METHODOLOGY REVIW – FINAL
ADDITIONAL MATERIAL REQUESTED AT THE
2023 SALMON METHODOLOGY REVIEW MEETING

Supplement to "A re-evaluation of preseason abundance forecasts for Sacramento River winter Chinook salmon" in response to requests from the SSC-SS

Tanya Rogers
Michael O'Farrell

Fisheries Ecology Division
Southwest Fisheries Science Center
National Marine Fisheries Service
National Oceanic and Atmospheric Administration
Santa Cruz, CA

October 23, 2023

Following the Salmon Methodology Review meeting, the Scientific and Statistical Committee – Salmon Subcommittee (SSC-SS) requested a supplemental report that:

1. Calculates a targeted suite of performance metrics identified by the analyst for comparisons of the control rule outputs as applied to the postseason estimates or each of the primary forecast alternatives considered.

2. Repeats the main analyses for forecast performance, and the supplemental analyses of control rule error, when both the training data and postseason estimates used for evaluating model performance are based on year-specific estimates of age structure rather than the average age structure across years.

We address each of the requests made by the SSC-SS in this Supplemental report.

Performance metrics

As a measure of baseline accuracy, we would argue that any model with an arithmetic-scale R^2, log-scale R^2, or *r* value that is negative should be eliminated from consideration, or at least given low weight. If a model cannot explain at least some of the variance in the data, and predictions do not show at least some positive correlation with the observed values, this suggests it is not a useful model.

In the case of the Gaussian Process (GP) models, we would also argue that any model producing conditional responses that are unreasonable or inconsistent with biology be eliminated, even if the R^2 values appear reasonable. Because the GP is a 'machine learning' type of method, this can occasionally happen, and so some sanity checks are required.

If we assume that over- and under-forecasts should be given equal weight, our preferred metrics of bias would be the mean error (ME) and *mean* log accuracy ratio (meanLAR). Our preferred metrics of accuracy would be mean absolute error (MAE), *mean absolute* log accuracy ratio (meanALAR), arithmetic-scale R^2, and log-scale R^2. Some of these metrics were not calculated in the initial report, but are provided here. The ME and MAE metrics quantify error directly in the units of interest (numbers of fish or impact rate percentage points). The ME calculation was adjusted so that (consistent with meanLAR) under-forecasts are negative and over-forecasts are positive. The log accuracy ratio (LAR) quantifies proportional accuracy and gives equal weight to over- and under- predictions, unlike percent error. The median LAR was used in the initial report because that was the metric used in Satterthwaite & Shelton (2023), but we feel that the mean and mean absolute would be more informative. In terms of the R^2 values, the log-scale R^2 will weigh low values more than high values, whereas the arithmetic-scale R^2 will give equal weight to all values. Both are potentially informative, but reflect different error weighting.

We note that RMSE is redundant to R^2 (which is calculated from RMSE), so it will always provide the same ranking. Because the percent error metrics (MPE, MAPE) give more weight to over-predictions than under-predictions, we feel they are not optimal for evaluating model performance.

The performance metrics, when applied to impact rates from the harvest control rule (as opposed to escapement), amount to giving no weight to abundance forecast errors when the preseason and postseason estimates are both above 3000. Performance in this case is based only on whether the model can accurately predict escapement at relatively low abundances (below 3000, where the harvest control rule begins to ramp down allowable exploitation rates). The log-scale R^2 is not relevant for the impact rates.

Updated tables with our preferred metrics are provided for the abundance forecasts (Tables 1, 2) and for impact rates (Table 3).

Ultimately, the "best" metrics to use, and the "best" model, depends on the modelling objective and what is deemed to be most important. We seek to produce forecasts that have low levels of bias, are precise, and thus, in combination are accurate. Thus, performance measures that address bias and precision are important for evaluating competing forecasts. The suite of performance measures in Tables 1-3 therefore represent our targeted suite of performance measures.

Performance with respect to control rule error (request 1)

The SSC-SS requested that performance metrics be calculated for the forecast-based impact rates relative to impact rates based on the postseason estimate. We have plotted the impact rates resulting from each model forecast and from the postseason estimates for return years 2012-2022 in Figure 1 and 2. Performance metrics for the impact rate errors are provided in Table 3. We note that in the period of 2012-2022, there were only 3 years (2016-2018) when postseason estimates were below 3000, so those years weigh heavily in this performance assessment.

In terms of impact rates, the *GP-1* model was better than the *GP-2* model by all metrics. This difference was driven mainly by the predictions in 2016 and 2017, which were closer to the postseason estimate in *GP-1* than *GP-2*. Impact rates from the two models were otherwise very similar. Both over-estimated impact rates in 2016 and 2017, and under-estimated impact rates in 2012 and 2018. The *GP-1* model was also better than the *Base* and *ETF* models by all metrics except *r*, which was highest in the *ETF* model.

Impact rates from the *ETF* model showed the most negative bias. The *ETF* model underpredicted impact rates in 2017-2019, more than any other model. In 2016, the *ETF* model overpredicted impact rates, but it was the closest of all the models to the 2016 postseason estimate. The *ETF* was the only one of the 4 models to produce a preseason escapement estimate less than 3000 in 2023. The *ETF* model thus seems to be the most 'conservative' in that it over-estimated impact rates the least, although in some cases leads to impact rates that are much lower than they would be based on the postseason estimate (e.g. 2017-2018).

