CSIRO
MARINE RESEARCH

# AN UPDATE OF THE GRAPHICS USED FOR EVALUATING THE PERFORMANCE OF CANDIDATE MANAGEMENT PROCEDURES FOR SOUTHERN BLUEFIN TUNA

**Paige Eveson**

**Prepared for the CCSBT 4th Stock Assessment Group Meeting
25-29 August 2003, Christchurch, New Zealand**

CCSBT-ESC/0309/23

# Table of Contents

## Abstract

A number of changes have been made to the graphical methods presented in the document CCSBT-MP/0304/05 (Eveson and Ricard 2003) for evaluating the performance of candidate management procedures (MP's). Changes were made to reflect the new operating model scenarios and the new evaluation criteria agreed upon at the April 2003 CCSBT-MP workshop in Queenstown, New Zealand. The graphs consist of: worm plots, which show the projected biomass and catch trajectories for a single MP and operating model scenario; comparison plots, which compare the performance statistics and robustness criteria obtained from multiple MP's for a single operating model scenario; and summary plots, which compare multiple MP's across all the reference case operating model scenarios. In addition, plots which compare the results obtained from a single robustness trial operating model scenario with those obtained from the corresponding reference case operating model scenario, and a plot which summarizes the results over all robustness trials, have been created. Each graph type is described in detail and an example given.

## Introduction

In this document, we provide an update of the graphical methods that have been created for evaluating and comparing candidate MP's. Graphics have been developed for the reference case operating model scenarios, as well as for the robustness trials that were agreed upon at the Second Meeting of the Management Procedure Workshop, April 2003, Queenstown, New Zealand. The updated graphics also take into account the new performance measures and robustness criteria that were proposed at the meeting.

We keep the same terminology used in the previous graphics document (Eveson and Ricard 2003); namely, we refer to the general principle for setting the TAC as a "decision rule" (e.g. keep the TAC at a constant *C*). The decision rule will most likely have tuning parameters (e.g. *C*), and a specific set of tuning parameters is referred to as a "version" of the rule. A decision rule plus a version define an MP.

Each MP is evaluated under eighteen reference case operating model scenarios, which include 3 steepness values (H30, H55, H80), 3 mortality vectors (M05, M10, M15) and 2 future catchability assumptions (Q0 and Q1), as well as 25 robustness trial operating model scenarios; for details refer to Anon. (2003). In addition, each MP and operating model is run using two simulation hierarchies (H1 and H3). Finally, each MP is run using the full posterior MCMC operating model (hieraracy H4). For hierarchies H3 and H4, we elected to always run 100 historical replicates and 1 projection replicate per historical replicate (only 1 historical and projection replicate needs to be run for H1 since this hierarchy assumes no errors).

All graphs are based on data in the .sum files produced by the CCSBT management procedures software "sbtprojv4". These files contain the median, 10th and 90th percentiles over all replicates of the 11 performance statistics outlined in items 29 and 31 of Anon. (2003). These files also contain the 3 TAC-related performance measures outlined in item 32 of Anon. (2003), and the evaluation of the robustness criteria proposed in item 33. Furthermore, the .sum files contain the median, 10th and 90th percentiles of the biomass projections for years 2002 to 2021 and the catch projections for years 2002 to 2020, as well as up to ten individual realizations of the biomass and catch trajectories (referred to as "bworms" and "cworms" respectively).

## Worm plots

The worm plots, which plot the biomass and catch trajectories for a particular MP under a particular operating model scenario, remain the same as described in Eveson and Ricard (2003) except that the 10th and 90th percentiles of the trajectories are now shown as shaded areas as opposed to error bars. The operating model scenario can be a reference case or a robustness trial. Again, the plotting function has the option to include anywhere from 0 to 10 bworms and cworms. Figure 1 gives an example of such a plot with 10 bworms and 10 cworms.

## Comparison plots

We again created graphs to compare the performance of multiple MP's under a single operating model scenario and hierarchy. The operating model can be either a reference case or a robustness trial, but usually a reference case since special plots have been created for the robustness trials (see below).

For each of the 8 original catch and biomass performance statistics (item 29 of Anon. 2003) and 3 additional ones (item 31 of Anon. 2003), a plot is created showing the median value along with error bars extending to the 10th and 90th percentiles. Again, the catch statistics labelled C 5yr.avg and C 20yr.avg have been rescaled such that they are relative to the 2001 catch level (i.e. the median, 10th and 90th percentile values have been divided by 15385.7 MT since this is the 2001 total catch value[1]). Thus, a value of 1.0 can be used as a reference point for how an MP performs with respect to current catch levels.

In addition, two TAC-related performance measures (those given in equations (13) and (14) of item 32 of Anon. 2003) are plotted. The first of these is the number of times the TAC changes go in opposite directions over all projection years and replicate runs; we plot it as the percent occurrence out of all possible cases, which is (# projection years – 2) *(# replicates). The second is the number of times the change in the TAC and the change in the spawning stock biomass go in opposite directions over all projection years and replicate runs; we again plot this as the percent occurrence but now the number of possible cases is (# projection years – 1) * (# replicates). Although a third TAC-related measure, referred to as the *A*-statistic, was proposed (equation (12) of Anon. 2003), this measure only makes sense when the TAC is set every year 3 years. We are currently setting the TAC yearly, so we have omitted this statistic for the time being.

