Groundfish Management Team Report on Updates to the Sablefish Trip Limit Model

Prepared for the October 12, 2023 Virtual Workshop of the Groundfish and Economics Subcommittees of the Scientific and Statistical Committee

Table of Contents

Section 1. Introduction

The Groundfish and Economics Subcommittees of the Scientific and Statistical Committee ("subcommittees") held a workshop on May 9, 2023, to review the sablefish trip limit model used by the Groundfish Management Team (GMT) to project landings under alternative trip limits in the Limited Entry Fixed Gear (LEFG) and Open Access (OA) fisheries. The GMT submitted a report for the May 9 workshop that described the fisheries and the models in detail, as well as potential model improvements the team explored prior to the workshop (Sablefish Daily Trip Limit Model Methodology Review 2023). The subcommittees provided their recommendations and requests from the May 9 workshop within the following three overarching categories, with some additional "other recommendations":

1) Alternative models should be evaluated using different metrics based on out-of-sample prediction performance.

2) Forecast methods should incorporate period-specific outcomes in a transparent and systematic way.

3) Consider the following when choosing explanatory variables to include in the models.

The GMT did not have adequate time to explore or accomplish all recommendations provided in the subcommittees' report, but we prioritized #1 and #2 of the overarching categories listed above to the best of our ability. The GMT performed similar updates and evaluations for both the LEFG and OA models north of 36° N. lat. (LEN and OAN, respectively), so the methods for those updates and evaluations are first described in a general sense in Section 2. The remainder of the report focuses more specifically on the two sectors separately and the results of model evaluations for each. Corresponding subcommittee recommendations from their report on the May 9 workshop are italicized throughout this report for reference.

Section 2. General Updates and Evaluation Methods

Recommendation: The subcommittees recommend that the models be evaluated based on out-of-sample performance. Cross-validation procedures that use part of the data to estimate a model and then compare it to the unused observations (e.g., leave-one-out, one-step-ahead) are potential methods.

Models were compared using a time-series cross validation (TSCV) procedure with an expanding window. Each cross validation fold uses a set of training data to estimate parameters of the models, then uses these parameterized models to predict observations in the test data. To do this, we used the `train()` function in the R package "`caret`" (R documentation). To specify a method of TSCV using the `trainControl()` function, the GMT set method = "timeslice", which calls `caret`'s internal function `createTimeSlices()` to cross validate time series data. The function `createTimeSlices()` creates the indices for rolling forecasting origin techniques that move the training and test sets in time, as described by Hyndman and Athanasopoulos (2013). The *initial window* parameter, which sets the amount of data in the first training set, was set at 18, which represents three years' worth of data (i.e., three years with six periods per year). The *horizon* parameter, which represents the amount of data to predict out-of-sample, was set at 12, which represents two years' worth of data. We chose two years, because that is the farthest out that the model is used for predictions (i.e., the biennial management cycle). We also turned off the *fixed window* parameter and allowed the training set window to grow in size with each run. We used those parameter settings for all model comparisons across both sectors. We then compared the resulting root mean squared error (RMSE) for each model to determine the model that best performs when predicting out-of-sample.

The GMT also performed leave-one-out cross validation (LOOCV) using the `train()` function by setting `trainControl()` method = "LOOCV". No other parameters are required to be set for LOOCV. For most model comparisons, the "best model" based on TSCV was the same as the "best model" based on LOOCV. In the few instances in which they conflicted, the GMT deferred to the TSCV "best model", given that TSCV better accounts for time-based predictions into the future. In this report, the GMT only presents the results of the TSCV for simplicity, with the exception of comparing models with different sets of years.

Recommendation: Add fixed effects for 2020 and 2021 to account for effects of the COVID-19 pandemic.

The GMT added fixed effects for 2020 and 2021 by creating a dummy variable called "COVID". For all 2020 and 2021 data points, the value for COVID is set at 1, and for data from all other years, the value for COVID is set at 0. The COVID dummy variable was then included in the model as a fixed effect. For example, the LEN linear model equation estimating pounds landed per vessel would be as follows:

$$Average\ lbs.\ per\ vessel \sim bimonthly\ trip\ limit\ +\ factor(COVID)$$

As shown in the time series cross validation (TSCV) results in the following sections, the COVID fixed effects variable generally improved model performance for both the LEN and OAN sectors. However, models with and without the COVID dummy variable seemed to generate identical forecasts for 2014-2020, because the COVID dummy variable cannot be estimated for any training data set that does not include, at a minimum, 2020. Therefore, using an initial window of 18 and a horizon of 12 means that models with and without the COVID fixed effects generated identical forecasts in the majority of the cross validation folds. This explains why in many cases, the TSCV results are very similar for models with and without the COVID variable that are otherwise identical. Therefore, it is difficult to fully interpret whether the COVID variable does or does not provide meaningful improvement to out-of-sample predictions solely based on the TSCV results, but the GMT does see logical merit in including the variable given the minimal difference in retrospective in-sample error and the significant difference in participation and landings during those years.

Recommendation: Estimate a pooled model (a single equation with all periods included in the data) with period specific fixed effects.

Rather than a separate linear model for each period, as is the status quo method, the GMT explored using period-specific fixed effects by including the bimonthly period as a factor variable in the model. For example, the linear model equation estimating pounds landed per vessel would be as follows:

$$Average\ lbs.\ per\ vessel \sim bimonthly\ trip\ limit\ +\ factor(PERIOD)$$

The GMT did not use cross validation to compare this approach to using a separate model for each period, but we did compare it to a baseline model that does not account for period effects at all. Based on the TSCV results, using period-specific fixed effects resulted in a lower RMSE and therefore performed better at predicting out-of-sample than the baseline model for both sectors, and the team concluded that this approach is preferable to the status quo method, given the subcommittees' recommendation.

Recommendation: The analysts should consider incorporating time components in the model. The simplest method is to include a time trend. Using lagged dependent variables to incorporate autocorrelation is another option.

In addition to the models described in Sections 3.2.1, 3.3.1, 4.2.1, and 4.3.1, we used TSCV to test two models with predictor variables introduced to account for trend or memory in the time-series, one using a lagged dependent variable (LDV Model) and one using a time trend (Linear Trend Model). The R package "`caret`" was not used to train and test these models, but rather, functions were written to split and cross validate the time-series data. However, the same initial window (18) and horizon (12) parameters were used, with a growing training set window at each fold.

For both sector's models predicting average pounds per vessel, these additional models were compared against a model with bimonthly trip limits, period fixed effects, and COVID fixed effects (Base + Covid Model), as well as a similar model that did not include COVID fixed effects (Base Model). Estimating equations for the three models compared for the LEN and OAN models predicting pounds per vessel are:

1. The *Base Model*: $AVG\,LB_t = \alpha + \sum_{j=2}^{6} \gamma_j PERIOD_t^j + \beta_1 TL.BIMON_t$
2. The Base + Covid Model: $AVG\,LB_t = \alpha + \sum_{j=2}^{6} \gamma_j PERIOD_t^j + \beta_1 TL.BIMON_t + \eta D_t^{covid}$
3. The *LDV Model*: $AVG\,LB_t = \alpha + \sum_{j=2}^{6} \gamma_j PERIOD_t^j + \beta_1 TL.BIMON_t + \phi AVG\,LB_{(t-1)}$
4. The *Linear Trend Model*: $AVG\,LB_t = \alpha + \sum_{j=2}^{6} \gamma_j PERIOD_t^j + \beta_1 TL.BIMON_t + \delta t$

where,
t is a time ordered index of observations, and
$PERIOD_t^j = 1$ if observation t occurs in period j and 0 otherwise.

Table 1 displays the RMSE results for model predictions generated by the TSCV to compare time-based variables. To evaluate each model's tendency to over or under predict, Figures 1 and 2 display the distributions of absolute forecast errors generated by the TSCV for the LEN and OAN sectors, respectively. For the LEN sector, the Linear Trend Model resulted in the lowest

RMSE and a median forecast error closest to zero. For the OAN sector, the Base + Covid Model resulted in the lowest RMSE, followed by the Base Model, the latter of which also had a median forecast error closest to zero. Both the LDV Model and the Linear Trend Model led to more overpredictions than the Base or Base + Covid models. The GMT did not have time to complete similar comparisons for the two sectors' participation models at the time of writing this report, but we may be able to provide those results at the workshop. We also did not include any time component models in the evaluations of risk to the sablefish north ACL, but if a time component is added to any of the models based on improved performance, it can be assumed that the risk to the sablefish north ACL would be lower than what is already identified in this report.

Table 1. Expanding window time-series cross validation results to compare time-based variables.

Estimating Equation	**RMSE**	
	LEN	**OAN**
$AVG\,LB_t = \alpha + \sum_{j=2}^{6} \gamma_j PERIOD_t^j + \beta_1 TL.BIMON_t$	639.27	359.41
$AVG\,LB_t = \alpha + \sum_{j=2}^{6} \gamma_j PERIOD_t^j + \beta_1 TL.BIMON_t + \eta D_t^{covid}$	638.43	342.62
$AVG\,LB_t = \alpha + \sum_{j=2}^{6} \gamma_j PERIOD_t^j + \beta_1 TL.BIMON + \phi AVG\,LB_{(t-1)}$	599.12	369.54
$AVG\,LB_t = \alpha + \sum_{j=2}^{6} \gamma_j PERIOD_t^j + \beta_1 TL.BIMON + \delta t$	598.18	392.84

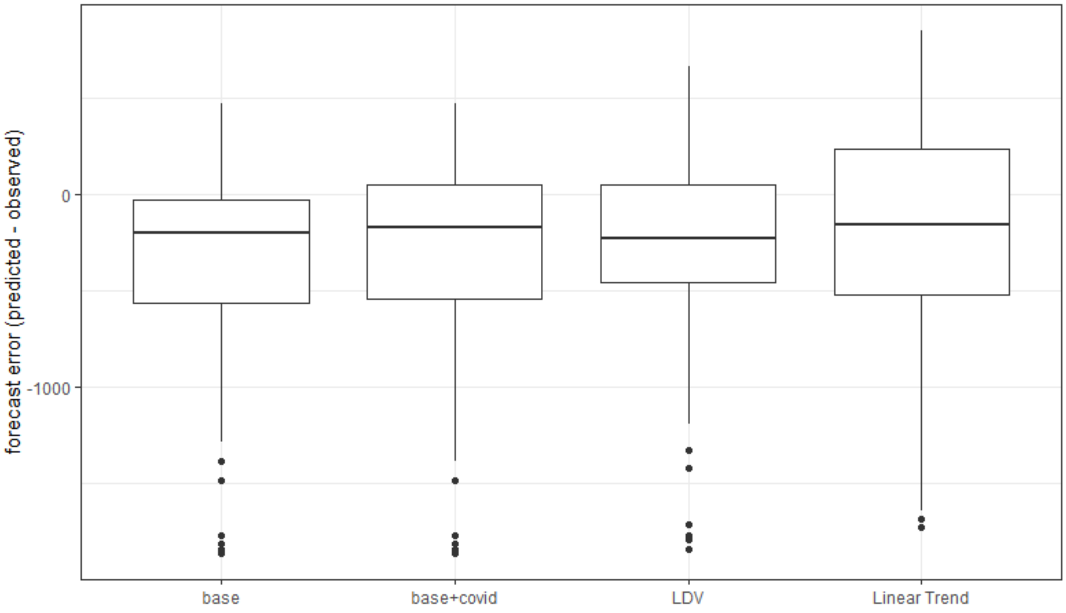


Figure 1. Distribution of time-series cross validation forecast errors for LEN sector.

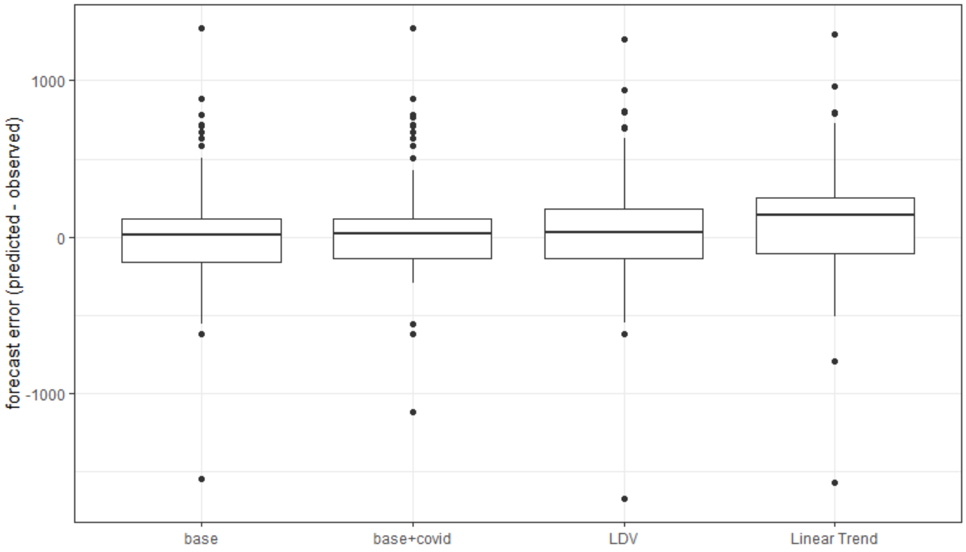


Figure 2. Distribution of time-series cross validation forecast error for OAN sector.

