Appendix D. Supplemental Catch Estimation Methodology Review Report.

This addendum to the *Methodology Review Panel Report: Catch Estimation Methodology Review* describes additional follow-up work by the SWFSC analytical team (the Team) to address a series of short-term questions developed by the methodology review panel (the Panel) at its March 2018 methodology review. The follow-up material was presented to the Panel at a day-long webinar held on 31 July 2018. The draft agenda from the webinar is attached as Appendix E. The list of participants is included in part (2) of Appendix C. The terms of reference for the supplemental review consisted of a series of requests from the list of short-term requests in the main *Methodology Review Panel Report*. The requests are provided below, along with corresponding descriptions of the Team's responses. The requests follow the same ordering as in the main *Methodology Review Panel Report* and do not necessarily follow the same order as their presentation during the webinar.

Request #1. As a diagnostic template, for each sampled stratum compare the posterior predictive distributions at the 68th, 95th, and 99th percentiles with the current observed species proportions (create fully stratified versions of tables 2 and 3 in the Grunloh et al. methods documentation). With each row, include sample sizes and associated landing weights with a graphical display to highlight problems and outliers (circle size proportional to landing weights).

Rationale: The Team [at the March meeting] provided broad-scale summary metrics (e.g., MSE and DIC) for evaluating the goodness-of-fit of the different model forms and structures. Fine-scale diagnostics are needed to help identify aspects of the data that are not adequately addressed by the different models. The diagnostic template will provide a mechanism for fine-scale exploration of goodness-of-fit.

Team Response: The Team prepared a set of tutorial slides about their modeling approach, a refresher on its background, and a description and examples of the tool they had developed for evaluating model diagnostics and displaying the results. Many of the slides presented during the webinar focused on the Beta-binomial model M4 that included fixed main effects in the set of parameters for *Species*, *Port*, and *Gear* and random effects with a single estimated variance parameter for the *Year* : *Quarter* interactions. Because the models produce results for a very large number of *Species – Port – Gear – Year – Quarter* combinations, it was infeasible to thoroughly consider results except at aggregated levels (e.g., *Species – Gear – Year*, *Species* – Year) and for a limited number of species. However, the slides included clickable links to a Github repository[1] where results are available in the form of pdf figures and corresponding .csv files. An example diagnostic plot is provided in Figure D1 and the corresponding .csv file is in Table D1.

The diagnostic plots depict the proportion of sample *%Species* observations that are covered by the model's posterior predictive highest density intervals, HDI (i.e. whether they match). In each diagnostic plot there is a vertical line indicating a reference posterior predictive HDI and a series of horizontal lines and endpoints (indicating results for different classifications of the data) that show the proportions of observed values covered by the HDI. This HDI approach (unlike the more familiar confidence intervals) can accommodate bimodal intervals. Deviations in either direction are treated equally and the approach does not take into account where inside or outside

[1] https://github.com/gasduster99/sppComp/tree/master/sscRuns

the interval the data occur (i.e., the diagnostic does not focus on model bias). Note that with very sparse sampling there will be many observed zeroes and possibly some samples that were purely composed of one species, particularly in cases in which a given species occurs only once or twice within a market category: time period combination. For such samples the diagnostic plot will produce a whisker ending at zero or one. Several of the examples of "poor" fits reflect this circumstance.

Some discussion focused on how to interpret diagnostic plots in which there was good correspondence between the observed distributions of *%Species* values and the posterior predictive HDIs for the 95% reference level, but poor correspondence for the 68% reference level (e.g., Figure D2). There was insufficient time to uncover the source(s) of the differences.

To help focus attention and provide a mechanism for ranking species-level results, the Team used a weighted Mean Absolute Deviation (MAD) statistic based on the product of the relative landings times the absolute difference between the observed predictive accuracy and the nominal 68^{th} (or 95^{th}) prediction accuracy. A given species would receive a low MAD score (better performance) either if it occurred in relatively low amounts of the market category landings or if there were small differences between the observed predictive accuracy and the nominal prediction accuracy. Conversely, a species would receive a high MAD score if it occurred in relatively low amounts of landings or if there were small differences between the observed predictive accuracy and the nominal prediction accuracy.

Request #2. The diagnostic template should be developed for each of the sensitivity runs (vary across a range of plausible time models and priors and limit to the top 2-3 market categories).

Rationale: Application of the diagnostics across a wide range of models will form a test of how well the diagnostics illustrate whether the models capture important structural features that are thought to be embedded in the data.

Team Response: The Team organized their response into two sets of sensitivity runs, one that compared different model forms to account for temporal variation (models M1 to M6, using diffuse normal priors for the beta parameters) and another that compared different prior distributions for model M4. They presented similar sets of comparative diagnostic plots for market categories 250, 253, and 269. For each market category they presented comparisons of results for the three "best" models, as measured by Δ-DIC.

Figure D3 (left panel) provides an example of the diagnostic plot for bocaccio in market category 253 (nominal bocaccio) comparing models M4, M5, and M6. These models all produced very similar Δ-DIC scores, ranging from 0 to 0.07, and the diagnostic plot also indicates relatively small differences among the three models. Also shown in Figure D3 (right panel) is the corresponding diagnostic plot for model M1, which had a very large Δ-DIC score (1409). The diagnostic plot indicates considerable under-coverage of the observations by the model's posterior predicted interval.

Figure D4 provides examples of diagnostic plots that illustrate the influence of the form assumed for the prior when used with model M4.

The Team's general conclusion regarding the influence of the distribution assumed for the prior was that the prior had little influence in a data-rich setting, the U4 prior often performed well in data-poor settings, and the U4 prior was "less stable" in a data-poor setting, perhaps due to flat likelihood surfaces leading to lack of convergence.

