3 November 2010

The impact of different performance measures on model selection for Fraser River sockeye salmon

Jonathan W. Cummings[1,3,*], Merran J. Hague[2],

David A. Patterson[2], and Randall M. Peterman[1]

[1] School of Resource and Environmental Management
Simon Fraser University
Burnaby, British Columbia, Canada V5A 1S6

[2] Fisheries and Oceans Canada
Cooperative Resource Management Institute
Simon Fraser University
Burnaby, British Columbia, Canada V5A 1S6

[3] Present address:
Rubenstein School of Environment and Natural Resources
University of Vermont
Burlington, Vermont, 05405

*Corresponding author:
peterman@sfu.ca

Submitted to the North American Journal of Fisheries Management (NAJFM)

Abstract

Uncertainties prevalent in fisheries systems result in deviations between management targets and observed outcomes. As an example of attempting to deal with such uncertainty, fishery managers of sockeye salmon (*Oncorhynchus nerka*) from the Fraser River, British Columbia, Canada use environmentally-based "management adjustment" (MA) models to forecast indices of in-river loss of adults as they migrate upstream to spawn. Forecasts of losses from MA models are directly incorporated into estimates of total allowable catch, resulting in harvest reductions that aim to increase the probability of achieving spawning escapement targets. However, the relative forecasting success of different MA models has not been assessed rigorously. Therefore, we used a suite of forecasting and hindcasting metrics to rank the performance of numerous MA models. We found that the rank of each model varied across sockeye salmon stock aggregates (i.e., run-timing groups) and depended on the performance measures chosen for evaluation. Although model selection in fisheries research is often determined solely by model-fitting criteria such as R^2 and AIC, in our case, models with the largest mean R^2 value and/or the smallest mean AIC_c often ranked poorly for other measures of model hindcast performance (i.e., mean raw error, mean absolute error, root mean square error). Although no single model performed best across all run-timing groups, failure to apply an MA produced the worst (in 3 of 4 run-timing groups) or second-worst (in 1 of 4) outcomes. We provide a framework for model selection based on the relative importance of different model selection criteria and their associated performance measures. We urge scientists and managers to work closely together to develop appropriate metrics to assess

model performance and objectively select forecast models that will best meet management objectives.

Keywords: Fraser River, sockeye salmon, management adjustment, model selection, performance measures, management objectives, retrospective analysis

Introduction

Fisheries managers are tasked with meeting society's competing demands, including opportunities for income, employment, cultural identification, productive ecosystems, recreation, and sustenance. However, given the variability in natural systems and variation in effectiveness of management efforts, there can be considerable discrepancies between target management objectives and realized outcomes at the end of a fishing season (Holt and Peterman 2006). Improvements to methods that quantify the complex system dynamics contributing to these sources of outcome uncertainty can therefore help managers meet both spawning and harvest objectives (Holt and Peterman 2008; Macdonald et al. 2010). For example, for fisheries on Fraser River sockeye salmon (*Oncorhynchus nerka*) in British Columbia (BC), Canada, managers use models that forecast in-river loss of upstream-migrating adults to help reduce one such source of uncertainty and increase the chance of meeting spawning escapement targets.

The Fraser River sockeye salmon fishery is the largest salmon fishery in Canada, with average annual catches of 5.5 million fish for over 50 years (Pacific Salmon Commission-PSC). These salmon are of great importance both as a fishable product and as a social and cultural resource to both First Nations and residents of British Columbia. However, in the last 16 years, both population abundance and catches have declined

(Peterman et al. 2010). In addition, there has been a recent increase in the frequency of large in-river losses of upstream-migrating adults. Those losses correspond with an increasing frequency of extreme environmental conditions during migration (Macdonald et al. 2010). Specifically, in years with extremely high river temperatures or flows, large associated in-river losses have created a challenge to fisheries managers to meet both spawning escapement targets and harvest allocation objectives (i.e., First Nations, commercial, and recreational catch) (Cooke et al. 2004; Patterson et al. 2007b).

Management adjustment (MA) models attempt to predict a proxy for this in-river loss (termed escapement discrepancy), which is defined as the ratio of upper-river escapement estimates to lower-river estimates of abundance, after accounting for in-river catch estimates. Lower-river escapement estimates are made at a hydroacoustic facility near Mission, BC, whereas upper-river escapement estimates are obtained from spawning ground surveys (locations in Figure 1). In the absence of direct mortality estimates, the estimated escapement discrepancies are used to represent a historical index of in-river loss. During the fishing season, these discrepancies are incorporated by fisheries managers into estimates of total allowable catch, thus potentially reducing available harvest for regulated fisheries in years when the forecast of loss is high (Macdonald et al. 2010). Underestimates of in-river loss can lead to conservation concerns with too few fish reaching spawning grounds due to excess catch, whereas overestimates of in-river loss can result in foregone catch. Therefore, management of the Fraser River sockeye salmon fishery would benefit from identifying MA models that produce the most precise and unbiased predictions of in-river loss.

For management purposes, returning Fraser River sockeye salmon stocks are assigned to four major management groups (run-timing groups) based on their historical return times to the river: (1) Early Stuart, (2) Early Summer, (3) Summer and (4) Late-run (Gable and Cox-Rogers 1993). Models specific to each run-timing group are used to predict appropriate harvest or management adjustments (i.e., MAs), which are then used by managers to account for in-river losses (Hague and Patterson 2007; Macdonald et al. 2010). The larger the predicted escapement discrepancy, the larger the associated MA value required (i.e., reduction in catch) to meet a given spawning escapement target. Forecasting appropriate MAs is difficult because estimates of in-river loss are not only affected by natural mortality resulting from extreme environmental conditions, but also by potential measurement errors in adult salmon abundance in both lower-river and spawning-ground escapement estimates, uncertain catch estimates, and unreported harvest (Macdonald et al. 2010).

Several MA models have been applied historically by biologists at the two relevant agencies responsible for Fraser River sockeye salmon management, Fisheries and Oceans Canada (DFO) and the PSC (Macdonald et al. 2010). The simplest model assumes that the MA should equal the average of historical annual escapement discrepancies between the lower-river (Mission) and up-river (spawning ground) abundances. In 2001, DFO and PSC biologists adopted MA models that were based on environmental conditions to better reflect the association between extreme freshwater migration conditions and in-river loss estimates; these models are supported by well-documented biological rationale (Macdonald et al. 2010). These models forecast escapement discrepancies for each run-timing group, both pre-season and in-season, as a function of (a) predicted Fraser River environmental conditions (water temperature and flow) for Early Stuart, Early Summer, and Summer runs, and (b) fish behavior (river entry timing) for the Late-run (Hague and Patterson 2007; Macdonald et al. 2010). However, despite the wide variety of these MA models that have either been proposed or applied in the past, there has been little comprehensive analysis that quantitatively compares their statistical performance.

Here we compare the performance of a suite of MA models using retrospective analysis, which is a cross-validation technique (Shao 1993) that uses historical data up to a given year to fit various forecasting models, and then iteratively re-fits the model with each additional year of data and compares annual forecasts to subsequently observed annual values. The performance of each model is then averaged over the entire period of analysis. In fisheries research, such retrospective methods have been previously applied to evaluate a variety of forecasting methods, such as models predicting salmon abundance (Wood et al. 1997, Peters et al. 2001; Holt and Peterman 2004; Haeseker et al. 2005; Haeseker et al. 2008) and annual harvests of Atlantic menhaden (*Brevoortia tyrannus*) (Hanson et al. 2006).

Our research objective was to develop a standardized framework to quantitatively evaluate new and existing MA models and, more generally, to explore how different model performance measures can influence the rank-order of model selection. Our framework, in the form of a retrospective analysis, will help to streamline the planning process for Fraser River sockeye salmon fisheries with respect to selecting MA models and will also quantify the influence of competing performance indicators on model choice. Specifically, we examined the efficacy of five MA models, plus a model-combining technique, using five performance measures that reflect different management objectives.