Recommendation: The analysts should explain the "upweighting" method used and evaluate alternative weighting schemes.

The GMT used an upweighting data approach to give "like" or similar year(s) more influence in the model. For example, to do this when predicting total 2024 landings or landings for the remainder of 2023, data from 2022 and 2023 are assigned a value of five while all other data years included in the model are assigned a value of one. This approach is used in the current DTL model, as well as other GMT trip limit models (with slightly different weighting schemes), in an effort to account for recent changes in a volatile fishery. For example, if running the model to predict 2024 landings, upweighting recent years (2022 and 2023) could minimize the influence of years when COVID impacted the fishery, and account for the unusually high recruitment of sablefish in recent years.

The GMT did not prioritize evaluating alternative weighting schemes at this time because the status quo data weighting method generally did not perform better than models without, as will be shown in the following sections, and we did not see the need to use an alternative data weighting method at this time given the inclusion of COVID fixed effects in some of the better performing models. Additionally, if a lagged dependent variable or linear time trend is included in the final model, there would no longer be a need to upweight recent data.

Section 3. Limited Entry North of 36° N. lat.

Section 3.1. Evaluating Data Prior to 2011

Recommendation: Another longer-term exploration could involve modeling pre-2011 data using fixed effects or other methods to account for that structural change.

In the LEN sector, there was a daily trip limit in place as well as a bimonthly limit prior to 2010. Prior to 2010, median vessel-level attainment of the bimonthly limit was 58.2 percent and mean attainment was 62.8 percent. The daily limit was removed in 2010, after which the vessel-level attainment of the bimonthly trip limit shows a statistically significant increase in median and mean attainment (66.4 percent and 65.9 percent, respectively). This suggests that, before 2010, when the daily trip limit was in place, vessels generally experienced lower bimonthly attainment due to daily trip limit constraints. Using the LOOCV method, the GMT explored whether a model that includes pre-2010 data could account for differences in bimonthly attainment before and after 2010 by using a dummy factor variable (".dl" is the qualifier) for all data prior to 2010 to capture the fixed effects (Figure 3). All models in Figure 3 are predicting average pounds per vessel using both weekly and bimonthly limits as predictor variables.

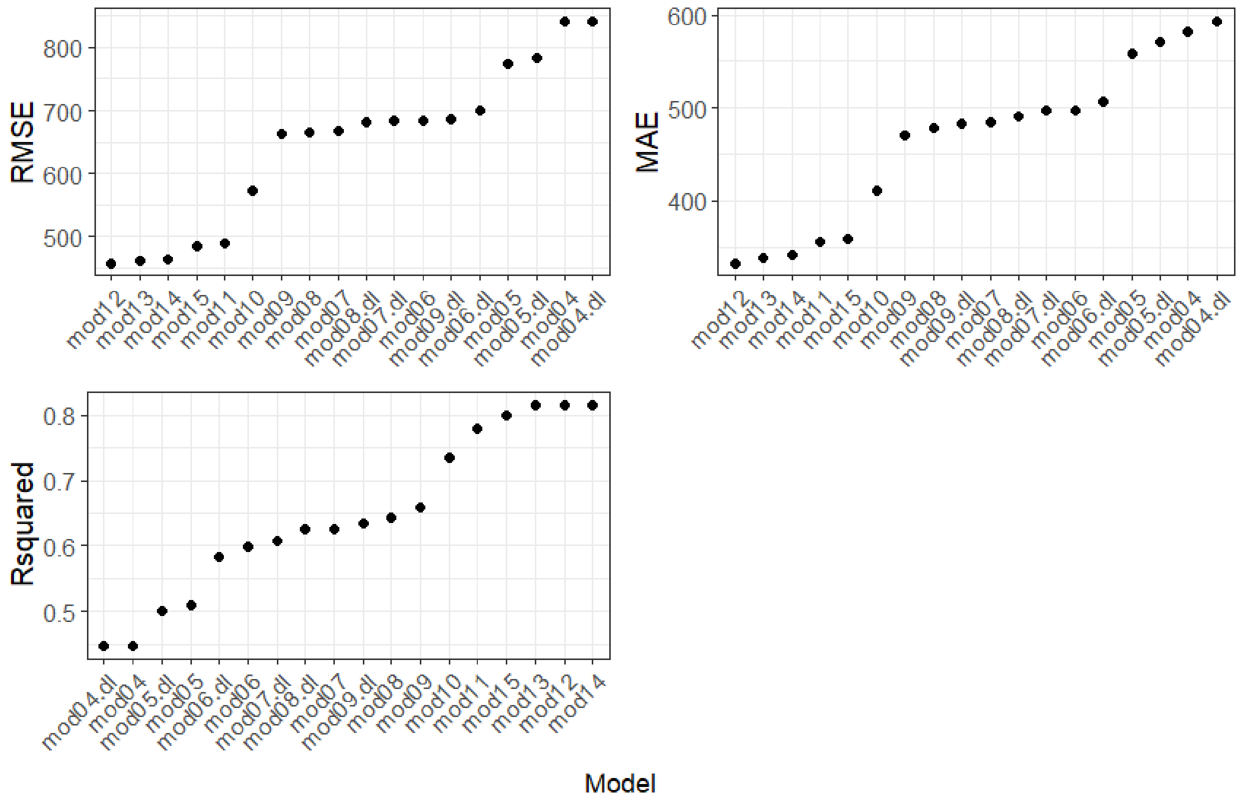


Figure 3. Metric results from the LOOCV used to compare LEN models with alternative starting years (2004-2015). In all models, the final year used is 2023. The order of the x-axis varies by plot. ".dl" indicates that a dummy factor variable for all data prior to 2010 is used.

The 2012-2023 model had the lowest RMSE and MAE and is therefore the model used for the remainder of the model evaluations (N = 69). The best model excludes 2010 and 2011 due to the exceptionally high bimonthly trip limits of 7,000 to 8,500 lbs. in 2010 and continuing through period 3 of 2011, before dropping back down below 5,500 lbs. Bimonthly trip limits then remained at or below 5,500 lbs. until period 6 of 2020. Additionally, accounting for the fixed effects prior to 2010 with a dummy factor variable did not improve the model performance compared to simply excluding all data prior to 2010.

As a result of the 2023 limited update assessment for sablefish, the sablefish north ACLs in 2025 and 2026 are expected to increase threefold, and trip limits will likely be increased to account for the higher landings targets. However, it is difficult to say at this time how much they are likely to be increased, and it is possible the Council will not want to increase trip limits proportional to the ACL increases. In the future, if the model has difficulty predicting landings under the higher 2025 and beyond trip limits, or the error appears to be a concern, the GMT could consider including 2011 and 2010 in order to capture very high trip limits in the training data.

Section 3.2. LEN Model Predicting Average Pounds per Vessel

Section 3.2.1. Time Series Cross Validation Results

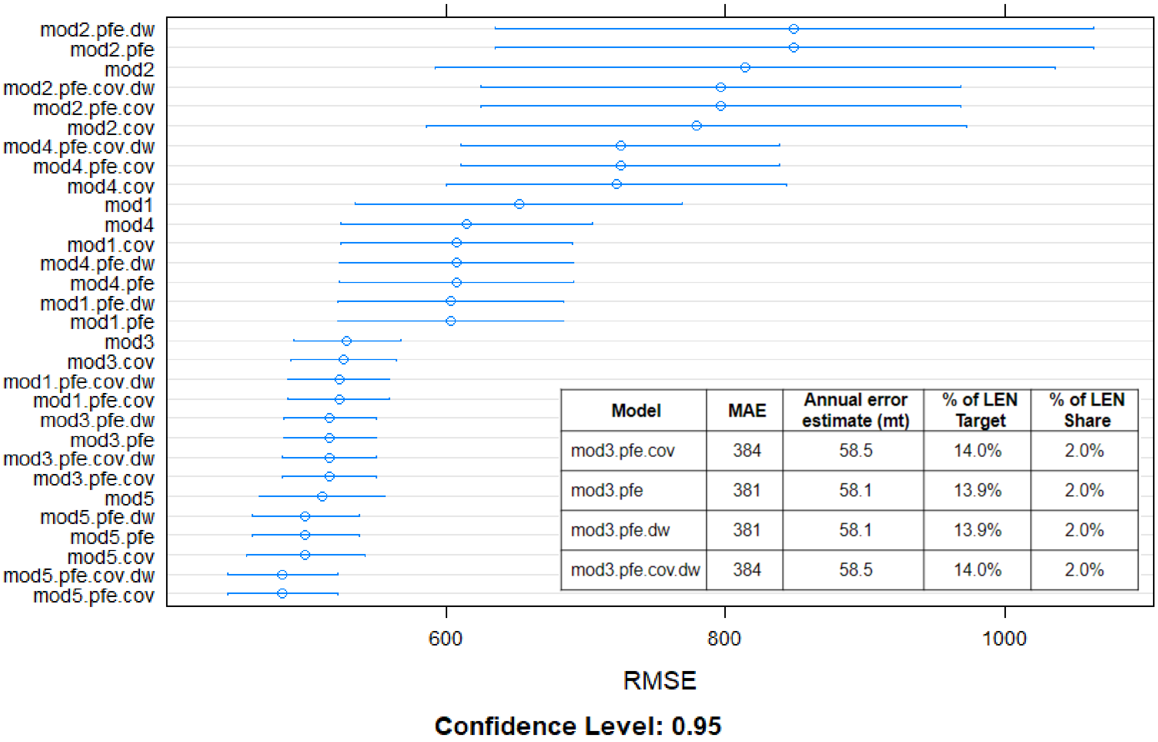
As noted above, the GMT used the RMSE from TSCV to compare models. For the LEN model that predicts average landings per vessel, Table 2 below lists the variables or methods used in each of the individual models shown in Figure 4. The qualifiers are used to identify the various components of each model in Figure 4. The model performs best when the weekly limit is not included, indicating that only the bimonthly limit appears necessary. Furthermore, given that the weekly limit is exactly half of the bimonthly limit, the effects of either limit are correlated. Therefore, to simplify model comparisons, the GMT only explored adding the Alaska sablefish Total Allowable Catch (TAC) and U.S.-Japan (i.e., yen) exchange rate to the model that uses the bimonthly limit but not the weekly limit. The table within Figure 4 shows the mean absolute error (MAE) from the TSCV runs of the top models that do not include the yen exchange rate (due to an inability to forecast that variable), as well as an estimate of mean annual error in metric tons, which is calculated as:

$$Mean\ annual\ error\ (mt)\ =\ (MAE\ \times\ 56\ \times\ 6)/2204.6$$

where 56 is the average annual number of vessels in the time series data (2012-2023), 6 is the number of periods within a year, and 2,204.6 is the conversion from pounds to metric tons. This mean annual error estimate is also shown in Figure 4 as a percentage of the 2023 LEN target and the 2023 LEN share. Note that there is very little difference in error among the top four models identified by the GMT, indicating little difference in risk with or without fixed effects and/or data weights. Based on out-of-sample prediction error, around 14 percent of the 2023 LEN landings target and around 2 percent of the 2023 LE share was over or underpredicted, on average, which equates to roughly 58 metric tons. See Section 3.5 below for a more detailed explanation of how the sablefish north commercial harvest guideline, which is divided into the LE and OA shares, was used as a proxy for risk to the sablefish north ACL.

Table 2. Variables and methods used to compare LEN models predicting landings per vessel.

Qualifier	Description
mod1	weekly limit + bimonthly limit (status quo model)
mod2	weekly limit
mod3	bimonthly limit
mod4	bimonthly limit + Alaska sablefish Total Allowable Catch (TAC)
mod5	bimonthly limit + yen exchange rate
.pfe	period fixed effects
.cov	2020 and 2021 as fixed effects (COVID)
.dw	data weights



Model	MAE	Annual error estimate (mt)	% of LEN Target	% of LEN Share
mod3.pfe.cov	384	58.5	14.0%	2.0%
mod3.pfe	381	58.1	13.9%	2.0%
mod3.pfe.dw	381	58.1	13.9%	2.0%
mod3.pfe.cov.dw	384	58.5	14.0%	2.0%

Figure 4. RMSE and MAE results for LEN models predicting average pounds per vessel, using TSCV. Error bars represent the 95% confidence interval.