The *Base* model produced the most accurate prediction in 2018, but otherwise performed poorly. Impact rates were overestimated in 2016-2017, and underestimated in 2014, 2019, and 2020. The *Base* model was the only model to produce a preseason escapement estimate more than 3000 in 2024.

The selection of the GP models was based on performance for predicting escapement. If model selection was instead based on predicting impact rates, then different models would have been selected than the 2 models that were presented here.

Analysis using year-specific estimates of age structure (request 2)

The SSC-SS requested repeating the forecast evaluation using year-specific estimates of age structure rather than the average age structure across years when estimating postseason abundances.

A plot comparing the postseason estimates using each method is provided in Figure 3. The predictions from the *Base* and *ETF* models remain unchanged, since the postseason estimates do not enter into these models. The GP model predictions will change, since they are fit to the postseason estimates.

As stated in the original report, when using E_3^0 based on year-specific (rather than average) estimates as the response variable, the *GP-1* model (predictors: DD12 and spawners) and a model with the same predictors plus empirical ETF had similar performance; however, the 3 predictor model produced unrealistic conditional relationships with the predictors and was sensitive to starting values of the hyperparameters, so this model would not be recommended. The *GP-2* model (predictors: DD12, spawners, hatchery releases) was the least biased of the models with positive R^2 values, but the *GP-1* model was the most accurate. For the year-specific estimates, we also found that using natural-origin (as opposed to total) spawners produced somewhat better fits. Model selection results are shown in Table 4. Thus, for the year-specific estimates, we present results from the *GP-1* and *GP-2* models but using natural-origin (as opposed to total) spawners. Conditional responses are shown in Figures 4 and 5. Forecasts are presented in Table 5 and Figure 6.

For all models, performance metrics using the year-specific estimates are provided for the escapement forecasts in Tables 6 and 7. Plots of impact rates are in Figures 7 and 8, and performance metrics for impact rates are in Table 8.

The relative performance of the *Base*, *ETF*, and GP models remains largely unchanged, except that the *GP-1* model outperforms the *GP-2* model in metrics of accuracy. However, the *GP-1* model is somewhat more negatively biased than the *GP-2* model, which was also the case when using average age structure. The impact rates from the GP models were closer to the postseason estimated impact rate in 2016 and 2017 compared to the model using average age structure, but impact rates were more underpredicted in 2012, which led to poorer performance metrics for both models than using the average age structure.

Table 1. Fit statistics for each model for leave-future-out forecasts (return years 2012-2022).

Model	ME	MAE	MeanLAR	MeanALAR	R^2	log R^2	*r*
Base median	-326.55	3688.55	0.06	0.93	-1.28	-0.61	-0.16
Base mode	-2917.18	3621.18	-0.83	1.26	-1.22	-1.48	-0.19
ETF median	3834.64	6706.82	0.01	1.11	-11.35	-1.42	-0.06
ETF mode	-1048.82	3971.55	-0.83	1.35	-1.71	-2.42	-0.05
GP-1	-801.49	2131.22	-0.09	0.66	0.45	0.36	0.72
GP-2	-14.44	1968.15	0.06	0.67	0.45	0.19	0.72

Table 2. Fit statistics for each model for leave-future-out forecasts (return years 2015-2022).

Model	ME	MAE	MeanLAR	MeanALAR	R^2	log R^2	*r*
Base median	-1138.63	3692.62	0	1.04	-0.68	-0.46	-0.11
Base mode	-3422.75	3950.5	-0.89	1.37	-1.10	-1.2	-0.18
ETF median	-1488.50	2460.75	-0.61	0.9	-0.04	-0.39	0.51
ETF mode	-3191.13	3290.12	-1.41	1.5	-0.60	-2.18	0.50
GP-1	-935.64	2071.68	-0.05	0.67	0.58	0.47	0.82
GP-2	-624.29	1689.4	0.02	0.68	0.66	0.31	0.83

Table 3. Fit statistics for each model for leave-future-out forecasts of *impact rates* from the harvest control rule (return years 2012-2022). Units of ME and MAE are percentage points.

Model	ME	MAE	MeanLAR	MeanALAR	R^2	*r*
Base median	0.75	2.83	0.06	0.18	-0.61	-0.26
ETF median	-1.94	2.58	-0.22	0.27	-0.84	0.68
GP-1	-0.03	2.04	0	0.14	0.03	0.47
GP-2	0.32	2.4	0.03	0.16	-0.26	0.26