Methods

Data

Fisheries and Oceans Canada provided historical spawning escapement estimates for sockeye salmon (T. Cone, personal communication, DFO Stock Assessment, Annacis Island, BC), and the PSC provided sockeye salmon abundance estimates at Mission and estimates of sockeye salmon catch upriver of Mission. Spawning ground abundance estimates were obtained through a variety of methods outlined in Schubert (2007), and Mission abundance and run-timing estimates were obtained using hydroacoustic sonar (Xie and Hsieh 1989; Xie 2000). Lower Fraser River temperatures (near Qualark, 165 km upstream from the river delta) were provided by the DFO Environmental Watch Program (Patterson et al. 2007a), and flow data (near Hope, 150 km upstream from the river delta) were from Environment Canada's Water Survey of Canada (http://www.wateroffice.ec.gc.ca/index_html).

MA Models

Because there is no long-term record of directly estimated in-river mortality for upstream-migrating Fraser River sockeye salmon, we indexed in-river loss using the ratio of estimates of up-river spawning escapement abundance (S) to lower-river potential spawning escapement abundance estimates (S_p). The latter, potential spawning escapement, estimates the number of fish escaping from lower-river fisheries by subtracting estimates of catch above Mission from the Mission escapement estimates.

The closer the ratio of S to S_p is to one, the smaller the associated discrepancy or in-river loss. In the MA models, the index of the escapement discrepancy response variable is $\log_e(D)$, where $D = S/S_p$ (Hague and Patterson 2007; Macdonald et al. 2010).

Our study evaluated MA models that contained four predictor variables: (1) Fraser River temperature (T) in degrees Celsius measured at Qualark, British Columbia, (2) Fraser River flow (Q) in cubic meters per second measured at Hope, British Columbia, (3) migration timing (R) in terms of the Hells Gate 50% date (the date by which 50% of the run-timing group has passed Hells Gate, 35 km upstream of Qualark), and (4) the average of the historically observed discrepancies ($\overline{D}$) (Macdonald et al. 2010). Five MA models were evaluated and are denoted using their respective predictor variables (Equations 1 – 5 in Table 1): (1) temperature-only (T), (2) discharge-only (Q), (3) temperature and discharge (T+Q), (4) up-stream migration timing of the run (R), (5) average historical escapement discrepancy ($\overline{\mathrm{D}}$). These models were all compared to each other and to the outcome from applying no management adjustment (NMA; i.e., the forecasted escapement discrepancy, $\log_e\left(\hat{D}\right)$, is 0, Eq. 6 in Table 1). Parameters for each run-timing group were estimated for each of the six candidate models by fitting time-series of $\log_e(D)$ to time-series of environmental and run-timing conditions using equations in Table 1.

Retrospective Analysis

Retrospective predictions were made for 1995 – 2007. For example, a and b parameters for a given model in Table 1 were initially estimated using data from 1977 - 1994 and the resulting model was then used to forecast $\log_e(D)$ in 1995. The observed

$\log_e(D)$ was later compared with that forecasted value. In the next iteration, the observed 1995 environmental and escapement discrepancy data were added to the time series, model parameters were re-fit, and $\log_e(D)$ was forecasted for 1996. These iterations were repeated for all remaining years of available data (up through 2007) and for all run-timing groups. The degree to which each model could correctly forecast the observed $\log_e(D)$ over the entire time series was then calculated using performance measures as described below. This process was repeated for each of the six management adjustment models and each run-timing group. Due to logistical problems with Mission escapement estimates for several years (Macdonald et al. 2010), Late-run MA models were initialized using data from 1977 – 1999, and were evaluated from 2000 – 2007. Models were fit using the linear modeling function lm() in the statistical software package R, version 2.6.0 (R Development Core Team 2009).

Model performance measures. -- As defined with equations shown later, we used five performance measures to rank the suite of MA models listed above: (1) mean raw error (MRE) in forecasts of $\log_e(D)$, (2) mean absolute error (MAE), (3) root mean square error (RMSE), (4) mean small-sample size Akaike information criterion (AIC_c), and (5) mean adjusted R^2 (R^2). These measures were selected to provide an assessment of model forecast skill (i.e., model bias (MRE) and accuracy (MAE, RMSE)), and hindcast skill (i.e., model fit (R^2, AIC_c)) (Burnham and Anderson 2002; Willmott and Matsuura 2005). These measures have been used in previous studies to evaluate performance of pre-season abundance forecasting models for sockeye, chum (*O. keta*), and pink (*O. gorbuscha*) salmon (Wood et al. 1997; Haeseker et al. 2005; Haeseker et al. 2008).

To facilitate interpretation of results, we converted model error, the difference between the predicted and observed values of $\log_e(D)$ based on equations 1-6 in Table 1, to "raw error" (E) measured on a linear scale using:

$$(7)\ E_{n,i} = \hat{D}_{n,i} - D_n$$

where $E_{n,i}$ is the raw error in year n of model i, $\hat{D}_{n,i}$ is the forecasted discrepancy, and D_n is the observed discrepancy in year n. By converting discrepancy to the linear scale, a positive error has the same absolute value as a negative error of the same magnitude. Thus, $E_{n,i}$ is a unitless measure of the extent to which the forecasted ratio of spawning ground abundance to Mission abundance reflects the actual ratio of S/S_p realized at the end of the season. Across all years, E_i measures the bias of a given MA model, i.

The three performance measures (MRE, MAE, RMSE) are all derived from annual E values. First, the MRE is the average bias for each model:

$$(8)\ MRE_i = \frac{\sum_{n=1}^{N} E_{n,i}}{N}$$

where MRE_i is the mean raw error for MA model i across all N evaluated years starting with $n = 1$ for 1995 (except for Late-run models, where it starts in 2000), and i corresponds to MA models 1-6 in Table 1.

An unbiased model (MRE = 0), in which positive raw errors exactly offset negative E over the entire time series, provides no indication of forecast precision.

Therefore, MAE (Eq. 9) and RMSE (Eq. 10) were also calculated to reflect the average magnitude of MA model residuals. Values of MAE or RMSE approaching zero are considered optimal. The MAE is the average absolute magnitude of MA model error, regardless of sign:

$$(9)\quad MAE_i = \frac{\sum_{n=1}^{N} \left| E_{n,i} \right|}{N}$$

The RMSE weights large errors more heavily than MAE. The model with the smallest RMSE results in the lowest variance in residuals:

$$(10)\quad RMSE_i = \sqrt{\frac{\sum_{n=1}^{N} \left(E_{n,i} \right)^2}{N}}$$

Finally, we calculated two measures of goodness of fit to assess how well models fit observed data: adjusted R^2 (Zar 2006) and AIC_c (Burnham and Anderson 2002). Because models were refit for each iteration of the retrospective analysis, a mean adjusted R^2 and a mean AIC_c across years were used for retrospective evaluation of each model *i*:

(11) $$R_i^2 = \frac{\sum_{n=1}^{N} R_{n,i}^2}{N}$$

(12) $$AIC_{c_i} = \frac{\sum_{n=1}^{N} AIC_{c_{n,i}}}{N}$$

In addition to AIC_c, for each model we report the number of parameters (K), difference between the AIC_c of a given model and that of the best model (Δ AIC_c), and AIC_c weight (*w*, the relative degree of support assigned to an individual model within each model set, calculated from the standard formula in Burnham and Anderson 2002). Models with R^2 and *w* values closest to one were considered top-ranked for these performance measures. We also calculated AIC_c fit to the entire dataset (including 2007), because this is the more commonly applied use of the statistic.

On the basis of these performance measures, a rank was given to each MA model for each run-timing group and each measure, where 1 = "best" and 6 = "worst". In addition, we averaged the ranks of a given model for all five performance measures for each sockeye salmon run-timing group to calculate an average rank as a measure of overall model performance.

Model Averaging

It can be useful to combine forecasting models to make a single, potentially more precise and less biased prediction by using all of the information contained in various candidate models (Link and Barker 2006). Thus, in addition to a retrospective analysis of individual MA models, we also explored the viability of applying a model-combining procedure. Model averaging has been proposed as a means of setting rebuilding targets for New England groundfish stocks (Brodziak and Legault 2005), estimating vessel impacts on Mississippi River fisheries (Gutreuter et al. 2006), and making hydrological predictions (Duan et al. 2007).

