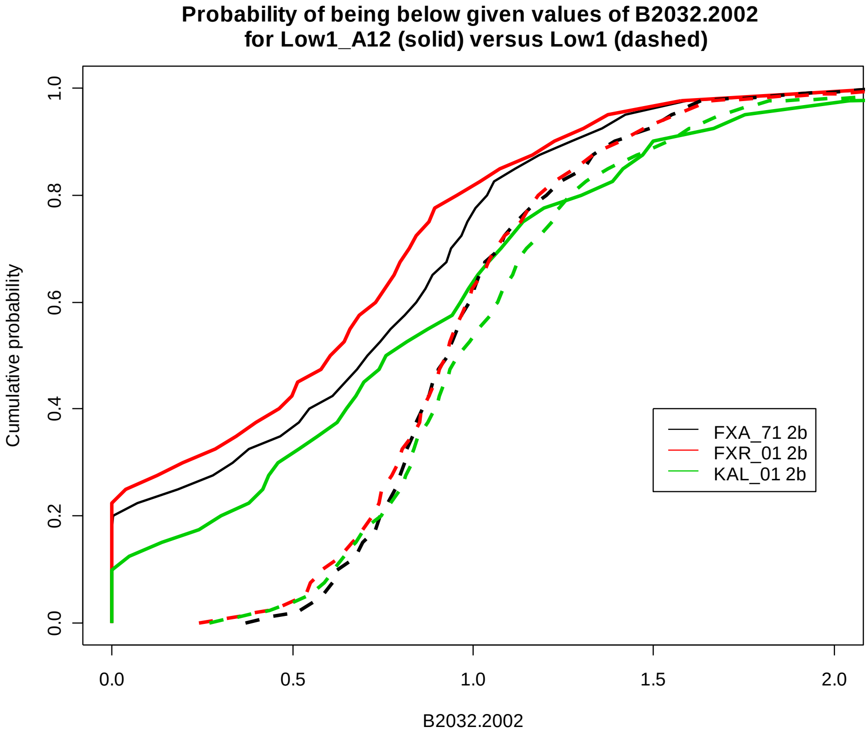


Figure 20: Comparison of the cumulative probability for the spawning biomass in 2032 relative to 2002 (B2032/B2002) for the Low1_A12 robustness test with the corresponding trial for the reference set for the FXA_71, FXR_01 and KAL_01 decision rule.