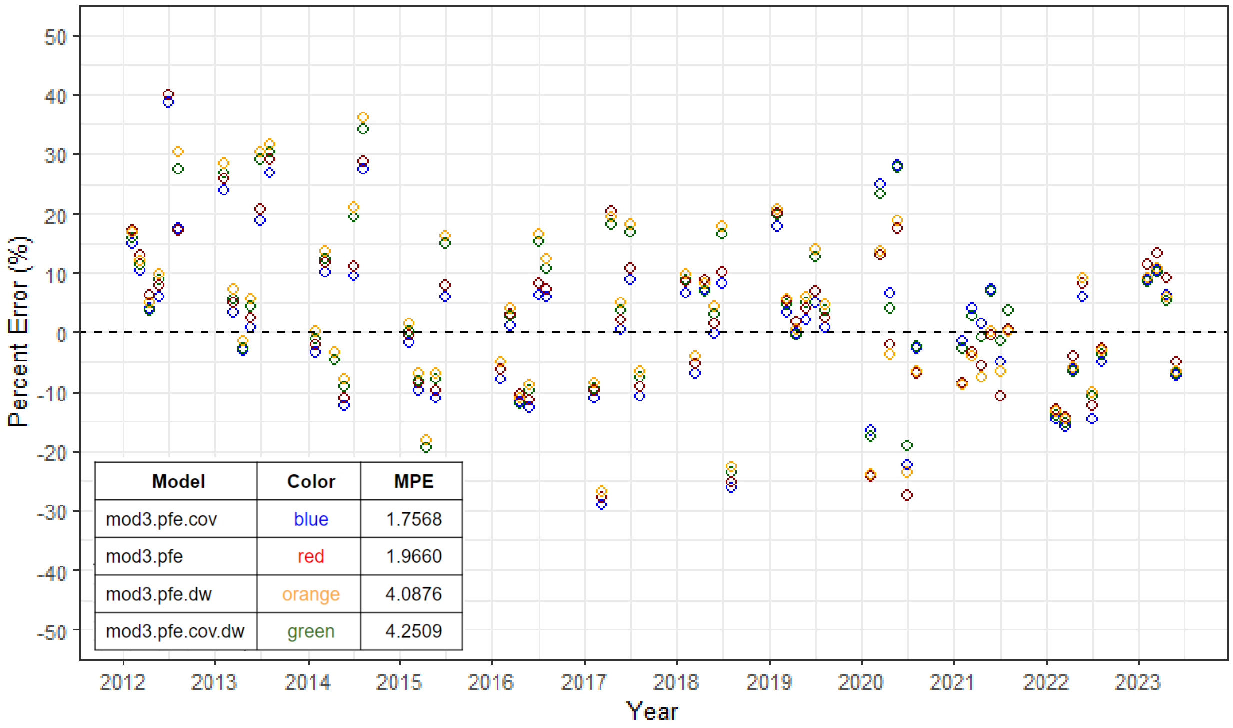
Based on the results in Figure 4, the model with the lowest RMSE is the model that uses the bimonthly limit and the yen exchange rate, along with period-specific fixed effects and COVID fixed effects. There is little difference when the status quo data weighting method is used. In general, including the yen exchange rate seemed to be an improvement compared to only using the bimonthly limit.

Section 3.2.2. Tendency to Over- or Under-predict

Recommendation: It is possible to quantify prediction bias, the tendency to over-predict versus under-predict. Measures such as Mean Forecast Error and Mean Percent Error will assess a model's tendency to over-or under-predict.

The GMT further evaluated the tendency to over- or under-predict based on the top four variations of mod3, which is the model that uses the bimonthly limit as a predictor of average pounds per vessel. We did not include the yen exchange rate variable (mod5) in the risk evaluations, because we have not, at this time, identified a way to predict future yen exchange rates. It may be possible to rely on forecasts made by other economic agencies or organizations, but we did not have sufficient time to narrow down any one source, given that different economic forecasts are made for different reasons and based on different assumptions. Given that the prediction error of mod5 is generally lower than that of mod3, the GMT assumes that the risk of exceeding the sablefish north allocations and ACL would be lower under mod5 than mod3.

Figure 5 below shows the in-sample percent prediction error across the time series for each of the top four models (excluding yen exchange rate) from the TSCV results. The model using bimonthly limit as an explanatory variable, along with period fixed effects and COVID fixed effects, had the lowest Mean Percent Error (MPE) of 1.76 percent. That model underpredicted a maximum of roughly 29 percent in 2017 and overpredicted a maximum of roughly 39 percent in 2012. For all models, the percent error has generally been lower in the last three years compared to prior years. Prior to 2015, the model was largely over-predicting average pounds per vessel but has since been more balanced between over- and under-predicting. While Figure 5 (and similar figures for the other models in this report) shows the percent prediction error of in-sample predictions, the GMT may be able to generate similar figures based on the out-of-sample forecast error from the TSCV tests in time for the workshop. We were not able to compile those results at the time of writing this report.



Model	Color	MPE
mod3.pfe.cov	blue	1.7568
mod3.pfe	red	1.9660
mod3.pfe.dw	orange	4.0876
mod3.pfe.cov.dw	green	4.2509

Figure 5. Percent prediction error of top four LEN average pounds models (excluding yen exchange rate) based on TSCV results. Mean Percent Error (MPE) for each of those top four models is shown in the table within the figure.

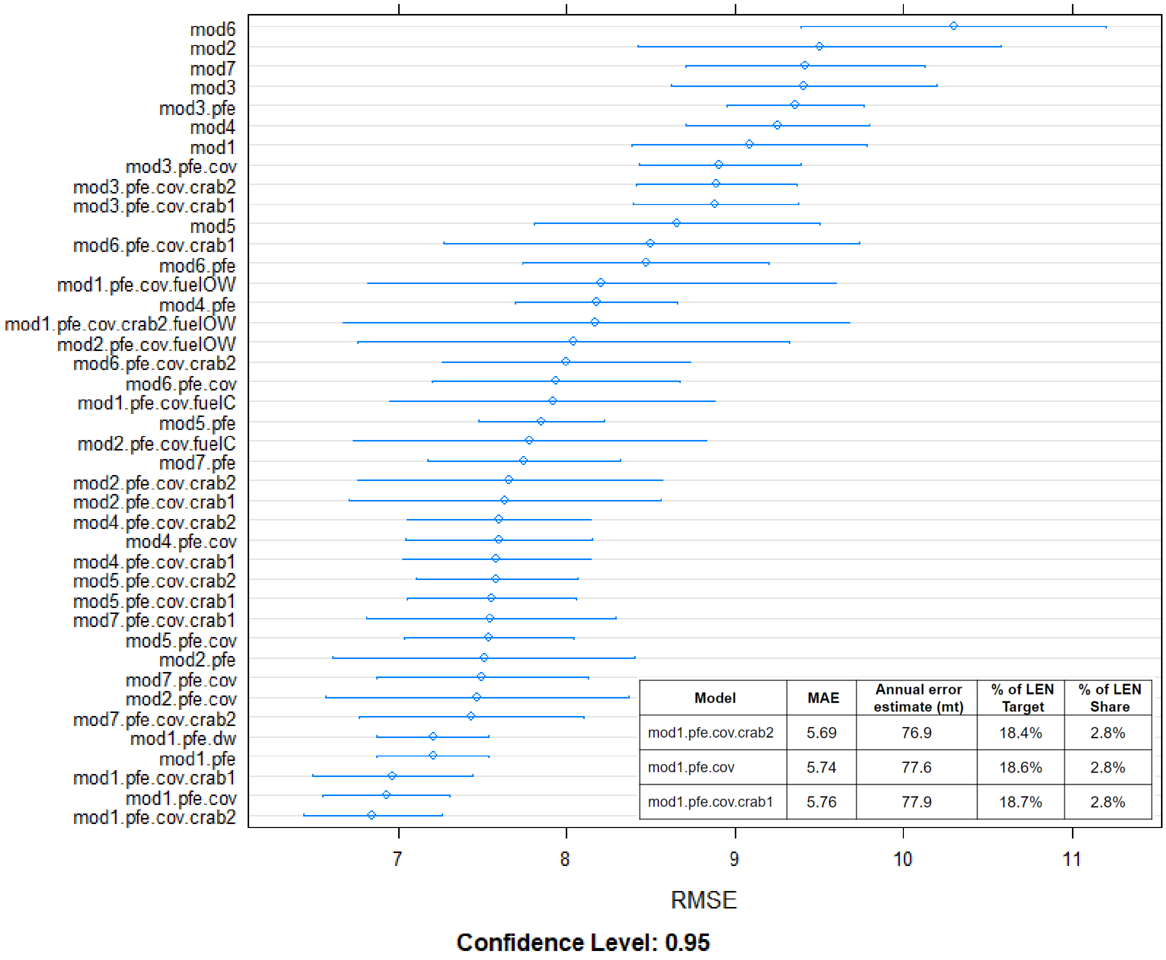
Section 3.3 LEN Model Predicting Number of Vessels

Section 3.3.1. Time Series Cross Validation Results

For the LEN model that predicts the number of vessels in the fleet, Table 3 below lists the variables or methods used in each of the individual models shown in Figure 6. Similar to the previously discussed model, the GMT only explored adding the Alaska sablefish TAC and yen exchange rate to the best performing model (mod1) to simplify comparisons.

Table 3. Variables and methods used to compare LEN models predicting landings per vessel.

Qualifier	Description
mod1	avg. sablefish price
mod2	avg. sablefish price + max. sablefish price
mod3	max. sablefish price
mod4	median sablefish price
mod5	median sablefish price + max. sablefish price
mod6	avg. sablefish price + Alaska sablefish TAC
mod7	avg. sablefish price + Yen exchange rate
.pfe	period fixed effects
.cov	2020 and 2021 as fixed effects (COVID)
.crab1	+ avg. crab price
.crab2	+ max. crab price
.fuelOW	OR and WA fuel prices (combined)
.fuelC	CA fuel prices north of 36° N. lat.
.dw	data weights



Model	MAE	Annual error estimate (mt)	% of LEN Target	% of LEN Share
mod1.pfe.cov.crab2	5.69	76.9	18.4%	2.8%
mod1.pfe.cov	5.74	77.6	18.6%	2.8%
mod1.pfe.cov.crab1	5.76	77.9	18.7%	2.8%

Figure 6. RMSE results for LEN models predicting average pounds per vessel, using TSCV. Error bars represent the 95% confidence interval.

Based on the results in Figure 6, the model with the lowest RMSE is the model that uses the average sablefish prices and maximum Dungeness crab prices, along with period-specific fixed effects and COVID fixed effects. In general, the addition of fuel prices or Alaska sablefish TAC did not provide substantial model improvements, and in the case of fuel prices, tended to result in a wider range of RMSE results across runs. Including the yen exchange rate in the model does have some influence on the model estimates, but including this factor does not outperform the model without this factor.

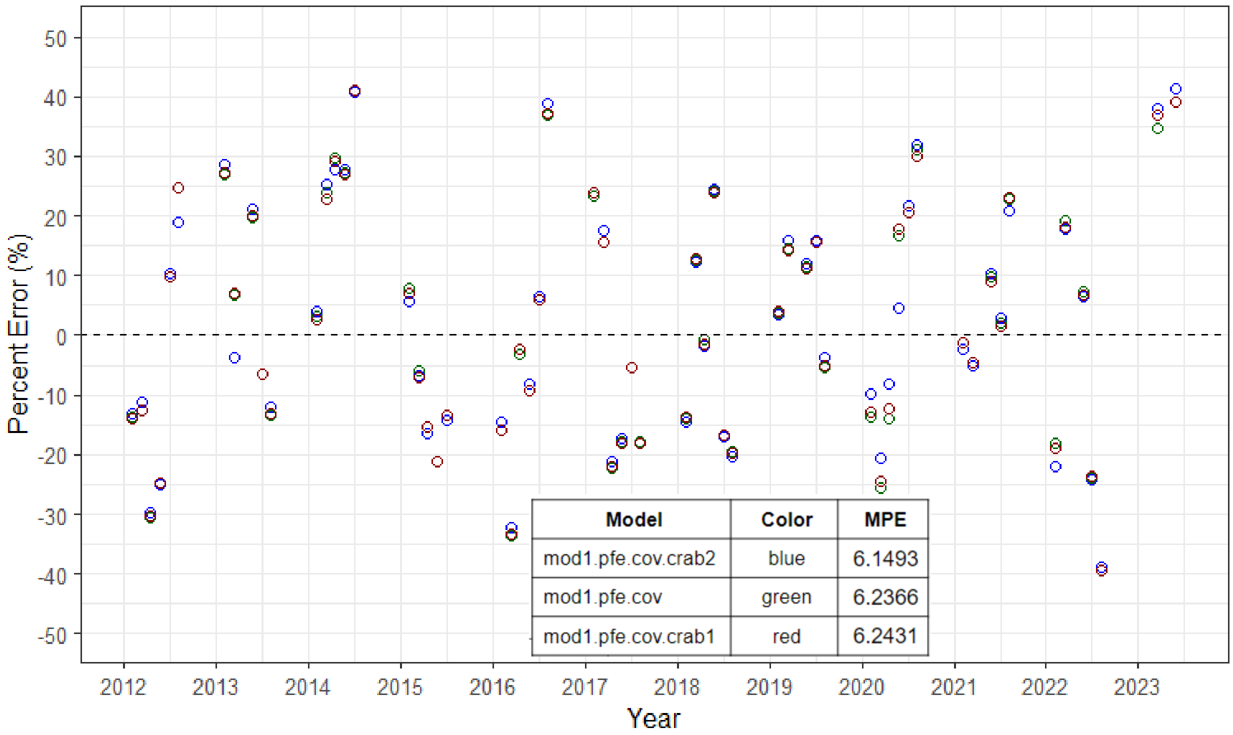
The MAE from the TSCV runs was roughly 5.7 vessels for the top three models (Figure 6), and we used the following calculation to estimate mean absolute error for a full year:

$$Mean\ annual\ error\ (mt)\ =\ (MAE\ \times\ 4967\ \times\ 6)/2204.6$$

where 4,967 is the average pounds landed per vessel per period to date in 2023, 6 is the number of periods within a year, and 2,204.6 is the conversion from pounds to metric tons. The mean absolute error for a full year, based on out-of-sample predictions, is estimated at roughly 76.9-77.9 mt, which equates to 18.4-18.7 percent of the 2023 LEN landings target and 2.8 percent of the 2023 LE share.