Request #3. Explore an alternative time block: an extension of 1983 and 1984 to the first time block.

Rationale: The panel [at the March meeting] expressed concerns about how the model would perform when applied to shorter time periods, as will occur when the model is used with data more recent than 1990. Results from the above recommendation could be compared to the results from the current two time blocks (1978-1982; 1983-1990) to explore how fits to data from the late period degrade when the model for the late period is based on fewer years of data. Also, comparisons of the two forms of blocking serve as a sensitivity evaluation of the selection of the block boundary, which was chosen on a fairly arbitrary basis.

Team Response: The Team presented slides comparing diagnostic plots for model M4 (with diffuse normal priors for the beta parameters) based on data from four different sets of time periods: (1) 1978 to 1982 (as in the results presented in March and in other sections of the July webinar presentation); (2) 1978 to 1983; (3) 1978 to 1984; and (5) 1978 to 1985. Examples are provided in Figure D5.

The Team concluded that results for model M4 are reasonably robust to the ending year of the time blocking, but performance was variable across the different market categories. The Team's analysis only considered the early time period of the available data, which starts with 1978.

The Team did not address the issue raised in the rationale for this request regarding how "fits to data from the late period degrade when the model for the late period is based on fewer years of data."

Request #4. Explore various two-way interactions (beyond the current explorations; e.g., Species : Port and Species : Gear).

Rationale: The Team [at the March meeting] did not have time to search across the multitude of possible interaction terms that they could have included in the model. From various anecdotal comments made during the review it seemed likely that the model would benefit from the inclusion of other interaction terms. Explorations with the diagnostic template may suggest potentially beneficial terms.

Team Response: The Team presented a series of slides with diagnostic plots that illustrated the effect of including *Species* : *Port* and *Species* : *Gear* interaction terms in model M4. The models with the interactions produced large reductions in DIC score compared to the model without these interaction terms. The "best" fits overall were for the model with *Species* : *Port* interactions. Examples of the comparative diagnostic plots are given n Figure D6.

The Team concluded that both interactions may be appropriate to include, but the effects are market category dependent. They suggested that models with the interactions should have "shrinkage" priors that would allow the model to predict species compositions for unsampled strata and reduce the effective number of parameters.

Request #5. Redo the modeling of the early time block without southern CA ports. Explore spatially and temporally (i.e., alternative time blocks).

Rationale: The available dataset does not have any sample data in the early time block from the southern CA ports. It was unclear how this lack of data influenced the model results. The requested analysis will clarify the situation.

Team Response: The Team produced a set of slides with diagnostic plots for market category 250 (unspecified rockfish) that compared results for models that either included or excluded ports south of Point Conception (e.g., Figure D7). The diagnostic plots were very similar, which implied that the *%Species* estimates were not greatly influenced by the lack of data from the south. The Team concluded that "making predictions in unobserved strata, does not affect predictions in observed strata".

Request #6. Compare alternative ComX outputs and the current time series of estimated catches.

Rationale: It would be informative to see the landings estimates corresponding to the additional models developed in response to the above requests. The landings estimates can be generated for a small set of illustrative species and do not need to be comprehensive.

Team Response: The Team produced the requested landings estimates for the models that they explored in connection with their responses targeted at (a) the model used to accommodate temporal variation (e.g., M1 to M6), (b) the assumed prior for the beta parameters (e.g., diffuse normal versus inverse gamma), (c) the use of interaction terms (*Species* : *Port* and *Species* : *Gear*), and (d) the assumed ending year for the early time-block. Examples are provided in Figure D8.

There was discussion about an unusual feature apparent in some of the landings plots, in which the median estimates were essentially zero whereas the mean values were relatively far from zero and sometimes fell above the 80% credibility interval (e.g., the lower two panels in Figure D8). Evidently the model's landings estimates can have unusual distributions.

During the concluding webinar discussions the Team proposed developing an additional diagnostic tool that might provide insights into how well the model is able to predict the distribution of the data. The procedure would involve overlaying on the landings plots some purely data-based estimates of species landings (with no data-borrowing as in CalCOM).

Request #7. Provide a summary table of species' sample sizes in each market category by time block.

Rationale: The requested information will assist in understanding where there are gaps in the available data that the model is filling in by means of its pooling structure.

Team Response: The Team produced a set of Excel workbooks that had summary information regarding port sample data on species compositions for (1) flatfish and skate market categories, (2) the major rockfish market categories, and (3) the remaining rockfish (and other) market categories. The workbooks included sample-by-sample summary information (e.g., Table 2), with each row representing one sampled fish and 0/1 flags to mark each unique market category sample and each unique species with a sample. Each workbook also had a worksheet with an interactive pivot table that could be queried to show the species that appeared in each market category for any combination of stratum levels and the number of unique market category samples for that combination of stratum levels (e.g., Table 3).

In discussions following the webinar it was discovered that the Excel workbooks inadvertently included a modest number of samples that have not been used in either previous CalCOM catch estimates or in the model-based approach used here, as they lack corresponding landings weight information. This inconsistency reflects a small fraction of all of the species composition

samples (116 of 21,275 for market categories 245 to 271). However, this did lead to some discrepancies between the spreadsheets and the sample sizes reported on the diagnostic plots in several instances.

Request #8. Provide self-test documentation (simulated data) for example models.

Rationale: Results from this analysis will provide a demonstration of model performance under best-case scenarios, where the data being analyzed exactly conform to the assumptions of the statistical model. The analysis will serve to verify (or refute) that the model performs as expected.

Team Response: The Team did not have enough time to produce any self-test documentation to demonstrate that the model works correctly with simulated data.

Conclusions and recommendations

The Panel and Team identified the following items as requiring additional attention.

