Identifying Scenario Spaces, their Outliers and Redundancies, with Application to Bowhead AWMP Evaluation Trials

Geof H. Givens and Marsha S. Huang*

May 24, 2000

Abstract

This paper describes how to characterize groups of simulation trial scenarios designed for testing Aboriginal Whaling Management Procedures (AWMPs). Specific features of the trials established for testing bowhead whale AWMPs are also examined.

Trial scenarios are determined by a very high-dimensional parameter vector. The dimensionality and potential for important interactions between parameters mean that it is very difficult to identify outlying or redundant scenarios, or to see which scenarios share common characteristics that are appropriate for certain scenario classes.

The four classes of scenarios identified by the Scientific Committee are reviewed, along with the appropriate relationships between these classes. We propose characterizing scenarios by a five-statistic summary based on population trajectory features rather than on the input parameter vector. This five-number summary can be used for identifying outlying or redundant scenarios, as we demonstrate in an example. We also discuss how to use this summary to ensure that scenarios intended for a common purpose do not stray from the region of parameter space useful for testing that purpose.

The examples in this paper are made in the context of trials established for testing AWMP management of the Bering-Chukchi-Beaufort Seas stock of bowhead whales, yet the ideas could be more widely applied to other management procedure development and testing contexts. Our results motivate a list of suggested changes to the bowhead AWMP trial framework.

INTRODUCTION

Aboriginal Whaling Management Procedure (AWMP) Strike Limit Algorithms (SLAs) will be tested by the International Whaling Commission (IWC) Scientific Committee (SC) using

*Geof H. Givens is Assistant Professor of Statistics and Marsha S. Huang is Graduate Research Assistant, both at the Department of Statistics, Colorado State University, Fort Collins, CO 80523.

computer simulation. Each candidate SLA will be subjected to replicated simulations of a wide variety of scenarios. For each scenario, a computer model will be used to simulate the dynamics of a whale stock and data arising from the hypothetical study thereof, and the SLA must set strike limits on the basis of available data without knowledge of the true stock status or dynamics. Replication is used because the simulated data provided to the SLA vary randomly between replicates of a given scenario.

Scenarios are determined by specific values for biological and other parameters underlying the simulation model and by certain modeling assumptions about population dynamics. Other variables that determine a scenario include parameterized assumptions about the natural environment and future hunting, including time series of future numbers of strikes 'needed' by the hunting community.

'Need' is determined by the IWC from time to time through a political process outside the purview of the SC. A successful SLA must be able to satisfy as much need as possible while meeting the IWC's overriding objectives of whale population increase for depleted stocks and avoidance of population extinction. For a more detailed review of IWC objectives and priorities for the AWMP development process, see IWC (1995; 2000a).

Let the vector of parameters which determine a scenario be written as $\boldsymbol{\theta}$. Let the identity of the population dynamics model itself be an element of $\boldsymbol{\theta}$. The SC has begun to select scenarios, i.e. instances of $\boldsymbol{\theta}$, which will be used to test candidate AWMPs. Recently, 125 trial scenarios were identified (IWC, 2000b) for testing SLAs intended for management of the Bering-Chukchi-Beaufort Seas stock of bowhead whales (hereafter 'bowheads'), and it was noted that more scenarios might be developed as required for comprehensive testing. Examples below relate to management of this bowhead stock, however the principles discussed here are more general.

Scenarios are parameterized by a very large number of variables. Table 1 shows the first portion of the primary input file to the simulation testing program, which includes specification of most key elements of $\boldsymbol{\theta}$. Eight hundred additional parameters of lesser importance are stored in another input file. It is very difficult to identify outlying scenarios because complex parameter interactions may exist and scenarios whose parameterization amounts to hidden extrapolation are easily concealed in such a high-dimensional parameter space.

This paper considers issues related to the selection of scenarios. Particularly, we address the question of how to characterize scenarios as typical or as outlying. For example, one might ask whether a particular scenario is redundant, or whether it lies outside some range of plausibility implied by the assumptions spanned by other scenarios. Such questions revolve around an idea we call a 'scenario space'. These issues are important in order to maintain the fairness and limit the complexity of the AWMP trial framework.

There are several classes of scenarios used to test any AWMP; the different classes have different purposes. It would be unfair to developers and potentially misleading to AWMP evaluators if scenarios from a parameter space region suited for one purpose were used to test AWMPs for a different purpose. Therefore, an important task of the SC is to understand where scenario class boundaries lie and to ensure that trials for a particular intended purpose

```
MANAGEMENT PARAMETERS         CASE  BE01                 YEAR OF 1ST GENERATED SURVEY         IGSURV 2002
Base Case: 2.5%, Need=67->201                            STRATEGIC SURVEYS?       0=No 1=Yes  STRATG    0
NAME OF PARAMETER FILE               FILNAM  BE01.PAR    DETERMINISTIC ABUNDANCE  0=No 1=Yes  OPTDET    0
No. OF TRIALS                        NTRIAL  100         CV OPTION: READ IN 0=ETA or 1=CVTRUE OPTCV     1
No. OF YEARS IN SIMULATION           NYEAR   100         FREQUENCY OF ABUNDANCE ESTIMATES     IFREQ     5
YEAR TO BEGIN POPULATION PROJECTION  ISTART 1848         FUTURE BIAS FORM 0=linear, 1=cyclic  OPTB      0
YEAR MANAGEMENT BEGINS               IYRMAN 2003         FIRST HISTORIC SURVEY BIAS           BIASH1  1.00
NUMBER OF YEARS OF QUOTA             IQUOTA    5         FINAL HISTORIC SURVEY BIAS           BIASH2  1.00
MAXIMUM AGE                          MAXAGE   35         BIAS IN YEAR 0                       BIAS0   1.00
STOCASTIC PD MODEL: 0=No, 1=Yes 2=+  ISTOCH    0         BIAS IN FINAL YEAR                   BIASF   1.00
MSY COMPONENT       0=1+, 1=Mat      OPTF      0         NUMBER OF DEGREES OF FREEDOM         DOF     19.00
DENSITY DEPENDENCE: 0=1+, 1=Mat      OPTDD     0         CV OF ABUNDANCE ESTIMATES            CVTRUE  0.25
TIME DEPENDENCE IN K 0:Const         OPTK      0         CV ESTIMATE: EXPECTATION VALUE (1st) CV1EST  0.25
TIME DEPENDENCE IN A 0:Const         OPTA      0             Age data:
TIME DEPENDENCE IN H 0:Const         OPTH      0         SAMPLE SIZE (OF CALVES & MATURE)     NSAMP   250
PROBABILITY OF EPISODIC EVENTS       ERATE   0.0         MINIMUM AGE FOR CALVES               AMINC     0
CORRELATION IN RECRUITMENT           RHO     0.000       MAXIMUM AGE FOR CALVES               AMAXC     1
VARIATION IN ENVIRONMENTAL IMPACT    SIGE**2 0.325       AGE WHEN 50% OVER CALF LENGTH        A50C      1
FUTURE CATCH SEX RATIO               CRATIO  0.5         DISPERSION PARAMETER FOR CALVES      GAMMAC  0.0
HISTORIC CATCH BIAS                  BIASC   1.0         BIAS IN A50C                         LAMDAC  1.0
YEAR OF FIRST CATCH                  ICAT1  1848         ERROR IN GAMMAC                      SIGAC   0.0
CATCH BIAS APPLIES UP TO             ICATB  1914         MINIMUM AGE FOR MATURE               AMINM    10
FORM OF NEED CURVE 0=Lin 1=Step      OPTN      0         MAXIMUM AGE FOR MATURE               AMAXM    20
INITIAL NEED LEVEL                   NEED0    67.0       AGE WHEN 50% LONGER THAN 12.9M       A50M     30
FINAL NEED LEVEL                     NEEDF   201.0       DISPERSION PARAMETER FOR MATURE      GAMMAM  0.0
M PARAMETERS        M1,M2,M3      2000.0  0.8  0.9       BIAS IN A50M                         LAMDAM  1.0
    Abundance data:                                      ERROR IN GAMMAM                      SIGAM   1.91
YEAR OF 1ST HISTORIC SURVEY          ISUR1  1978
```

Table 1: First portion of the primary input file for bowhead AWMP simulation testing.

do not stray beyond the appropriate parameter space region.

Types of scenarios planned by the SC

Bowhead AWMP scenarios are currently classified as belonging to one of four classes (IWC, 2000b):

Initial Exploration Trials: These nine scenarios are used to assess the merits of performance statistics and to provide a framework for developers working on AWMP SLAs. These scenarios manipulate three areas of uncertainty: the $MSYR_{1+}$, future need trajectory, and data quality. Many potential scenario variables are fixed for these trials.

Evaluation Trials: These 30 scenarios are to be used for formal comparisons of candidate SLAs. SLA comparison on these scenarios should occur before initiating robustness trials.

Robustness Trials: These 95 scenarios are used to examine SLA performance for a full range of plausible scenarios. Robustness trials would be applied to one or more SLA candidates that are chosen on the basis of good performance in evaluation trials.

Cross-Validation Trials: These trial scenarios are intended to be held aside from all prior SLA development so that any SLA which successfully emerges from all other trials can be subjected to a subsequent independent test. There are currently no bowhead cross-validation trials explicitly specified.