Section 3.3.2. Tendency to Over- or Under-predict

Figure 7 below shows the in-sample percent prediction error across the time series for each of the top three models from the TSCV results. The best model from the TSCV also has the lowest Mean Percent Error (MPE) of 6.15 percent, and that model underpredicted a maximum of roughly 39 percent in 2022 and overpredicted a maximum of roughly 41 percent in 2014 and 2023. Note that the percent error for the model predicting the number of vessels is generally larger than the model predicting the average pounds per vessel, even for the best performing model.



Model	Color	MPE
mod1.pfe.cov.crab2	blue	6.1493
mod1.pfe.cov	green	6.2366
mod1.pfe.cov.crab1	red	6.2431

Figure 7. Percent prediction error of top three LEN models predicting the number of vessels, based on TSCV results. Mean Percent Error (MPE) for each of those top three models is shown in the table within the figure.

The models with the lowest RMSE and lowest MPE from the analysis above are as follows:

$$Average\ lbs.\ per\ vessel \sim bimonthly\ trip\ limit\ +\ factor(PERIOD)\ +\ factor(COVID)$$

$$Number\ of\ vessels \sim avg.\ sablefish\ price\ per\ lb.\ +\ maximum\ D.\ crab\ price\ per\ lb.\ +\ factor(PERIOD)\ +\ factor(COVID)$$

While maximum Dungeness crab prices appeared to improve model performance, setting predicted maximum prices is more uncertain to do, since there is large variability in maximum prices. Therefore, setting a maximum price point estimate potentially increases the risk of getting our predictions wrong, which is largely mitigated when setting predicted average prices. The maximum crab prices in the data set (2012-2023) range from \$4.81 per pound up to \$29.38 per pound. Until a better method is developed to set predicted maximum price, given the uncertainty, the GMT focused on only including average sablefish prices and did not include maximum Dungeness crab prices as an input variable. Therefore, the GMT considers the following model predicting vessel participation to be the best performing model that is also usable in management at this time:

$$Number\ of\ vessels \sim avg.\ sablefish\ price\ per\ lb.\ +\ factor(PERIOD)\ +\ factor(COVID)$$

The status quo model has been using average sablefish prices in management, and its use has already proven useful to the Council's decision making. Industry continues to indicate the importance of sablefish markets, and specifically sablefish price per pound, as a driving factor for participation, especially when balancing recent low sablefish prices on the West Coast with opportunities in other fisheries those vessels also participate in (e.g., Alaska sablefish and Pacific halibut). While the status quo method of setting predicted sablefish prices is not perfect, the GMT did not have time to thoroughly develop a better alternative. There are many different factors that could be driving sablefish price fluctuations, and additional thought is needed to develop a model-based approach to forecasting, given that such a model would also need inputs that are known into the future. In the absence of that, the GMT considers the status quo method of setting predicted sablefish prices as the best available method at this time, and we provide some analysis to evaluate the status quo method in the following section.

Section 3.3.3. Price Forecasting Evaluation

Recommendation: If variables that require forecasting are included in the models, then the forecast method must also be included in the model evaluation.

Based on the status quo method of setting predicted prices, the GMT first compares either the current year's prices to date when running the model inseason or the most recent year's prices with prior years (Figure 8). 2023 prices appear to be most similar to 2020-2022 prices. Assuming we are running the model at the end of period 4 (e.g., late August, early September) for the Council to set trip limits at the September Council meeting, we would then set predicted prices for period 5 and 6 based on average period 5 and 6 prices from 2020-2022. Given the uncertainty in price predictions, we create low, average, and high price scenarios by multiplying our average predicted prices in each period by 0.9 and 1.1 for the low and high scenarios, respectively. If we are setting predicted prices in the biennial management measure process for a full two-year period, we would simply use the average prices by period from the most recent years that appear most similar. There is greater uncertainty in that case, but the Council can and does adjust trip limits inseason as sablefish prices change. Prediction error based on setting uncertain future sablefish prices does not impact allocations in other fisheries.

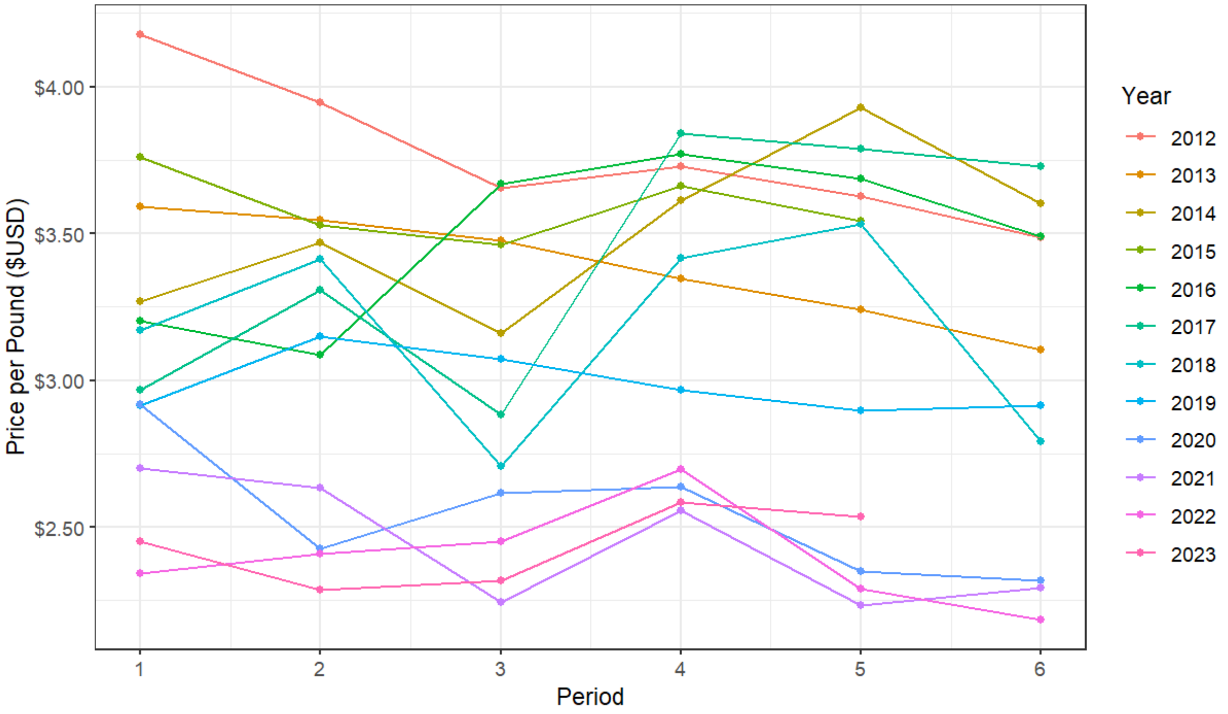


Figure 8. Average inflation adjusted sablefish price per pound by period in the LEN sector, 2012-2023.

Figure 9 provides a retrospective example of sablefish price prediction error if the price data from years 2013-2015 were used to set predicted low, average, and high prices for 2016 and 2017, based on the trends shown in Figure 8. We selected these years as an example, because price trends in 2013-2015 are relatively similar to each other as well as to 2016, but prices fluctuated very widely in 2017. Thus, this case scenario is meant to demonstrate an extreme example of the risk when recent years' prices do not fully reflect the following two years. The predicted prices were compared to actual prices in 2016 and 2017 to calculate the error. The MPE across all predictions shown in Figure 9 is 1.8 percent. The method tended to overpredict 2016 and 2017 prices in periods 1 and 2 and to underpredict in periods 4-6, with a wide range of prediction error in period 3, possibly because period 3 tends to see the lowest LEN participation.

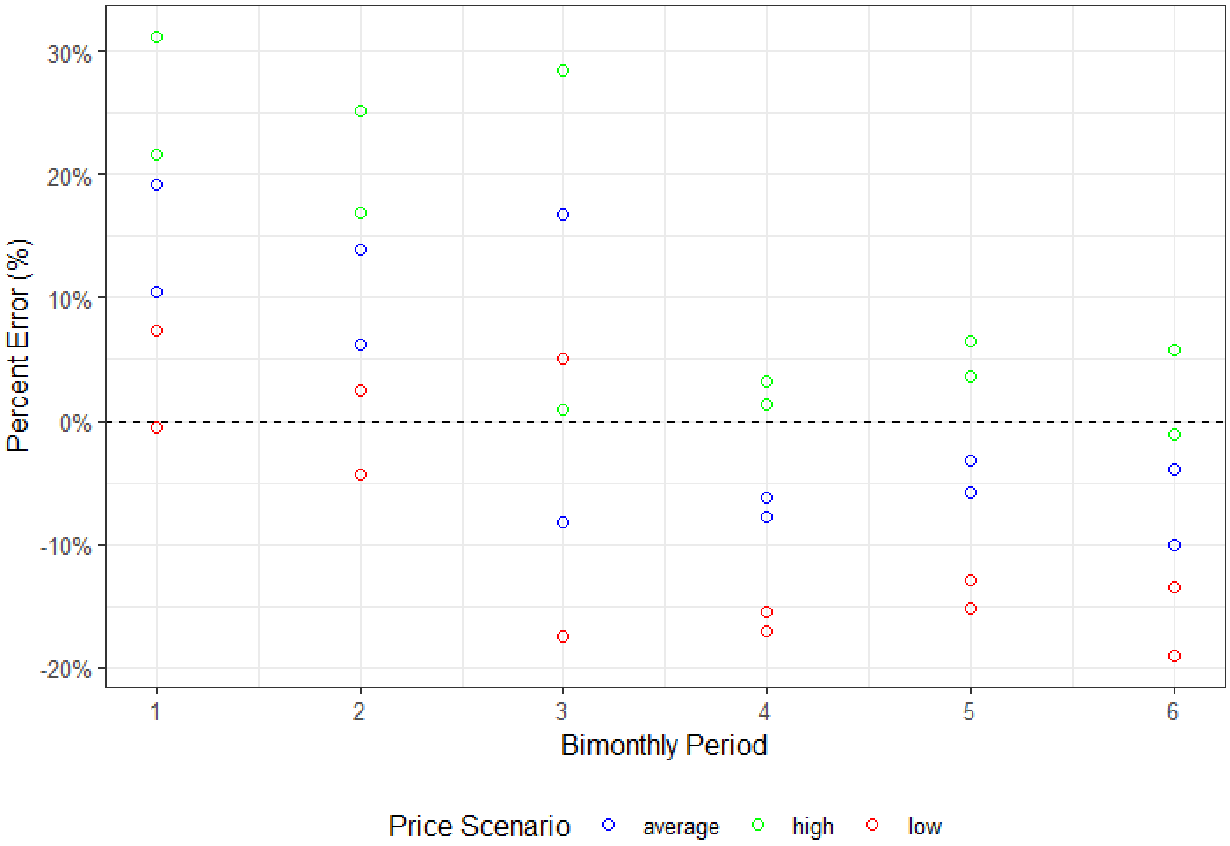


Figure 9. Example of percent error of sablefish price predictions using 2013-2015 prices to predict 2016-2017 prices. There are five data points for each period, representing the percent error under the low, average, and high price scenarios for both 2016 and 2017.

There is greater risk of underpredicting participation (and therefore total landings) when prices are underpredicted, because higher prices tend to entice greater participation. If prices are underpredicted most in periods 4-6 when setting trip limits in the biennial management measures process, the GMT can use actual prices within the biennium to make better inseason predictions and adjust trip limits as needed. Continuing with the example of using 2013-2015 prices to predict 2016-2017 prices, total landings by period in 2016 and 2017 were underpredicted a minimum of 1.2 percent and a maximum of 39.9 percent, compared to actual total landings during those years (Figure 10). The MPE of the total landings predictions in Figure 10 is -3.81 percent.