- Resolve whether or not to weight composition samples by the landed catch amounts. The current version does not weight the sample %*Species* observations, which means the model treats sample data from a 10,000 pound landing as being equivalent to sample data from a 100 pound landing. In constrast, CalCOM weights the sampled %*Species* by the sampled landing weight.
- Attempt to identify some of the major source(s) of the discrepancies between the model-based estimates of landings and the CalCOM estimates (based on data borrowing). Although such discrepancies are to be expected, some of the major noted discrepancies, such as difference between widow rockfish catches early in the time period, are substantial enough to warrant greater concern and investigation. Two potential sources are CalCOM's data-borrowing and its weighting by the landings.
- Investigate the effects on model performance of having increased number of market categories over time. There could be conflicting effects from having increased numbers of market categories. With more market categories there are likely to be fewer samples per market category (causing a loss of accuracy), but the purity of the samples may improve due to having fewer component species (causing an increase in accuracy).
- As an additional diagnostic tool, compute posterior predictive distributions of the landings for the sampled strata and compare these to the sampled data expanded to the sampled landings (i.e., with no data borrowing). If possible, develop a form of Q-Q plot diagnostic that compares the two sets of estimated landings. In addition, this exercise could compare landings-weighted estimates and unweighted estimates.

The Team was very industrious, responsive, and successful at addressing the short-term requests made by the Panel at the March review meeting. While the Team does not yet have a final model that the Panel can endorse, the Team has taken a large and important step forward. The Team indicated that their target is to develop and finalize the model in time for producing revised landings estimates for California that could be used in the 2021 round of groundfish stock assessments.

Appendix D Figures

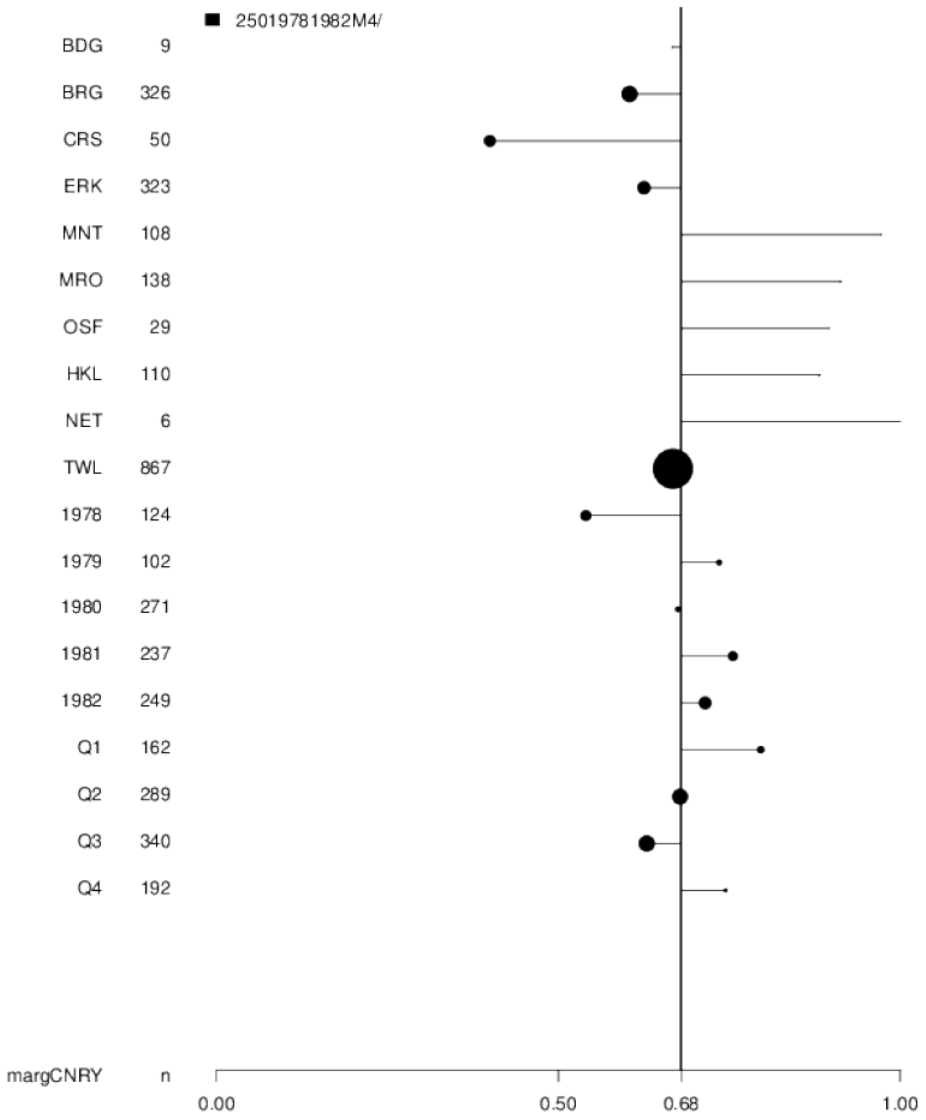

Figure D1. A summary ("marginal") diagnostic plot for canary rockfish in market category 250 (unspecified rockfish) during 1978 to 1982 for the M4 beta-binomial model. The nominal prediction interval is 68%, represented by the vertical line. Each row represents for the given stratum level the summation across levels of all the other strata (i.e., along the margin of the multiway cross-classification). *Ports* are the top rows (BDG to OSF), followed by *Gears* (HKL to TWL), followed by *Years*, and then by *Quarters*. For the horizontal lines pointing to the right the model prediction intervals cover more than 68% of the observed values (over-coverage). For horizontal lines pointing to the left the model prediction intervals cover less than 68% of the observed values (under-coverage). The diameter of the circle represents the magnitude of the market category landings for the given stratum level.