The SC agreed that none of the trial classes should include any implausible scenarios (IWC, 2000b). In particular this means that robustness trials should be limited by plausibility arguments. There is no benefit to be gained by evaluating the performance of a SLA

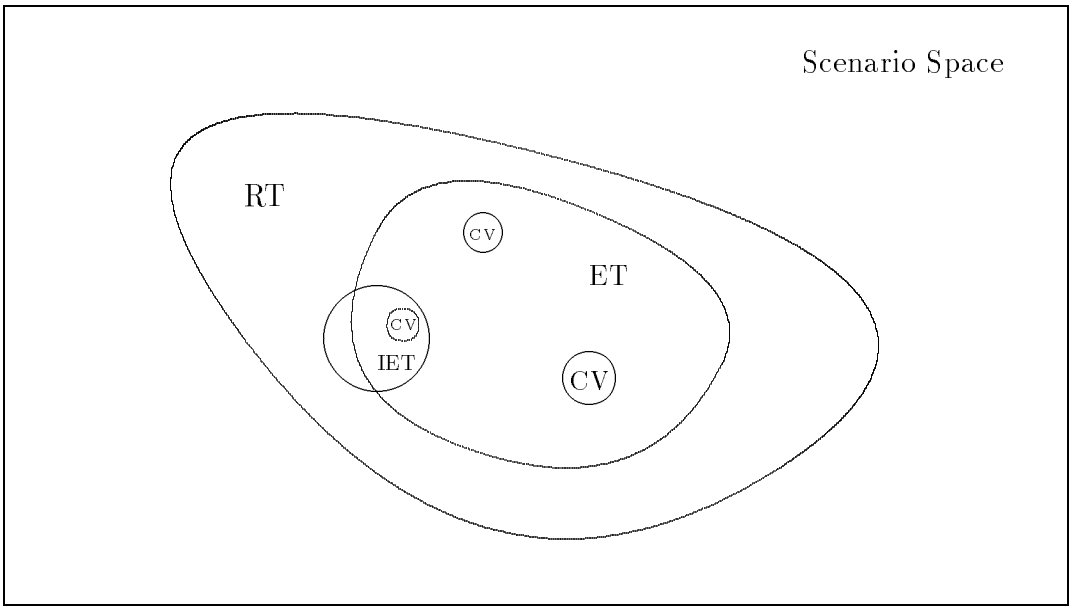

Figure 1: Venn diagram of proposed relationships between the portions of potential scenario space explored by Robustness (RT), Evaluation (ET), Initial Exploration (IET), and Cross-validation (CV) Trials.

on implausible trials. Any modifications, tunings, or preferences arising from such simulations would reduce the likelihood of satisfying IWC objectives to the greatest extent possible because performance on plausible scenarios would be degraded in order to achieve satisfactory performance on implausible objectives. Several authors have addressed the issue of plausibility in similar contexts (Butterworth *et al.*, 1996; Butterworth, 1995; Hilborn, 1996).

The SC also discussed desired relationships between trial scenario classes. It agreed that evaluation trials should not be limited to only the most plausible regions of the robustness scenario space. The SC had difficulty agreeing on the relationship between cross-validation trials and the other scenario classes. It agreed that some such trials should be designed during the 2000 SC meeting, and that some previously identified trials in other classes would be designated then as being useful for cross-validation (IWC, 2000b).

Our focus on quantifying similarities and differences between scenarios allows us to address the issue of relationships between scenario classes without explicitly considering plausibility.

Fig. 1 illustrates relationships that we believe should hold for the scenario classes. Robustness trials should be the most diverse class of scenarios, but should not explore all corners of potential scenario space. The spaces for evaluation trials and initial exploration trials should be subspaces of the space for robustness trials. The goal is that evaluation and comparison of SLAs should be based on a relatively small set of highly plausible trials. The evaluation trials should resemble the initial exploration trials, but not be limited to the types of scenario variation allowed in initial exploration trials.

The cross-validation trial space should be contained in the space of evaluation trials. The purpose of cross-validation trials is to check the SLA selection process. Therefore,

cross-validation trials are essentially alternative evaluation trials. Cross-validation trials should not be allowed outside of the evaluation trial space (eg. in the larger robustness trial space). Use of such a trial to cross-validate a preference decision is irrelevant and potentially misleading. For this reason, we disagree with the SC's approach to treat some robustness trials as cross-validation trials, unless all such trials it chooses for cross-validation happen to clearly resemble evaluation trials.

Given decisions about the appropriate relationships between scenario classes, a major challenge still remains: identifying the scenario space encompassed by a finite collection of trials of a specific class. Quantitative delineation of scenario spaces would allow assessment of whether appropriate relationships exist between the relevant trial spaces. This would also help the SC assess the fairness of the trials, particularly those used for cross-validation.

IDENTIFYING SCENARIO SPACE BOUNDARIES

Characterizing Scenario Spaces by Trajectory Shapes

We believe it is most effective and simplest to characterize scenarios and scenario spaces by the shape of their population trajectories. This approach reduces the risk of failing to notice hidden extrapolations and interactions between elements of $\boldsymbol{\theta}$.

For any single bowhead scenario, the 100 'replicate' simulated population trajectories vary due to random sets of biological input parameters, random simulated future data, and, in some cases, the use of a stochastic population dynamics model. Properties of single trajectories may therefore be too noisy to use as scenario descriptors. We based most of our descriptive statistics on the pointwise 5^{th}, 50^{th} and 95^{th} percentile trajectories for a given scenario.

Recall, however, that trajectory shape depends on the SLA applied. We used three strike limit regimes: zero strikes, constant strike limit of 100, and an actual SLA which reacts to available data. It seemed unfair for the third choice to be a SLA under serious consideration by the SC. Therefore, we tentatively used the 'Potential Biological Removal' SLA prototype (Givens, 1998) based on certain aspects of US domestic marine mammal management policy (Wade and Angliss, 1997). The SC might usefully recommend a more relevant choice that would be more reactive to different scenarios; our choice tends to be conservative and unresponsive to scenario variation.

The Bowhead Evaluation Trial Scenario Space

The bowhead AWMP evaluation trials BE01–BE19 (30 trials) constitute a class of scenarios. We seek a method for determining whether any future proposed trial is typical of the scenario space spanned by these, or whether the introduction of the new trial would represent an unfair extrapolation of that space. We furthermore seek a method for examining the 30 trials in the evaluation trial space to determine whether all are necessary or whether some

are redundant. Finally, we want the method to be useful for identification of scenarios that are unusual compared to the rest of the scenario space members.

Our method is based on population trajectory shape. Figs 2–5 show the 100 replicates of each trial for two SLAs: zero strikes and strikes equal to need. The results for constant strike limit of 100 and 'Potential Biological Removal' are intermediate to the graphs shown[1]. Clearly there are visible similarities among the trajectories in these graphs: they are fairly smooth, not too steep, have roughly one or fewer major bends, and generally show only relatively low-frequency and low-amplitude within-scenario variation.

For the bowhead example we derived five measures of trajectory shape useful for identifying scenario spaces. Fig. 6 summarizes these indicators of the boundaries of the bowhead evaluation trial scenario space.

1. **Pointwise 90% Band Boundaries:** A new member of the scenario space must not have a pointwise 95% trajectory that lies anywhere above all such bands for existing members, nor a pointwise 5% trajectory that lies anywhere below all such existing bands. In the context of Fig. 6, this amounts to requiring that any member of the scenario space must not have a pointwise 95% or 5% trajectory which lies anywhere in the shaded areas.

2. **Tolerable Early Increase:** A new member of the scenario space must have a pointwise median population trajectory increase from year 0 through 20 which does not exceed the largest such increase observed for existing members of the space. The increase also must not be less than the smallest such increase observed for existing members of the space. For a stock size of 7,500 in year 0, this requirement is shown in Fig. 6 as a triangular window in which year 20 stock size must fall. Trajectory behavior in years 0–19 is not explicitly regulated.

   We have chosen to express population increase as a number of whales rather than as a percentage. The reason for this is that trajectories very near zero will exhibit huge decrease percentages; when these huge rates are applied to larger stocks, the result is an excessively permissive bound. The reverse problem is not as severe due to the nature of bowhead trials.

3. **Tolerable Later Increase:** A new member of the scenario space must have a pointwise median population trajectory increase from year 60 through 80 which does not exceed the largest such increase observed for existing members of the space. The increase also must not be less than the smallest such increase observed for existing members of the space. For a stock size of 7,500 in year 60, this requirement is shown in Fig. 6 as a triangular window in which year 80 stock size must fall. Trajectory behavior in years 60-79 is not explicitly regulated. We have chosen the later increase bounds to be expressed as a number of whales for the same reason as stated in 2.

[1] All the examples in this paper use simulation software and scenario specifications from (IWC, 2000b), available as the May 17, 2000, distribution from `www.stat.colostate.edu/~geof/iwcsoftware.html`, which were the latest available at the time of writing.

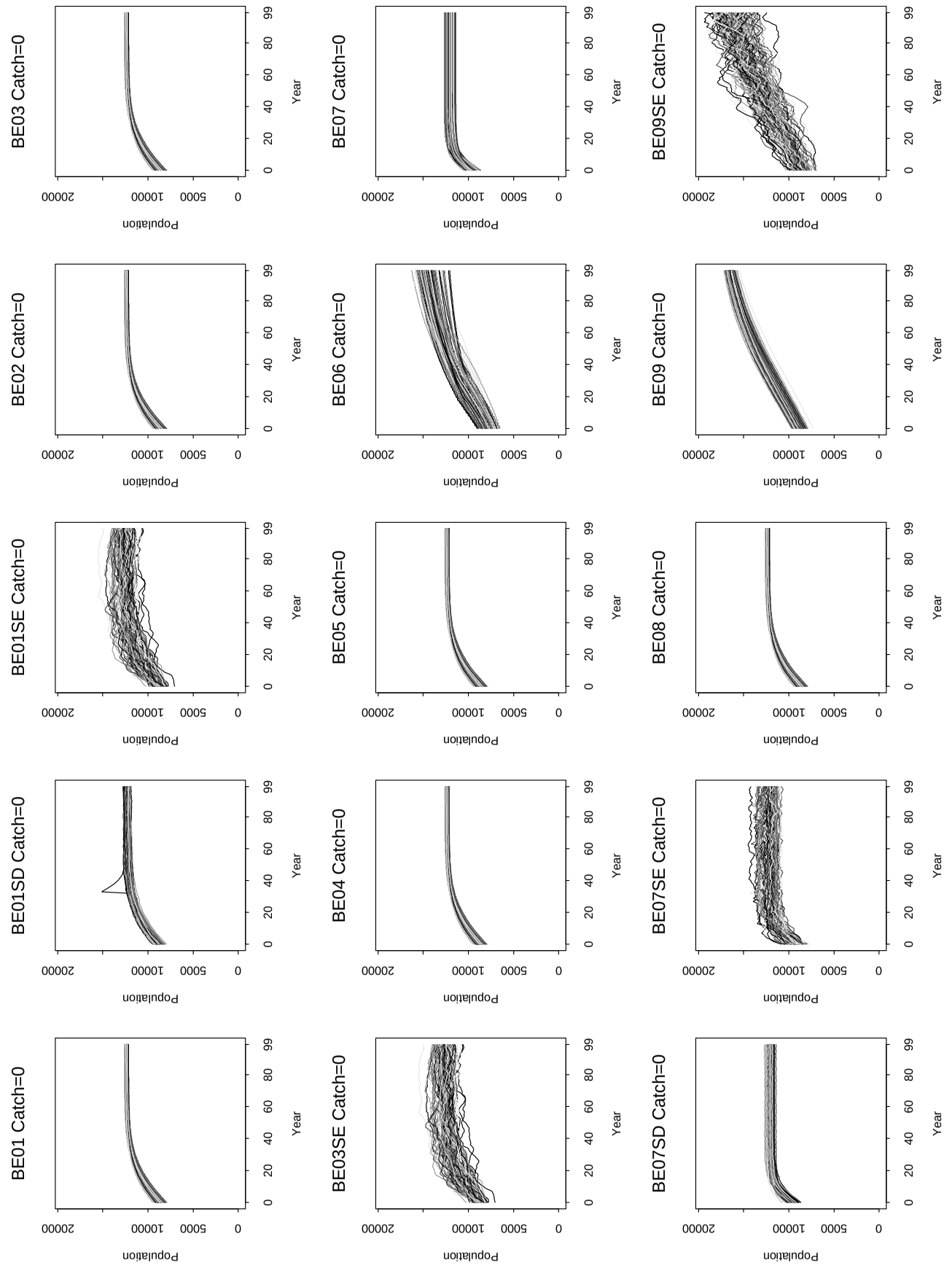

Figure 2: The 100 replicate population trajectories for the first 15 bowhead evaluation trials when catch equals zero.