Finally, 6 robustness criteria that an MP can either pass or fail (item 33 of Anon. 2003) are included in the comparison plots. All of the robustness criteria are displayed on a single plot, in what we have termed a "stoplight" plot. If an MP passes the criteria, then its "stoplight" is left white (i.e. it is shown as an open circle); if it fails, then its "stoplight" is coloured black (i.e. it is shown as a black filled circle). So when looking at a stoplight plot, remember that **black is bad**. Note that the third robustness criteria, MedianB2022>MedianB2021>MedianB2020, was only deemed

[1] The problem of the total catch in weight differing depending on the operating model has been resolved.

relevant for models with steepness 0.3, so for models with other steepness values the stoplights are drawn as open circles with an "x" through them to indicate that they are not applicable. Also note that the fourth to sixth robustness criteria are quite complicated in that the fourth criteria (namely whether Pr[B2020>B1980]>0.9) is only to be calculated if the operating model can meet this criteria when the TAC is set at 75% of current catch. Otherwise, the fifth and sixth criteria are to be calculated instead. Because none of the reference case operating models meet the criteria under 75% of current catch, this would mean that only the fifth and sixth criteria would ever be relevant. We have calculated the fourth criteria anyway and left it as one of the stoplights being plotted because there are some MP's that under certain operating models can achieve this objective (MP's that are either more conservative than 75% current catch or else are "smarter" in the way they set the TAC).

Figure 2 shows an example of a comparison plot. The comparison plots now have an option to make the y-axis scale for each panel fixed at values preset by the software, or else variable with the range being determined by the minimum and maximum values being plotted. Fixed axes are useful for comparing between comparison plots for different operating models, however variable axes are useful when just looking at the plot for one operating model since the differences between MP's are more obvious. Variable axes were chosen in the figure shown.

Note that on all of the comparison plots, the results from running an MP that sets the TAC to zero in all years is included as a reference (labelled "no catch"). The no catch case shows the maximum conservation (i.e. biomass) performance that can be attained under the operating model scenario, and serves as a realistic basis for evaluating the conservation level achieved by other MP's.

## Summary plots

The comparison plots contain a lot of information that can be difficult to process without careful inspection. For example, we often have a selection of MP's whose performance we want to compare across operating model scenarios; this would currently require comparing the output from multiple graphs. Furthermore, from the graphs presented so far, it is difficult to evaluate the trade-off between catch performance and conservation (i.e. biomass) performance. Consequently, we have created summary graphs to assist with such assessments. Although some of the summary plots are similar to those presented in Eveson and Ricard (2003), a number of modifications have been made and some new graphics have been added, so for completeness we describe them all in detail.

The first type of summary graph concentrates on just one measure of catch performance and one measure of biomass performance, as specified by the user. By default, the average catch over a 20 year projection period (C 20yr.avg) and the ratio of spawning stock biomass in 2020 to 1980 (B2020:1980) are used, but any of the 11 performance statistics outlined in items 29 and 31 of Anon. (2003) can be used. Figure 3 shows a summary graph comparing multiple MP's across all reference case operating model scenarios; the top two plots are for the 9 reference case models with constant future catchability (Q0) and the bottom two plot are for the 9 reference case models with increasing future catchability (Q1). The plots on the right- and left-hand sides contain the same information, just expressed in different ways. On the left-hand side is what we term a "star plot", with each star corresponding to a particular MP.

The centroid of the star represents the mean catch performance value (for the catch statistic chosen) and mean biomass performance value (for the biomass statistic chosen) across the 9 reference models under consideration. It is marked with a circle whose radius is proportional to the inter-annual variability in the catch (AAV), so that a smaller circle is desirable. Each star has nine rays, where each ray extends to the catch, biomass coordinate obtained under one of the model scenarios. The star plot does not allow one to distinguish which ray corresponds to which scenario, which is necessary if one wishes to compare performance of MP's under a particular scenario. This gives rise to the plots on the right-hand side of the page, where the results from a given model scenario are connected by a line and labelled accordingly.

The above summary plots were described using the median values of the chosen catch and biomass performance statistics. Using the median does not allow one to see how much uncertainty exists in the performance statistics. As such, the plotting function has the option to use the 10th percentiles for the statistics rather than the medians (this of course will not make any difference when considering hierarchy 1).

We have created a second type of summary graph that attempts to summarize the results of the robustness criteria (see Figure 4). The top panel shows the percent of robustness criteria that are passed by each of the reference case operating models, as represented by the percent of the box shaded grey – no shading means 0% of the robustness criteria have been passed, fully shaded means 100% of the criteria have been passed, et cetera. So when looking at these plots, remember that **grey is good**. For operating models with steepness 0.3, the number of criteria that may be passed is 6, whereas for all other models it is 5 (omitting the Median2022>Median2021 >Median2020 criteria).

Looking at the same information in a different way, the lower panel shows the percent of the reference case models (both Q0 and Q1) that pass each of the robustness criteria. No shading means 0% of the models pass the criteria, fully shaded means 100% of the models pass the criteria, et cetera. So, once again, when looking at these plots, remember that **grey is good**. For the criteria Median2022>Median2021> Median2020, the relevant number of models that can be passed is 6 (those with steepness 0.3), whereas for all other criteria the number of models that can be passed is 18.

As with the comparison plots, all of the summary plots include the results from running an MP that sets the TAC to zero in all years as a reference (labelled "no catch").

## Robustness Plots

In addition to the 18 reference case operating models, 26 robustness trial models a (including the MCMC case) have been added to evaluate how sensitive the results are to certain parameters and assumptions. We have created some graphs to help determine which of the robustness trials have substantial effects on the performance statistics and robustness criteria obtained for a set of MP's. These graphs look exactly like the comparison plots described above except each panel now includes the results from two operating models, one being a robustness trial and the other being a reference case. Figures 5 and 6 give examples. By default, a robustness model is compared to the corresponding reference case model (i.e. H30M10Q0_Psi is compared to H30M10Q0). It is not clear what the corresponding reference model for comparison should be for H__M10Q0_CC and H__M__Q0 (the mcmc model), but we have chosen to make the default H55M10Q0 for both of these. Note, however, that the user has the ability to choose any reference model for comparison with any robustness model.