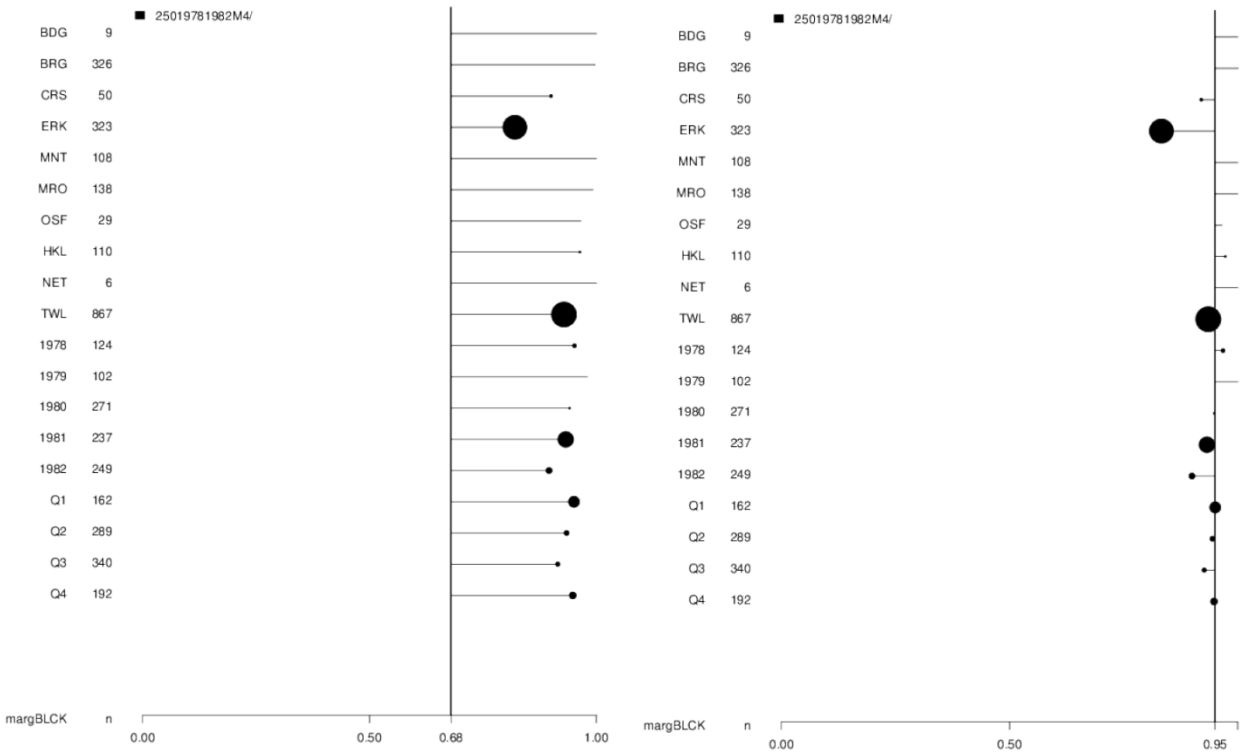

Figure D2. Marginal diagnostic plots for black rockfish landed in the market category 250 (unspecified rockfish) during 1978 to 1982 for the M4 beta-binomial model. The left panel, which is for the 68% nominal prediction interval, indicates over-coverage by the model. The right panel, for the 95% nominal prediction interval, indicates good coverage by the model's posterior predictive highest density intervals of the observed distributions of %*Black*.