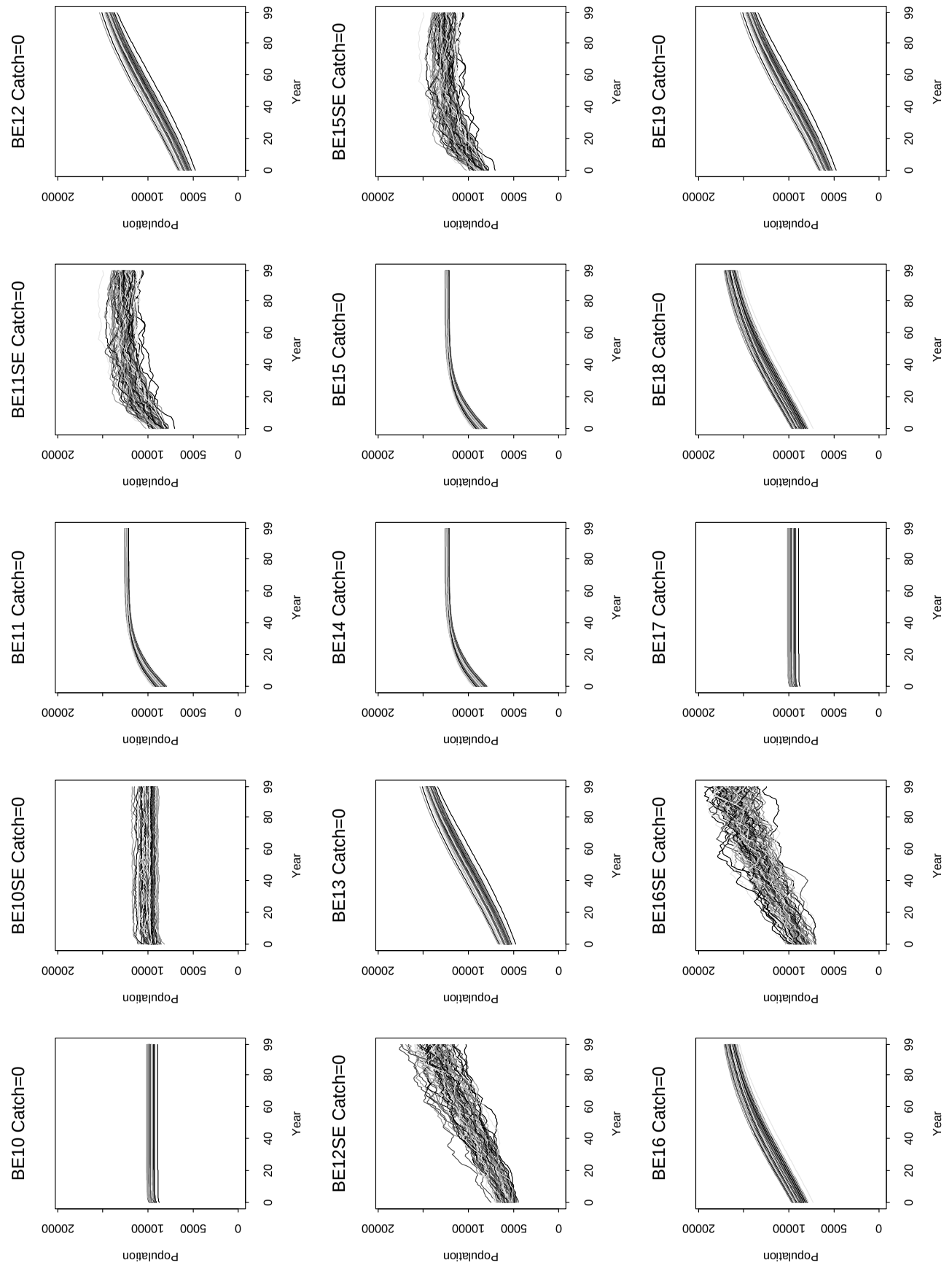

Figure 3: The 100 replicate population trajectories for the second 15 bowhead evaluation trials when catch equals zero.

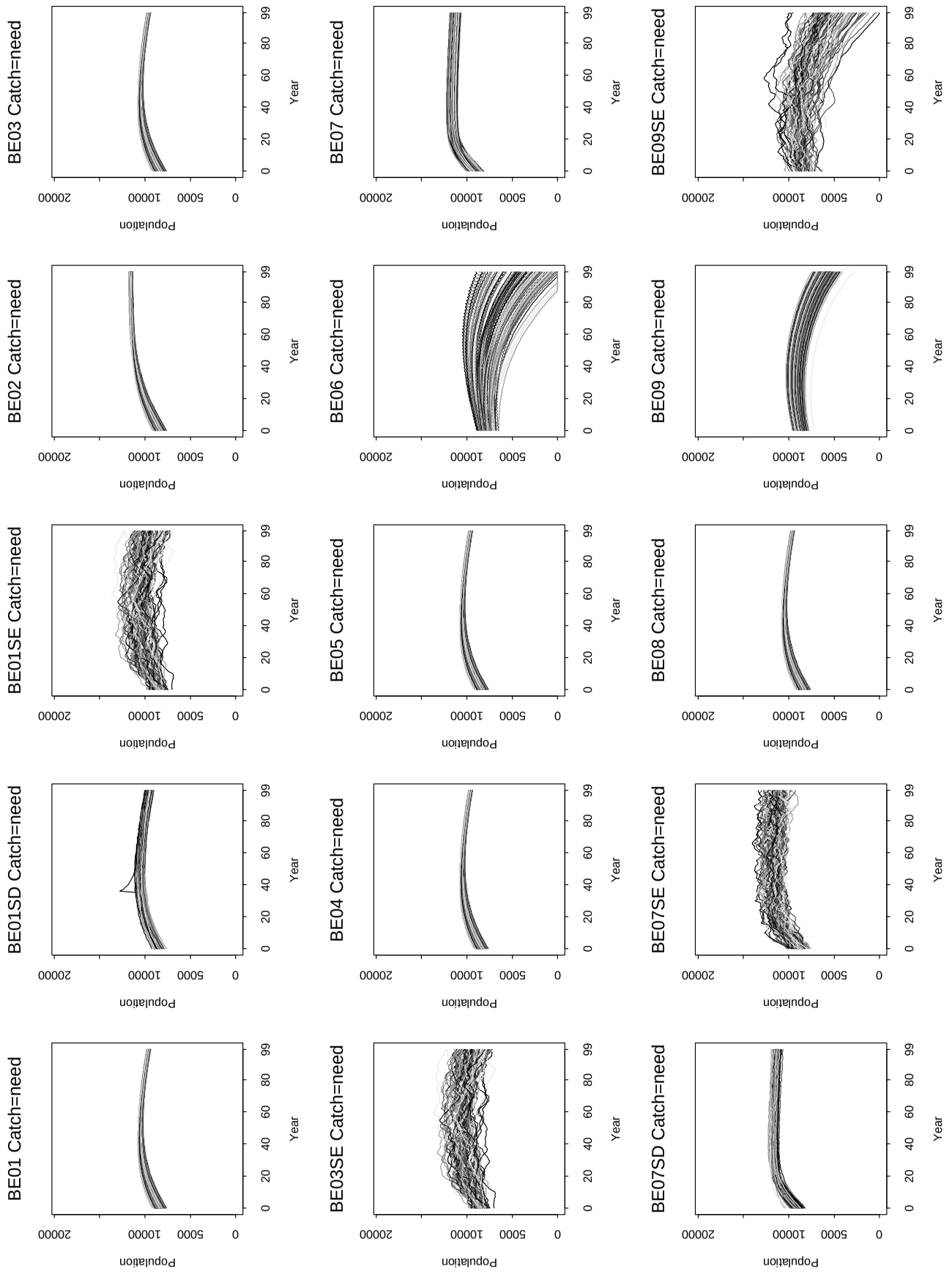

Figure 4: The 100 replicate population trajectories for the first 15 bowhead evaluation trials when catch is set equal to need.