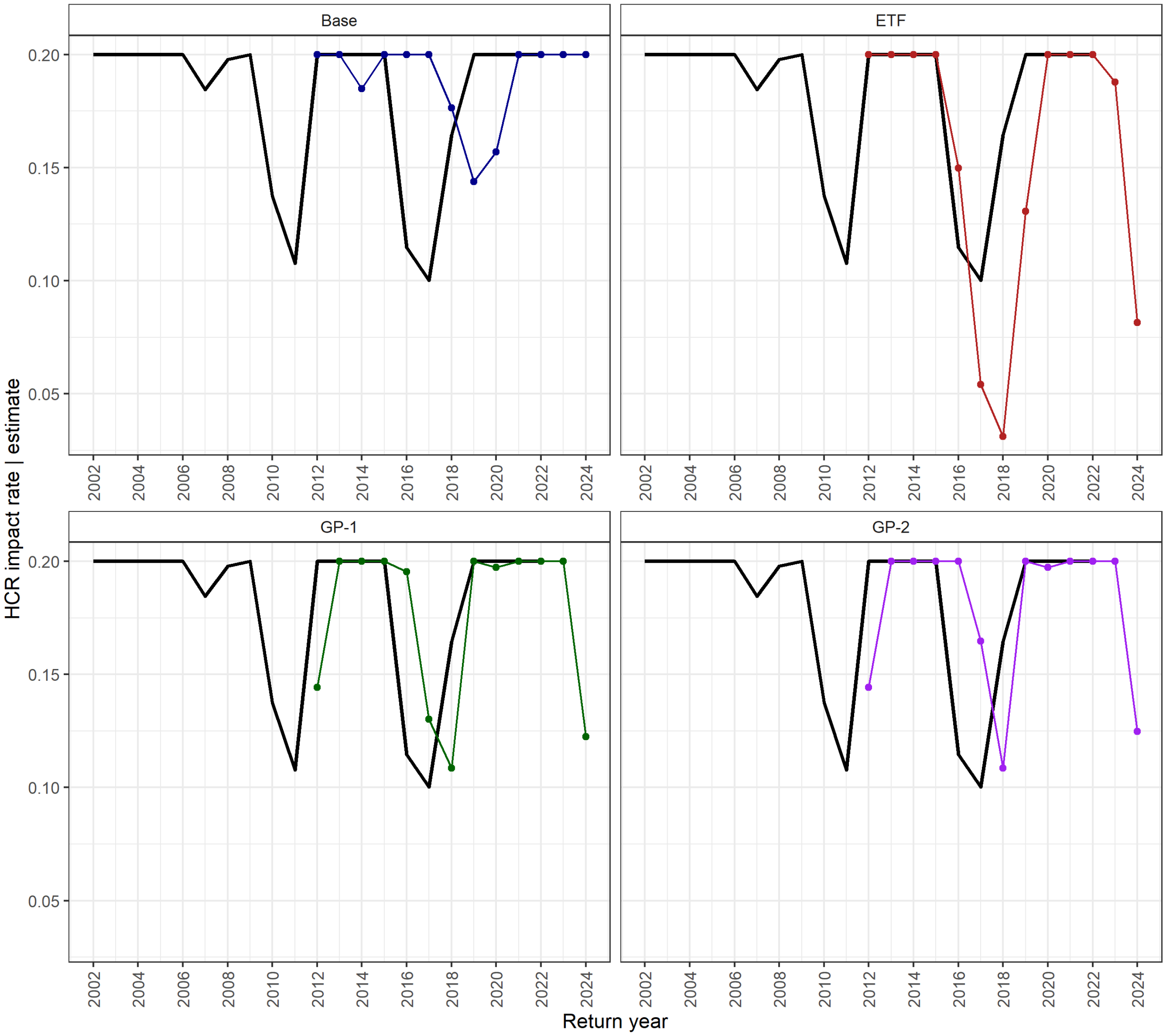

Figure 1. Harvest control rule impact rates given postseason escapement estimates (black line) and preseason leave-future-out forecasts (colored lines) from each model.

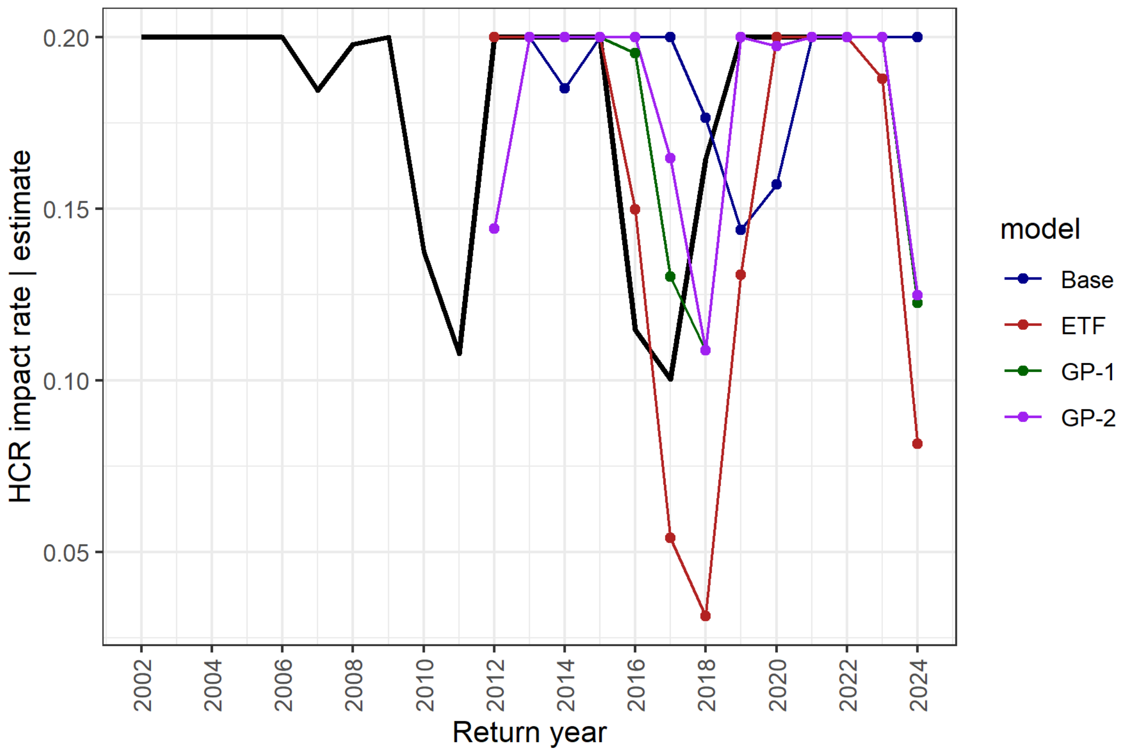

Figure 2. Same as Figure 1, but with forecast-based impact rates overlaid.