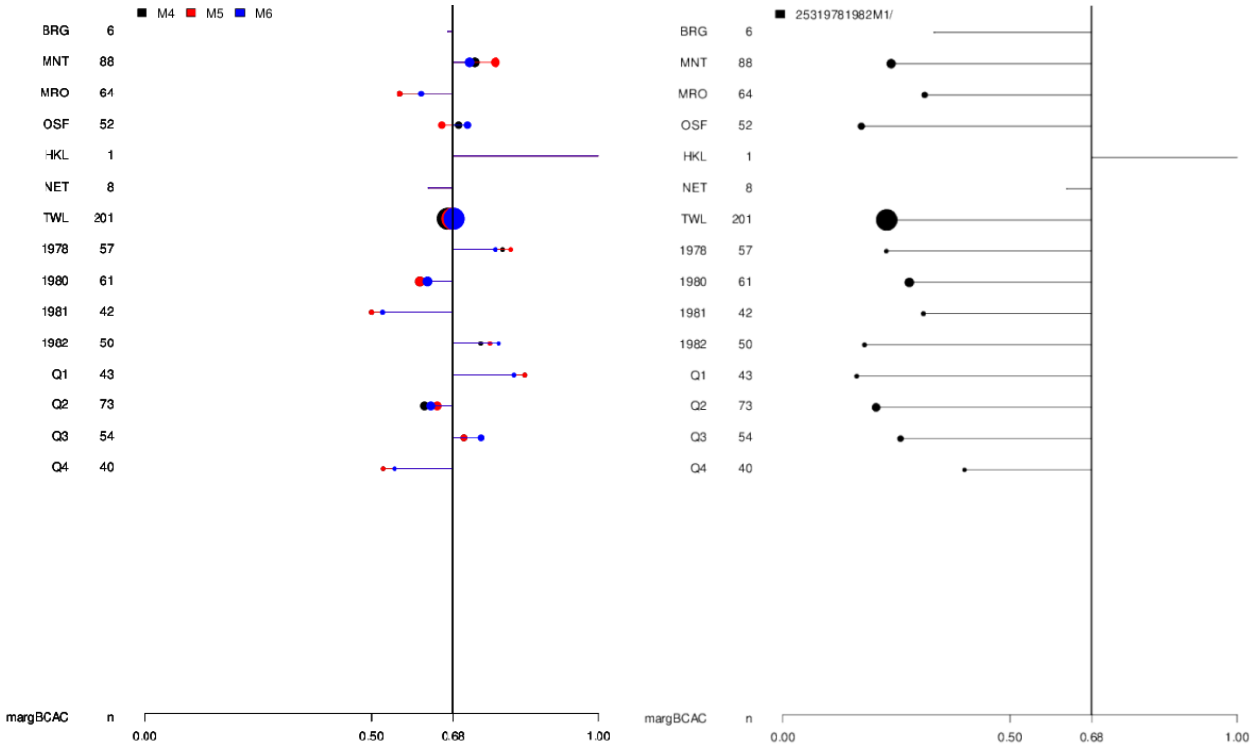

Figure D3. Marginal diagnostic plots for bocaccio in market category 253 (nominal bocaccio) during 1978 to 1982. The left panel compares the coverages for models M4 (in black), M5 (in red), and M6 (in blue). The right panel shows the coverages for Model M1, which had a very large DIC score compared to the other three models.

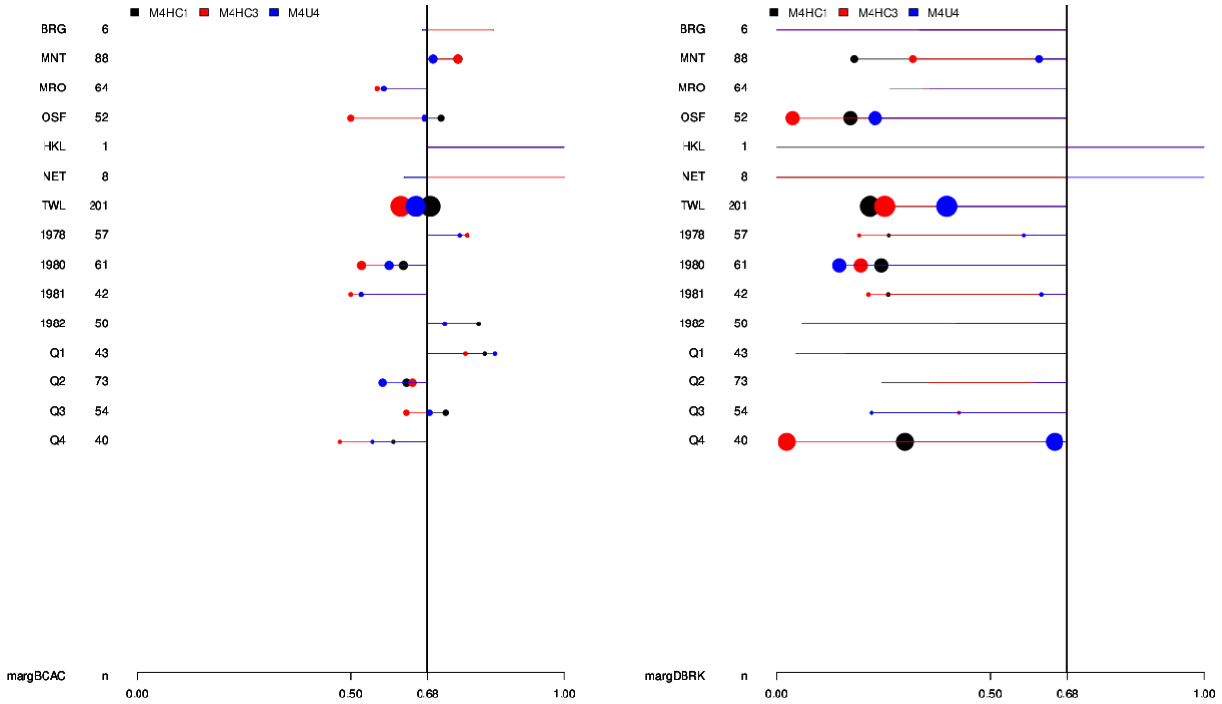

Figure D4. Examples of marginal diagnostic plots that illustrate the sensitivity to the assumed prior distributions for the beta parameters when used with model M4. Model M4HC1 (in black) uses a half-Cauchy(10), M4HC3 (in red) uses a half-Cauchy(10^3), and M4U4 (in blue) uses a uniform(0,10^4). The left panel compares the coverages for bocaccio; the right panel shows the coverages for darkblotched rockfish. Both are for market category 253 during 1978 to 1982.