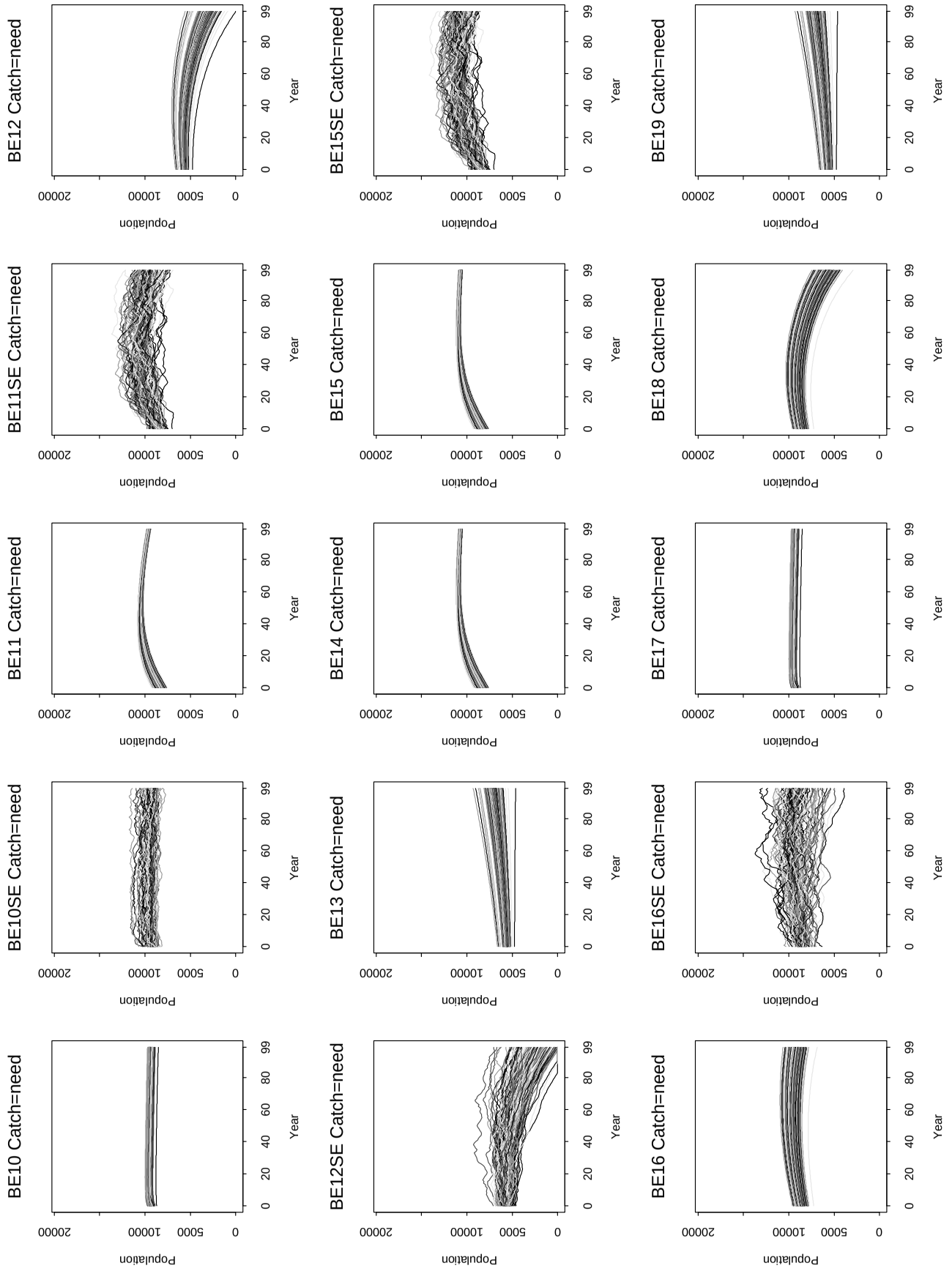

Figure 5: The 100 replicate population trajectories for the second 15 bowhead evaluation trials when catch is set equal to need.

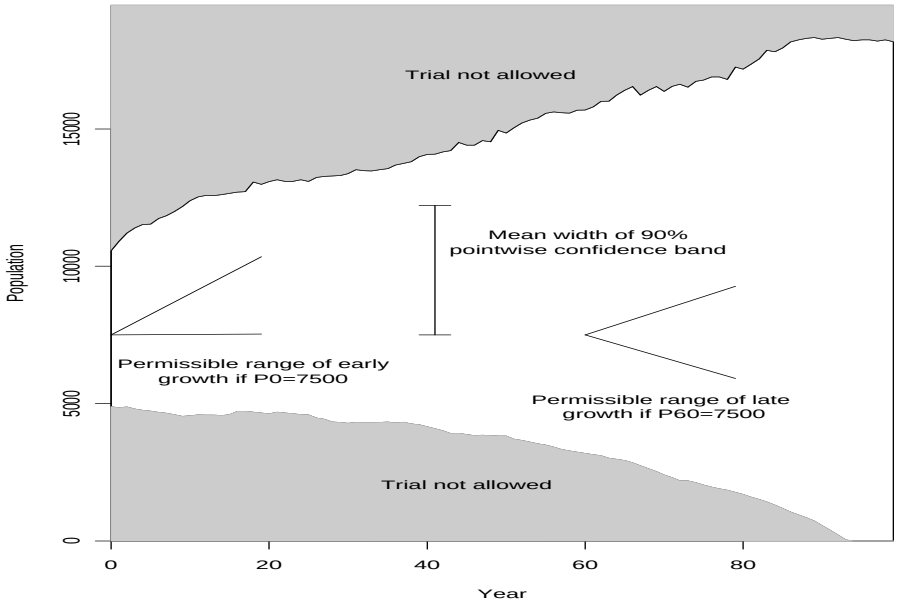

Figure 6: Graphical representation of the proposed statistics for quantifying the bowhead evaluation trial scenario space. The limitation on large-scale trajectory variation (i.e. curvature) is not shown here.

4. **Limitation on Within-Scenario Variation:** A new member of the the scenario space must have a mean (over time) 90% pointwise range (pointwise 95%tile minus pointwise 5%tile) that does not exceed the largest such mean range of existing members. This requirement is shown in the center of Fig. 6 as a vertical window showing the maximum allowable 90% range. This bound applies to the 100-year average; its position in the plot is merely for convenience.

5. **Limitation on Large-Scale Median Trajectory Variation:** Pointwise median trajectory variation over time is measured by an analogue to the 'anti-curvature statistic, N11' used elsewhere in AWMP analysis (IWC, 2000a), namely:

$$\sum_{t=1}^{\min(98,t_e-2)} \left|(P_{i,t}^{0.5} - M_{i,t})\right| / \max(5675, M_{i,t})$$

where $P_{i,t}^{\alpha}$ is the pointwise α^{th} percentile stock size in simulation year t of scenario i, t_e is the first year that the pointwise median simulated stock size drops to zero, and $M_{i,t} = (P_{i,t+1}^{0.5} + P_{i,t-1}^{0.5})/2$. A new member of the scenario space must have a pointwise median trajectory with a value of this anti-curvature statistic no greater than the largest value obtained by the pointwise median trajectories of existing members.

The number 5,675 was chosen because it was the smallest median initial stock size among the bowhead evaluation trials. For the bowhead example, the maximum permissible value of the anti-curvature statistic is 1419.5; this value arose from the BE06 trial with a constant annual catch of 100. However, the BE06 trial is later shown to be highly atypical of other evaluation trials, and should probably be eliminated. Omitting

BE06, the maximum permissible value of the anti-curvature statistic would be 339.9 from the BE12SE trial with a constant annual catch of 100.

Since this anti-curvature statistic is based on the pointwise median trajectory, it is most sensitive to long-term trends and is relatively unaffected by stochastic noise in individual trajectories. It might seem important to include a limitation on individual trajectory variation, too. However, such variation is confounded with stochastic population dynamics model usage: stochastic scenarios exhibit high-frequency low-amplitude variation in addition to the variation and trends typical of deterministic scenarios. We would not want to eliminate stochastically modeled trials on this basis alone. Therefore, we believe a limitation on individual trajectory variation is unwarranted.

Uses of a 5-Statistic Summary

Finding unusual scenarios

Five statistics—motivated by the list above—can be used to find scenarios that seem unusual relative to a class of scenarios. This task is likely to be extremely difficult to tackle by direct examination of $\boldsymbol{\theta}$ because the parameter space used to define trials is very high dimensional. (Consider sorting through two files like Table 1 to assess the degree of similarity between the trajectories they produce.) The problem can be reduced to only five dimensions using:

$$T_{i,1} = \frac{1}{100}\sum_{t=0}^{99} I_{\{P_{i,t}^{0.05}<LB_t \text{ or } P_{i,t}^{0.95}>UB_t\}} \quad (1)$$

$$T_{i,2} = (P_{i,20}^{0.5} - P_{i,0}^{0.5})/20 \quad (2)$$

$$T_{i,3} = (P_{i,80}^{0.5} - P_{i,60}^{0.5})/20 \quad (3)$$

$$T_{i,4} = \frac{1}{100}\sum_{t=0}^{99} \left(P_{i,t}^{0.95} - P_{i,t}^{0.05}\right) \quad (4)$$

$$T_{i,5} = \sum_{t=1}^{\min(98,t_e-2)} \left|(P_{i,t}^{0.5} - M_{i,t})/\max(5675, M_{i,t})\right| \quad (5)$$

where LB_t is the pointwise minimum of $P_{j,t}^{0.05}$ across all j indexing current members of the scenario class, UB_t is the pointwise maximum of $P_{i,t}^{0.95}$ across all j indexing current members of the scenario class, and $I_{\{A\}}$ is 0 if A is false and 1 if A is true.

Each scenario in a collection may be placed in a five-dimensional space with coordinate axes corresponding to these statistics. This space is sufficiently low-dimensional that it can be visually inspected for unusual or outlying scenarios.

Fig. 7 shows the 30 evaluation trials plotted in a 4-dimensional space excluding $T_{i,1}$ (because $T_{i,1}$ is always zero for existing members of the space). There are three symbols plotted for each scenario: one corresponding to each of the three catch regimes mentioned

previously. The symbols 's' and 'd' represent scenarios that use stochastic and deterministic population dynamics models, respectively. One trial (BE06 with catch=100) is not shown because it falls far beyond the limits in Fig. 7: it has $T_{BE06,2} = 16.1$, $T_{BE06,3} = -83.2$, $T_{BE06,4} = 4713.8$, and $T_{BE06,5} = 1419.5$. Figs 8 and 9 show marginal dotplots of the four statistics when catch equals zero and catch equals need, respectively.

These figures are intended to help identify unusual scenarios. One could be misled by comparing several scenario parameter sets directly (say $\boldsymbol{\theta_1}$ versus $\boldsymbol{\theta_2}$) because two very different values of $\boldsymbol{\theta}$ may produce effectively redundant scenarios. One attractive feature of these graphical methods is that the complexity and dimensionality of these figures do not increase as the number of scenarios or the number of variables changed between scenarios increases. In the remainder of this subsection, we make several observations about the evaluation trial framework based on these figures.

Trial BE06 is clearly a potential outlier since it cannot even be plotted adequately for one catch and is quite atypical for the other two catch regimes. Closer inspection of this trial shows that the low value assumed for MSYR sometimes leads to negative values for what IWC (2000a) denotes as A and z (the resilience and degree of compensation parameters controlling density dependence). In addition to very unusual population trajectories, this causes very erratic simulated behavior of replacement yield and hence of ideal quotas (H). See the discussion section for further details.

The set of evaluation trials does an excellent job of capturing a range of values for $T_{i,2}$, average early population growth. Both deterministic and stochastic trials evenly span a range of values.

$T_{i,4}$ and $T_{i,5}$ appear to be rather collinear. This means that the current evaluation trial space does not include trajectories which exhibit high curvature with low variation or low curvature with high variation. While this is not necessarily a problem, it does indicate a portion of scenario space not explored by the evaluation trial class.