With 26 robustness models that means there are 26 comparison plots to consider for a set of MP's. We have attempted to summarize the information into a single graph (Figure 7). On the x-axis are all 26 robustness trial models, along the y-axis are all of the evaluation measures (the 11 catch and biomass performance statistics, the 2 TAC-related performance measures, and the 6 robustness criteria). The plot shows the percent of MP's (as chosen by the user and indicated in the subtitle) whose evaluation measures differ substantially between each robustness model and its corresponding reference model. The percent of the box that is shaded grey represents the percent of MP's whose evaluation measures differ substantially between the robustness model and the reference model, so **grey means take heed**. The definition of a substantial difference depends on the type of evaluation measure. For the 11 catch and biomass performance statistics, a substantial difference is defined as more than a 10% change in the median value relative to the reference case (i.e. $|\text{median}_{\text{robust}} - \text{median}_{\text{reference}}| / \text{median}_{\text{reference}} > 0.1$) or more than a 20% change in the range from the 10th to 90th percentile (i.e. $|(90^{th} - 10^{th})_{\text{robust}} - (90^{th} - 10^{th})_{\text{reference}}| / (90^{th} - 10^{th})_{\text{reference}} > 0.2$). For the TAC-related measures, these are already expressed as percent occurrences, so a substantial difference is defined as an absolute change of more than 10% (i.e. $|\text{TAC\_measure}_{\text{robust}} - \text{TACmeasure}_{\text{reference}}| > 0.1$). For the 6 robustness criteria, a substantial change is simply whether or not the result changes between the robustness model and the reference model (i.e. the robustness model passes the criteria where the reference model fails, or vice versa).

Looking up a column for a particular robustness model, a lot of grey is a flag that the results are sensitive to the change being tested by that robustness trial and one should take a closer look at the comparison plots to better understand the differences.

Note there is an "x" in the box for robustness criteria Median2022> Median2021> Median2020 when the robustness model assumes a steepness value other than 0.3 because this criteria is only relevant when steepness equals 0.3.

## Discussion

A number of graphics have been proposed to: 1) help compare and evaluate candidate MP's under the reference case operating model scenarios, and 2) evaluate the sensitivity of the results to changes in key model parameters and assumptions (through plotting the results from the robustness trial operating models).

With respect to 1), the summary plots that have been proposed are intended to help identify general trends and key features. They mask a lot of important information which can only be obtained by looking at the more detailed comparison plots. For example, the summary star plots (Figure 3) do not show the amount of variability in the catch and biomass performance statistics being plotted. We attempted to deal with this issue by adding the option to plot the lower 10th percentiles as opposed to the median values, but one still must look to the comparison plots (Figure 2) for a more complete picture. Neither the summary star plots nor the comparison plots can be used to assess the trend in the biomass and catch trajectories. For example, two MP's may lead to the same average catch over the 20-year projection period but one may achieve this with a constant catch level and another with a steadily increasing trend in catch. The worm plots (Figure 1) are important for investigating these issues.

With respect to 2), it is important to keep in mind in interpreting the robustness plots (Figures 5-7) that the results are dependent on the MP's being considered; the parameter or assumption being tested by a particular robustness trial may have little effect on the results for some MP's but large effects for other MP's. Also, the definitions of substantial differences that were used to create the summary plot in Figure 7 are somewhat arbitrary and open to discussion.

## Literature Cited

Anonymous. 2003. Report of the Second Meeting of the Management Procedure Workshop. CCSBT. Queenstown , New Zealand, 7-9 & 12, 14-15 April 2003.

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.