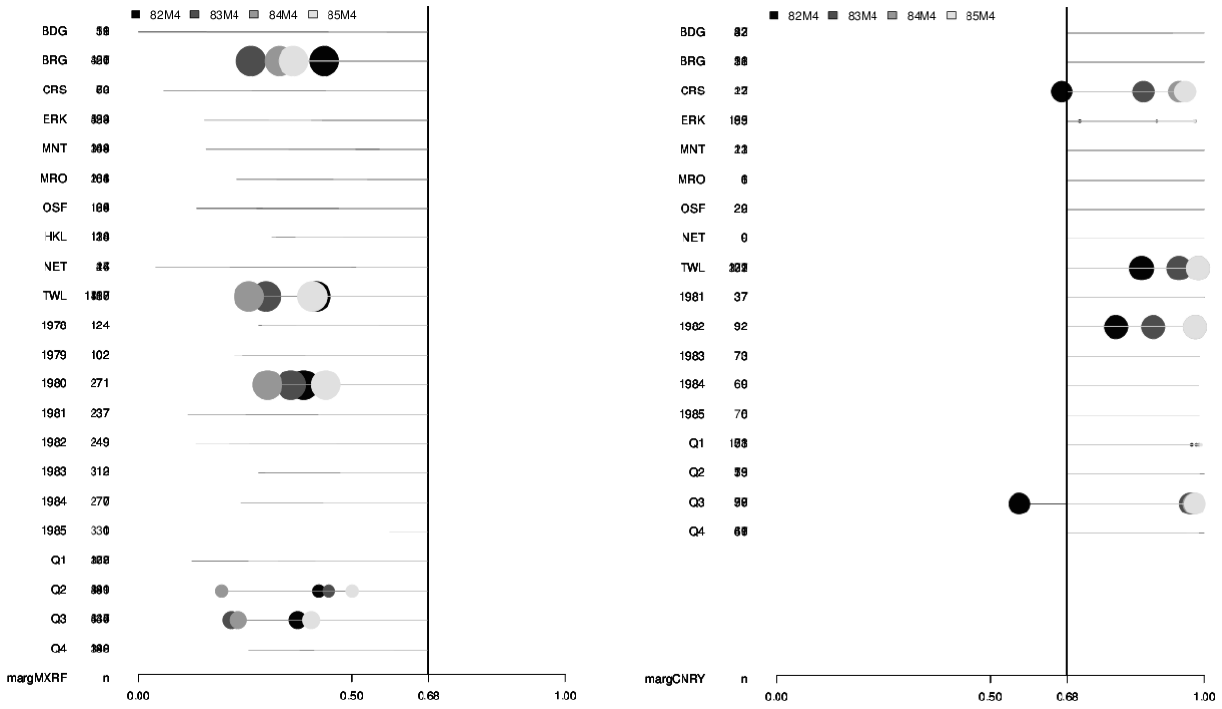

Figure D5. Examples of marginal diagnostic plots that illustrate the sensitivity to the number of years included in the underlying data set used to inform the model. The left panel compares the coverages for Mexican rockfish in market category 250 (unspecified rockfish); the right panel shows the coverages for canary rockfish in market category 269 (nominal widow rockfish).

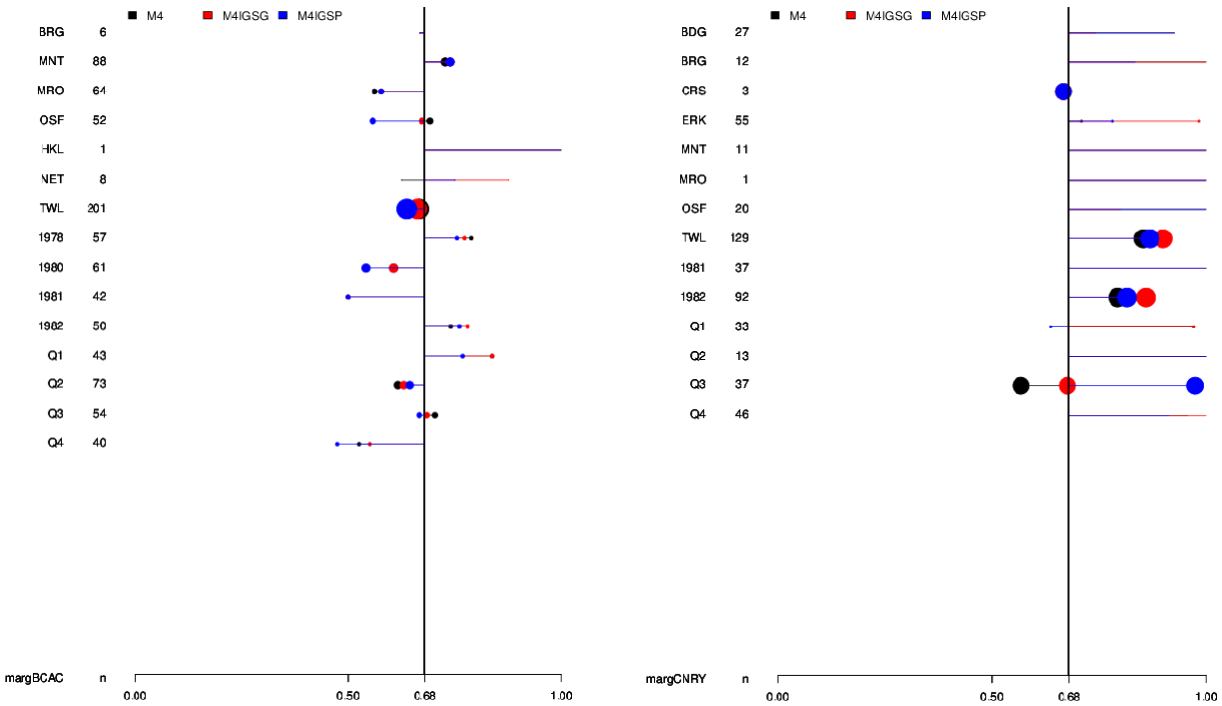

Figure D6. Examples of marginal diagnostic plots that illustrate the influence of adding interaction terms for *Species* : *Gear* (M4IGSG) and *Species* : *Port* (M4IGSP) to model M4. The models with the added interaction terms have an inverse gamma prior for the beta parameters. The left panel compares the coverages for bocaccio in market category 253 (nominal bocaccio); the right panel shows the coverages for canary rockfish in market category 269 (nominal widow rockfish).