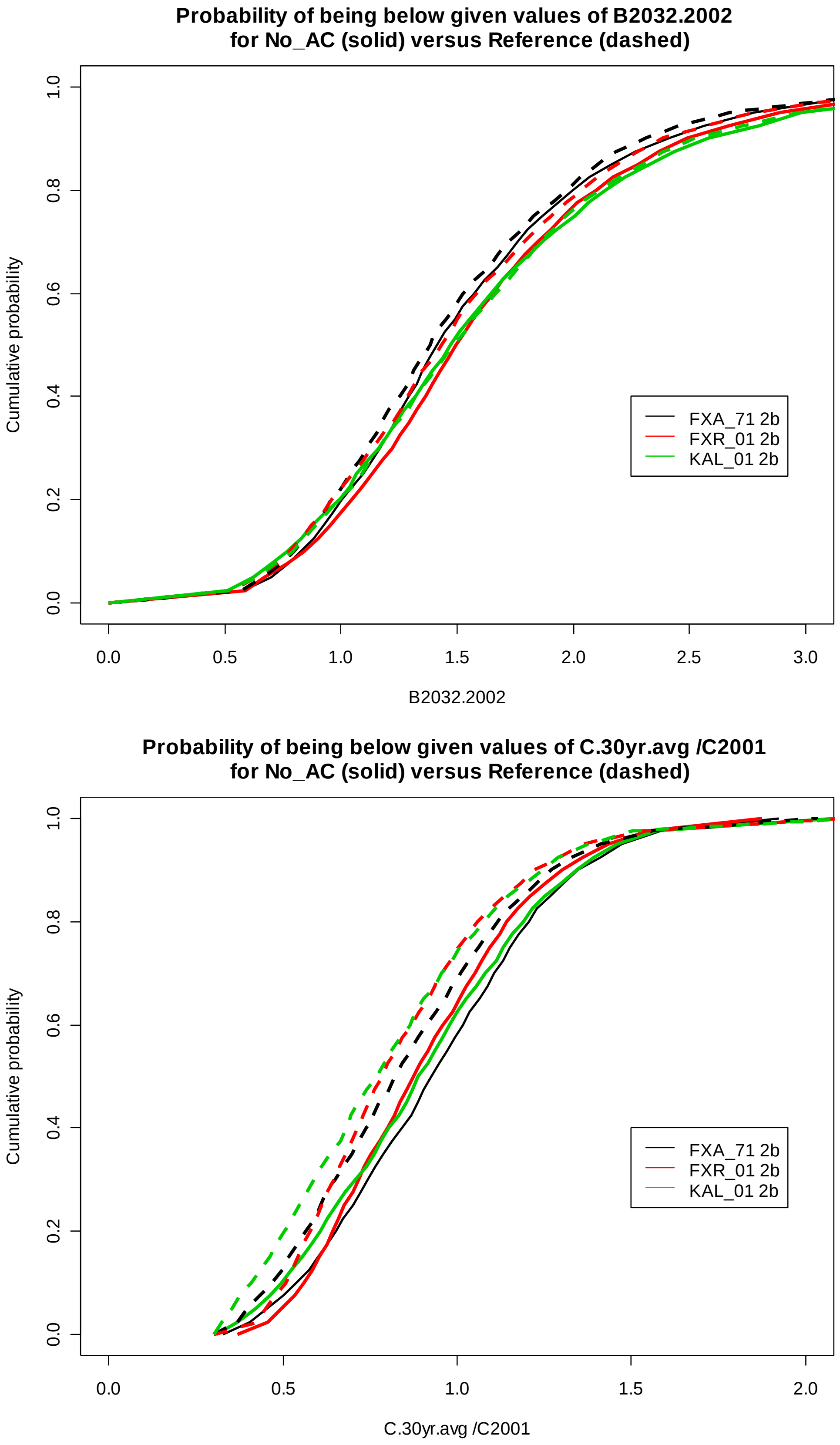


Figure 21: (Upper panel) Comparison of the cumulative probability for the spawning biomass in 2032 relative to 2002 (B2032/B2002) for the no-AC robustness test with the reference model for the FXA_71, FXR_01 and KAL_01 decision rules. (Lower panel) Same but for average 30-year catch instead of B2032/B2002.