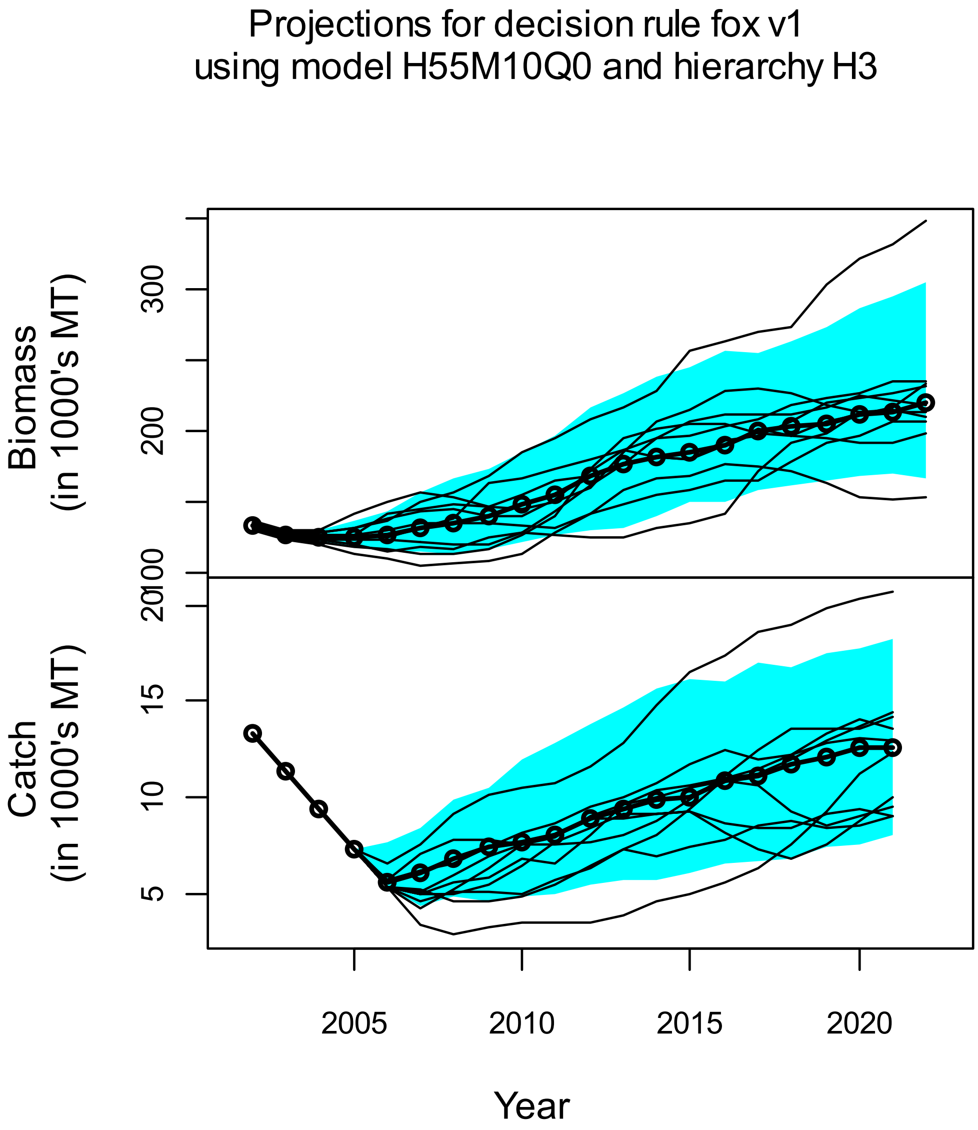

Projections for decision rule fox v1
using model H55M10Q0 and hierarchy H3
Biomass
(in 1000's MT)
300
200
100
Catch
(in 1000's MT)
20
15
10
5
2005
2010
2015
2020
Year

Figure 2.

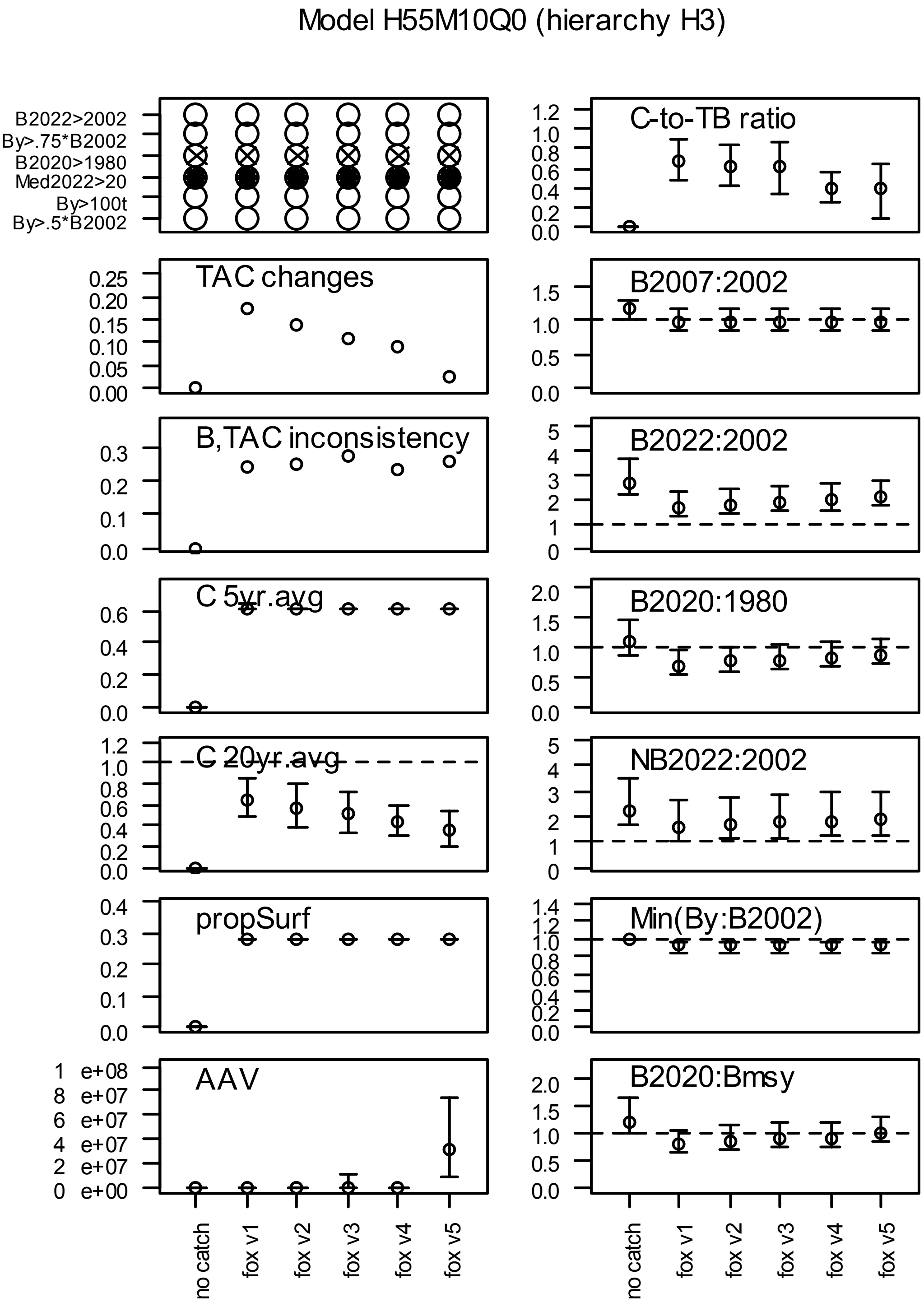

Model H55M10Q0 (hierarchy H3)
B2022>2002
By>.75*B2002
B2020>1980
Med2022>20
By>100t
By>.5*B2002
C-to-TB ratio
TAC changes
B2007:2002
B,TAC inconsistency
B2022:2002
C 5yr.avg
B2020:1980
C 20yr.avg
NB2022:2002
propSurf
Min(By:B2002)
AAV
B2020:Bmsy
no catch
fox v1
fox v2
fox v3
fox v4
fox v5

Figure 3.