Table 4. Selection of predictors for *GP* model using *year-specific estimates* based on fit statistics for leave-future-out cross validation. Table shows all models with R^2 and *r* values >0, sorted by R^2. The best values and selected best models are highlighted in bold. *This model produces unrealistic conditional responses.

	Predictors	ME	MAE	MeanLAR	MeanALAR	R^2	log R^2	*r*
GP-1	**logDD12** **logspawnersnat**	-634.61	1906.57	-0.14	0.62	**0.54**	0.30	0.77
*	logDD12 ETF.official logspawners	-806.48	**1851.44**	-0.26	**0.57**	0.53	**0.36**	0.77
	logDD12 logspawners	-1056.51	1970.00	-0.19	0.63	0.50	0.31	**0.78**
	logDD12 ETF.official logspawners logyJhat	-604.52	2043.30	-0.18	0.61	0.45	0.30	0.72
	logDD12 logyJhat	523.09	2204.54	0.14	0.69	0.42	0.14	0.73
	logDD12 logspawners logyJhat	214.78	2206.84	0.06	0.71	0.40	0.12	0.73
GP-2	**logDD12** **logspawnersnat** **logyJhat**	**137.04**	2288.90	0.04	0.73	0.36	0.10	0.71
	logDD12 ETF.official logspawnersnat	-1287.71	2132.72	-0.34	0.62	0.25	0.28	0.64
	logDD12 ETF.official logyJhat	668.08	2283.28	**-0.01**	0.64	0.19	0.25	0.78
	logDD12 ETF.official logspawnersnat logyJhat	-972.99	2429.31	-0.24	0.66	0.15	0.23	0.55

Table 5. Postseason escapement estimates and leave-future-out forecasts from the GP models fit using *year-specific estimates*.

Brood year	Mgmt year	Return year	E_3^0 yearly	E_3^0 mean	*GP-1* median	*GP-2* median
1999	2001	2002	9042	8488		
2000	2002	2003	9732	9070		
2001	2003	2004	6329	5962		
2002	2004	2005	19518	18046		
2003	2005	2006	19569	18862		
2004	2006	2007	1857	2612		
2005	2007	2008	3022	2947		
2006	2008	2009	4483	4142		
2007	2009	2010	1344	1436		
2008	2010	2011	501	694		
2009	2011	2012	3523	3255	998	1279
2010	2012	2013	6436	5946	8449	8961
2011	2013	2014	3163	3060	5259	7658
2012	2014	2015	3990	3709	3377	3193
2013	2015	2016	843	865	2734	3284
2014	2016	2017	526	507	940	1931
2015	2017	2018	2280	2112	649	654
2016	2018	2019	8757	8119	6172	6173
2017	2019	2020	7471	6918	2887	2887
2018	2020	2021	11467	10883	9429	12761
2019	2021	2022	3997	6369	4580	5179
2020	2022	2023			3963	4839
2021	2023	2024			1028	1089

Table 6. Fit statistics for each model for leave-future-out forecasts (return years 2012-2022) using *year-specific estimates*.

Model	ME	MAE	MeanLAR	MeanALAR	R^2	log R^2	*r*
Base median	-391.09	4097.45	0.05	0.99	-1.40	-0.73	-0.29
Base mode	-2981.73	3637.55	-0.84	1.25	-1.21	-1.60	-0.29
ETF median	3770.09	6898.45	0.00	1.15	-10.09	-1.48	-0.05
ETF mode	-1113.36	3972.64	-0.83	1.34	-1.55	-2.52	-0.05
GP-1	-634.61	1906.57	-0.14	0.62	0.54	0.30	0.77
GP-2	137.04	2288.90	0.04	0.73	0.36	0.10	0.71

Table 7. Fit statistics for each model for leave-future-out forecasts (return years 2015-2022) using *year-specific estimates*.

Model	ME	MAE	MeanLAR	MeanALAR	R^2	log R^2	*r*
Base median	-1119.75	4214.25	0.01	1.12	-0.89	-0.61	-0.34
Base mode	-3403.88	3932.38	-0.88	1.35	-1.07	-1.32	-0.40
ETF median	-1469.62	2831.88	-0.60	0.98	-0.16	-0.54	0.40
ETF mode	-3172.25	3276.75	-1.39	1.49	-0.59	-2.35	0.39
GP-1	-1070.53	1792.41	-0.13	0.60	0.64	0.49	0.86
GP-2	-408.62	1989.11	0.03	0.72	0.61	0.27	0.80

Table 8. Fit statistics for each model for leave-future-out forecasts of *impact rates* from the harvest control rule (return years 2012-2022) using *year-specific estimates*. Units of ME and MAE are percentage points.