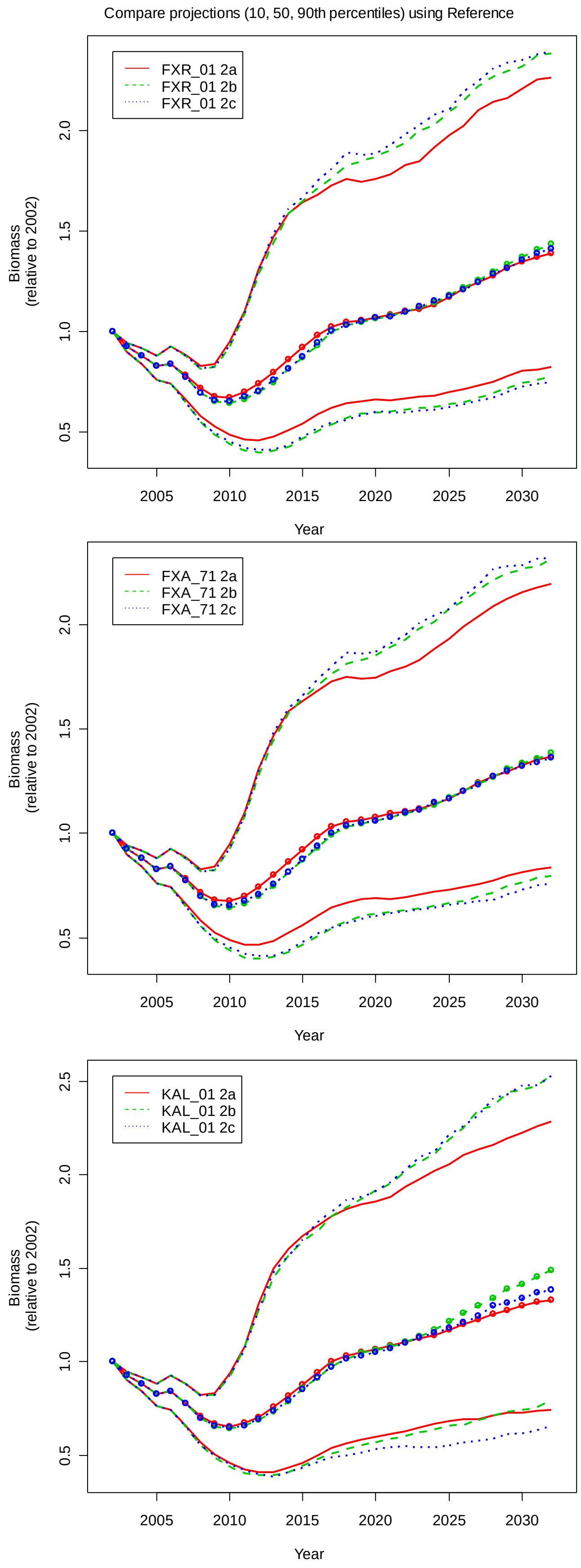


Figure 22: Comparison of the time trajectories for the median,10th and 90th percentiles for spawning biomass for a 1, 3 and 5 year interval between TAC changes (upper panel for FXR_01, middle panel for FXA_71 and lower panel for KAL_01). Note in the figure legends that 'a', 'b' and 'c' refer to the 1, 3 and 5 year changes respectively.