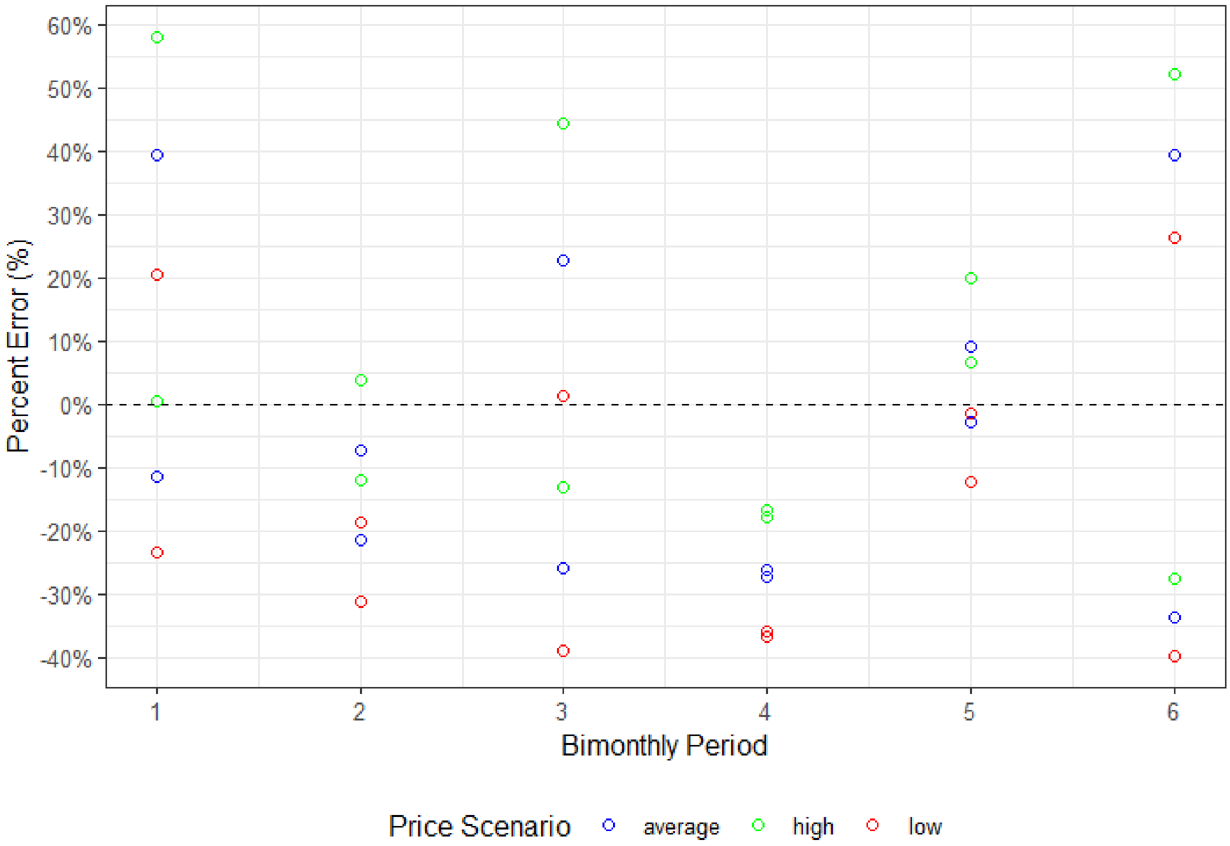


Figure 10. Example of percent error of total LEN sablefish landings using 2013-2015 prices to predict 2016-2017 prices.

Section 3.4. Best Performing Models for LEN

In summary, the GMT has identified the following models as the best performing models, at this time, to predict LEN landings for management purposes:

$$Average\ lbs.\ per\ vessel \sim bimonthly\ trip\ limit\ +\ factor(PERIOD)\ +\ factor(COVID)$$

$$Number\ of\ vessels \sim avg.\ sablefish\ price\ per\ lb.\ +\ factor(PERIOD)\ +\ factor(COVID)$$

Section 3.5. Risk to the Sablefish North Annual Catch Limit

Recommendation: The analysts should consider the consequences of poor prediction performance and evaluate model performance in terms of the risk of exceeding the ACL. Measures such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) measure the magnitude of expected prediction error.

The allocation scheme for sablefish north of 36° N. lat. is shown in Figure 11 below, which is in Section 6.3.2.1 of the Pacific Coast Groundfish Fishery Management Plan. After tribal, research, and recreational deductions, the commercial harvest guideline is divided into the LE share (90.6 percent) and the OA share (9.4 percent). The LE share is divided into the LE trawl share (58 percent) and the LE fixed gear share (42 percent). The LE trawl share is made up of a 100 mt at-sea set-aside and the shorebased Individual Fishing Quota (IFQ) program allocation (i.e., the remainder of the LE trawl share). The LE fixed gear share is reduced by 19 percent to account for discard mortality to determine the landed LE share. Within the landed LE share, 85 percent is allocated to the primary sablefish fishery, which is a catch share program, and the remaining 15 percent is allocated to the DTL fishery (i.e., LEN target). The OA share is reduced by 19 percent to account for discard mortality to determine the landed OA share, and the remaining 81 percent is allocated to the DTL fishery.

Therefore, 85 percent of the overall commercial harvest guideline is made up of allocations that are attributed to some catch share program. Catch share programs are designed so that the risk of exceeding the allocation is low, given that each permit holder to the program is allocated a permit-level share or quota of the overall allocation. Therefore, any risk of exceeding the sablefish north commercial harvest guideline is likely to come from the DTL fishery, or to a smaller extent, bycatch in the at-sea Pacific whiting fishery. Mortality and attainments associated with off-the-top deductions (i.e., tribal, research, and recreational) are difficult to predict given that they are set-asides rather than allocations. Therefore, the GMT considered the risk to the sablefish north commercial harvest guideline as a proxy for risk to the ACL. The Council could consider action to keep total mortality within the ACL if the commercial harvest guideline is at risk of being exceeded.

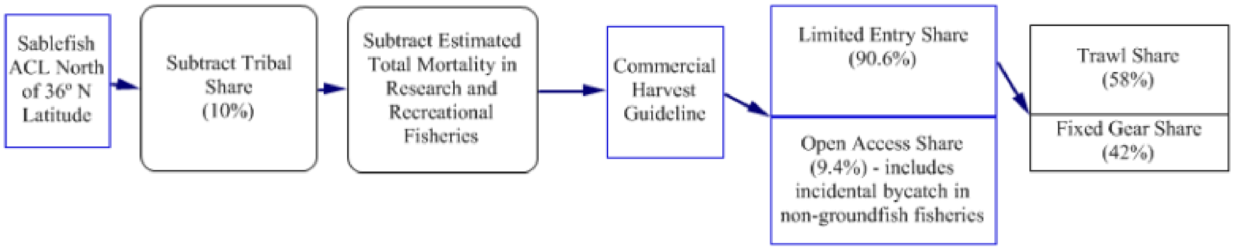


Figure 11. Allocation scheme for sablefish north of 36° N. lat., described in section 6.3.2.1 of the Pacific Coast Groundfish Fishery Management Plan.

Table 4 summarizes the total annual in-sample prediction error for 2017-2022 using the combination of the best pounds model (i.e., mod3.pfe.cov), which includes bimonthly limit, period fixed effects, and COVID fixed effects and the vessel model (i.e., mod1.pfe.cov), which includes average inflation adjusted sablefish price, period fixed effects, and COVID fixed effects. The risk to the sablefish north commercial harvest guideline is greater when the model under predicts (e.g., 2017, 2018 and 2022) because in the event that the model predicts 100 percent attainment, the prediction error could result in higher than 100 percent realized attainment. However, the magnitude of the prediction error as a percentage of the LE Share is small (less than 2 percent), and therefore, the overage might not risk exceeding the ACL if all other sectors don’t fully attain their allocation.

Given that the model has historically been both under- or over-predicting, there is reason for the Council to be precautionary when interpreting model projections, especially at levels above 90 percent. The hypothetical LEN Target attainment column in Table 4 is intended to demonstrate that, for example, if the model had predicted 98 percent attainment in 2017 (i.e., the actual attainment), the prediction error for that precise year (-32.3 mt) could have instead led to a hypothetical realized attainment of 111 percent. In reality, the model projections provided to the Council in 2017 when making their decision were lower than 98 percent, which is why attainment remained within the target. Furthermore, the Council has the opportunity to adjust trip limits at every Council meeting throughout the season, and attainment projections may change as the season progresses based on actual participation, price fluctuations, or actual landings. The Council can respond as needed to keep landings within the Target.

Table 4. Sum annual prediction error of the LEN model (combined average pounds per vessel and number of vessels) and the retrospective impacts of that annual prediction error on the LEN target and the LE share, 2017-2022. Under-predictions are highlighted in gray.

Year	Sum annual prediction error (mt)	Under- or over-prediction	Actual LEN Target attainment	Hypothetical[a/] LEN Target attainment	LE Share (mt)	Total prediction error as % of LE Share
2017	-32.3	under	98%	111%	4,252	0.76%
2018	-30.9	under	84%	96%	4,434	0.70%
2019	25.6	over	65%	56%	4,537	0.56%
2020	5.1	over	56%	54%	4,636	0.11%
2021	3.3	over	51%	50%	5,586	0.06%
2022	-78.7	under	92%	116%	5,320	1.48%

a/ Hypothetical attainment if the model had predicted what is in the actual target attainment column. The inverse of the sum annual prediction error is added to the actual annual landings. In other words, if the model underpredicted that year, the absolute value of the prediction error is added to the actual landings, but if the model overpredicted that year, it is subtracted from the actual landings for that year.

Section 4. Open Access North of 36° N. lat.

Section 4.1. Evaluating Data Prior to 2011

Recommendation: Another longer-term exploration could involve modeling pre-2011 data using fixed effects or other methods to account for that structural change.

In the OAN sector, prior to 2011 the weekly limit was about 25 to 33 percent of the bimonthly limit, whereas after 2011, the weekly limit increased to exactly half of the bimonthly limit. Prior to 2011, the median vessel-level attainment of the bimonthly trip limit was 45.5 percent, and mean attainment was 52.3 percent. With the 2011 increase in the weekly limit to exactly half of the bimonthly limit, vessel-level attainment of the bimonthly limit shows a statistically significant increase in median and mean attainment (55.2 percent and 58.3 percent, respectively). This suggests that, before 2011, when the weekly limit was lower than 50 percent of the bimonthly limit, vessels generally experienced lower bimonthly attainment due to weekly limit constraints.

Using the LOOCV method, the GMT explored whether a model that includes pre-2011 data could account for differences in bimonthly attainment before and after 2011 by using a dummy factor variable (".wl") for all data prior to 2011. The LOOCV method generates three different performance metrics: RMSE, MAE, and R-squared (Figure 12). The subcommittees recommended using metrics other than R-squared to compare models, so the GMT looked closely at RMSE and MAE when comparing models using cross validation. Even so, the model with the lowest RMSE and MAE did not have the highest R-squared value, so the GMT calculated an aggregate metric by adding RMSE and MAE, then subtracting R-squared. The 2007-2023 model with a pre-2011 dummy variable had the lowest RMSE, MAE, and aggregate score, and is therefore the model used for the remainder of the model evaluations (N = 95). Models using the years 2007-2023 without the pre-2011 dummy variable are also included in the TSCV for comparison. It's possible that the large amount of error in the models that use 2004 and/or 2005 data (as well as 2006 to some extent) is due to the fact that trip limits prior to 2007 were very high with minimal variability. Similar to the LEN model, if 2025 and beyond trip limit increases result in poor model performance, the GMT could consider adding additional years prior to 2007 to account for extremely high trip limits.

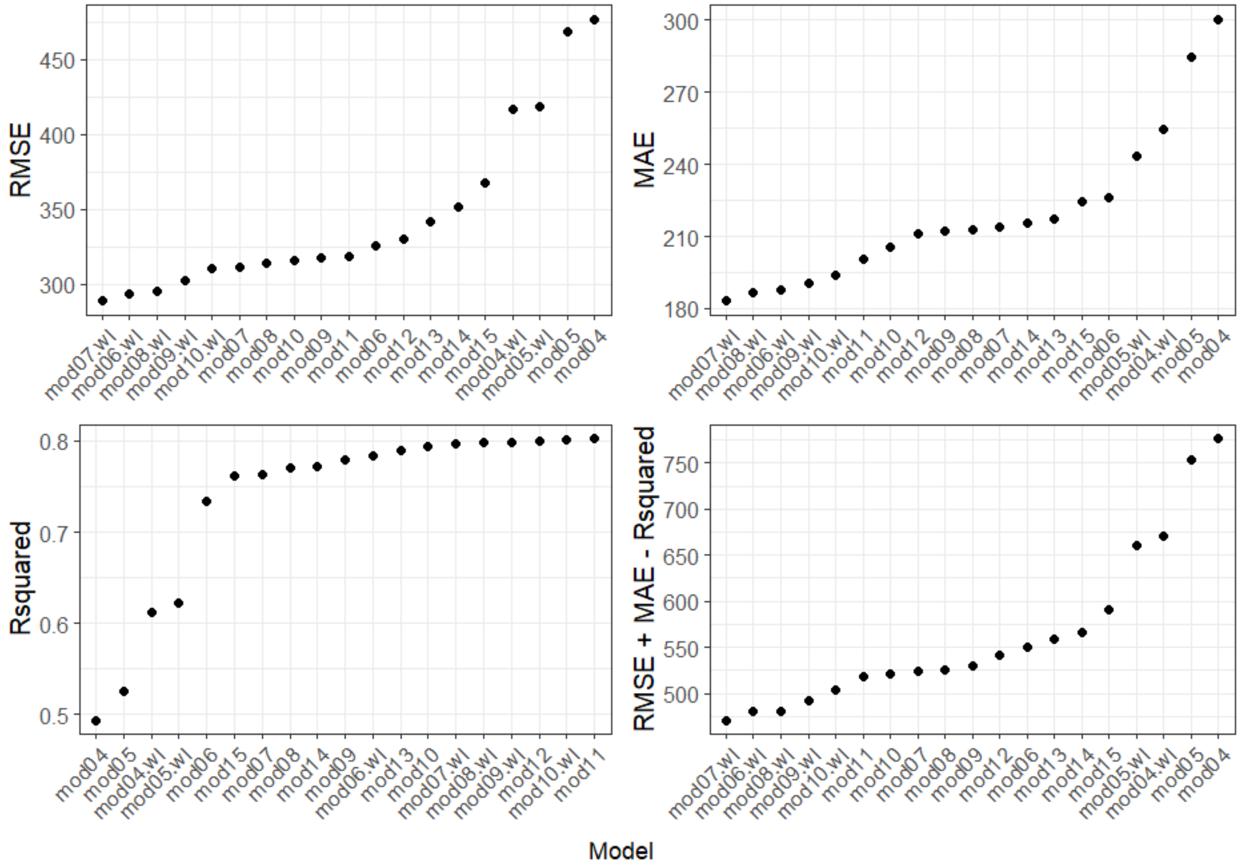


Figure 12. Metric results from the LOOCV used to compare OAN models with alternative starting years (2004-2015). In all models, the final year used is 2023. The order of the x-axis varies by plot.