The evaluation trial class also exhibits some interdependencies between early and late population growth ($T_{i,2}$ and $T_{i,3}$). There are no evaluation trials that exhibit (i) low early growth and high later growth, or (ii) high early growth and negative later growth. Perhaps the absence of (i) is justifiable on the basis of plausible dynamics and carrying capacity. However, environmental change or different values for MSYL, current depletion, and carrying capacity could generate such trajectories. The absence of (ii) might be explained if such trajectories have been relegated to the class of robustness trial scenarios.

Figs 8 and 9 suggest that the three trials based on BE07 are qualitatively different than other trials, in that they provide significantly higher early and late growth. These dotplots also show that the trials based on a stochastic population dynamics model with demographic and environmental variation (denoted by 'SE' in the trial labels) are not homogeneous. The same holds for trials using a stochastic model with only demographic variation (denoted by 'SD' in the trial labels). Trials with the deterministic model are more homogeneous and more distinct from trials with the other two models.

Finally, we note that $T_{i,4}$ and $T_{i,5}$ are confounded with whether the dynamics model is stochastic or deterministic. Although the stochastic model produces trajectories with

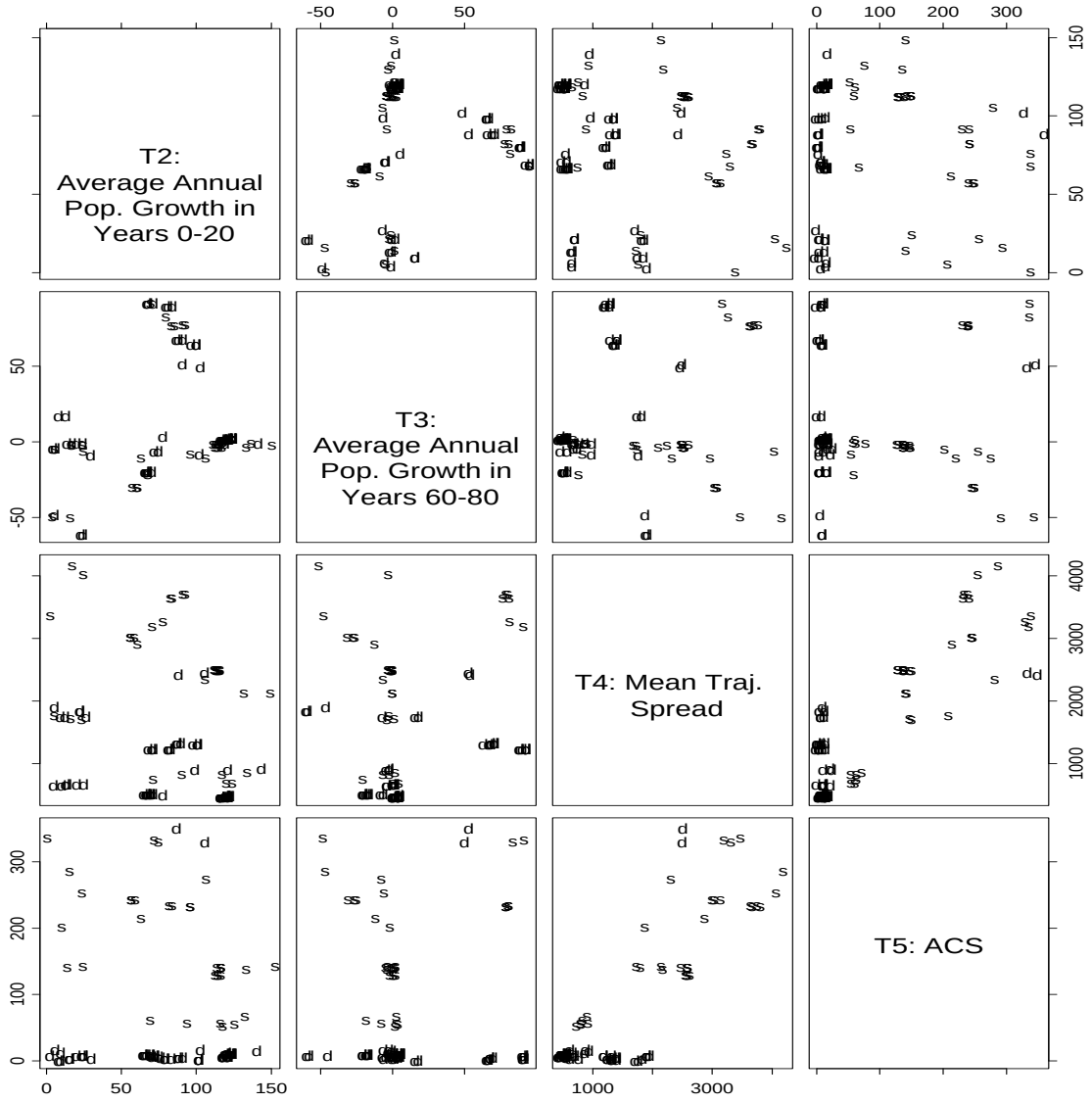

Figure 7: Graphical representation of the trajectory features of the bowhead evaluation trials. Each trial was run with catch taken according to the three SLAs mentioned in the text; thus each scenario is represented by three points on each panel. Points are labeled 'd' or 's' according to whether the scenario employs a deterministic or stochastic population dynamics model. Trial BE06 with constant catch of 100 is excluded here for reasons given in the text.

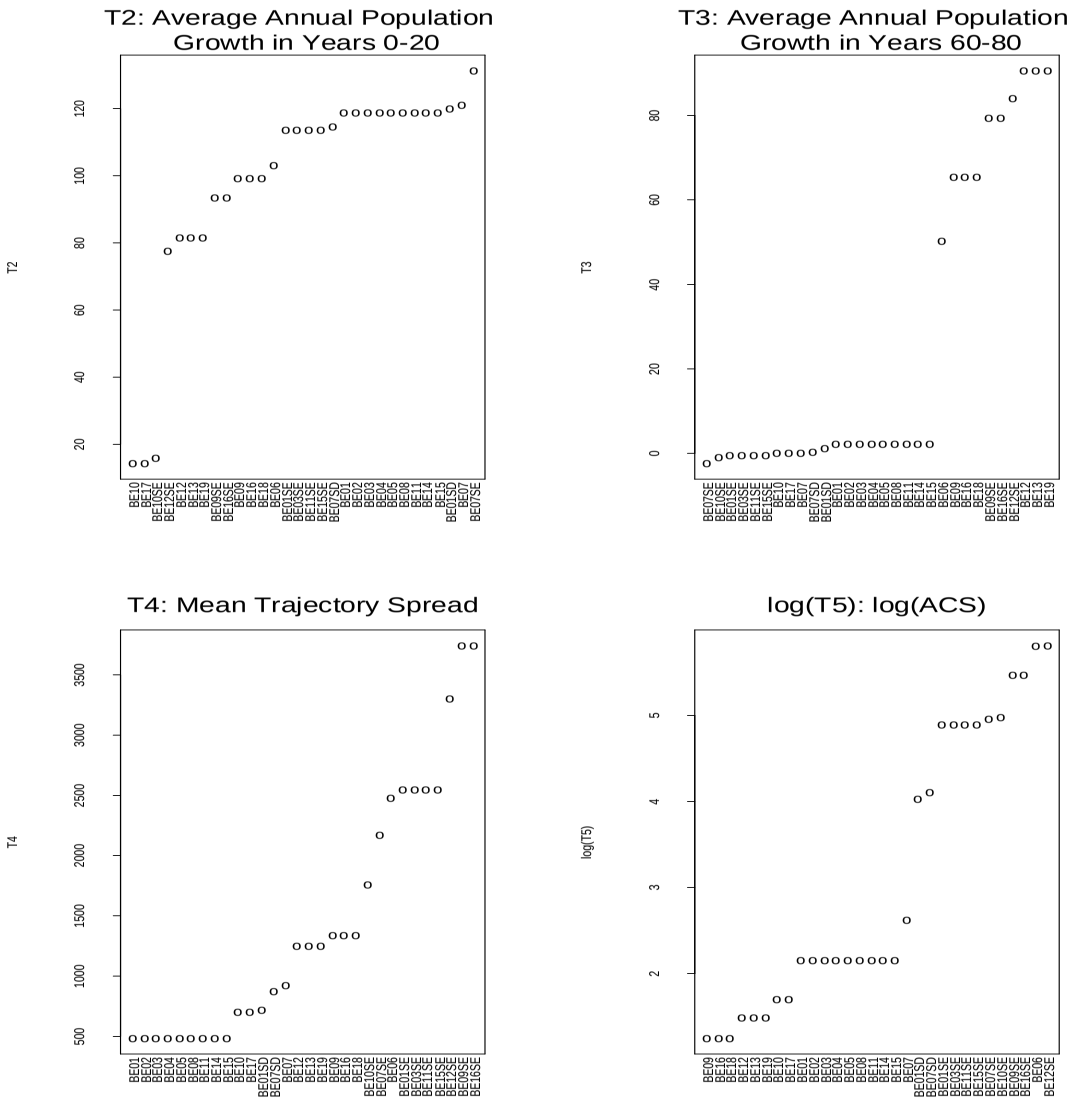

Figure 8: Dotplots showing the distribution of the trajectory shape summary statistics for all bowhead evaluation trials when catch equals zero. Note that the anti-curvature statistic is shown on a log scale.

