DRAFT | Peer Review Purposes Only | Not for Citation

Probabilistic Length-at-Date Model and Application to Spring-run Chinook Salmon on Tributaries of the Sacramento River, California

Authors

Noble Hendrix, QEDA Consulting; Sean Canfield, California Department of Water Resources; Brett Harvey, California Department of Water Resources

Acknowledgments

The data used in this modeling effort are rich and extensive; in some cases, data collection has been ongoing since the late 1990s. This work would not be possible without the field staff collecting daily rotary screw trap data and the data stewards that have managed these data over time. We specifically acknowledge the current data stewards, Anna Allison, Nicolas Bauer, Drew Huneycutt, Jeanine Phillips and Corey Fernandez, and the field staff participating in data collection, who assisted the effort to curate and make compatible the data used in this analysis, and provided valuable insights to contribute to the modeling process.

We also thank the Spring-run Juvenile Production Estimate Core Team, Modeling Advisory Team, and Interagency Review Team for their useful comments and advice.

Work by QEDA Consulting was supported by the California State Water Contractors organization.

Contents

1 Introduction 1

2 Probabilistic Length-at-Date Model 3

2.1 Objectives 3

2.2 Model Structure 3

2.2.1 Finite Mixture Model 3

2.2.2 Modeling the Parameters of the Fork Length Distribution . 5

2.2.3 Modeling the Proportion of Each Component 6

3 Summary of Genetic and Catch Data 8

3.1 Genetic Data 8

4 Estimation 10

4.1 Bayesian Estimation 10

4.2 Data and Likelihoods 11

4.3 Priors 12

4.4 Software 12

5 Application of the Probabilistic Length-at-Date Model to Sacramento River Tributaries 13

5.1 Definitions and Filtering Data 13

5.2 Fitting to Genetic Data 14

6 Application of the Probabilistic LAD Models 15

6.1 Predicting Probability of Run Type 15

6.2 Out-of-Sample Prediction 17

6.3 Spring-run Proportion in Historical Catch Data 19

7 Future Work 22

7.1 Expansion of the Modeling Framework with Hierarchical Modeling 22

7.2 Modeling Feather River 23

7.3 Use of Catch Data 24

8 References 25

Tables

Table 1. Genetic Data Analyzed for Each Site by Genetic Designation Tables-1
Table 2. Posterior Distributions for PLAD Parameters Fit to Juvenile Chinook Salmon Captured in Battle Creek Tables-1
Table 3. Genetic Run Assignments and Probabilistic Length-at-Date Predictions Tables-2
Table 4. Genetic Run Assignments and Run Assignments Based on Fisher at Battle Creek in 2025 .... Tables-2
Table 5. Individuals Genetically Identified to Winter-run in Battle Creek and Probabilistic Length-at-Date Predictions Tables-2

Figures

Figure 1. Hypothetical Finite Mixture Distribution Over Finite Time Interval 1
Figure 2. Hypothetical Finite Mixture Distributions with Variable Component Proportions.. 2
Figure 3. Hypothetical Variation in Proportions of the Three Generic Run Types 3
Figure 4. Genetic Samples Categorized by Run Type for Each Site 4
Figure 5. Runs of Juvenile Chinook Captures in Rotary Screw Traps and Identified to Run Type 5
Figure 6. PLAD Model fits to Juvenile Chinook Salmon Captured in Rotary Screw Traps and Identified to Run Type 6
Figure 8. Prediction of Run Assignment by Fork Length at Four Dates in the Sampling Season at Battle Creek 8
Figure 9. Battle Creek Juveniles Genetically Identified to Run Type during the 2025 Migration Season... 9
Figure 10. Proportion of Spring-run in Battle Creek....................................... 10
Figure 11. Hypothetical Catch Distribution, Initial PLAD Fit, and Updated PLAD Fit 11

Acronyms and Abbreviations

Term	Definition
°C	degrees Celsius
CDFW	California Department of Fish and Wildlife
DWR	California Department of Water Resources
ESU	evolutionarily significant unit(s)
GT-seq	genotyping-in-thousands by sequencing
ITP	Incidental Take Permit
JAGS	Just Another Gibbs Sampler
JPE	juvenile production estimate
LAD	length-at-date
mL	milliliter
mm	millimeter
MCMC	Markov Chain Monte Carlo
NMFS	National Marine Fisheries Service
PLAD	probabilistic length-at-date
ROC	receiver operating characteristic curve
RST	rotary screw trap
spring-run	spring-run Chinook salmon

1 Introduction

In 2024, the California Department of Fish and Wildlife (CDFW) issued an Incidental Take Permit (ITP) to the California Department of Water Resources (DWR) for operation of the State Water Project. As part of ITP Condition of Approval (COA) 7.9.3, DWR agreed to lead an interagency Core Team in the development of a modeling approach for calculating an annual juvenile production estimate (JPE) for spring-run Chinook salmon (*Oncorhynchus tshawytscha*) (spring-run) produced in the Sacramento River watershed, and then use this approach to calculate a JPE annually beginning with the 2026-27 outmigration season (i.e., the 2026 brood year that begins outmigration in the fall of 2026 and finishes in the spring of 2027). Central to this endeavor is an extensive network of rotary screw traps (RSTs) used to monitor outmigration of juvenile Chinook salmon in the Sacramento River and its tributaries, although the majority of these RSTs programs existed prior to the JPE effort, and some of these sites were never intended to provide estimates of outmigrant absolute abundance (CDFW email communication December 15, 2025). One of the key challenges of using the RST data for a spring-run JPE is the co-occurrence of multiple runs of Chinook salmon at most sites, often making the run type of outmigrating juveniles difficult to distinguish. This chapter describes the development of a probabilistic length-at-date (PLAD) model that can be tailored to specific monitoring sites to predict the run assignment of captured Chinook salmon based on the length of juvenile salmon, and the date and location of capture.