Model	ME	MAE	MeanLAR	MeanALAR	R^2	*r*
Base median	0.69	2.77	0.06	0.18	-0.62	-0.27
ETF median	-2.00	2.66	-0.23	0.28	-1.01	0.64
GP-1	-0.52	2.20	-0.03	0.15	-0.20	0.44
GP-2	0.04	2.55	0.01	0.18	-0.41	0.23

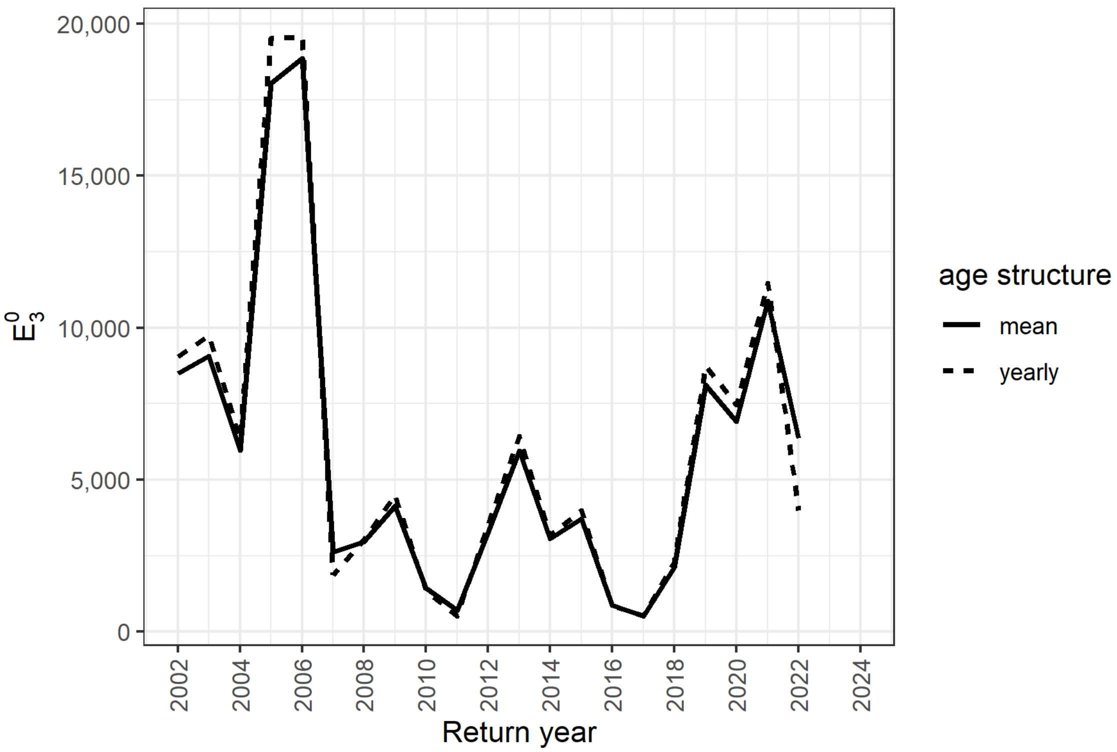

Figure 3. Comparison of postseason estimates using mean vs. year-specific estimates of age structure.