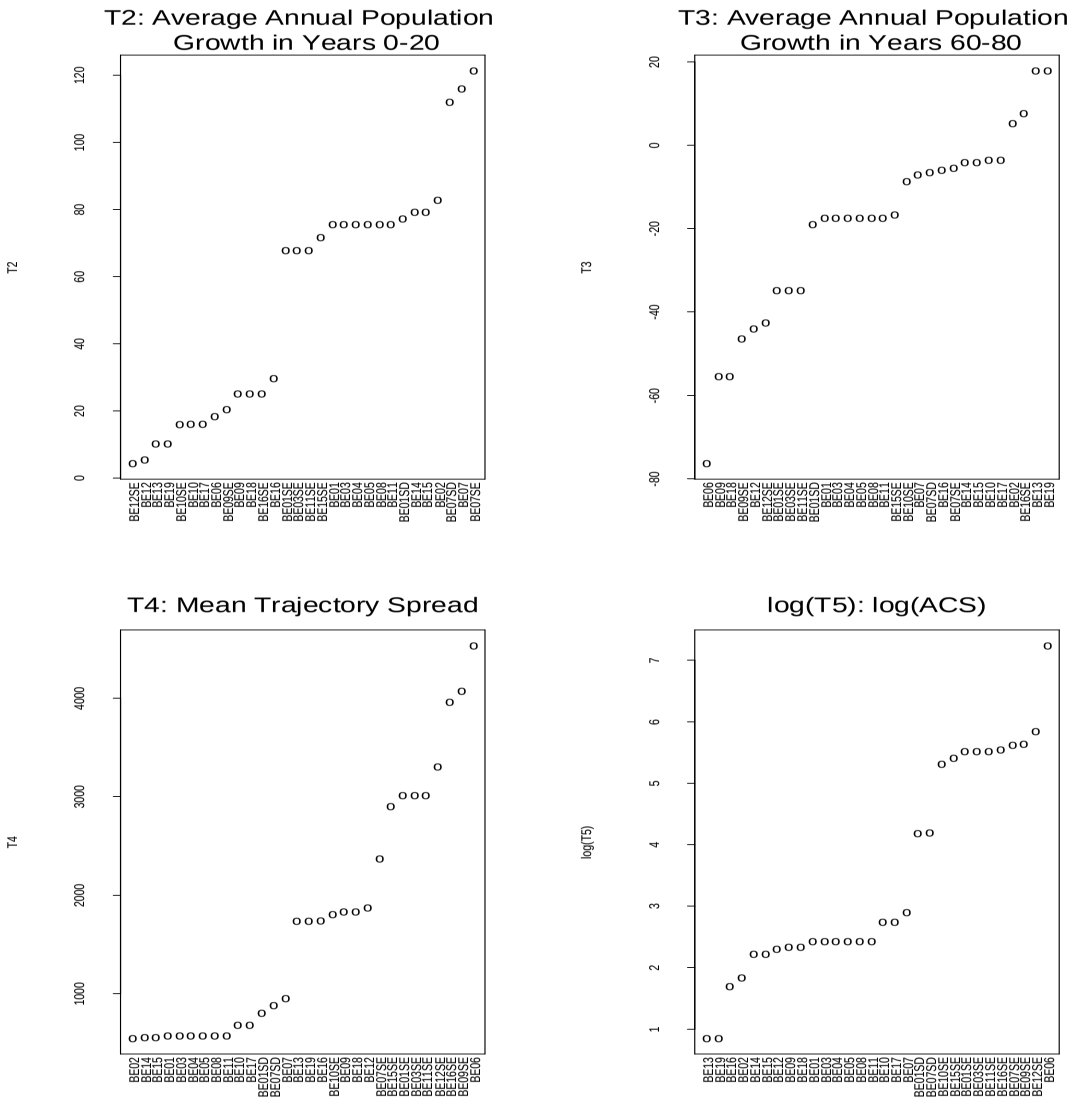

Figure 9: Dotplots showing the distribution of the trajectory shape summary statistics for all evaluation when catch is set equal to need. Note that the anti-curvature statistic is shown on a log scale

higher values on these statistics, the other characteristics of the stochastic trajectories are reassuringly similar to those of the deterministic trajectories.

Adding to a pre-established scenario space

There may be some scenarios which the SC believes clearly belong to a scenario space, and others whose membership in that space is debated. For instance, debate might focus on whether a certain trial with pessimistic productivity assumptions should belong to the set of evaluation trials or to the set of robustness trials.

We propose that an initial scenario space be identified using a set of non-controversial scenarios. This space, and the statistics listed above, can then be used to determine whether more controversial scenarios belong or by how much they lie outside the current space. Just because a controversial scenario lies outside the current space does not mean it cannot be added to the space. The SC might add the scenario to the space despite the scenario's atypicality if it is important or if it is very close to the existing space. In such a case, the scenario space would be redefined to account for the inclusion of the new trial.

As an artificial example, consider the question of whether the nine bowhead initial exploration trials could be added to the class of evaluation trials (i.e. whether they are interior to the evaluation trial scenario space). Table 2 shows the values of the five statistics for each initial exploration trial and each catch regime. The allowable ranges of each statistic are shown at the bottom.[2] Instances where trials seem to lie outside the evaluation trial scenario space are boldfaced in this table.

The results in Table 2 show that BIET7a is more extreme than members of the evaluation trial space because it is too pessimistic about stock growth under high catch. BIET03 and BIET10 are too optimistic about early stock productivity. BIET13 exhibits more between-trajectory variation than scenarios in the evaluation trial space. Many of these trials are also atypical for other reasons. Overall, five initial exploration trials seem to fit in the evaluation trial space and four do not; this is not in conflict with the scenario class relationships we proposed in Fig. 1.

Finding redundant scenarios

The trajectory summary statistics described above can also be used to help identify redundant trials in a scenario class. Consider the members of the bowhead evaluation trial scenario class. Relative to the boundaries of this class, all these trials have $T_{i,1} = 0$ by definition. We omit this measure and begin by standardizing the other four measures, denoting

$$T^s_{i,k} = (T_{i,k} - \text{mean}(T_{i,k}))/\ (\text{st.dev.}(T_{i,k})) \tag{6}$$

[2]The limit for the anti-curvature statistic is a poor one because it is based on the atypical BE06 trial. If that trial is removed as suggested in the discussion section, the limit would be 339.9. Omitting BE06 from the evaluation trial scenario space results in only small changes to the rest of the table.

Trial Name	$T_{i,1}$ Out of Bounds	$T_{i,2}$ Early Growth	$T_{i,3}$ Late Growth	$T_{i,4}$ Mean Spread	$T_{i,5}$ ACS
BIET01					
catch= 0	0.00	120.9	2.6	504.1	8.4
catch=PBR	0.00	123.1	3.2	525.3	12.3
catch=100	0.00	78.4	5.1	535.6	6.6
BIET03					
catch= 0	**0.09**	**157.9**	20.0	2138.1	98.2
catch=PBR	0.00	**156.4**	21.4	2107.0	109.9
catch=100	0.00	92.5	-8.8	1927.6	387.1
BIET06					
catch= 0	0.00	95.5	71.4	1434.2	3.5
catch=PBR	0.00	85.2	73.9	1441.6	7.4
catch=100	0.00	30.1	36.4	1784.9	1.5
BIET07					
catch= 0	0.00	95.5	71.4	1434.2	3.5
catch=PBR	0.00	85.2	73.9	1441.6	7.4
catch=100	0.00	17.6	-68.7	2037.6	12.6
BIET07a					
catch= 0	**0.03**	76.6	92.6	1209.2	4.2
catch=PBR	**0.04**	65.0	**93.6**	1201.7	9.7
catch=100	**0.88**	**-8.0**	**-132.4**	1193.6	21.2
BIET10					
catch= 0	**0.09**	**157.9**	20.0	2138.1	98.2
catch=PBR	0.00	**156.4**	21.4	2107.0	109.9
catch=100	0.00	105.1	9.3	2083.6	129.8
BIET11					
catch= 0	0.00	125.6	17.0	1982.3	17.6
catch=PBR	0.00	119.8	20.1	1919.8	22.0
catch=100	0.00	72.1	22.9	1552.0	20.1
BIET12					
catch= 0	0.00	125.6	17.0	1982.3	17.6
catch=PBR	0.00	119.8	20.1	1919.8	22.0
catch=100	0.00	61.1	-19.7	2311.6	16.3
BIET13					
catch= 0	0.00	147.5	31.9	2087.0	38.1
catch=PBR	0.00	137.2	34.8	2052.8	41.6
catch=100	0.00	72.3	-0.2	**5246.9**	14.1
BE Allowable	0.00	(1.7, 149.9)	(-83.2, 93.4)	$\leq$4713.8	$\leq$ 1419.5

Table 2: Five-statistic summaries of bowhead initial exploration trial scenarios, for comparison with allowable limits of the evaluation trial scenario class. The allowable boundary on $T_{i,5}$ would be 339.9 if BE06 were omitted from the space of evaluation trial scenarios.

MSYR	NEED		
	Low	Moderate	High
Low	13 19	12, 12SE, 16, 16SE	9 18, 9SE
Moderate	2	14 15, 15SE	1 3 4 5 8 11, 1SD, 7SE 1SE 3SE 11SE, 7 7SD, 6
High	None	None	10 17, 10SE

Table 3: Cross-classification of evaluation trials according to their assumed levels of MSYR and need. Groups of trials enclosed within small boxes are trials whose trajectories seem nearly indistinguishable according to Figs 10 and 11.

for $k = 2, 3, 4, 5$ where the mean and standard deviation are taken across the 30 scenarios in the evaluation trial class. We then define the distance between trials i and j to be

$$d(i,j) = \sqrt{\sum_{k=2}^{5} \left(T^s_{i,k} - T^s_{j,k}\right)^2}. \tag{7}$$

Fig. 10 plots $d(i,j)$ versus i and j for zero catch; Fig. 11 shows the same for the case when catch is set equal to need. In these figures, the area of each circle is proportional to the corresponding $d(i,j)$, so two trials sharing a very small circle have very similar trajectories and two trials sharing a very large circle have very different trajectories. Thus, the small circles in this figure indicate possibly redundant trials.

These figures clearly show that there are many groups of redundant trials. Table 3 summarizes Figs 10 and 11 by cross-classifying trials according to their assumed levels of MSYR and need. Groups of trials enclosed within small boxes are trials that seem nearly indistinguishable on the basis of the two figures. This table shows several interesting features of the evaluation trial space. First, there are no trials providing a severe test of need satisfaction in a case where need should be easily satisfied: the bottom left two cells in the table are empty. Second, there is an abundance of redundant trials, especially for cases with high need. The most egregious redundancy is perhaps that evaluation trials 1, 3, 4, 5, 8, and 11 are virtually indistinguishable. If one reduced all redundant trial groups to a single trial each, the total number of evaluation trials would be nearly halved.