One option for combining models is to use model averaging based on information theoretic criteria (Burnham and Anderson 2004; Brodziak and Legault 2005; Gutreuter et al. 2006). We therefore also weighted annual forecasts produced by each model by the retrospective annual AIC_c weights (Eq. 14) to produce a single combined MA forecast for each year.

$$(14)\quad \log_e\left(\hat{D}_w\right)_n = \sum_{i=1}^{6}\left[w_{n,i} \bullet \log_e\left(D\right)_{n,i}\right]$$

where $w_{n,i}$ is the AIC_c weight in year n for model i (summing to 1 over all models), and $\hat{D}_W$ is the new forecasted weighted discrepancy in year n for model i.

The AIC_c-weighted models were evaluated using MRE, MAE, and RMSE and were then ranked against the six individual models for each run-timing group from the retrospective analysis using each performance measure. Here, model ranks ranged from 1 (best) to 7 (worst) for each performance measure.

Jack-knife analysis

We also conducted a jack-knife analysis (Shao and Dongsheng 1995) to determine the sensitivity of model rankings and performance to removal of each year's forecast. Raw errors from single years of the 13-year retrospective evaluation period (8 years for the Late run-timing group) were sequentially removed with subsequent replacement, and performance measures were re-estimated, eventually producing 13 (8 for Late-run) replicates of model ranking. We then compared the top-ranked model for each performance measure from the retrospective analysis to the top-ranked model for each performance measure from each jack-knife replicate, recording the number of jack-knife replicates that selected a different top-ranked model.

Results

Retrospective Model Performance

Interannual variability in retrospective forecasts of escapement discrepancies for each MA model differed from the observed discrepancies for each sockeye salmon run-timing group (Figure 2). For the Early Stuart and Summer run-timing groups, the historic model ($\overline{\mathrm{D}}$) was least biased (see MRE in Figure 3), but did a poor job of tracking interannual variability (Figure 2). That is, the historic model, $\overline{\mathrm{D}}$, tended to underestimate the loss when it was large and overestimate it when it was small. In contrast, the environmental models, temperature (T), discharge (Q), and temperature-plus-discharge (T+Q), displayed temporal variation more similar to the observed discrepancies (Figure 2), but produced a positive bias on average (i.e., underestimated the true discrepancies by forecasting $\hat{D}$ to be closer to 1 than the realized D at the end of the season) (Figure 2 and MRE in Figure 3).

Examination of values of each performance measure provides additional insight into the magnitude of differences across models within each run-timing group (Figure 3). In many instances, differences between model ranks were due to only minor disparities in actual values of performance measures (Figure 3). For example, for the Early Stuart, the difference between the second-worst and best-ranked MA models was less than 0.04 using MAE.

There was considerable variation across adjusted R^2 values for *top-ranked* models in each run-timing group, ranging from a mean R^2 of 84% for the R model for the Late run-timing group to a mean R^2 of only 22% for the $\overline{D}$ model for the Early Summer run-timing group (Figure 3). Interestingly, the models with the highest mean R^2 value and/or highest AIC_c weight (Figure 3; Table 2) often ranked poorly based on MRE, MAE, or RMSE performance measures (Figure 3). Due to trends in model performance across the years evaluated for the retrospective analysis for Early Stuart sockeye, the top-ranked model using the mean retrospective AIC_c was different than the top-ranked model using the AIC_c fit to the entire data set (D versus T+Q) (Table 2).

Based on the average model rank across performance measures, no single model performed best across all run-timing groups (Figure 4). However, one clear result was that failure to apply a management adjustment, the "No MA" (NMA) model, had the worst average rank (lowest mean rank) in three of the four run-timing groups, and the second-worst rank in the fourth group, the Summer run (Figure 4). This result is also evident from individual performance measures shown by model (Figure 3). The $\overline{D}$ model ranked best for the Early Stuart run, the $\overline{D}$ and T+Q models tied for highest rank for the Early Summer run, the T and T+Q models tied for best for the Summer run, and the Q model was best for the Late run (Figure 4).

For the Early-Summer, Summer, and Late run-timing groups, the top-ranked model varied across performance measures (Figure 5B, C, D). In contrast, the top-ranked model was consistent across most performance measures for the Early Stuart run-timing group (Figure 5A). For example, for the Early Summer run (Figure 5B), the $\overline{\mathrm{D}}$ model best explained the observed variance (R^2 and AIC_c), but the T+Q model ranked best for MAE and RMSE, and the Q model ranked first for MRE. For the Summer run-timing group (Figure 5C), the T+Q model had the best MAE and RMSE ranks. However, likely because of the T+Q model's additional parameters relative to the T model, the T model performed best for adjusted R^2 and AIC_c. The Q model in the Late run-timing group ranked best for the hindcast performance measures (MRE, MAE, RMSE), but the R model ranked best using forecast criteria (R^2 and AIC_c) (Figure 5D). In contrast, the Early Stuart $\overline{\mathrm{D}}$ model consistently ranked first for four performance measures (MRE, MAE, RMSE, AIC_c), while the T model consistently ranked second (Figure 5A).

There was considerable interannual variability in model performance based on the frequency (number of years) that an individual model ranked first in terms of yearly raw error for a given run-timing group (*E* from Eq. 7) (Figure 6). Using the Early Stuart run-timing group as an example, although the $\overline{\mathrm{D}}$ model ranked best overall (average rank = 1.4; Figures 4 and 5A), it produced the smallest raw error in only 2 of the 13 years of the retrospective evaluation (Figure 6A). In contrast, the T+Q model, which was ranked third overall (average rank = 3.4; Figure 4), had the smallest raw error in 4 of 13 years (Figure 6A).

AIC_c-Weighted Model

None of the AIC_c weighted average models ranked higher than any of the other candidate models when rank was averaged across the MRE, MAE and RMSE performance measures. The AIC_c-weighted MA model was ranked 3^{rd} for the Early Summer and Late run-timing group, 4^{th} for Early Stuart group, and 5^{th} for the Summer group. The AIC_c-weighted model was never top-ranked for any performance measure.

Jack-knife analysis

The jack-knife analysis showed that overall model performance based on our retrospective analysis was relatively insensitive to variability in year-to-year performance (Table 3). The top-ranked models across each performance measure remained generally consistent for the Early Summer, Summer, and Late run-timing groups. In the one group in which the top-ranked model was sensitive to removal of particular years of data (Early Stuart), the MAE and RMSE performance measures selected different top-ranked models in 46% and 54% of the jack-knife replicates, respectively. This sensitivity of ranks of Early Stuart MA models to individual years of data was likely due to the similarity in values of MAE and RMSE among five top-ranked models for this group, as shown in Figure 3.

Discussion

Our retrospective analysis provides a framework for evaluating alternative forecasting models across a range of hindcasting and forecasting performance measures. Because model rankings sometimes varied considerably as a function of the performance measures selected, our results emphasize the importance of carefully choosing the measures to be used in model selection. Model performance measures should not be chosen simply on the basis of statistical tradition, but instead should be consistent with the stated management objectives. For example, use of model rankings based only on

AIC_c or R^2 fit to the entire dataset (as is often the case) for management of the Early Summer run would result in the selection of the $\overline{D}$ model, (i.e., the historical average discrepancy model). However, for managers who place high priority on objectives that specifically aim to avoid extreme errors in achieving escapement targets, a model that minimizes MAE or RMSE, i.e., the T+Q (temperature and flow) model, would be preferred. We emphasize that both scientists and managers should carefully work together to determine which performance measures should be used for model selection. Even the apparently subtle difference between using a measure of long-term bias (MRE) to rank models instead of a measure of year-to-year deviation (MAE) can lead to different model choices.

Although the best management-adjustment (MA) models varied among run-timing groups and performance measures, one key finding from our study is the strong evidence that MA forecasts made from some combination of environmental or biological data out-performed the approach of applying no adjustment at all; i.e., the "No MA" model was consistently ranked low across all run-timing groups. This result further validates the decision to apply environmentally-based MA forecasts to inform the management of Fraser River sockeye salmon fisheries. This result is also consistent with the biological rationale linking escapement discrepancies to river conditions that contribute to natural mortality along the freshwater migration route (Macdonald et al. 2010).