Section 4.2. OAN Model Predicting Average Pounds per Vessel

Section 4.2.1. Time Series Cross Validation Results

Table 5 below lists the various model components and their respective qualifiers used to evaluate the OAN model that predicts average pounds landed per vessel. To simplify model comparisons, the GMT only added Alaska sablefish TAC and yen exchange rate variables to the models that use bimonthly limit as a predictor of average pounds per vessel (mod3), since using the bimonthly limit resulted in generally lower RMSE than models using weekly limit or both (mod1 and mod2, respectively). Based on the results in Figure 13, the model with the lowest RMSE in the TSCV is the model that includes bimonthly trip limits, yen exchange rates, period fixed effects, and COVID fixed effects. While including yen exchange rates appears to improve model performance, the GMT did not include that variable in the remaining analysis of error and risk due to an inability to forecast that variable at this time.

Table 5. Variables and methods used to compare OAN models predicting landings per vessel.

Qualifier	**Description**
mod1	weekly limit + bimonthly limit
mod2	weekly limit
mod3	bimonthly limit
mod4	bimonthly limit + Alaska sablefish TAC
mod5	bimonthly limit + yen exchange rate
.wl	Pre-2011 dummy variable
.pfe	period fixed effects
.cov	2020 and 2021 as fixed effects (COVID)
.dw	data weights

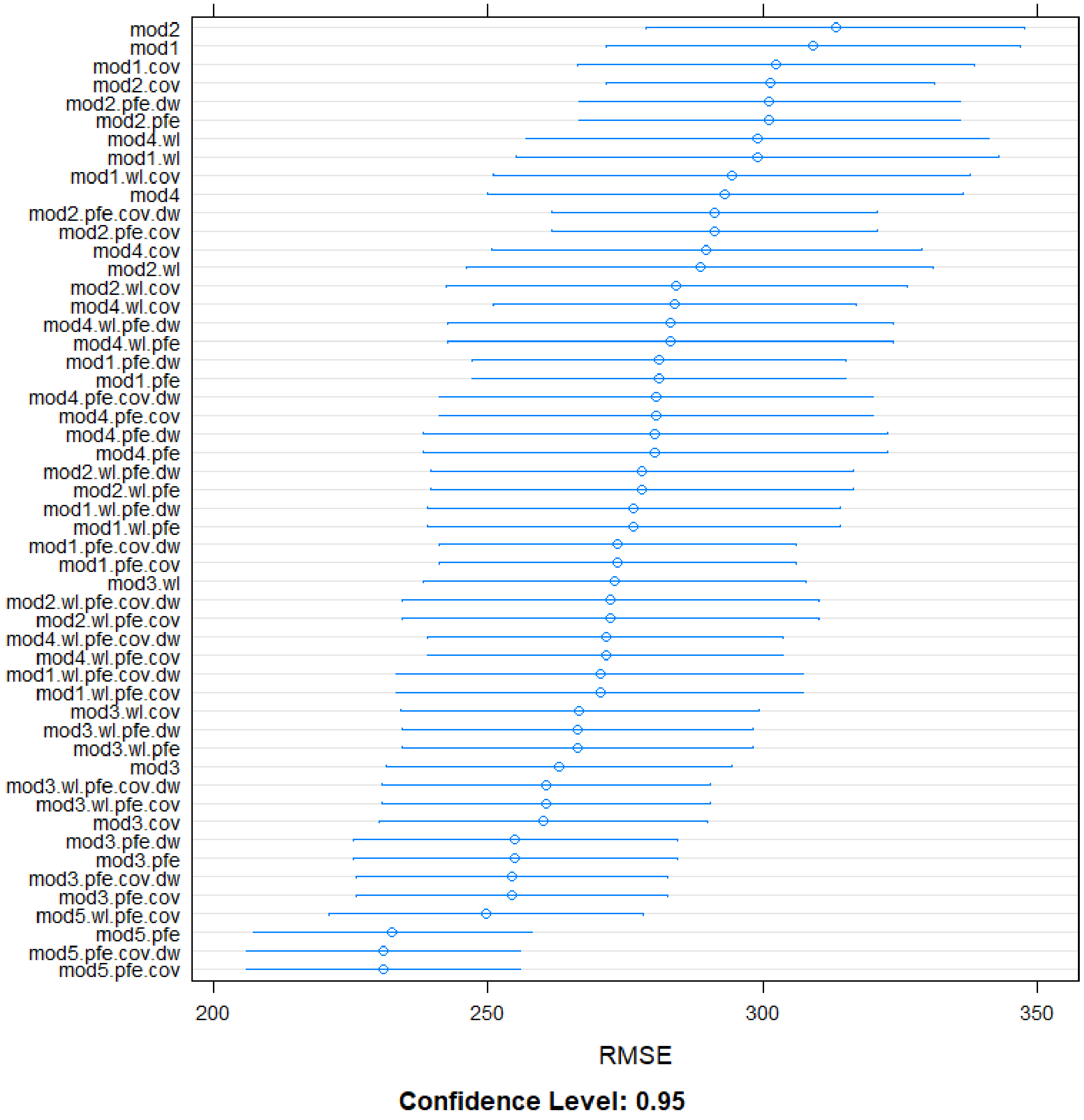


Figure 13. RMSE results for OAN models predicting average pounds per vessel, using TSCV. Error bars represent the 95% confidence interval.

The MAE from the TSCV runs was roughly 202.5 lbs. per vessel per period for the top four models (Table 6), and we used a similar equation to the one described in Section 3.2.1, but instead of using the average number of vessels participating each year in the entire time series, we multiplied the MAE by the average number of vessels within each period, given that OAN participation by period fluctuates much more dramatically than LEN participation. In doing this, we did not multiply by 6 periods, as in Section 3.2.1. The mean absolute error for a full year, based on out-of-sample predictions, is estimated at roughly 47.5-47.6 mt, which equates to 6.9 percent of the 2023 OAN landings target and 6.6-6.7 percent of the 2023 OA share.

Table 6. MAE results for OAN models predicting average pounds per vessel, using TSCV, annual absolute error estimates, and annual error estimates as a percent of the OAN target and OA share.

Model	MAE	Annual error estimate (mt)	% of OAN Target	% of OA Share
mod3.pfe	202.5	47.5	6.9%	6.6%
mod3.pfe.cov	202.8	47.6	6.9%	6.7%
mod3.pfe.cov.dw	202.8	47.6	6.9%	6.7%
mod3.pfe.dw	202.5	47.5	6.9%	6.6%

Section 4.2.1. Tendency to Over- or Under-predict

Figure 14 below shows the percent prediction error across the time series for each of the top four models (excluding yen exchange rate) from the TSCV results. The model using bimonthly limits and period fixed effects had the lowest MPE of 1.88 percent. That model underpredicted a maximum of roughly 29 percent in 2022 and overpredicted a maximum of roughly 47 percent in 2010.

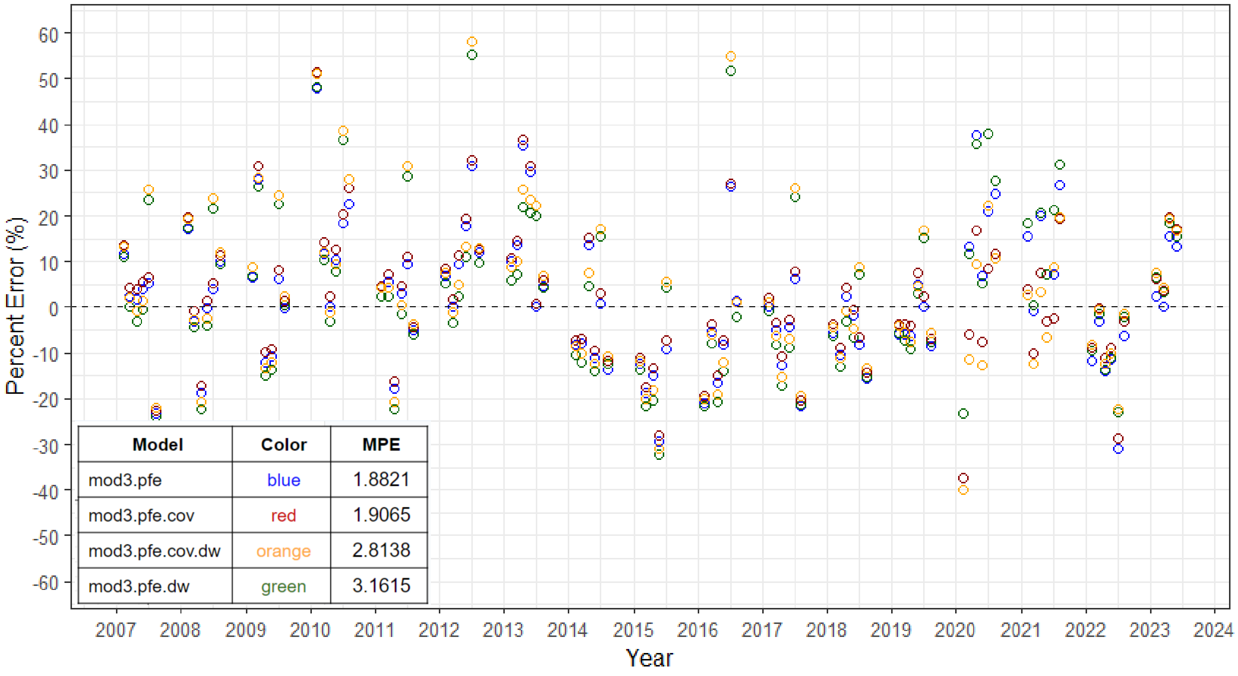


Figure 14. Percent prediction error of top selected OAN average pound models based on TSCV results. Mean Percent Error (MPE) for each of those top four models is shown in the table within the figure.

Section 4.3. OAN Model Predicting Number of Vessels

Section 4.3.1. Time Series Cross Validation Results

Table 7 below lists the various model components and their respective qualifiers used to evaluate the OAN model that predicts the number of vessels in the fleet. Note that the period 4 peak adjuster variable is included in this model evaluation and tends to perform better than the period fixed effects. However, for consistency across models/sectors and based on the SSC's advice, the GMT did not include the period 4 peak adjuster variable in the remaining analysis or in our concluding "best" model for management. Notably, the top performing models included northern California fuel prices as a predictor variable. This makes sense given that OA vessels tend to be smaller, land lower volumes than LE vessels, and generally have a wider variety of fishing portfolios, and therefore, fuel prices are likely to play a larger role in deciding whether to participate in the OA fishery and to target sablefish. However, again, the GMT did not identify a usable method of forecasting fuel prices at this time, so it is not considered as an input variable any further. When the fuel prices are not included, the models using average sablefish prices (mod1) tend to perform better than models using maximum price (mod3) or a combination of both (mod2). To simplify model comparisons, we added the Alaska sablefish TAC and yen exchange rate variables to only the model using average sablefish prices (mod1); however, those variables did not appear to improve model performance as much as for the LEN model.

Table 7. Variables and methods used to compare OAN models predicting landings per vessel.