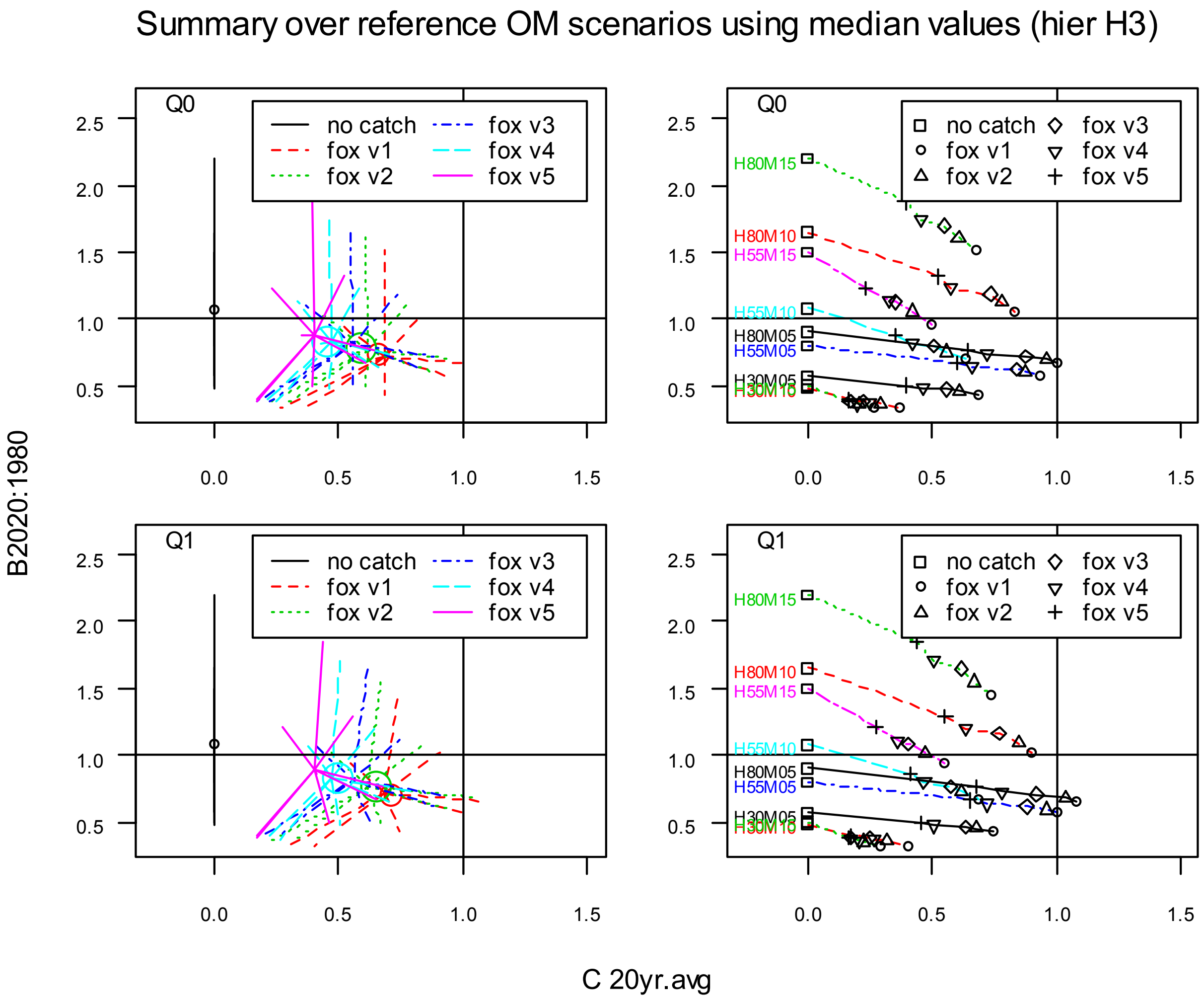
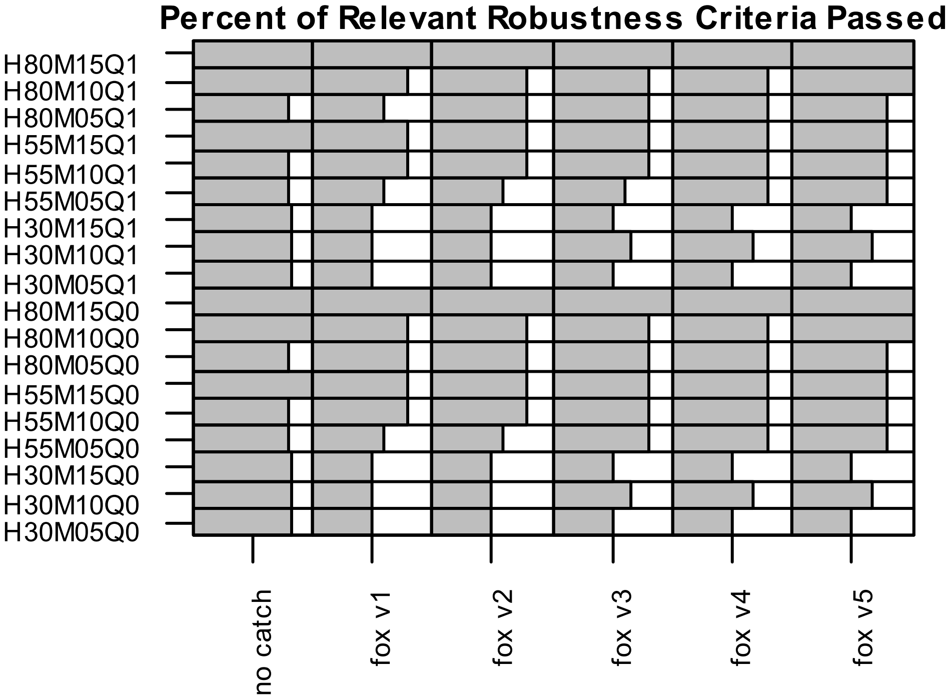

Summary over reference OM scenarios using median values (hier H3)
Q0
Q1
no catch
fox v1
fox v2
fox v3
fox v4
fox v5
H80M15
H80M10
H55M15
H55M10
H80M05
H55M05
H30M05
H30M10
0.5
1.0
1.5
2.0
2.5
0.0
B2020:1980
C 20yr.avg

Figure 4.