In the model selection field, where model rank is sensitive to interannual variation and/or where there are multiple competing candidate models, a weighted-average approach, which combines forecasts from the entire model suite, is thought to yield better results (Raftery and Zheng 2003). Such model averaging minimizes the influence of any individual forecast, and therefore reduces the effect of large rare errors associated with a single model by averaging errors across multiple forecasts; this could reduce the RMSE (Raftery and Zheng 2003). In other words, AIC_c-weighted models may tend to be more robust to interannual variability. However, in the case of the MA models for Fraser River sockeye salmon, model averaging rarely improved model performance assessed using forecast criteria (i.e. MRE, MAE, RMSE).

This result emphasizes the differences in model selection that can arise from using hindcast vs. forecast performance measures. An alternative model-combining technique that we did not explore, which includes a moving time window, may yield improved results over the AIC_c weighted averaging. Such approaches are often used in meteorological forecasting by selecting models that perform best based on observed climatic conditions over a specified period (e.g., Eckel and Mass 2005). This approach would also be suitable if there are observed time trends in model performance. While our jack-knife analysis demonstrated that MA model selection was relatively insensitive to individual years used in the analysis, this does not discount the possibility of non-stationarity in model performance over time.

Additional improvements in forecasts could come from examining models beyond the suite of models considered in this analysis and from examining different approaches to quantifying uncertainties in forecast errors. For example, our treatment of forecast errors assumes that outcome uncertainty, which reflects the degree to which spawning targets are achieved in practice (Holt and Peterman 2006), is independent of forecasts by MA models. However, if instead the sign and magnitude of deviation from such targets is correlated with forecasting error, then a future model that accounted for such an association could be an improvement. There is also a potentially wide range of additional MA models that could explain some of the historical variation in past escapement discrepancies or better account for changing discrepancies that result from different future environmental conditions. For instance, new MA models might have different structural forms, environmental predictor variables, and/or explicit consideration of observation errors in the latter variables. Furthermore, the absolute and relative performance of the existing suite of models may also be subject to change over time. For example, summer river temperatures have been increasing since the 1950s (Patterson et al. 2007b) and the temperature and flow regimes of the Fraser River are expected to continue to change with warming climate (Morrison et al. 2002; Ferrari et al. 2007). Changes in MA model fits and their predictive power may occur as a result of these shifting environmental baselines. One benefit of the retrospective evaluation approach used here is that as scientists or managers identify new modeling approaches, an expanded and/or updated list of models can easily be evaluated using this same framework.

Quantifying uncertainty in Mission escapement estimates is another critical area for research, and salmon biologists have attempted to estimate it in various ways for years, but without success. Nevertheless, by definition, the escapement discrepancies forecasted by the MA models inherently include uncertainties in lower- and upper-river escapement estimates. However, because these measurement errors have not been explicitly quantified (and likely vary in size and direction due to differences in run size, environmental conditions, and survey techniques applied to different systems), managers have accepted current discrepancy estimates as "true" for the purpose of managing the fishery. This is why we have clearly referred to the MA forecasts as "escapement discrepancies" rather than "in-river mortality".

Many studies use only one or two performance measures to rank models (e.g., Willmott 1982; Willmott and Matsuura 2005). More specifically, most studies in ecology, including past studies of MA models for Fraser River sockeye salmon, select models from a large set based only on R^2 and/or AIC performance criteria (Keefer et al. 2008; Buhle et al. 2009; McGowan and Ryan 2009; Macdonald et al. 2010). Such studies are a useful first step in the identification of candidate models reflecting biological relevance and good agreement with historical data, but as we show here, this practice only captures certain characteristics of model performance that are relevant to fisheries management choices. Clearly, multiple performance measures need to be considered in fisheries analyses (e.g., Haeseker et al. 2005; Adkison 2009) because of the competing management objectives typically faced by fisheries managers (de la Mare 1998; Hilborn 2007). When clear objectives are combined with appropriate affiliated performance measures, model selection through retrospective analysis can be used to provide scientific advice to managers to help increase the probability of achieving fishery management objectives. Future research should explore methods for also incorporating into the model selection process available information from multiple, unequally weighted management objectives and utility functions (Keeney 1977).

Acknowledgments

Thanks to Duncan Knowler for feedback on this manuscript. Insightful suggestions and feedback on preliminary analyses were provided by Steve Macdonald

(Fisheries and Oceans Canada), Ian Guthrie (Pacific Salmon Commission, or PSC) and Mike Lapointe (PSC). Support for this work was provided by the Southern Boundary Restoration and Enhancement Fund of the PSC, Fisheries and Oceans Canada's Fraser River Environmental Watch program, and the Natural Sciences and Engineering Research Council of Canada (the latter provided to Randall M. Peterman).

References

Adkison, M. D. 2009. Drawbacks of complex models in frequentist and Bayesian approaches to natural-resource management. Ecological Applications 19(1):198-205.

Brodziak, J., and C. M. Legault. 2005. Model averaging to estimate rebuilding targets for overfished stocks. Canadian Journal of Fisheries and Aquatic Sciences 62(3):544-562.

Buhle, E. R., K. K. Holsman, M. D. Scheuerell, and A. Albaugh. 2009. Using an unplanned experiment to evaluate the effects of hatcheries and environmental variation on threatened populations of wild salmon. Biological Conservation 142(11):2449-2455.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference, Second edition. Springer-Verlag, New York.

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference - understanding AIC and BIC in model selection. Sociological Methods and Research 33(2):261-304.

Cooke, S. J., S. G. Hinch, A. P. Farrell, M. Lapointe, S. R. M. Jones, J. S. Macdonald, D. A. Patterson, M. C. Healey, and G. Van Der Kraak. 2004. Abnormal migration timing and high en route mortality of sockeye salmon in the Fraser River, British Columbia. Fisheries 29(2):22-33.

de la Mare, W. K. 1998. Tidier fisheries management requires a new MOP (management-oriented paradigm). Reviews in Fish Biology and Fisheries 8(3):349-356.

Duan, Q. Y., N. K. Ajami, X. G. Gao, and S. Sorooshian. 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. Advances in Water Resources 30(5):1371-1386.

Eckel, F. A., and C. F. Mass. 2005. Aspects of effective mesoscale, short-range ensemble forecasting. Weather and Forecasting 20(3):328-350.

Ferrari, M. R., J. R. Miller, and G. L. Russell. 2007. Modeling changes in summer temperature of the Fraser River during the next century. Journal of Hydrology 342(3-4):336-346.

Gable, J., and S. Cox-Rogers. 1993. Stock identification of Fraser River sockeye salmon : methodology and management application. Pacific Salmon Commission Technical Report No. 5:36 p.

Gelman, A. and J. Hill. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge, UK.

Gutreuter, S., J. M. Vallazza, and B. C. Knights. 2006. Persistent disturbance by commercial navigation alters the relative abundance of channel-dwelling fishes in a large river. Canadian Journal of Fisheries and Aquatic Sciences 63(11):2418-2433.

Haeseker, S. L., R. A. Peterman, and Z. M. Su. 2008. Retrospective evaluation of preseason forecasting models for sockeye and chum salmon. North American Journal of Fisheries Management 28(1):12-29.

Haeseker, S. L., R. M. Peterman, Z. M. Su, and C. C. Wood. 2005. Retrospective evaluation of preseason forecasting models for pink salmon. North American Journal of Fisheries Management 25(3):897-918.

Hague, M. J., and D. A. Patterson. 2007. Quantifying the sensitivity of Fraser River sockeye salmon (*Oncorhynchus nerka*) management adjustment models to uncertainties in run timing, run shape and run profile. Canadian Technical Report of Fisheries and Aquatic Sciences 2776:1-55 + vii.

Hanson, P. J., D. S. Vaughan, and S. Narayan. 2006. Forecasting annual harvests of Atlantic and Gulf Menhaden. North American Journal of Fisheries Management 26(3):753-764.

Hilborn, R. 2007. Defining success in fisheries and conflicts in objectives. Marine Policy 31(2):153-158.

Holt, C. A., and R. M. Peterman. 2004. Long-term trends in age-specific recruitment of sockeye salmon (*Oncorhynchus nerka*) in a changing environment. Canadian Journal of Fisheries and Aquatic Sciences 61(12):2455-2470.

Holt, C. A., and R. M. Peterman. 2006. Missing the target: uncertainties in achieving management goals in fisheries on Fraser River, British Columbia, sockeye salmon (*Oncorhynchus nerka*). Canadian Journal of Fisheries and Aquatic Sciences 63(12):2722-2733.