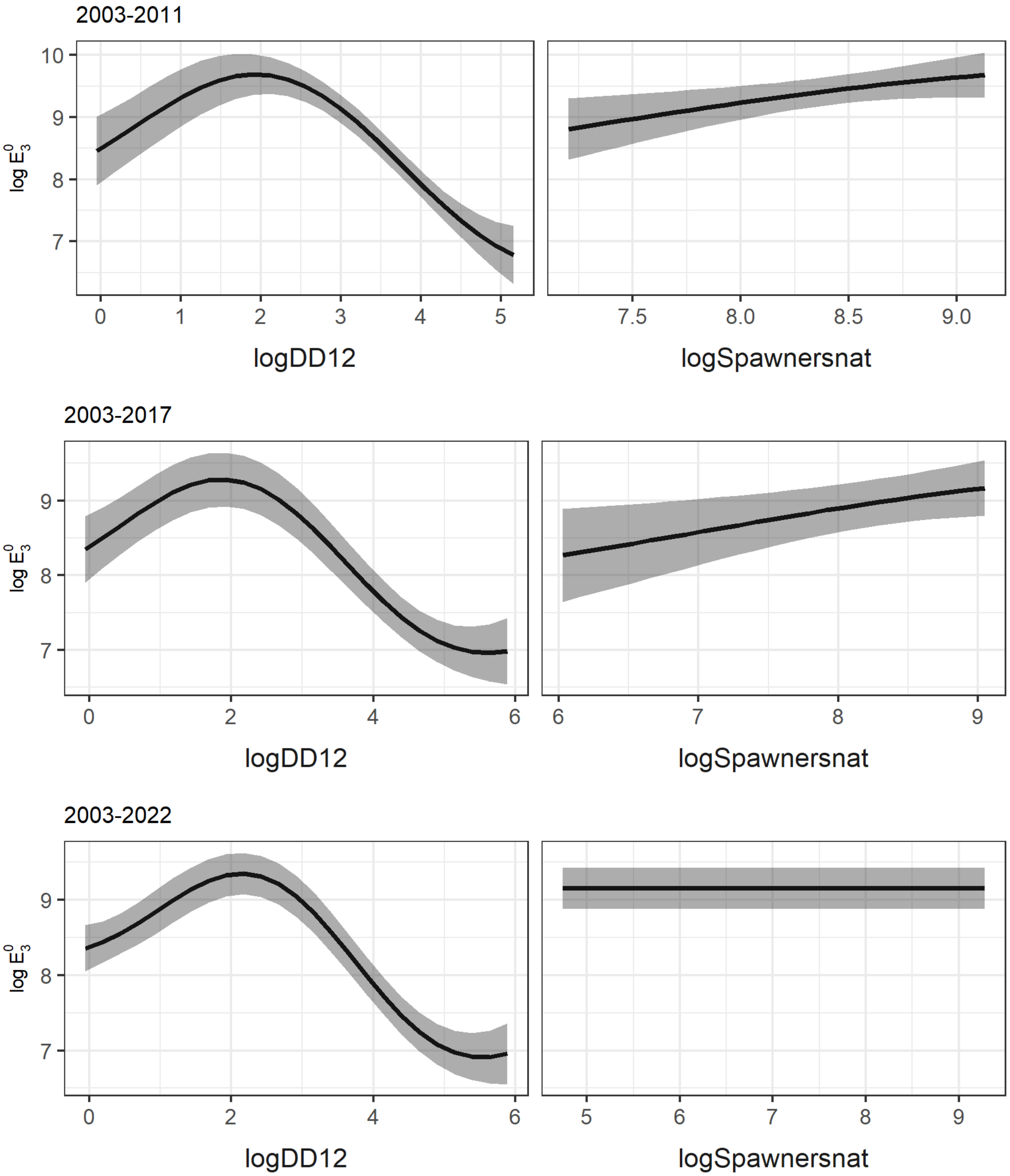

Figure 4. Conditional effects of each predictor in the *GP-1* model using *year-specific estimates* using differing amounts of training data (through return year 2011, 2017, and 2022), with other predictors fixed to their mean value (interactions among predictors are present but not shown). logDD12 is the temperature covariate.

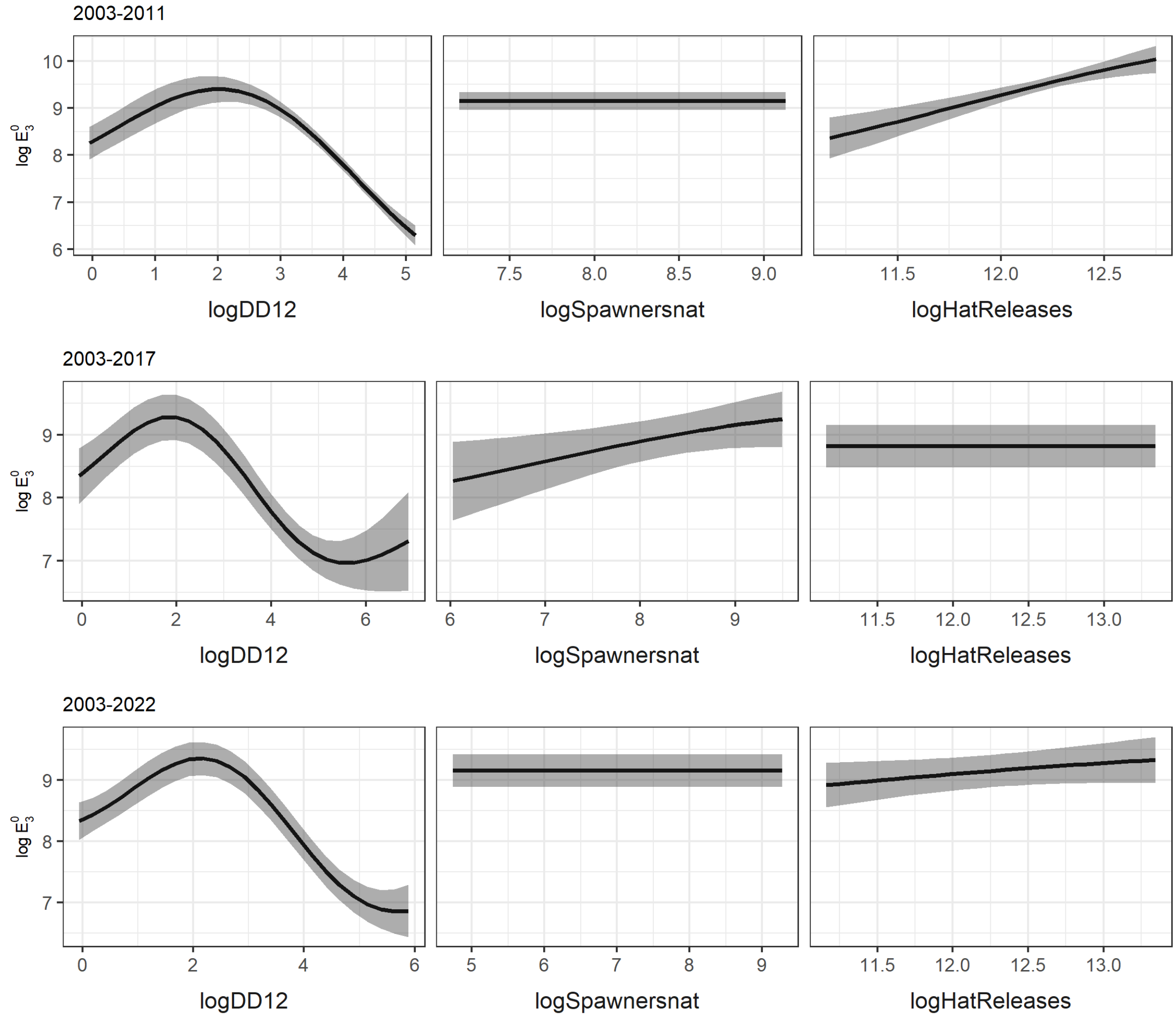

Figure 5. Conditional effects of each predictor in the *GP-2* model using *year-specific estimates* using differing amounts of training data (through return year 2011, 2017, and 2022), with other predictors fixed to their mean value (interactions among predictors are present but not shown). logDD12 is the temperature covariate.