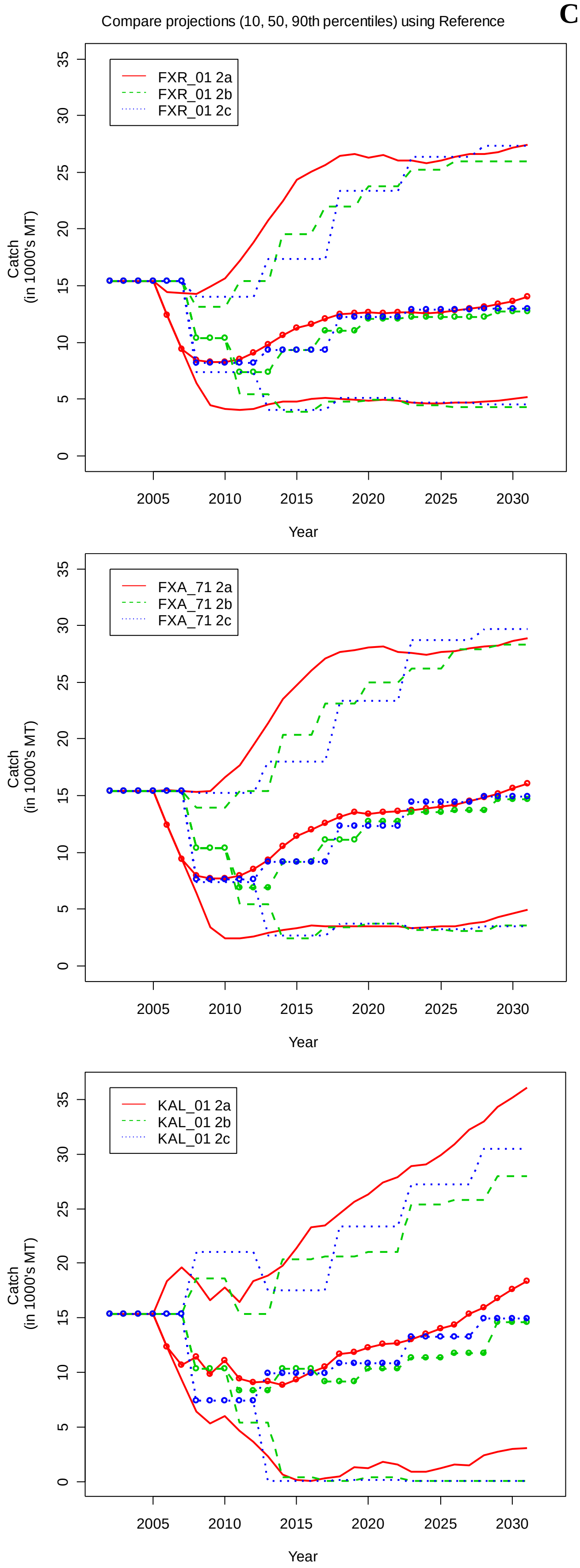
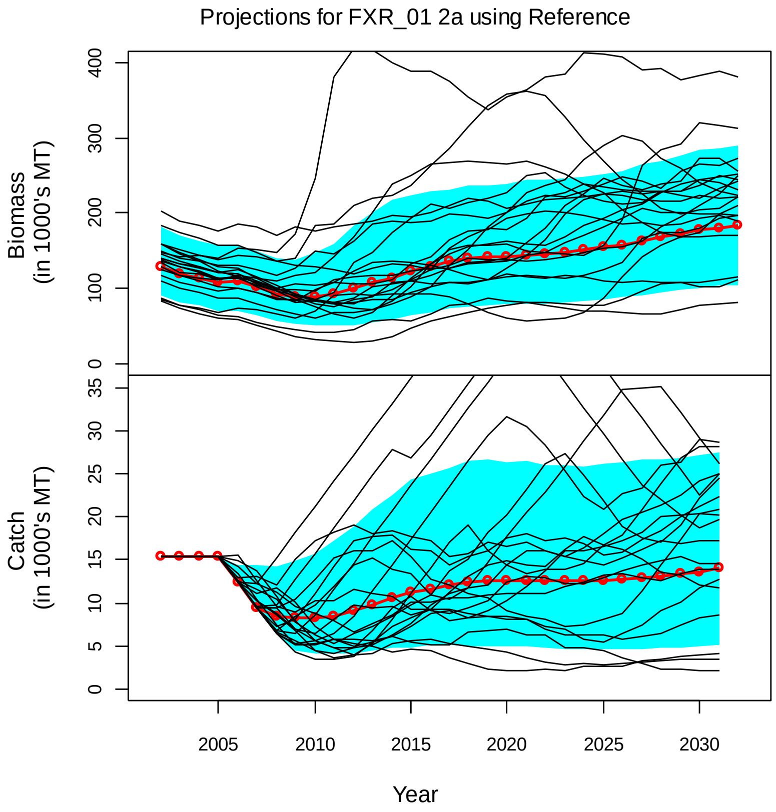


Figure 23: Comparison of the time trajectories for the median, 10th and 90th percentiles for catch for a 1, 3 and 5 year interval between TAC changes (upper panel for FXR_01, middle panel for FXA_71 and lower panel for KAL_01). Note in the figure legends that 'a', 'b' and 'c' refer to the 1, 3 and 5 year changes respectively.