Holt, C. A., and R. M. Peterman. 2008. Uncertainties in population dynamics and outcomes of regulations in sockeye salmon (*Oncorhynchus nerka*) fisheries: implications for management. Canadian Journal of Fisheries and Aquatic Sciences 65(7):1459-1474.

Keefer, M. L., C. A. Peery, and C. C. Caudill. 2008. Migration timing of Columbia River spring Chinook salmon: effects of temperature, river discharge, and the ocean environment. Transactions of the American Fisheries Society 137(4):1120-1133.

Keeney, R.L. 1977. A utility function for examining policy affecting salmon on the Skeena River. Journal of Fisheries Research Board of Canada 34(1):49-63.

Link, W. A., and R. J. Barker. 2006. Model weights and the foundations of multimodel inference. Ecology 87(10):2626-2635.

Macdonald, J. S., D. A. Patterson, M. J. Hague, and I. C. Guthrie. 2010. Modeling the influence of environmental factors on spawning migration mortality for sockeye salmon fisheries management in the Fraser River, British Columbia. Transactions of the American Fisheries Society 139:768-782.

McGowan, C. P., and M. R. Ryan. 2009. A quantitative framework to evaluate incidental take and endangered species population viability. Biological Conservation 142(12):3128-3136.

Morrison, J., M. C. Quick, and M. G. G. Foreman. 2002. Climate change in the Fraser River watershed: flow and temperature projections. Journal of Hydrology 263(1-4):230-244.

Patterson, D. A., J. S. Macdonald, K. M. Skibo, D. P. Barnes, I. Guthrie, and J. Hills. 2007a. Reconstructing the summer thermal history for the lower Fraser River, 1941 to 2006, and implications for adult sockeye salmon (*Oncorhynchus nerka*) spawning migration. Canadian Technical Report of Fisheries and Aquatic Sciences 2724:1-43.

Patterson, D. A., K. M. Skibo, D. P. Barnes, J. A. Hills, and J. S. Macdonald. 2007b. The influence of water temperature on time to surface for adult sockeye salmon carcasses and the limitations in estimating salmon carcasses in the Fraser River, British Columbia. North American Journal of Fisheries Management 27(3):878-884.

Peters, C. N., D. R. Marmorek, and R. B. Deriso. 2001. Application of decision analysis to evaluate recovery actions for threatened Snake River fall chinook salmon (*Oncorhynchus tshawytscha*). Canadian Journal of Fisheries and Aquatic Sciences 58(12):2447-2458.

Peterman R.M., D. Marmorek, B. Beckman, M. Bradford, N. Mantua, B.E. Riddell, M. Scheuerell, M. Staley, K. Wieckowski, J.R. Winton, C.C. Wood. 2010. Synthesis of evidence from a workshop on the decline of Fraser River sockeye. June 15-17, 2010. A Report to the Pacific Salmon Commission, Vancouver, B.C., 123 pp. + 35 pp. of appendices. 31 August 2010. Available from www.psc.org.

R Development Core Team. 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raftery, A. E., and Y. Y. Zheng. 2003. Discussion: performance of Bayesian model averaging. Journal of the American Statistical Association 98(464):931-938.

Schubert, N. D. 2007. Estimating the 1995 Fraser River sockeye salmon (*Oncorhynchus nerka*) escapement. Canadian Technical Report of Fisheries and Aquatic Sciences 2737: ix + 71 p.

Shao, J. 1993. Linear model selection by cross-validation. Journal of the American Statistical Association 88(422):486-494.

Shao, J., and T. Dongsheng. 1995. The jackknife and bootstrap. Springer-Verlag, New York.

Willmott, C. J. 1982. Some comments on the evaluation of model performance. Bulletin of the American Meteorological Society 63(11):1309-1313.

Willmott, C. J., and K. Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research 30(1):79-82.

Wood, C. C., D. T. Rutherford, D. Peacock, S. Cox-Rogers, and L. Jantz. 1997. Assessment of recruitment forecasting methods for major sockeye and pink salmon stocks in northern British Columbia. Canadian Technical Report of Fisheries and Aquatic Sciences 2187:85 p.

Xie, L. S., and W. W. Hsieh. 1989. Predicting the return migration routes of the Fraser River sockeye salmon (*Oncorhynchus nerka*). Canadian Journal of Fisheries and Aquatic Sciences 46(8):1287-1292.

Xie, Y. B. 2000. A range-dependent echo-association algorithm and its application in split-beam sonar tracking of migratory salmon in the Fraser River watershed. IEEE Journal of Oceanic Engineering 25(3):387-398.

Zar, J. H. 2006. Biostatistical analysis 3rd edition. Prentice-Hall, New Jersey.

Table 1. Description of predictor variables and structures of management adjustment models that were used to produce the forecast of logged escapement discrepancies ($\log_e\left(\hat{D}\right)$) for each major Fraser River sockeye salmon run-timing group. The rationale behind variable selection and structure of these models is described in detail in Macdonald et al. (2010). Environmental variables for each model are used as an abbreviation for the full model name within the text.

Environmental Variable	**Description**	**MA model equation**
T	Mean 31-day water temperature (°C) measured near Qualark, BC centered on the date that 50% of a run-timing group has passed Hells Gate, BC.	(1) $a+b_1T+b_2T^2$
Q	Mean 31-day water discharge (m^3s^{-1}) measured near Hope, BC centered on the date that 50% of a run-timing group has passed Hells Gate, BC.	(2) $a+b_1Q+b_2Q^2$
T + Q	Combined 31-day temperature and discharge.	(3) $a+b_1T+b_2T^2+b_3Q+b_4Q^2$
R	Julian date at which 50% of a run-timing group has passed Hells Gate, BC.	(4) $a+b_1R$
$\overline{D}$	The average historical escapement discrepancy.	(5) $\frac{\sum_{y=1977}^{Y} D_y}{Y}$
NMA	No management adjustment, i.e., zero escapement discrepancy.	(6) 0

Table 2. AIC_c values and their components for each run-timing-specific MA model averaged over all years of the retrospective analysis. K = number of parameters; AIC_c = mean small-sample Akaike information criterion; Δ AIC_c = mean delta AIC_c; *w* = mean AIC_c weight; AIC_{cALL} = AIC_c fit to all years, including 2007.

Run group	Model	K	AIC_c	Δ AIC_c	*w*	AIC_{cALL}
Early Stuart	T	4	39.01	0.49	0.33	52.21
	Q	4	42.84	4.31	0.05	57.97
	T+Q	6	40.85	2.33	0.13	51.27
	R	3	42.71	4.19	0.05	58.72
	$\overline{D}$	3	38.52	0.00	0.43	52.59
	NMA	1	47.69	9.16	0	71.05
Early Summer	T	4	28.57	6.10	0.04	41.02
	Q	4	27.50	5.03	0.07	37.41
	T+Q	6	30.64	8.16	0.01	35.99
	R	3	29.12	6.65	0.04	43.99
	$\overline{D}$	3	22.47	0.00	0.85	35.49
	NMA	1	38.89	16.42	0	60.48
Summer	T	4	9.80	0.00	0.72	8.35
	Q	4	19.02	9.22	0.01	26.71
	T+Q	6	15.92	6.12	0.03	11.49

	R	3	17.95	8.15	0.01	29.91
	$\overline{D}$	3	13.75	3.96	0.10	22.59
	NMA	1	13.17	3.37	0.13	26.09
Late	T	4	41.34	24.36	0	41.34
	Q	4	26.62	9.64	0.12	26.62
	T+Q	6	45.23	28.25	0.03	45.23
	R	3	17.11	0.13	0.85	17.12
	$\overline{D}$	3	31.65	14.66	0.01	31.65
	NMA	1	40.28	23.29	0	40.28

Table 3. Percentage of years from the jack-knife analysis in which a model other than the top-ranked model from the retrospective analysis was ranked first, by performance measure and by run-timing group of Fraser River sockeye salmon.

Performance measure	**Early Stuart**	**Early Summer**	**Summer**	**Late**
MRE	15	23	17	13
MAE	46	8	0	0
RMSE	54	23	8	13

Figure captions

Figure 1. Map of the Fraser River watershed in British Columbia, Canada. Locations of key spawning areas for each sockeye salmon run-timing group are indicated, along with the Mission sampling location in the lower river.