Percent of Relevant Robustness Criteria Passed
H80M15Q1
H80M10Q1
H80M05Q1
H55M15Q1
H55M10Q1
H55M05Q1
H30M15Q1
H30M10Q1
H30M05Q1
H80M15Q0
H80M10Q0
H80M05Q0
H55M15Q0
H55M10Q0
H55M05Q0
H30M15Q0
H30M10Q0
H30M05Q0
no catch
fox v1
fox v2
fox v3
fox v4
fox v5

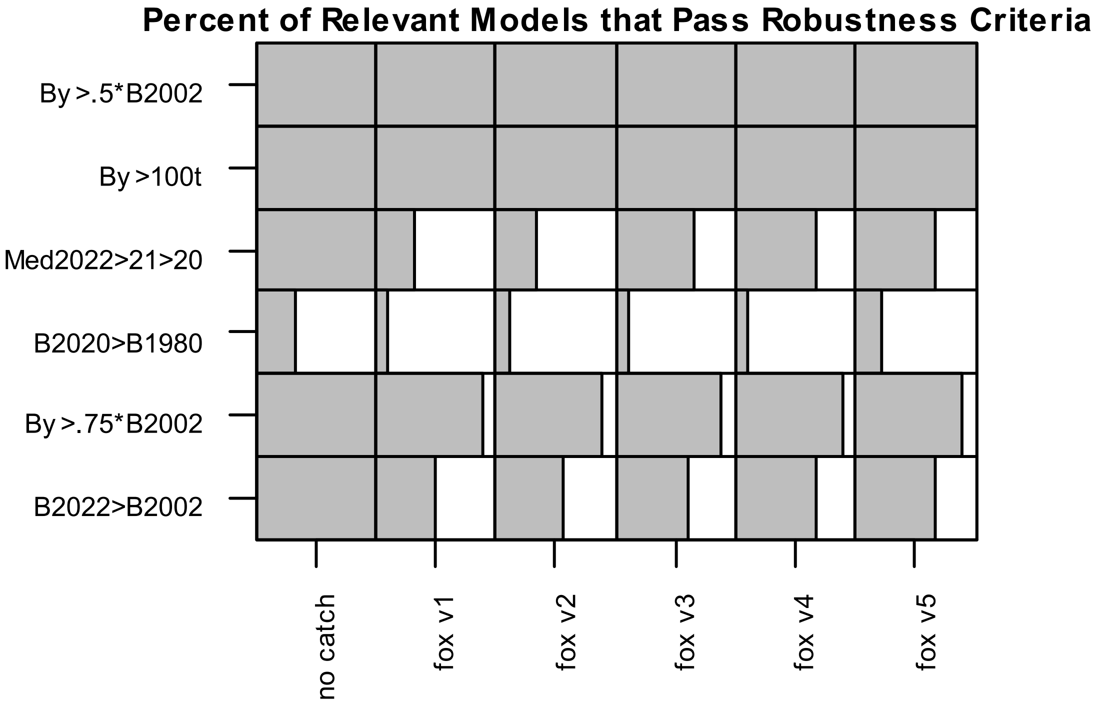

Percent of Relevant Models that Pass Robustness Criteria
By >.5*B2002
By >100t
Med2022>21>20
B2020>B1980
By >.75*B2002
B2022>B2002
no catch
fox v1
fox v2
fox v3
fox v4
fox v5

Figure 5.