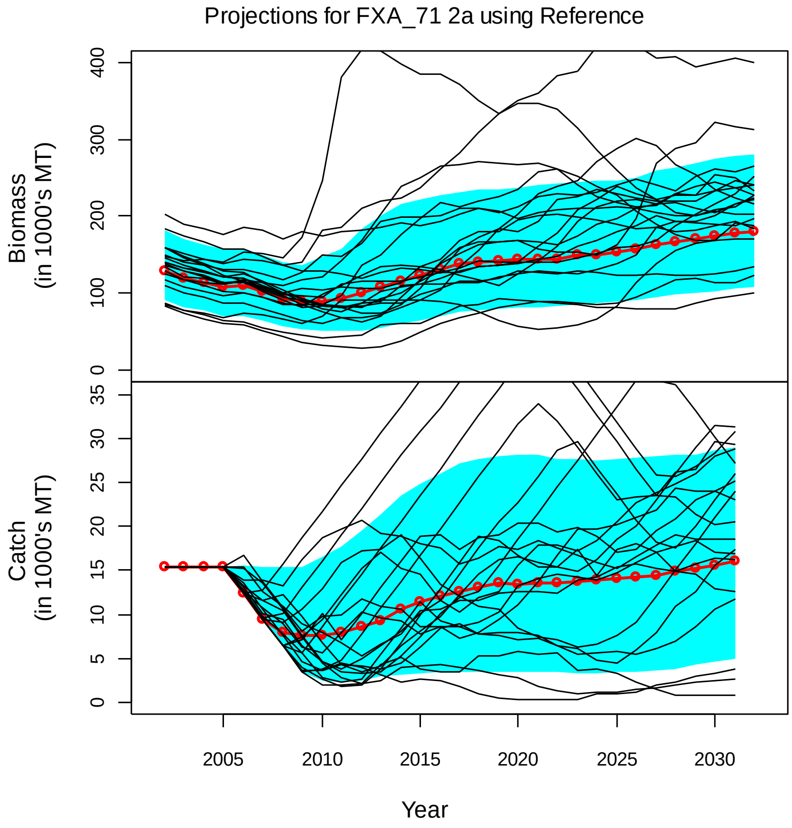


Figure 24: Comparison of worm plots for the FXR_01, FXA_71 and KAL_01 decision rules for the 1.1 tuning level with an annual interval between TAC changes.

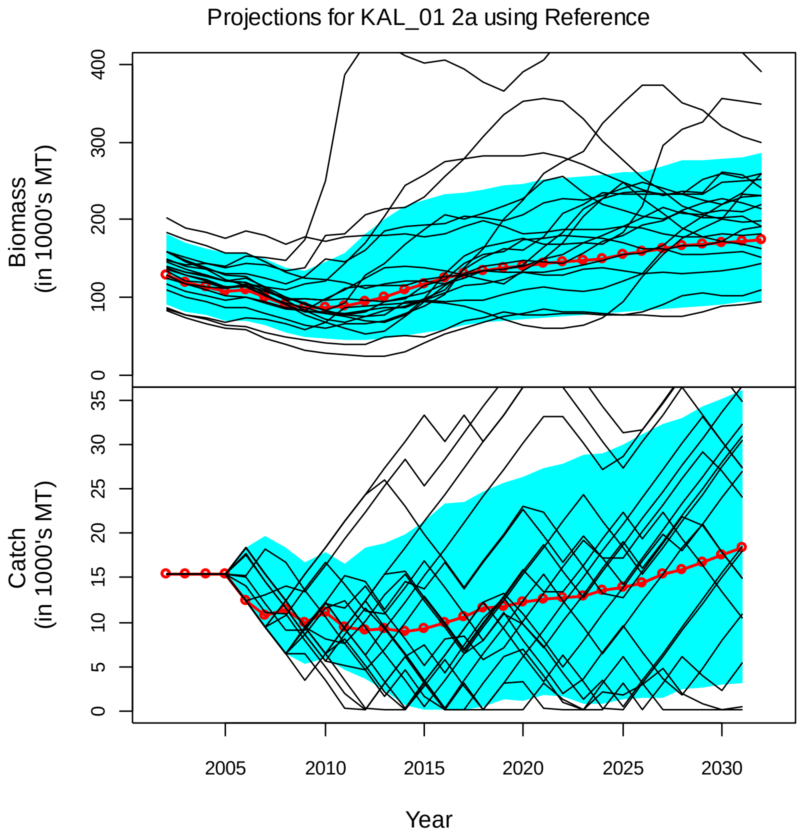


Figure 24 (continued): Comparison of worm plots for the FXR_01, FXA_71 and KAL_01 decision rules for the 1.1 tuning level with an annual interval between TAC changes.