Figure 2. Time-series of observed (D, bold solid line) and retrospectively forecasted ($\overline{D}$) escapement discrepancies for each Fraser River sockeye salmon run-timing group, by model. Table 1 defines the symbols for the five management adjustment models that were used to create these forecast time series. $D = 1$ indicates no in-river loss of upstream-migrating adult sockeye salmon, whereas $D = 0$ indicates 100% in-river loss.

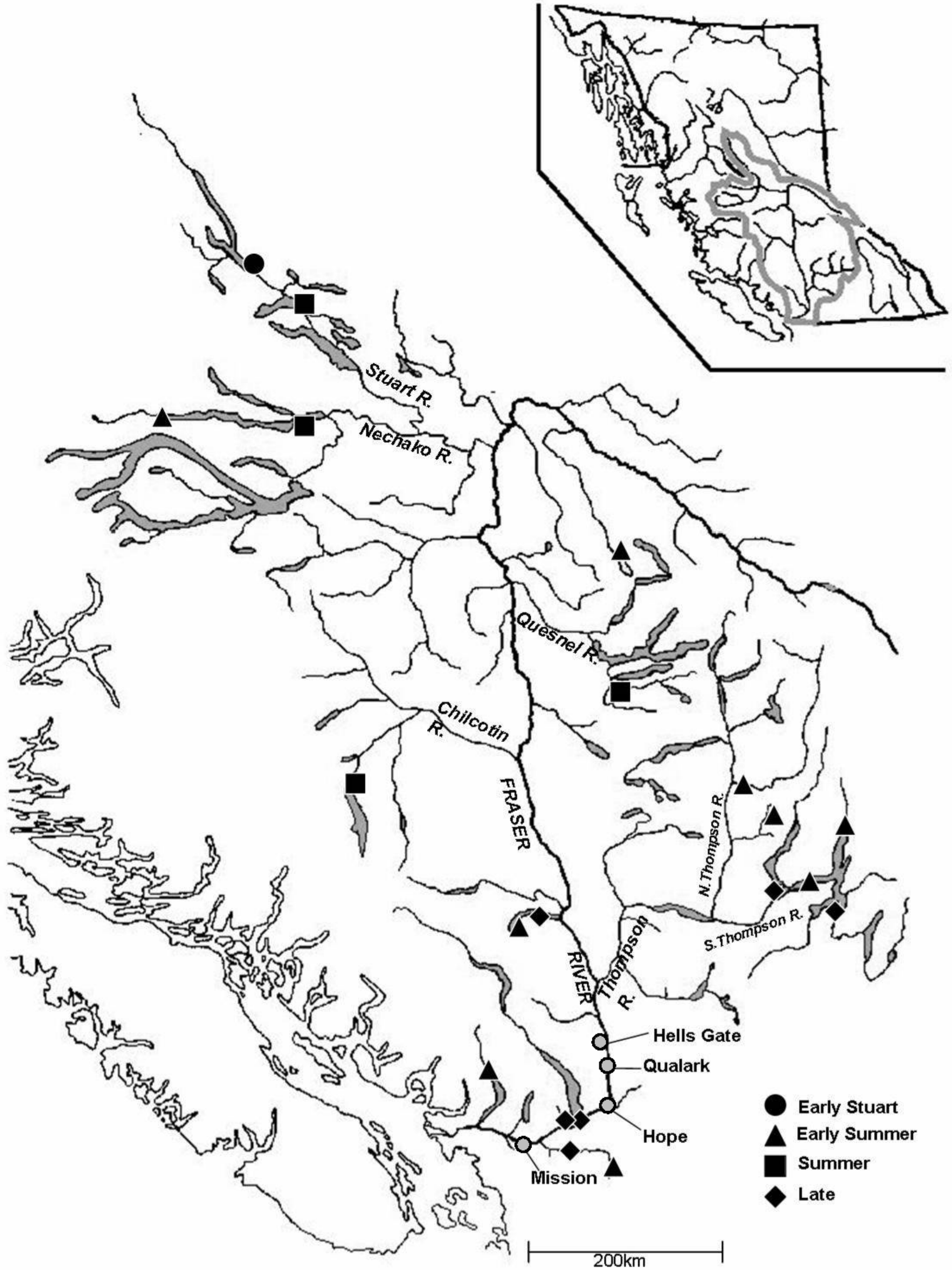
Figure 3. Values of performance measures (X axis) for each run-timing group (A-D) for each management adjustment model defined in the legend and Table 1, averaged over the 13-year retrospective evaluation period. Table 1 describes each management adjustment model used to create the forecasts. MRE = mean raw error, MAE = mean absolute error, RMSE = root mean square error, R^2 is the coefficient of determination, and w is the AIC_c weight. Negative adjusted R^2 values can result from models that fit the data so poorly that, "on average, the residual error variance is larger than the variance of the data" (Gelman and Hill 2007).

Figure 4. Mean rank of management adjustment models for each run-timing group, averaged over individual rankings calculated for each of the five performance measures. Best = 1, worst = 6. Models are defined in Table 1.

Figure 5. Ranks of management adjustment models within each run-timing group for each performance measure. Best = 1, worst = 6. These ranks were averaged across all performance measures to produce the average ranks in Figure 4.

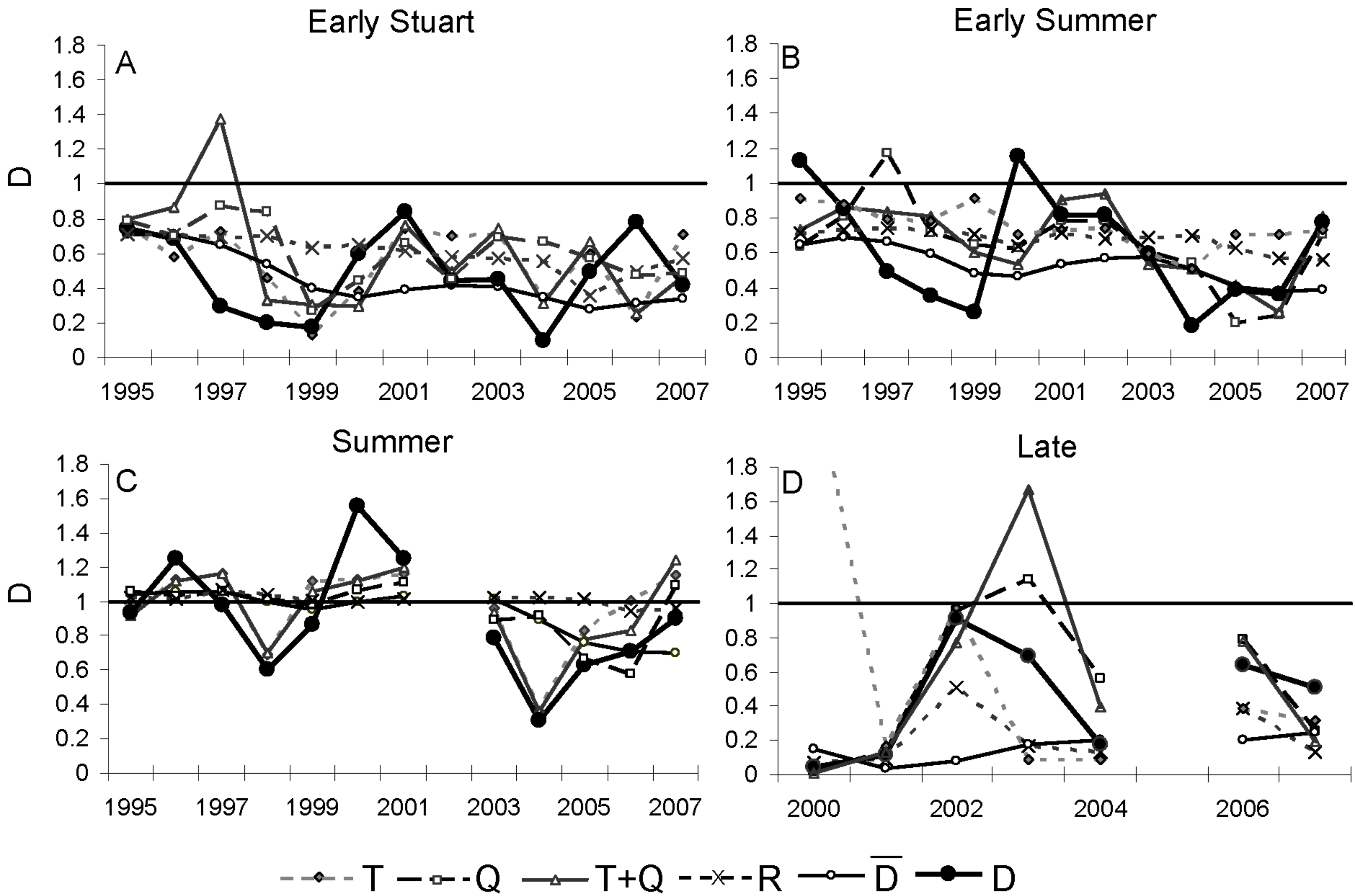
Figure 6. Frequency of best performance, shown as the number of years for which each management adjustment model was best, i.e., produced the smallest absolute raw error (E, equation 7) relative to the E of other models. For comparison, the number above each bar is the average model rank from Figure 4, with best = 1 and worst = 6.