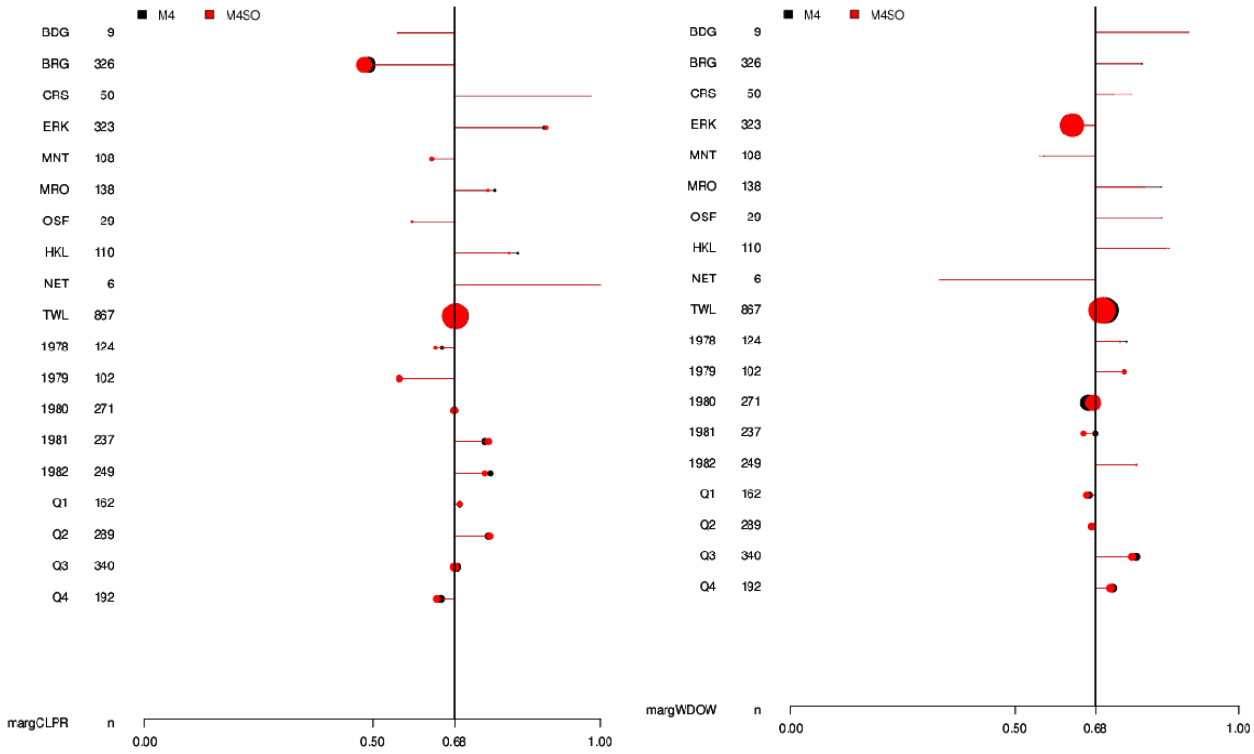

Figure D7. Marginal diagnostic plots for chilipepper rockfish (left panel) and widow rockfish (right panel) in market category 250 (unspecified rockfish) during 1978 to 1982 using M4 beta-binomial models that either include the southern ports (in black) or exclude them (in red).