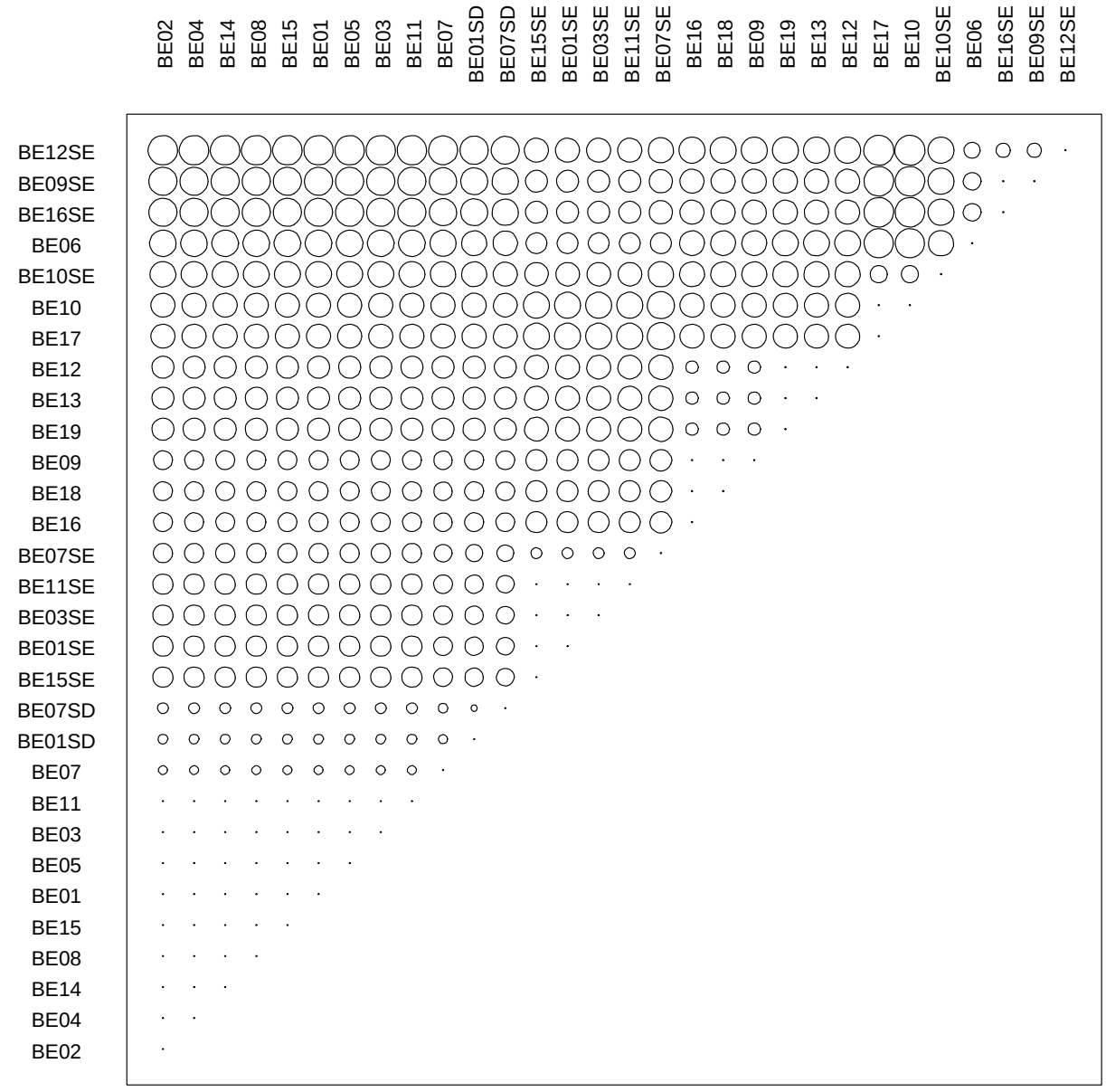

Figure 10: Distances between trials (as described in the text) when catch equals zero are proportional to the area of the circles.

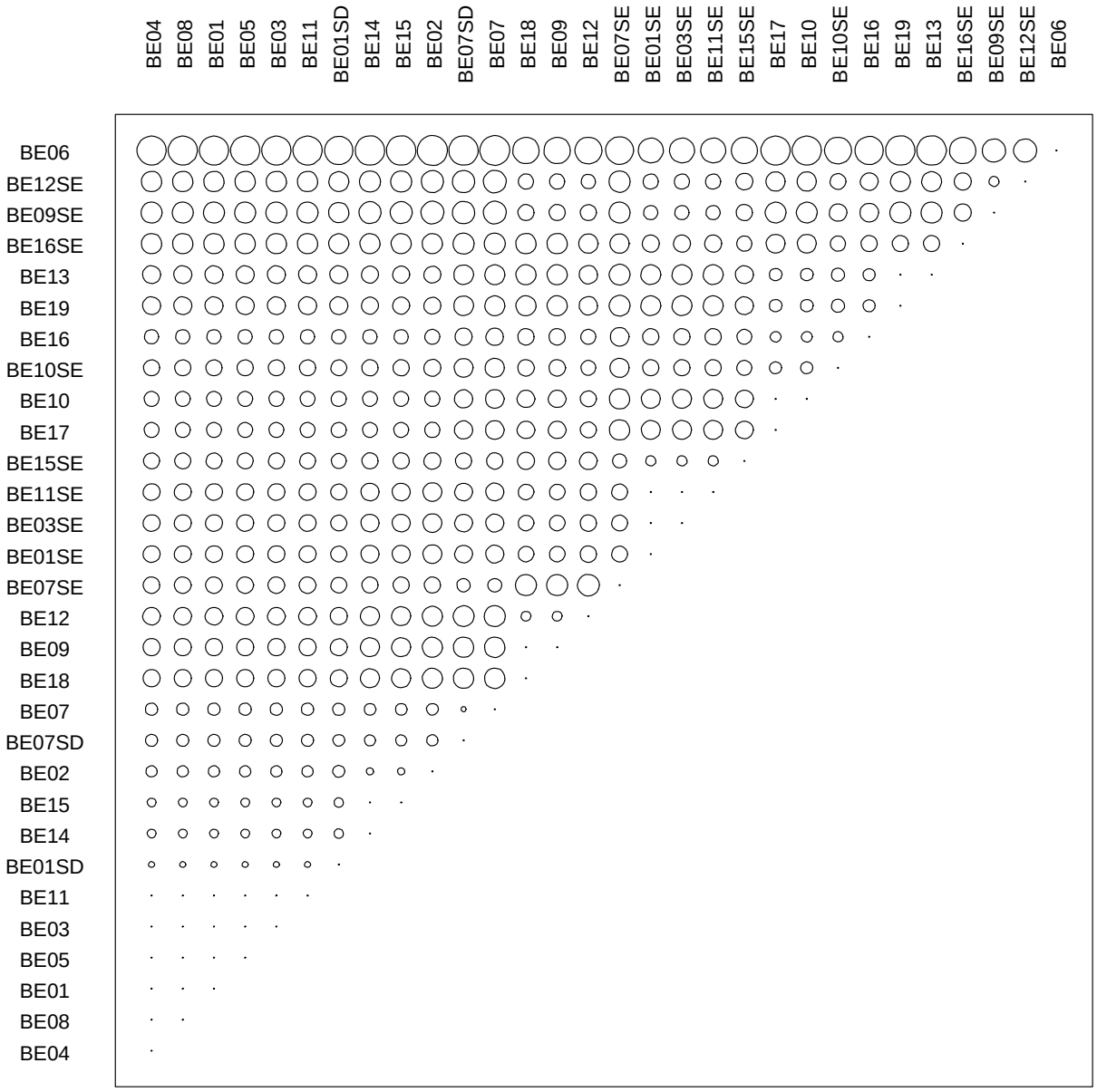

Figure 11: Distances between trials (as described in the text) when catch is set equal to need are proportional to the area of the circles.

DISCUSSION

An Examination of H Trajectories

IWC (2000a) defines ideal strike limits, H, for use with certain SLA optimisation strategies. Figs 12 and 13 show the time trajectories of H for the first three replicates of each bowhead evaluation trial. There are several trial scenarios for which the behavior of H appears absurd. This appears to be caused by an error in the simulation code, rather than poor trial design. For a given year, the simulation code calculates replacement yield (upon which H depends) as the number of recruits minus the number of deaths, divided by the weighted mean survival. In the case of stochastic dynamics modeling, however, such a calculation is affected by the random birth and death process realizations that occur during the given year. Replacement yield should instead be calculated on the basis of *expected* outcomes of the birth and death processes: expected recruits minus expected deaths, divided by the expected weighted mean survival. Punt (1999) and Butterworth (in IWC (2000c)) have previously suggested such a change. This problem in the calculation of H prohibits effective application of some SLA optimisation strategies.

Trial BE06 has an additional problem. This trial sets a low value for MSYL, with the result that the density dependence parameters denoted as A and z (the resilience and degree of compensation parameters controlling density dependence) by IWC (2000a) are sometimes negative. The result is erratic, flip-flopping recruitment, which in turn effects replacement yield and H. This is a problem with the trial specification itself.

Suggested Changes to the Evaluation Trials

Considering the analyses above, the following changes to the bowhead evaluation trial framework are suggested:

- Correct the calculation of replacement yield and H.
- Eliminate trial BE06 from the evaluation trial scenario class. A corrected version of it could be moved to the robustness trial scenario class.
- Change final need for BE01, BE01SD, and BE01SE to be 134. This is a more reasonable baseline and reduces the overabundance of high-need trials.
- Eliminate BE14. If the BE01 trial is changed, then BE14 is identical and can therefore be dropped.
- Add BE13SE, a stochastic version of BE13 to better balance the stochasticity and need assumptions.
- Move BE18 and BE19 from the evaluation to the robustness trial scenario class. The survey frequency factor should be investigated primarily as a robustness concern; this reduces the number and redundancy of evaluation trials.