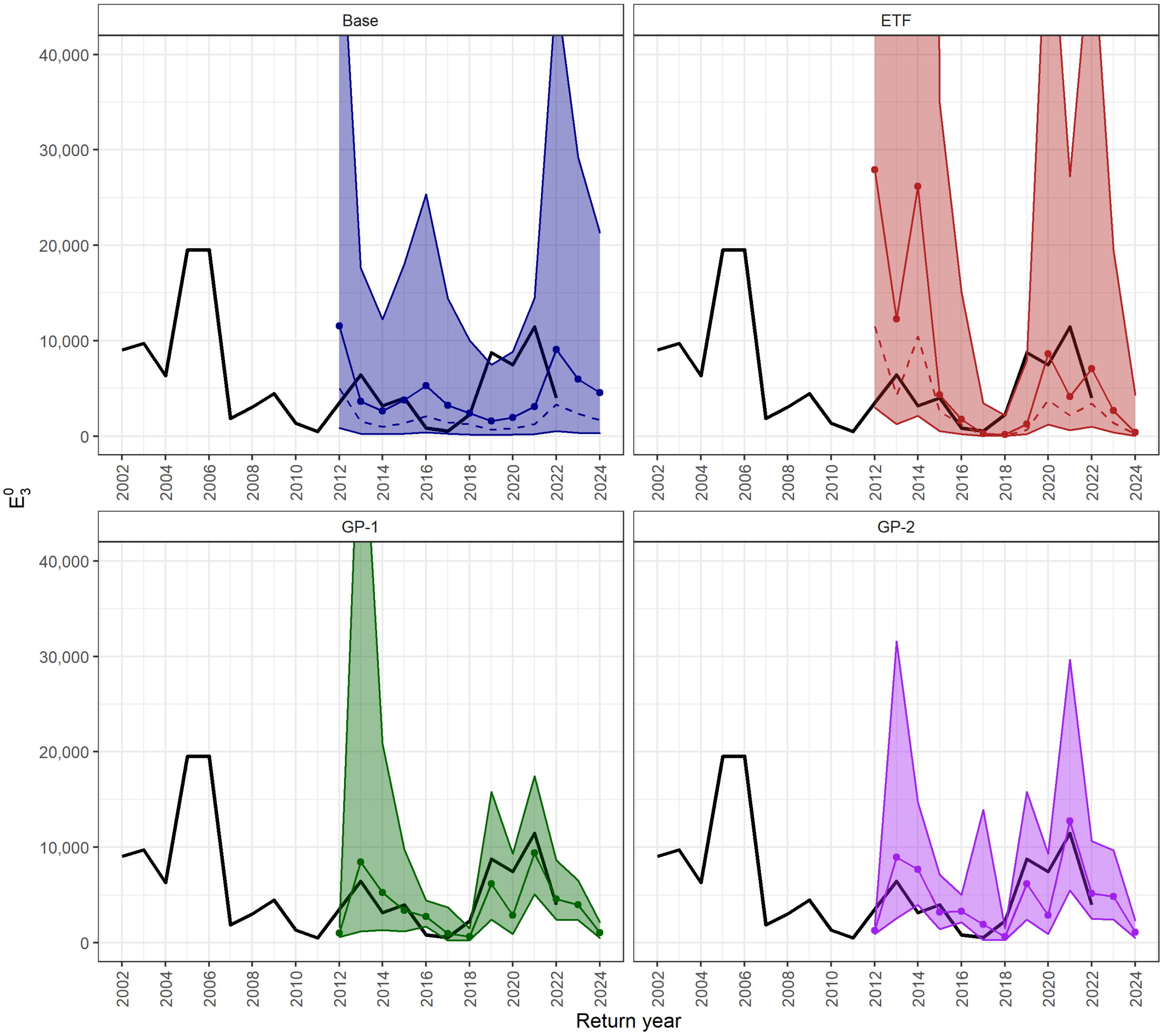

Figure 6. Postseason escapement estimates (black line) and leave-future-out forecasts (colored lines) from each model using *year-specific estimates*. Bands are 95% credible intervals (cropped in the *Base* and *ETF* models). Solid lines are medians and dashed lines (where plotted) are modes.