Qualifier	**Description**
mod1	avg. price
mod2	avg. price + max. price
mod3	max. price
mod4	median price
mod5	median price + max. price
mod6	avg. price + Alaska sablefish TAC
mod7	avg. price + yen exchange rate
.p4p	Period 4 peak adjuster fixed effect
.pfe	period fixed effects
.cov	2020 and 2021 as fixed effects (COVID)
.crab1	+ avg. crab price
.crab2	+ max. crab price
.dw	data weights

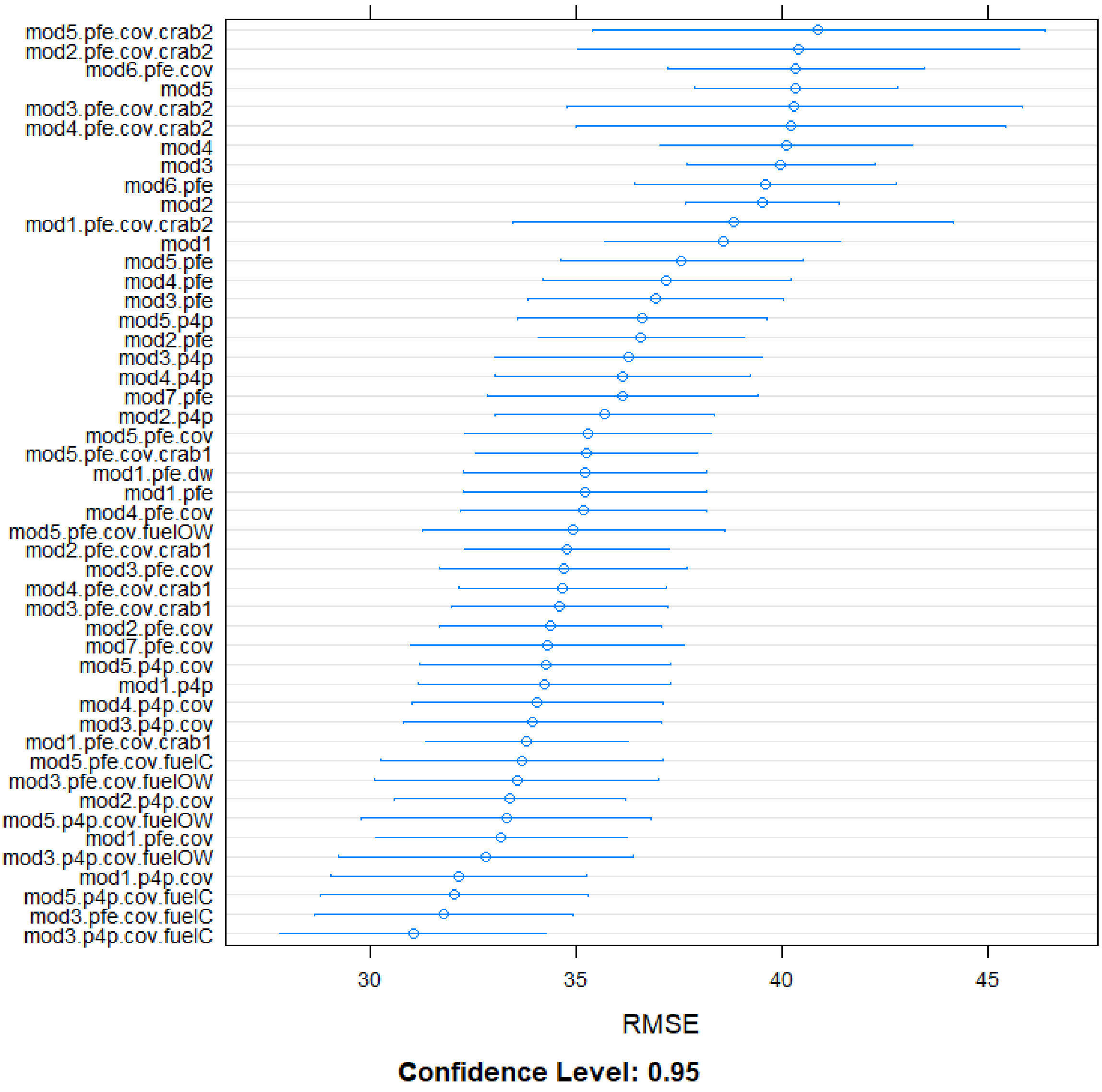


Figure 15. RMSE results for OAN models predicting average pounds per vessel, using TSCV. Error bars represent the 95% confidence interval.

Based on the results in Figure 15, the model with the lowest RMSE is the model that uses maximum sablefish price, the period 4 peak adjuster, COVID fixed effects, and northern California fuel prices. However, the GMT concluded that the best performing models that are most usable in management are 1) the model using average sablefish prices, period fixed effects, and COVID fixed effects; 2) the model using average sablefish prices, period fixed effects, COVID fixed effects, and average Dungeness crab prices; and 3) the model using average and maximum sablefish prices, period fixed effects, and COVID fixed effects.

The MAE from the TSCV runs was 21.6-25.7 vessels per period for the top three models (Table 8), and we used a similar equation to the one described in Section 3.3.1, where for OAN, the average pounds landed per vessel to date in 2023 has been 2,775 lbs. The mean absolute error for a full year, based on out-of-sample predictions, is estimated at roughly 163.1-194.1 mt, which equates to 23.7-28.2 percent of the 2023 OAN landings target and 22.8-27.2 percent of the 2023 OA share.

Table 8. MAE results for OAN models predicting average pounds per vessel, using TSCV, annual absolute error estimates, and annual error estimates as a percent of the OAN target and OA share.

Model	MAE	Annual error estimate (mt)	% of OAN Target	% of OA Share
mod1.pfe.cov.crab1	25.7	194.1	28.2%	27.2%
mod2.pfe.cov	21.6	163.1	23.7%	22.8%
mod1.pfe.cov	25.3	191.1	27.8%	26.7%

Section 4.3.2. Tendency to Over- or Under-predict

Figure 16 compares the percent prediction error of those three models. The model with the lowest RMSE in the TSCV was not the same model with the lowest MPE in Figure 16. The TSCV assesses out-of-sample error, and setting predicted Dungeness crab prices in addition to sablefish prices would be a new method that has not previously been used in management nor thoroughly developed, so the GMT concluded that the model using average sablefish prices, period fixed effects, and COVID fixed effects is the best model to be considered for use in management, and this is the model for which we evaluate risk to the sablefish OA share and commercial harvest guideline.

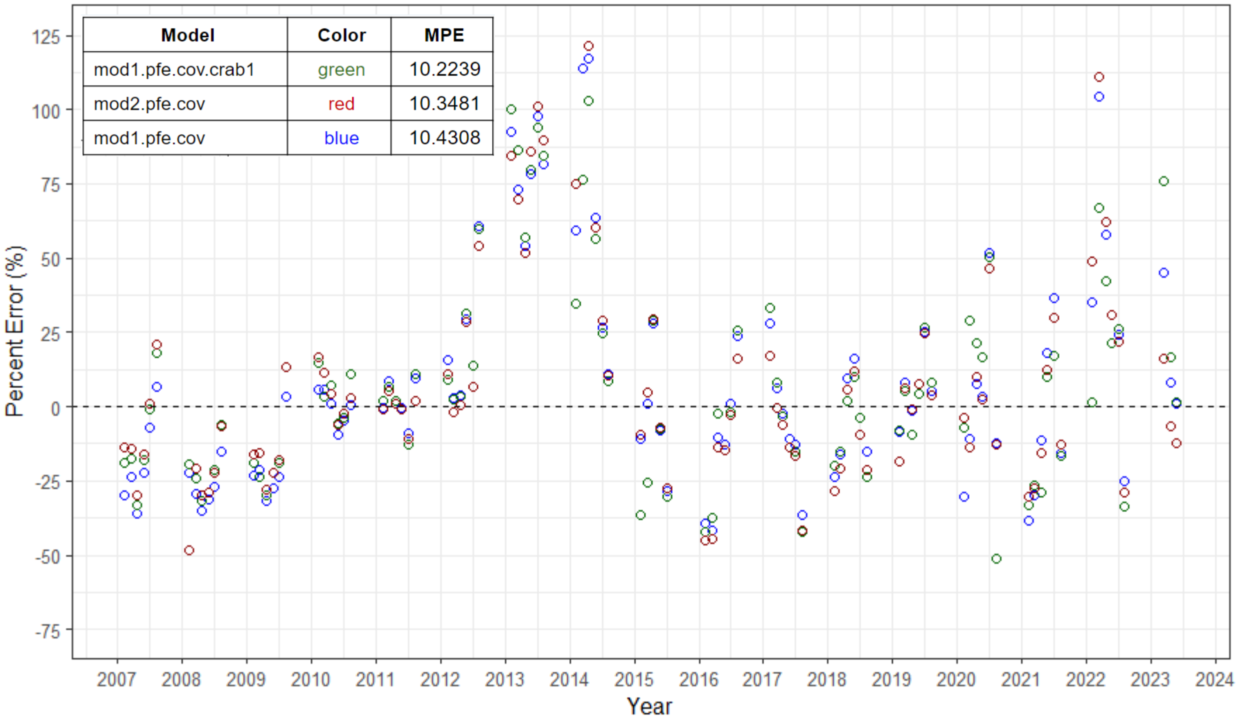


Figure 16. Percent prediction error of top select OAN models predicting the number of vessels, based on TSCV results. Mean Percent Error (MPE) for each of those top three models is shown in the table within the figure.

Recommendation: Season structure (e.g., number of open days for salmon or crab) may be a good predictor of sablefish effort and catch. However, the analysts indicated that the timing of the regulatory cycle may make using these variables in forecasts difficult.

Because of the volatility of the OAN participation model and the fleet's cross-over with the Dungeness crab fishery, the GMT gave further consideration to the feasibility of using Dungeness crab season start dates as an inseason predictor variable in the OAN participation model. Each state sets their own season start dates based on meat recovery percentage as well as other factors, and the start dates vary year-to-year. The Washington commercial Dungeness crab fishery, as an example, could open as early as December 1 or as late as February 15, and then the season typically closes in September. Washington generally has a later start date than the other states because of the tribal head start. Given the timing of season start dates, the earliest the Council could consider trip limit adjustments based on a model that uses the Dungeness crab season start date as a predictor is March. The crab season start date likely has the most influence on period 1 OAN participation, so including the start date in the model at the March meeting would likely not provide much benefit for the remainder of the year (i.e., periods 2-6). Dungeness crab prices, and generally market dynamics, are likely the more influential driver for periods 2-6 after the crab season dates have already been set.

Section 4.3.3. Comparison of Period-specific Effects

Recommendation: The analysts could also allow slope coefficients to vary by period by including an interaction between the slopes and period-specific dummy variables. This would leave few *degrees of freedom and may lead to overfitting, however, which again highlights the importance of using out-of-sample prediction in model selection.*

The GMT explored allowing the slope coefficients to vary by period using an interaction term, but results of TSCV runs indicated that this approach did not perform better than either the period 4 peak adjuster used in the status quo model or simply using period-specific fixed effects with a factor variable (Figure 17). Additionally, compared to using period-specific effects, allowing the slope coefficients to vary by period may be unnecessarily complex for use in setting trip limits.

Table 9. Variables and methods used to compare period-specific effects in OAN models predicting participation.

Qualifier	Description
mod1	avg. sablefish price
mod3	max. sablefish price
.pfe	period fixed effects
.p4p	period 4 peak adjuster fixed effects
.dum	Dummy variable for period 4 Period 4 = 1 Periods 1-3 and 5-6 = 0
.slo	interaction between period and price variables

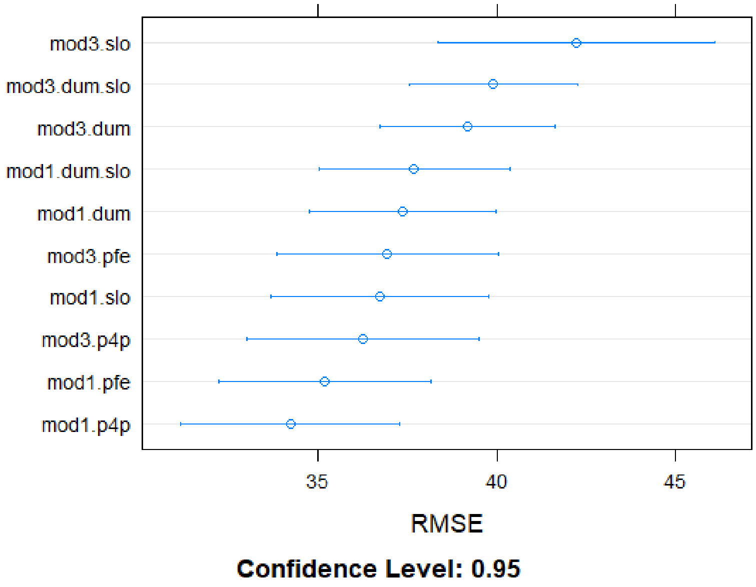


Figure 17. RMSE results of TSCV to compare OAN models using different methods to account for period-specific effects.

Section 4.3.4. Price Forecasting Evaluation

Recommendation: If variables that require forecasting are included in the models, then the forecast method must also be included in the model evaluation.

The same method used to set future predicted prices in the LEN model is used in the OAN model. Sablefish prices in the OAN sector in 2023 have been trending slightly lower than prices in 2022 and 2021, but those are the most similar recent years, so the GMT would use prices from those years to set predicted prices (Figure 18). The GMT will typically indicate to the Council when very recent prices are diverging from the general trend, such as in this case of 2023 prices tracking notably low, and how that might impact model projections.