Fig. 1

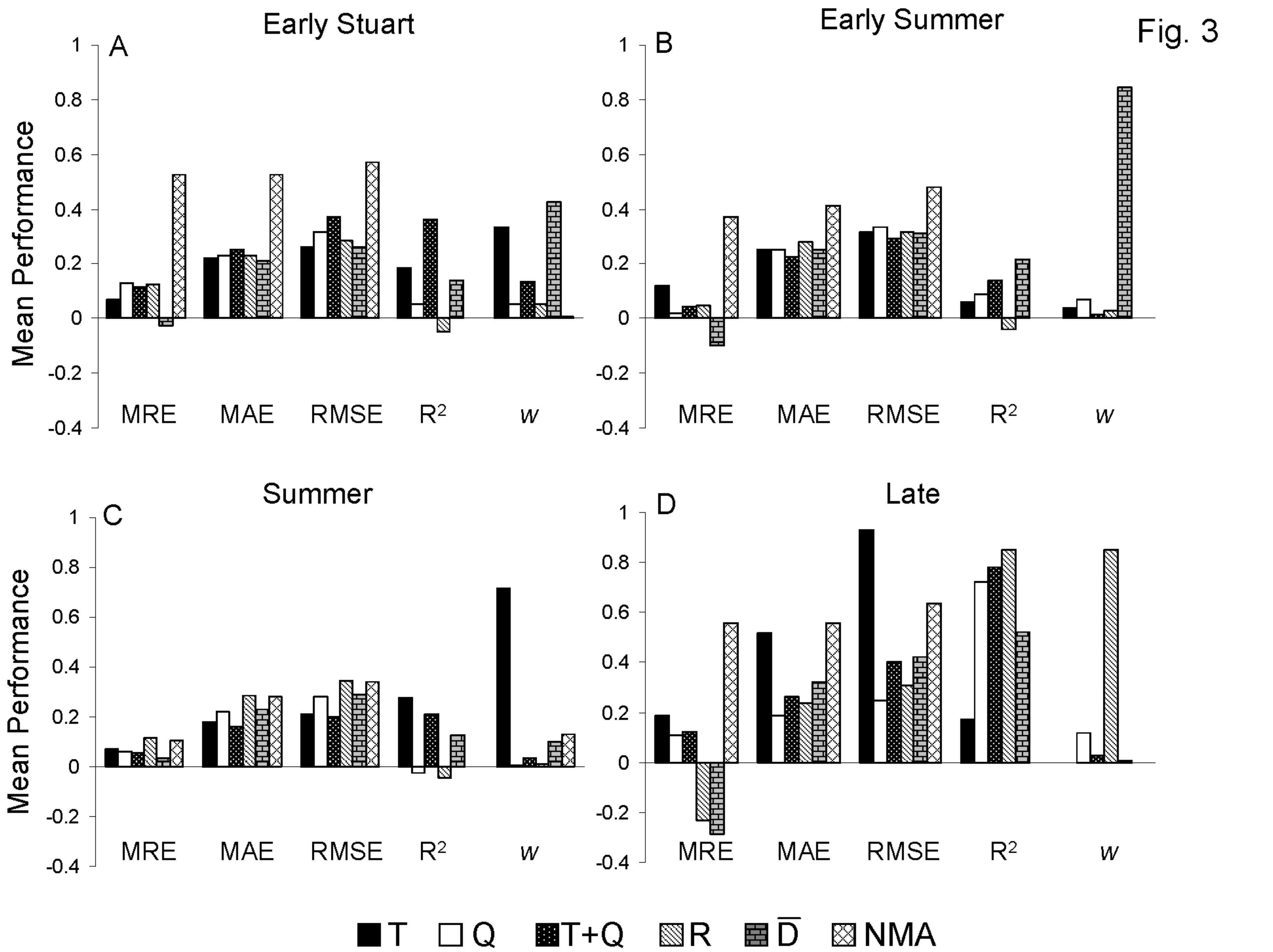


CAN285112_0034

Fig. 2

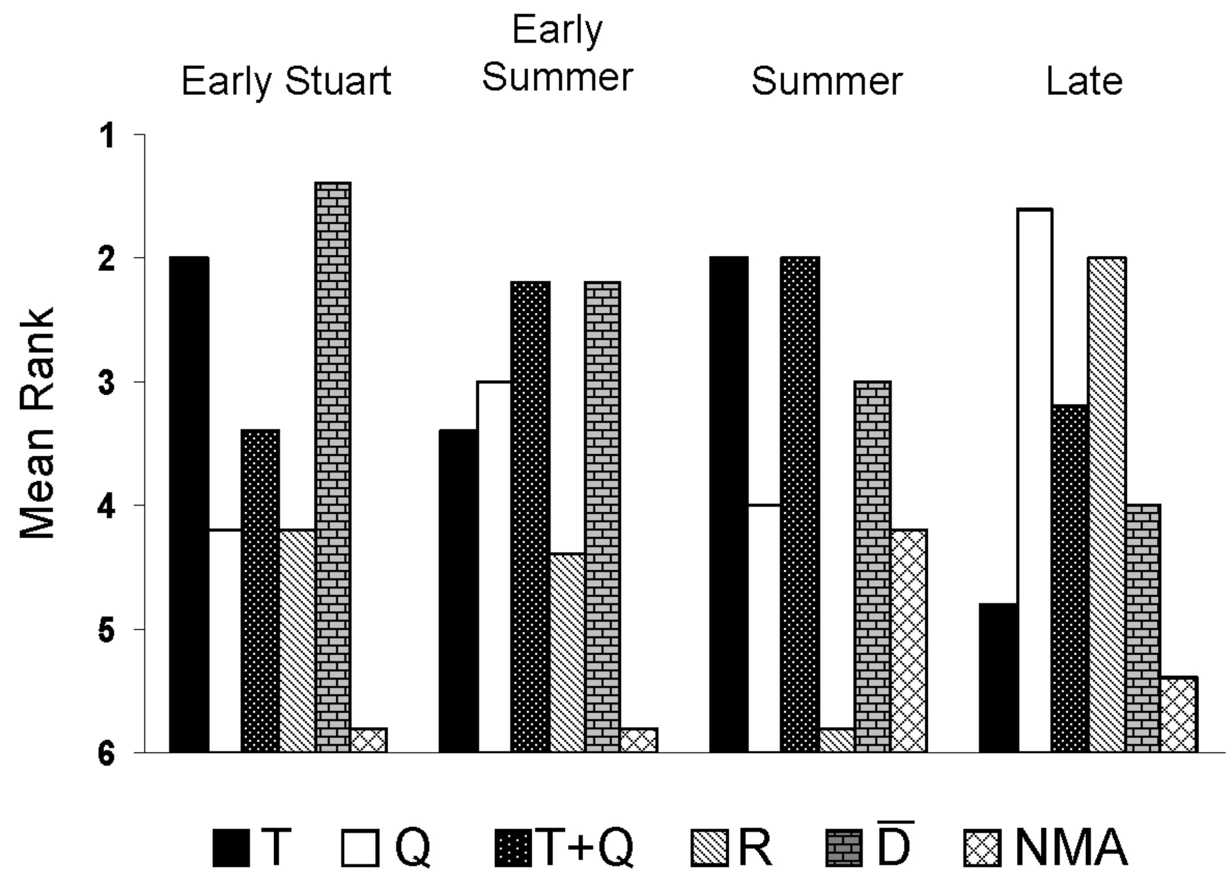


CAN285112_0035