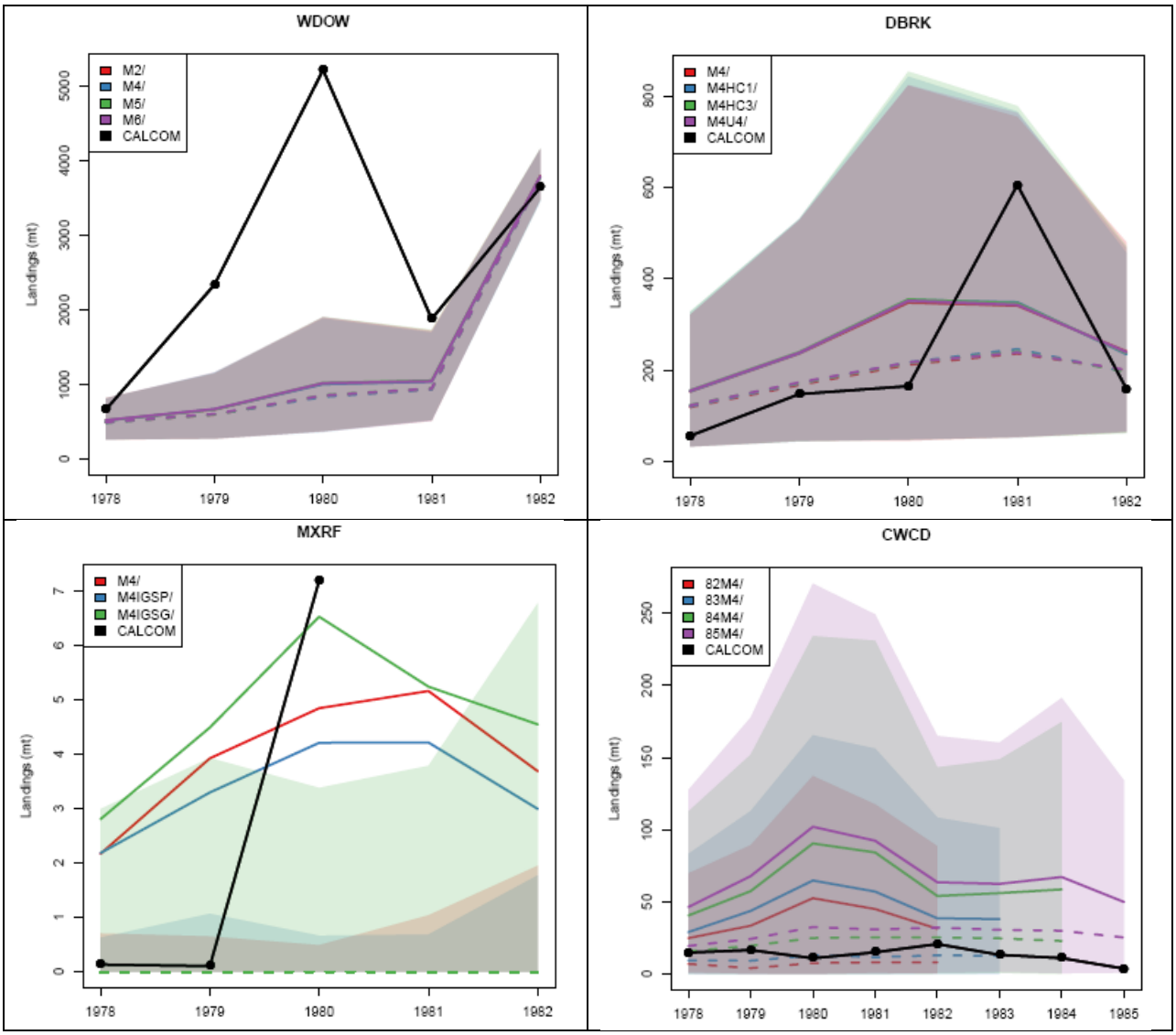

Figure D8. Examples of plots comparing annual landings for individual species (aggregated across all gear types) for different forms of the beta-binomial model (in color) and as estimated using the data-borrowing approach (CalCOM, in black). The solid lines (in color) are the mean estimates, the dashed lines are the median estimates, and the shaded regions show the 80% credible intervals.

Appendix D Tables

Table D1. Contents of the.csv file corresponding to the marginal diagnostic plot for canary rockfish displayed in Figure 1. Column "n" is the number of sampled landings. associated with the given row, column "landing" is the number of pounds of landings from market category 250 (unspecified rockfish) assigned to the marginal species being fit (here, canary rockfish) associated with the given row, and coverage is the proportion of the observed sample *%Species* (by weight) for the given row that are covered by the model's posterior predictive highest density intervals computed from 10,000 random draws. ***>>> DBS: Confirm that this is correct.***

	margCNRY	n	landing	coverage
2	BDG	9	66834	0.666666666666667
3	BRG	326	1975794	0.604294478527607
4	CRS	50	1398777	0.4
5	ERK	323	1577283	0.625386996904025
6	MNT	108	20852	0.972222222222222
7	MRO	138	100136	0.91304347826087
8	OSF	29	7637	0.896551724137931
9	HKL	110	132906	0.881818181818182
10	NET	6	0	1
11	TWL	867	5014407	0.667820069204152
12	1978	124	1259855	0.540322580645161
13	1979	102	626077	0.735294117647059
14	1980	271	615166	0.675276752767528
15	1981	237	1146147	0.755274261603376
16	1982	249	1500068	0.714859437751004
17	Q1	162	848186	0.796296296296296
18	Q2	289	1937730	0.678200692041522
19	Q3	340	1978133	0.629411764705882
20	Q4	192	383264	0.744791666666667

Table D2. Example of the sample-level summary information in the Excel workbook *Panel.sampledata.mcs245to271.xlsx*, which contained information for species compositions samples for the major rockfish market categories.

sample_no	mark_cat	live_fish	port_complex	gear_grp	clust_no	species	fish_no	year	quarter	time. period	unique. sample	unique. species
197800057	250	N	MRO	HKL	1	BCAC	1	1978	1	1978-82	1	1
197800057	250	N	MRO	HKL	1	BCAC	2	1978	1	1978-82	0	0
197800057	250	N	MRO	HKL	1	BCAC	3	1978	1	1978-82	0	0
197800057	250	N	MRO	HKL	1	BCAC	4	1978	1	1978-82	0	0
197800057	250	N	MRO	HKL	1	BCAC	5	1978	1	1978-82	0	0
197800057	250	N	MRO	HKL	1	BCAC	6	1978	1	1978-82	0	0
197800057	250	N	MRO	HKL	1	BCAC	7	1978	1	1978-82	0	0
197800057	250	N	MRO	HKL	1	BCAC	8	1978	1	1978-82	0	0
197800057	250	N	MRO	HKL	1	BCAC	9	1978	1	1978-82	0	0
197800057	250	N	MRO	HKL	1	BCAC	10	1978	1	1978-82	0	0
• • •	• • •	• • •	• • •	• • •	• • •	• • •	• • •	• • •	• • •	• • •	• • •	• • •
201780118	257	N	CRS	TWL	1	DBRK	13	2017	4	2010-17	0	0
201780118	257	N	CRS	TWL	1	DBRK	8	2017	4	2010-17	0	0
201780118	257	N	CRS	TWL	1	DBRK	9	2017	4	2010-17	0	0
201780118	257	N	CRS	TWL	1	DBRK	10	2017	4	2010-17	0	0
201780118	257	N	CRS	TWL	1	DBRK	11	2017	4	2010-17	0	0
201780118	257	N	CRS	TWL	1	DBRK	12	2017	4	2010-17	0	0
201780118	257	N	CRS	TWL	1	DBRK	14	2017	4	2010-17	0	0
201780118	257	N	CRS	TWL	1	DBRK	15	2017	4	2010-17	0	0
201780118	257	N	CRS	TWL	1	DBRK	16	2017	4	2010-17	0	0
201780118	257	N	CRS	TWL	1	DBRK	17	2017	4	2010-17	0	0

Table D3. Example pivot table for the Excel workbook *Panel.sampledata.mcs245to271.xlsx*, which that contained summary information for species compositions samples for the major rockfish market categories.

mark_cat	250
gear_grp	TWL
port_complex	ERK
live_fish	N
quarter	2

Sum of unique.sample	**Column Labels**			
Row Labels	**1978-82**	**1983-90**	**1991-99**	**Grand Total**
ARRA	4	9	5	18
BANK	3	21	11	35
BCAC	59	88	0	147
BLCK	11	9		20
BLGL		3	0	3
• • •	• • •	• • •	• • •	• • •
VRML	0	0		0
WDOW	21	0	0	21
YEYE		0	0	0
YMTH		0	0	0
YTRK	0	1	0	1
Grand Total	**117**	**164**	**27**	**308**