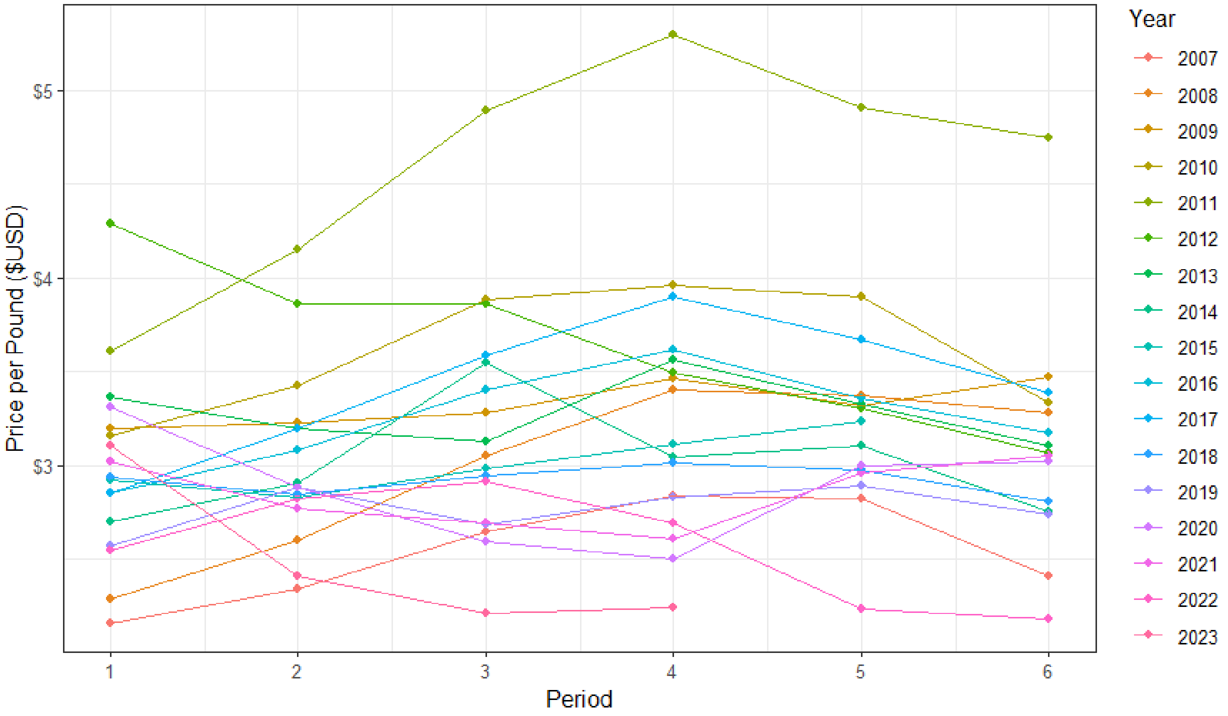


Figure 18. Average inflation adjusted sablefish price per pound by period in the LEN sector, 2012-2023.

To provide an example of the risk of getting our price predictions wrong, we used 2013-2015 sablefish prices to predict 2016 and 2017 prices, since 2016 and 2017 prices were generally higher than the prior three years, and higher prices attract more vessels into the fleet. Note that those are the same years used in the LEN price prediction evaluation in Section 3.3.3. The MPE across all predictions shown in Figure 19 is -3.46 percent. Prices were most under-predicted, up to roughly 25 percent, in period 4 when the most participation in the OAN fleet tends to occur. Figure 20 shows the prediction error of total fleetwide landings using 2013-2015 prices as inputs to the model to set predicted 2016 and 2017 prices. While the OAN price setting method severely over-predicted total OAN landings in period 1 of 2017, up to 187 percent, the maximum under-prediction was 29 percent in period 6 of 2017. If the model severely over or under predicts landings in period 1, the Council can adjust trip limits later in the year based on actual period 1 landings. The MPE across all fleetwide landings predictions shown in Figure 20 is 8.4 percent.

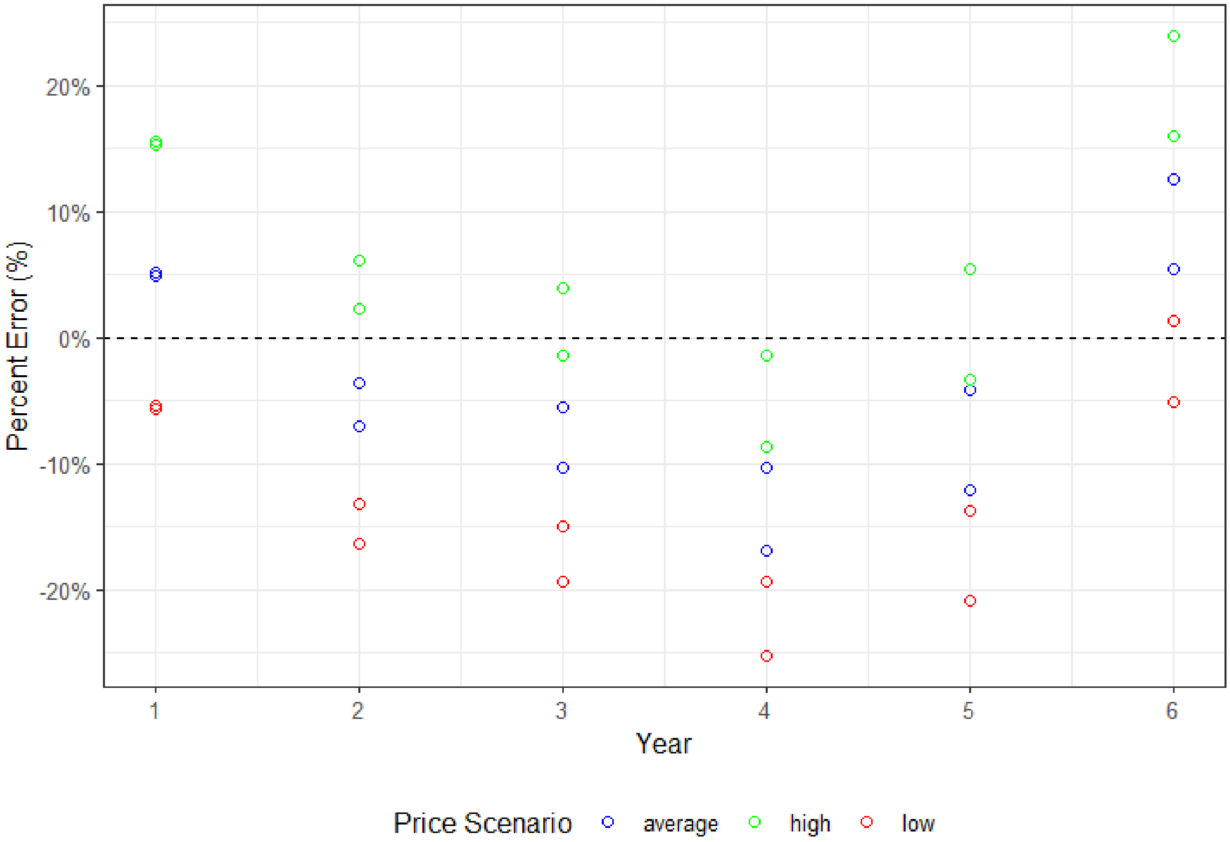


Figure 19. Example of percent error of sablefish price predictions using 2013-2015 prices to predict 2016-2017 prices. There are five data points for each period, representing the percent error under the low, average, and high price scenarios for both 2016 and 2017.

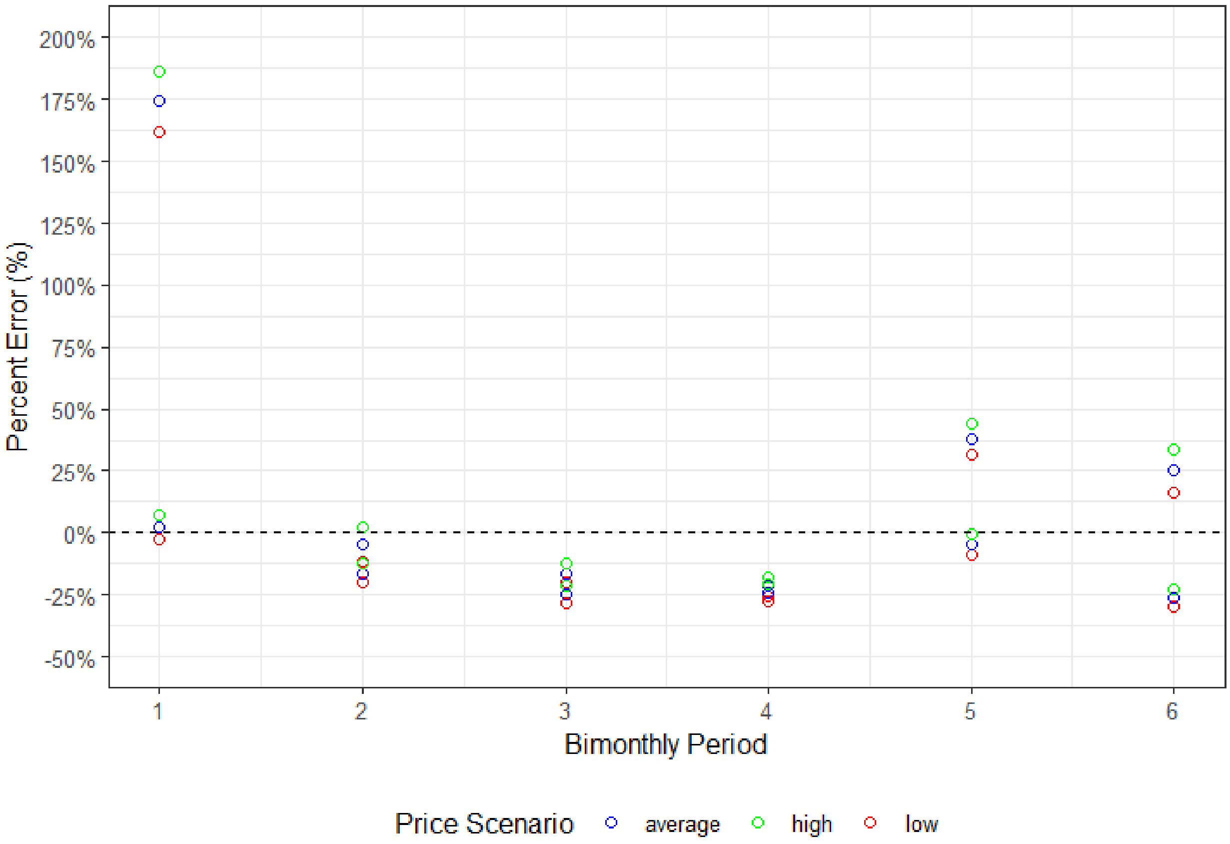


Figure 20. Example of percent error of total OAN sablefish landings using 2013-2015 prices to predict 2016-2017 prices.

Section 4.4. Best Performing Models for OAN

In summary, the GMT has identified the following models as the best performing models, at this time, to predict OAN landings for management purposes:

$$Average\ lbs.\ per\ vessel \sim bimonthly\ trip\ limit\ +\ factor(PERIOD)$$

$$Number\ of\ vessels \sim avg.\ sablefish\ price\ per\ lb.\ +\ factor(PERIOD)\ +\ factor(COVID)$$

Section 4.5. Risk to the Sablefish North Annual Catch Limit

Recommendation: The analysts should consider the consequences of poor prediction performance and evaluate model performance in terms of the risk of exceeding the ACL. Measures such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) measure the magnitude of expected prediction error.

Table 10 summarizes the total annual in-sample prediction error for 2017-2022 using the combination of the best performing LEN average pounds and number of vessels models. This prediction error is based on the current best performing models in this methodology review; however, within those years, the predictions were made with a slightly different model. That means that, while the current best model over-predicted landings in 2022, the model used to set trip limits at the time likely under-predicted landings leading to the target being exceeded by 2 percent.

In general, the OA fishery is harder to predict, because it has the greatest influx of new entrants, which can flow in or out of the fishery based on their fishing portfolios. While the total prediction error as a percent of the OA share ranges from 3 percent to 16 percent, the OA share is only 9.4 percent of the sablefish north commercial harvest guideline, which, as stated above, is used as a proxy for risk to the sablefish north ACL. Similar to the LEN model, the risk to the sablefish north ACL is greater when the model under predicts. In 2017 and 2018, the model under-predicted landings, so the hypothetical OAN target attainment was higher than the actual OAN target attainment. In other words, if the current best model had been used in 2017 to set trip limits and the model projected 91 percent attainment, the target could have been exceeded by 107 percent. This, again, indicates the need for the Council to be precautionary when the DTL models predict attainment levels above 90 percent, especially when projecting for an entire year. As with the LEN fleet, the Council has the opportunity to adjust trip limits at every Council meeting throughout the season, and attainment projections may change as the season progresses based on actual participation, price fluctuations, or actual landings. The Council can respond as needed to keep landings within the target.

Table 10. Sum annual prediction error of the OAN model (combined average pounds per vessel and number of vessels) and the retrospective impacts of that annual prediction error on the OAN target and the OA share, 2017-2022. Under-predictions are highlighted in gray.

Year	Sum annual prediction error (mt)	Under- or over-prediction	Actual OAN Target attainment	Hypothetical[a/] OAN Target attainment	OA Share (mt)	Total prediction error as % of OA Share
2017	-69.9	under	91%	107%	441	15.8%
2018	-38.7	under	76%	84%	460	8.4%
2019	16.2	over	73%	69%	471	3.4%
2020	36.3	over	37%	29%	481	7.6%
2021	20.6	over	44%	41%	580	3.6%
2022	15.4	over	102%	99%	552	2.8%

a/ Hypothetical attainment if the model had predicted what is in the actual target attainment column. The inverse of the sum annual prediction error is added to the actual annual landings. In other words, if the model underpredicted that year, the absolute value of the prediction error is added to the actual landings, but if the model overpredicted that year, it is subtracted from the actual landings for that year.

Section 5. Cumulative Risk to the Sablefish North ACL

The estimated range of annual mean absolute error for both the LEN and OAN sectors, based on out-of-sample predictions from the best models identified for management by the GMT, is 106-269 mt, which equates to 1.4-3.5 percent of the total 2023 sablefish north commercial harvest guideline. Given this estimated margin of error, the prediction error from these models poses very little conservation risk to the stock.

References

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.