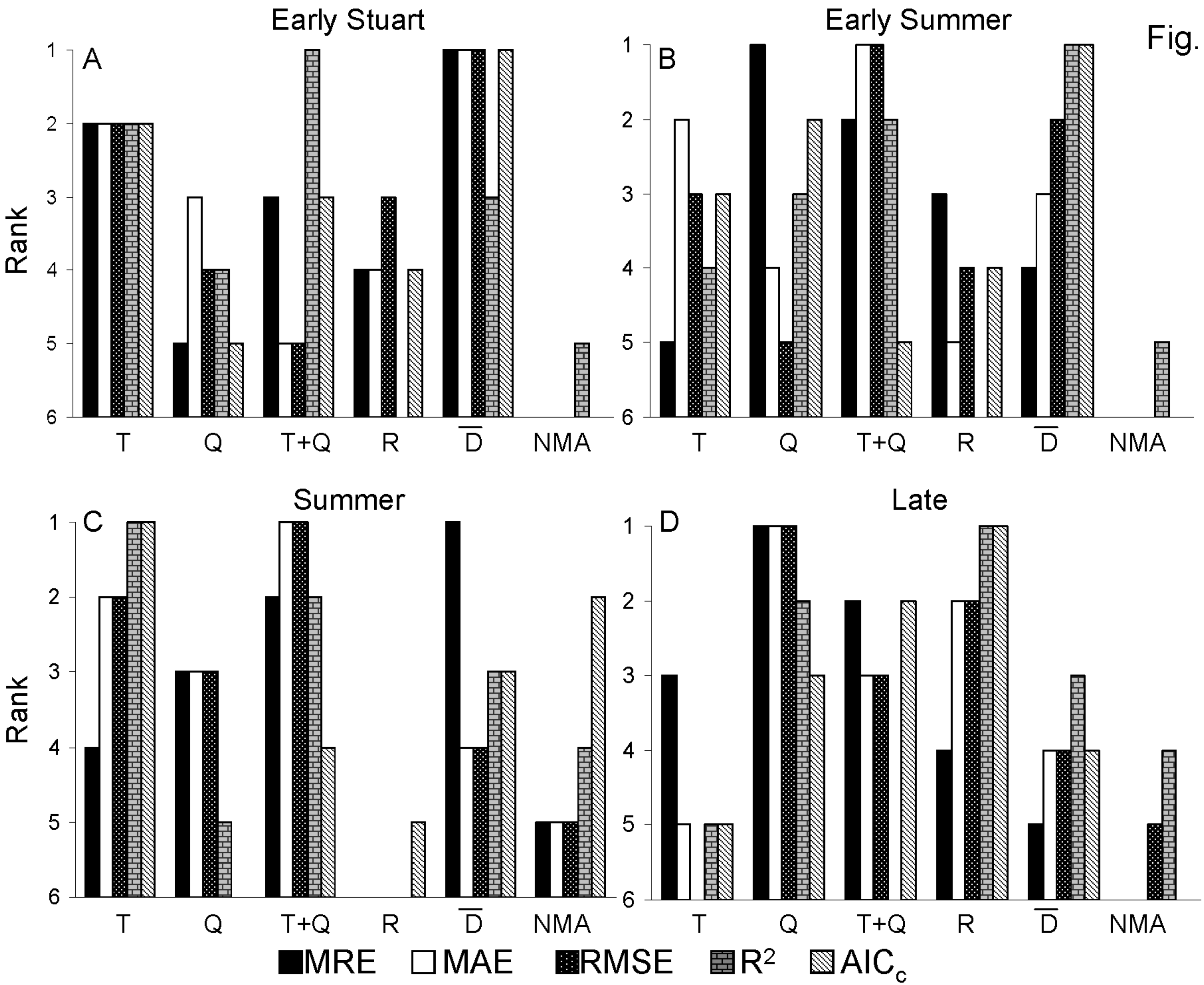
CAN285112_0036

Fig. 4



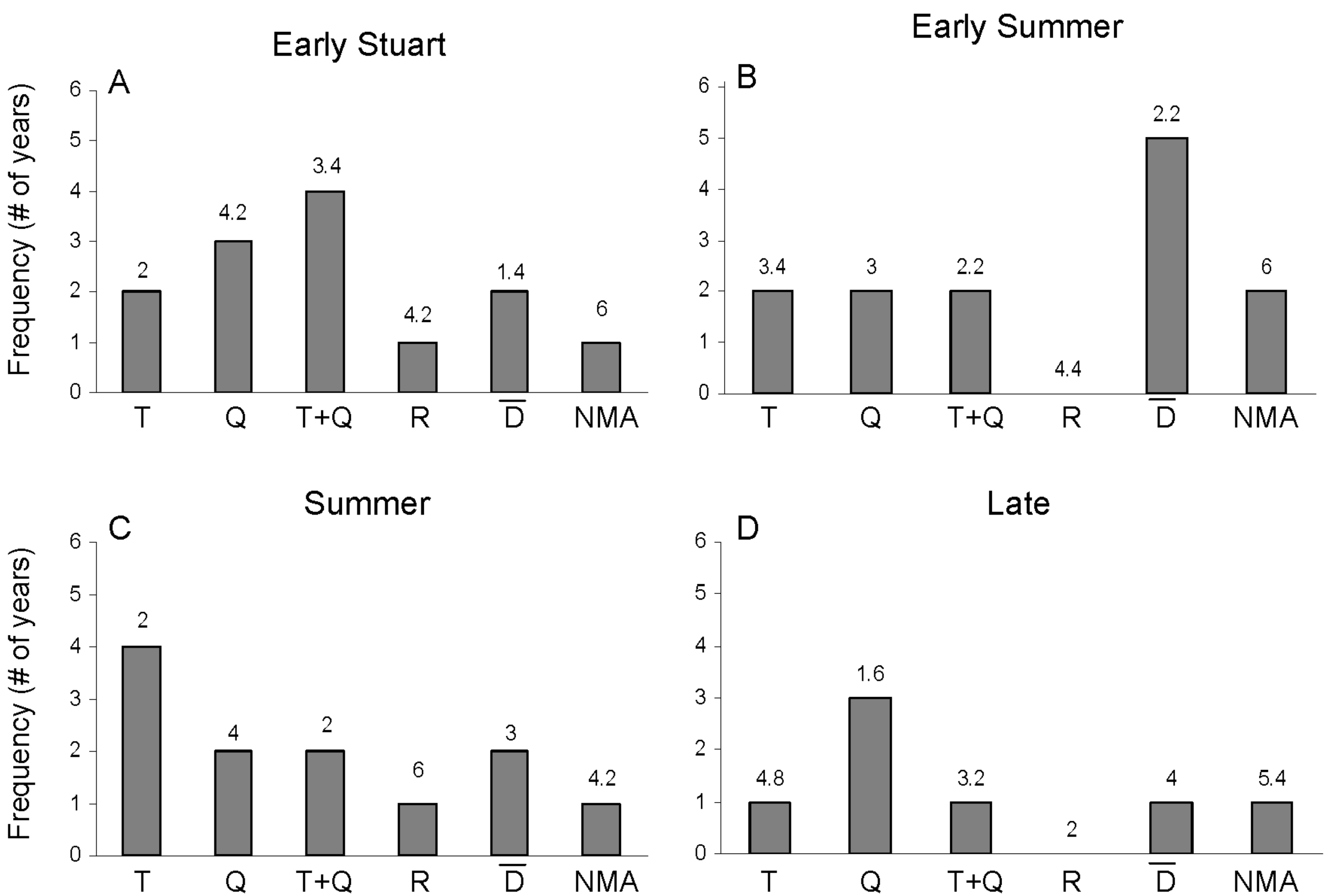
CAN285112_0037

Fig. 5



CAN285112_0038

Fig. 6



CAN285112_0039