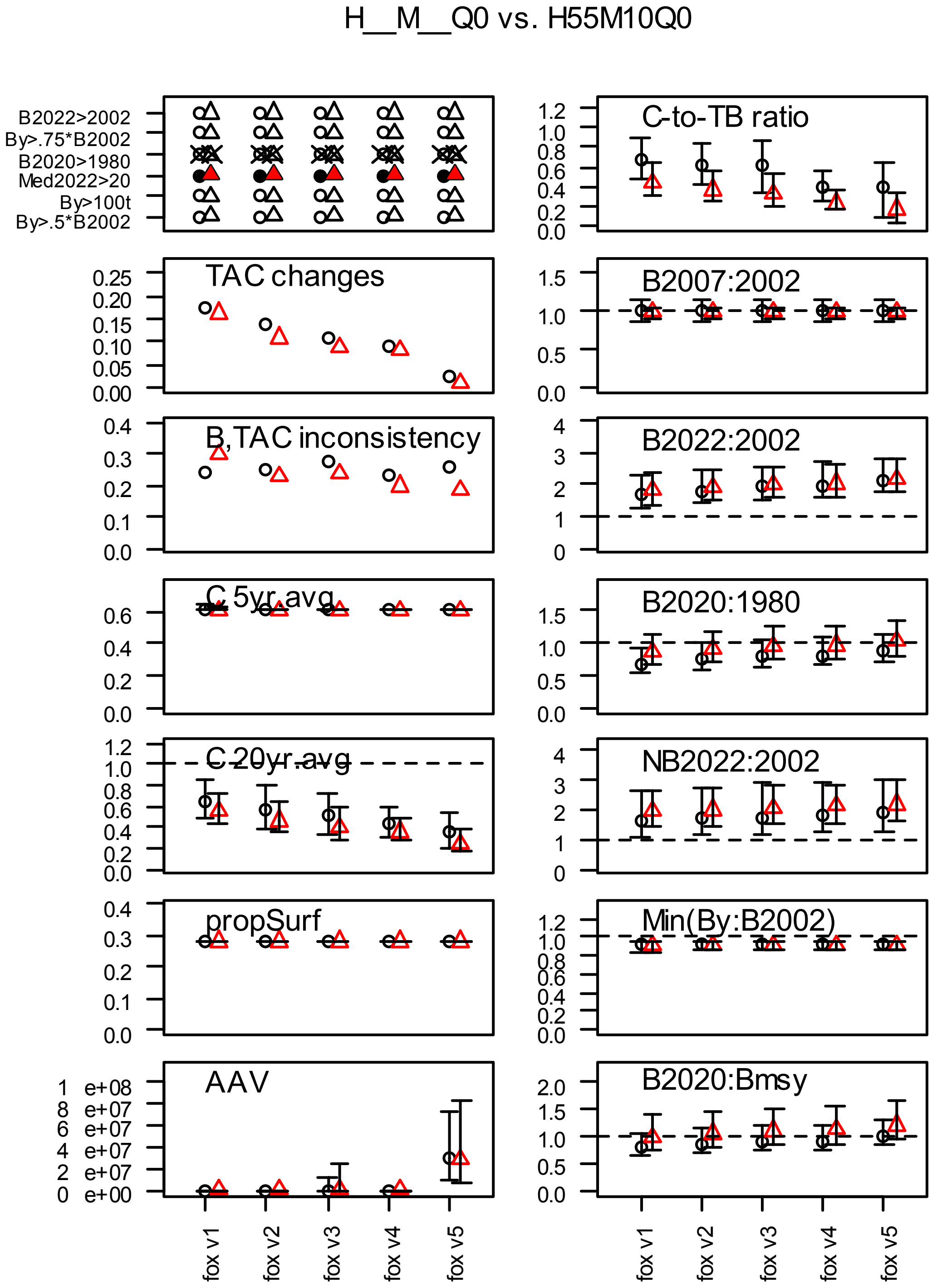

H__M__Q0 vs. H55M10Q0
B2022>2002
By>.75*B2002
B2020>1980
Med2022>20
By>100t
By>.5*B2002
C-to-TB ratio
TAC changes
B2007:2002
B,TAC inconsistency
B2022:2002
C 5yr avg
B2020:1980
C 20yr avg
NB2022:2002
propSurf
Min(By:B2002)
AAV
B2020:Bmsy
1 e+08
8 e+07
6 e+07
4 e+07
2 e+07
0 e+00
fox v1
fox v2
fox v3
fox v4
fox v5

Figure 6.