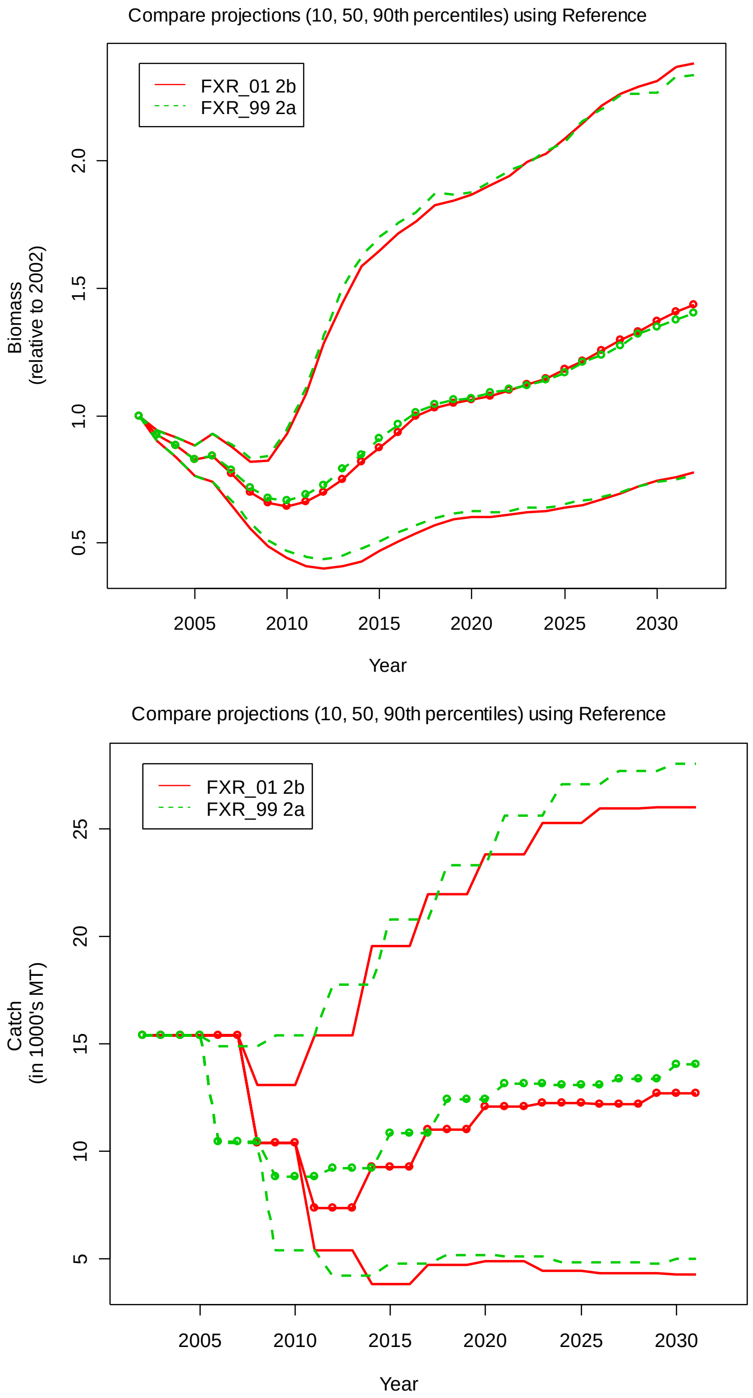


Figure 25: Comparison of the time trajectories for the median, 10th and 90th percentiles for spawning biomass and catch for a 2006 versus 2008 initial start year with a 3-year frequency change for the FXR_01 decision rule.