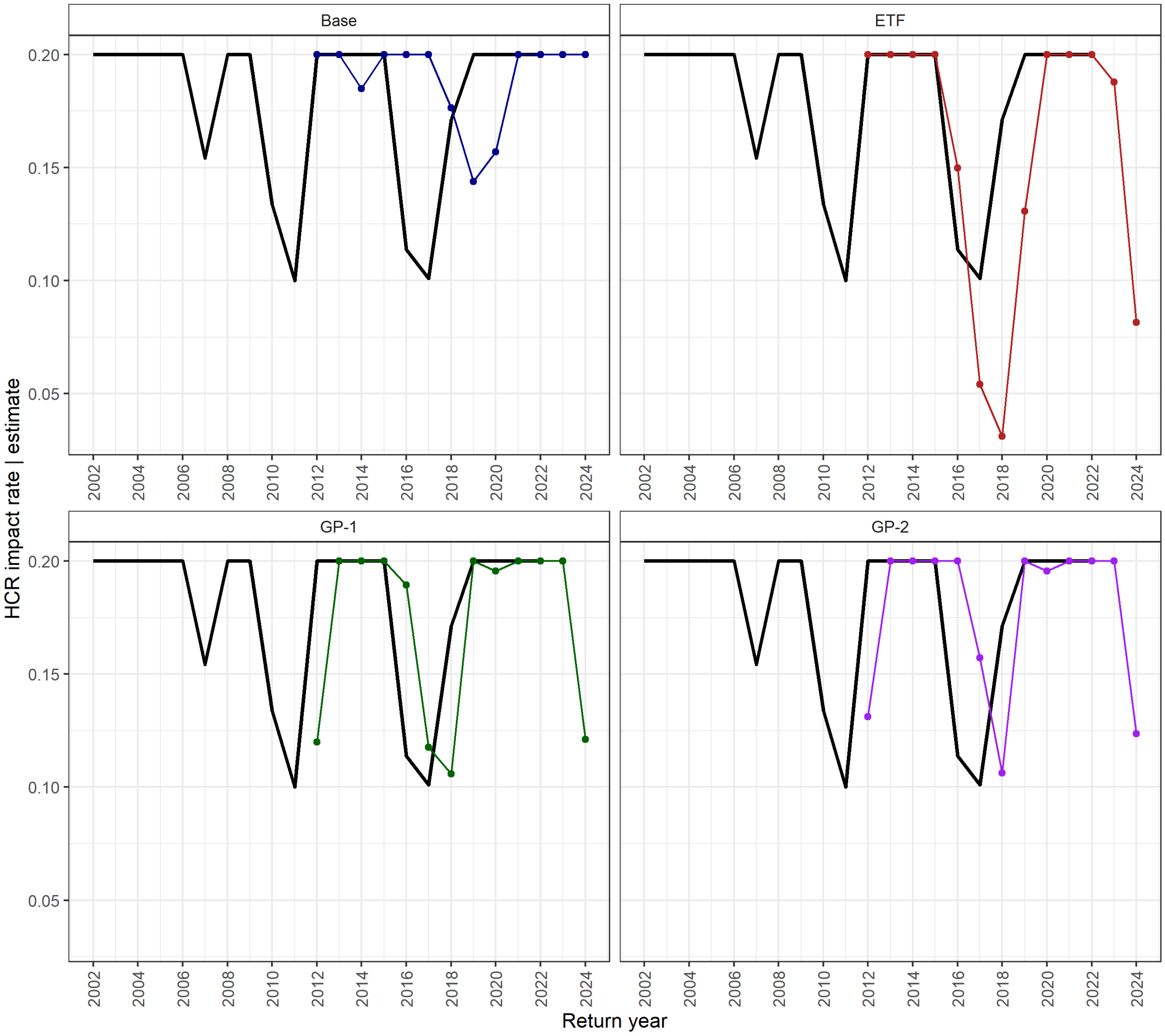

Figure 7. Harvest control rule impact rates given postseason escapement estimates (black line) and preseason leave-future-out forecasts (colored lines) from each model using *year-specific estimates*.

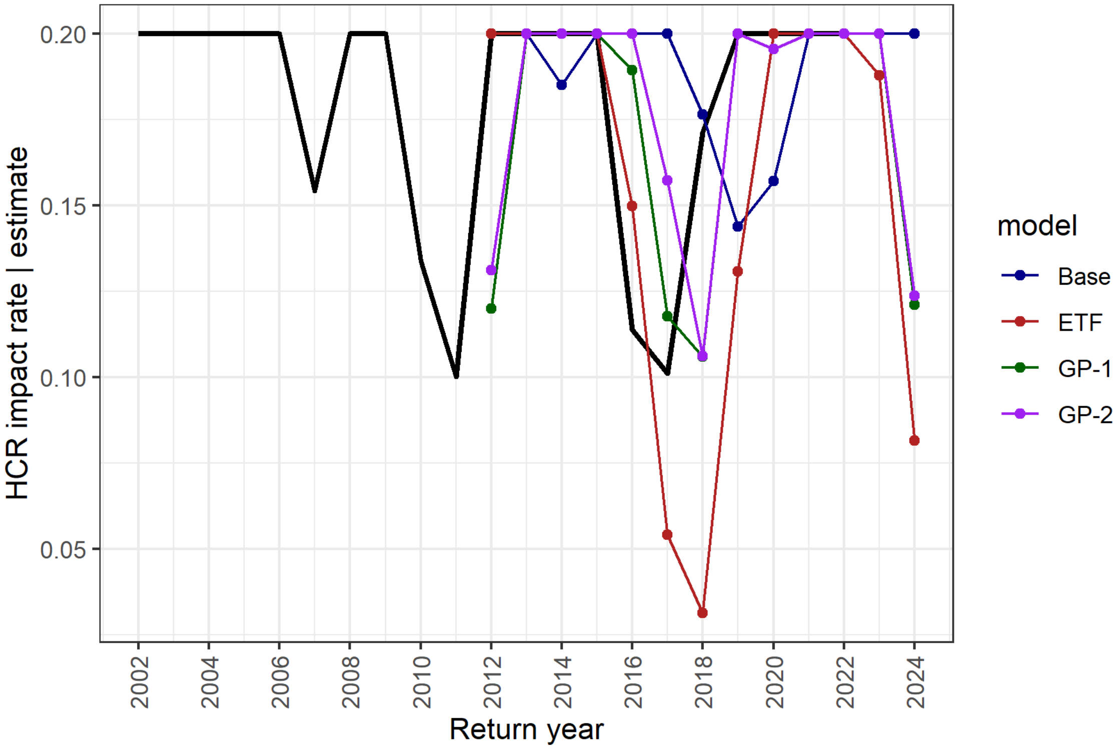

Figure 8. Same as Figure 4, but with forecast-based impact rates overlaid.