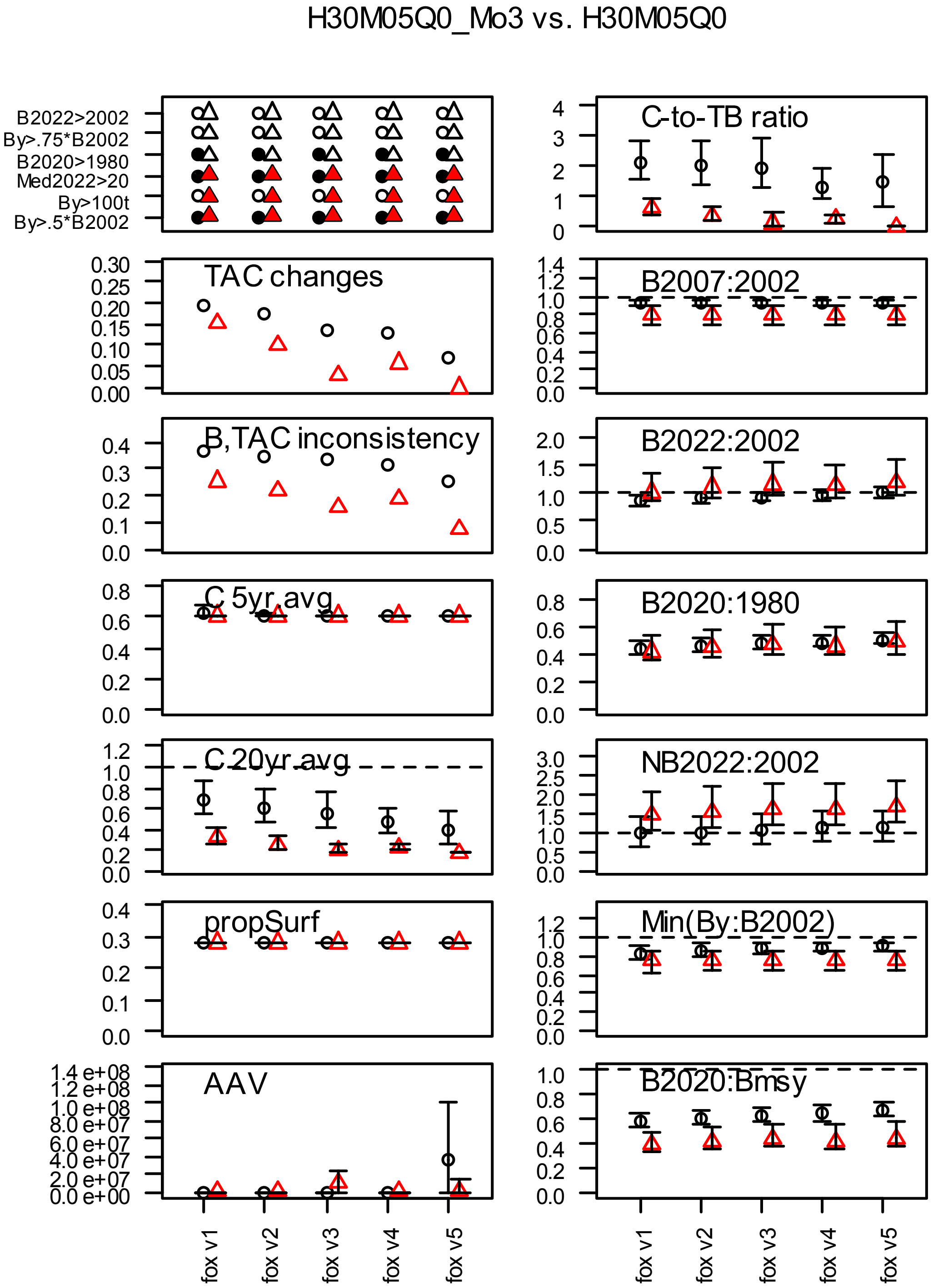
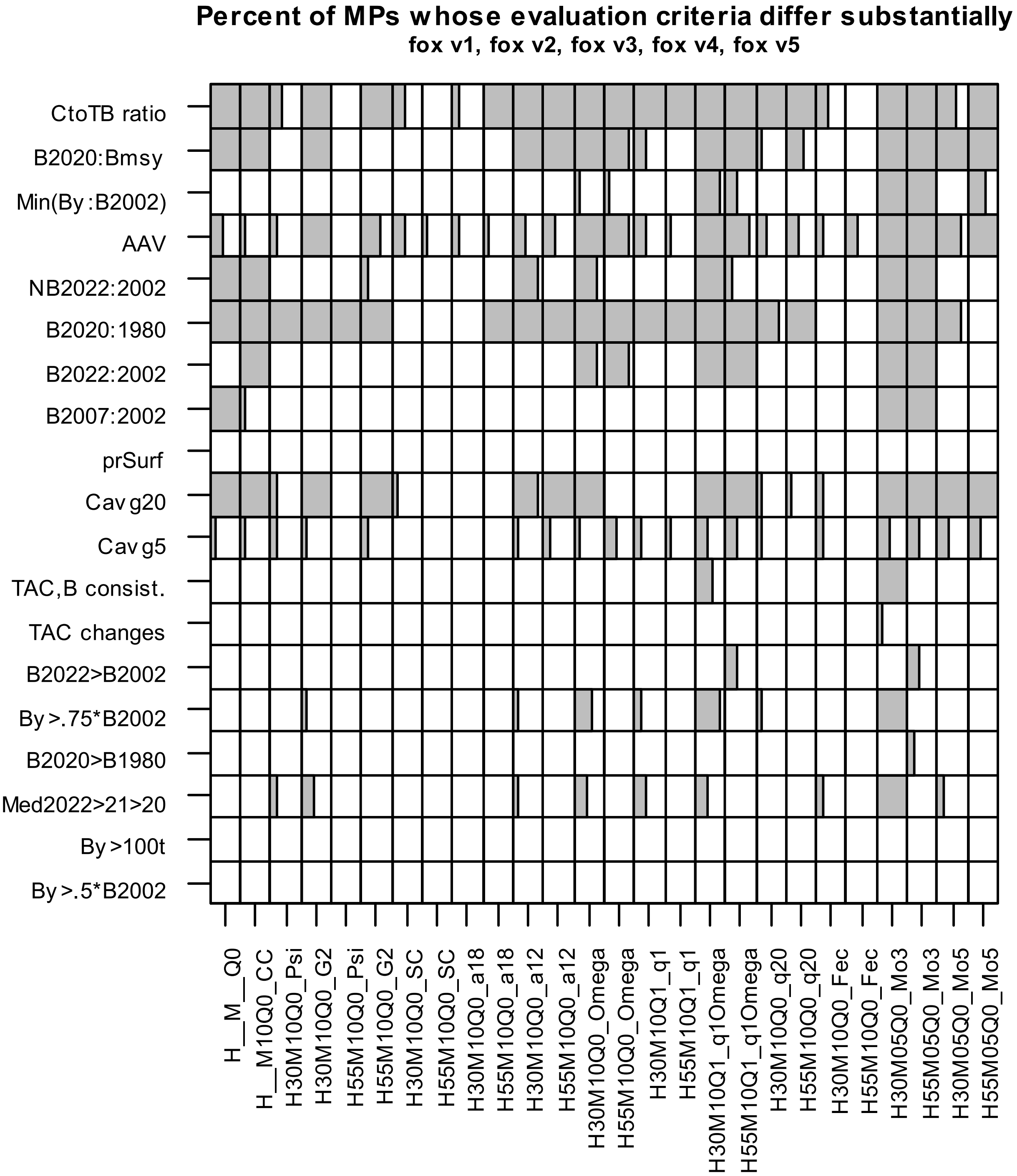

H30M05Q0_Mo3 vs. H30M05Q0
B2022>2002
By>.75*B2002
B2020>1980
Med2022>20
By>100t
By>.5*B2002
C-to-TB ratio
TAC changes
B2007:2002
B,TAC inconsistency
B2022:2002
C 5yr avg
B2020:1980
C 20yr avg
NB2022:2002
propSurf
Min(By:B2002)
AAV
B2020:Bmsy
fox v1
fox v2
fox v3
fox v4
fox v5

Figure 7.


Percent of MPs whose evaluation criteria differ substantially
fox v1, fox v2, fox v3, fox v4, fox v5
CtoTB ratio
B2020:Bmsy
Min(By:B2002)
AAV
NB2022:2002
B2020:1980
B2022:2002
B2007:2002
prSurf
Cavg20
Cavg5
TAC,B consist.
TAC changes
B2022>B2002
By>.75*B2002
B2020>B1980
Med2022>21>20
By>100t
By>.5*B2002
H__M__Q0
H__M10Q0_CC
H30M10Q0_Psi
H30M10Q0_G2
H55M10Q0_Psi
H55M10Q0_G2
H30M10Q0_SC
H55M10Q0_SC
H30M10Q0_a18
H55M10Q0_a18
H30M10Q0_a12
H55M10Q0_a12
H30M10Q0_Omega
H55M10Q0_Omega
H30M10Q1_q1
H55M10Q1_q1
H30M10Q1_q1Omega
H55M10Q1_q1Omega
H30M10Q0_q20
H55M10Q0_q20
H30M10Q0_Fec
H55M10Q0_Fec
H30M05Q0_Mo3
H55M05Q0_Mo3
H30M05Q0_Mo5
H55M05Q0_Mo5