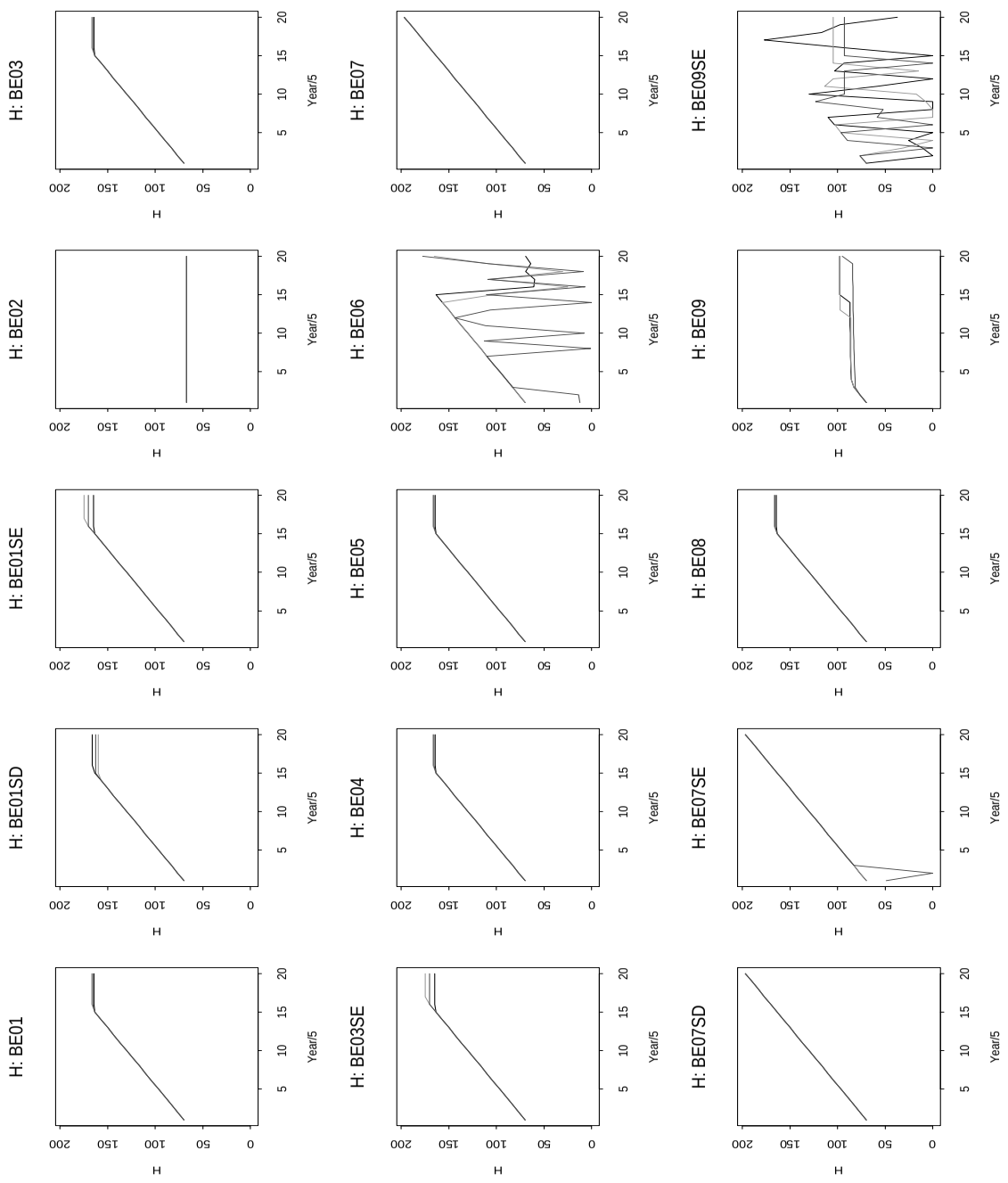

Figure 12: First three time trajectories of ideal strike limits, H, for first 15 bowhead evaluation trials.

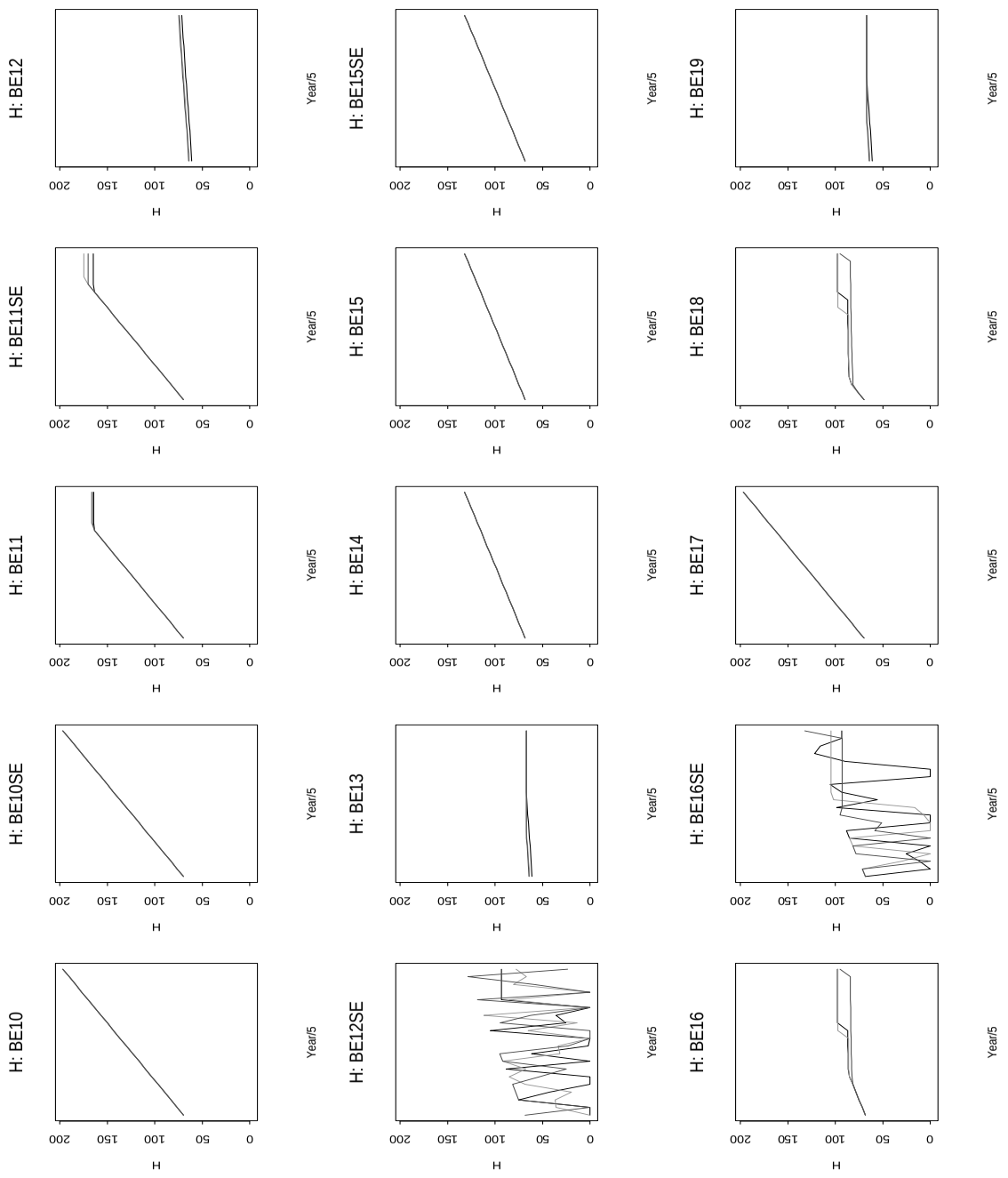

Figure 13: First three time trajectories of ideal strike limits, H, for second 15 bowhead evaluation trials.

- Eliminate BE15 (but not BE15SE). It and the revised BE01 (and BE14) are redundant.
- Eliminate BE05. It is intermediate to other trials and provides little information beyond those.
- Change BE08 to assume final need of 67 and MSYR=0.04. This removes redundancy and improves balance by shifting a trial from high to low need. It also allows examination of whether need satisfaction in an 'easy' case is impeded by infrequent surveys.
- Eliminate BE07 (but not BE07SD or BE07SE). BE07 and BE07SD are redundant.
- Change BE17 to assume final need of 134. This improves balance, removes redundancy (with BE10), and facilitates comparison with BE01.

The result of all these changes would be the net reduction of six evaluation trial scenarios and the addition of three robustness trial scenarios.

Future Work

Our work does not fully examine the relationships between trial classes: the code for many bowhead robustness trials is not ready for distribution, and no bowhead cross-evaluation trials have been identified. The two most important questions about trial design relate to such trials. First, it is important to understand what differences exist between the classes of robustness and evaluation trials; this question relates to the issue of fair AWMP evaluation. Second, it is important to structure cross-validation trials to resemble the trial class that is being cross-validated. We argue that cross-validation trials should resemble evaluation trials, not robustness trials. This question relates to the issue of fair AWMP cross-validation.

The SC agreed that "pending the results of the visualisation of scenario space analysis..., additional cross-validation trials may be specified..." (IWC, 2000b). Thus, the SC has shown interest in seeing these analyses repeated to clarify the role and suitability of cross-validation trials. Once robustness and cross-validation trials have been coded, analyses similar to ours can begin to shed light on these broader aspects of trial design.

ACKNOWLEDGEMENTS

This work was partially supported by the North Slope Borough (Alaska), the State of Alaska (through the Alaska Department of Community and Regional Affairs), and the National Oceanic and Atmospheric Administration (through the National Marine Mammal Laboratory to the Alaska Eskimo Whaling Commission). We thank the participants in the 1999 IWC Second Workshop on the Development of an AWMP for their helpful suggestions about some of the ideas presented here. We are also grateful to Cherry Allison for her intense effort to develop and test the IWC's bowhead AWMP simulation testing software, comprising

nearly 15,000 lines of code and related files, in time for analyses to be completed for the 2000 SC meeting.

REFERENCES

Butterworth, D. S. (1995). A brief note on plausible hypotheses, the interpretation of implementation simulation trials, and all that! Paper SC/47/MG5 presented to the IWC Scientific Committee, May, 1995.

Butterworth, D. S., Punt, A. E., and Smith, A. D. M. (1996). On plausible hypotheses and their weighting, with implications for selection between variants of the revised management procedure. *Rep. int. Whal. Commn*, 46:637–640.

Givens, G. H. (1998). AWMP development and diverse prototypes. *Rep. int. Whal. Commn*, 48:483–495.

Hilborn, R. (1996). Uncertainty, risk, and the precautionary principle. In Pikitch, E. K. *et al.*., editor, *Fisheries Management: Global Trends*. American Fisheries Society, Bethesda MD.

International Whaling Commission (1995). Chairman's report of the 46th annual meeting, Appendix 4. *Rep. int. Whal. Commn*, 45:42–3.

International Whaling Commission (2000a). Report of the Scientific Committee, Annex ?: Report of the Standing Working Group on the Development of an Aboriginal Subsistence Whaling Management Procedure (AWMP). *Journal of Cetacean Research and Management*, 2 (Suppl.):to appear.

International Whaling Commission (2000b). Report of the Second Workshop on the Develpment of an Aboriginal Subsistence Whaling Management Procedure (AWMP). Available from the IWC Secretariat.

International Whaling Commission (2000c). Report of the Scientific Committee, Annex ?: Report of the Sub-Committee on Aboriginal Subsistence Whaling. *Journal of Cetacean Research and Management*, 2 (Suppl.):to appear.

Punt, A. E. (1999). An examination of a stochastic population dynamics model for the Bering-Chukchi-Beaufort Seas stock of bowhead whales. Paper SC/51/AS1 presented to the IWC Scientific Committee, May, 1999.

Wade, P. R. and Angliss, R. P. (1997). Guidelines for assessing marine mammal stocks: Report of the GAMMS workshop April 3-5, 1996, Seattle, Washington. US Department of Commerce, NOAA Technical Memorandum NMFS-OPR-12, 93p.