California's Central Valley supports four run types of Chinook salmon: fall-run, late-fall-run, winter-run, and spring-run (Yoshiyama et al. 1998). Historically, Central Valley Chinook salmon stocks were one of the most productive systems for Pacific salmon. Spring-run were a dominant component of the mixture of stocks (Yoshiyama et al. 1998, Williams 2006), but now spring-run are listed as threatened, and winter-run are listed as endangered under the federal Endangered Species Act (National Marine Fisheries Service [NMFS] 2005). Each stock is managed under different evolutionarily significant units (ESUs), although fall-run and late-fall-run are listed within the same ESU.

Managers in the Central Valley would like to limit anthropogenic impacts to the natural-origin component of the ESUs, particularly to the endangered winter-run and spring-run, which requires identifying natural-origin Chinook salmon sampled in fish monitoring programs. In the 1970s, length-at-date (LAD) criteria were developed to serve this need, and continue to be used in monitoring programs throughout the Central Valley. Although there have been several modifications to the LAD model since its inception (Harvey 2011), the current versions of the model are similar to the original formulation in several ways. Namely, the existing LAD model has static boundaries that define run assignment via a growth function and upper and lower boundaries that represent early and late spawn timing, respectively.

The LAD model is known to have high misassignment errors (Harvey et al. 2014; Brandes 2021). As an alternative to the LAD model, methods of genetic run identification have been under continuous development since the 1990s and are increasingly applied across the Central Valley, including genetics sampling and testing specifically implemented to support run identification for the Spring-run JPE Program (Canfield 2025). However, genetics cannot be applied to every salmon detected in monitoring programs and there is still a need for an improved method for assigning a run type to fish that are not genetically tested. There is also a need for more dependable run assignment for monitoring data collected in years prior to genetic testing, which constitutes the majority of years of data currently used for the spring-run JPE models.

To fill this need, we developed a probabilistic length-at-date (PLAD) model that uses genetic run assignment from the subset of genetically tested salmon to predict the run type of salmon that were not genetically tested along with uncertainty in that run assignment. The PLAD models reported on here were specifically developed to assign run type to juvenile salmon sampled at RSTs operating at multiple sites in the Sacramento River and its tributaries. Run assignment for these historical catch data was coupled with the BT-SPAS-X model to estimate historical spring-run production (refer to Chapters 4 and 5 for application of PLAD with BT-SPAS-X at tributary and mainstem RST sites). One way of viewing the original LAD model is as a mathematical function: the size, date, and location are inputs, with the LAD function returning the run as the output. In these terms, our objective at its most basic was to develop a more accurate LAD function.

2 Probabilistic Length-at-Date Model

2.1 Objectives

The PLAD model was developed to meet several objectives with the goal of providing estimates of run assignment and statements of the uncertainty around those run assignments. The first was to calculate the probability of run assignment given a fish’s size, time, and location of capture rather than a binary assignment. The second was to develop PLAD assignment models such that they can vary spatially. There may be unique dynamics in portions of the spring-run range and the PLAD models should reflect those dynamics. The PLAD model explicitly incorporates uncertainty in run assignment via probabilistic modeling. Finally, the PLAD model workflow can be updated as new information becomes available through an iterative process of prediction and estimation. We begin by describing the basic model structure and delve into the topic of probabilistic modeling and estimation in later sections.

2.2 Model Structure

2.2.1 Finite Mixture Model

The PLAD model was developed to categorize individuals into mutually exclusive run types with a probability of assignment to that run. Given an individual sampled from the population, say via an RST at a site s and at time t, the run type of individual *j* ($R_{s,t,j}$) is an unknown or latent state. There are i = 1, …, N possible components or runs that individual *j* could belong to, and there are specific covariates that are used to predict the run type the of individual *j*. An individual fish’s fork length, its date of capture, and the location of capture are useful in assigning the individual to a specific run. This is a type of finite mixture analysis and the PLAD model uses this statistical framework to develop the method. We suppress the use of the site subscript *s* in the development of the model for clarity; however, the models we developed are site-specific and PLAD model results for all currently modeled RST sites in the Sacramento River valley are shown in Appendix A. In addition, biweekly spring-run assignment probabilities from completed site-specific PLAD models are shown in Chapter 4 Appendix B for tributary RST sites and Chapter 5 Appendix B for mainstem sites.

The finite mixture model begins with describing the process for identifying an individual to each of the categories of run type. A categorical random variable (Cat) is used to describe the run of individual *j* captured at time *t* ($R_{t,j}$) using probabilities that individual *j* belongs to each of the i= 1:N categories $p_{t,j,1:N}$ = ($p_{t,j,1}$,…, $p_{t,j,N}$). The distribution Cat() is used to define the actual outcome of the run assignment for individual *j* ($R_{t,j}$), and it is equivalent to a multinomial random variable with a single observation (i.e, Multinomial(1, $p_{t,j,1:N}$)).

Equation 1.

$$R_{t,j} \sim \text{Cat}(p_{t,j,1:N})$$

We can incorporate individual-level covariates to help improve the prediction of run type assignment. Here we use the fork length (FL in the equation) for individual j, and information on the capture at time t to model the change in the probabilities for each run type over time t. The probabilities, $p_{t,j,1:N}$ of run type are calculated as the probability of the observed fork length for individual j given the distribution of fork lengths for each run type.

At any given time t, there is a mixture of fork length distributions, with the components of the mixture arising from the lengths of each run type of juvenile Chinook salmon. Thus, the population of fork lengths may be described as a finite mixture distribution, in which multiple components combine to form an overall distribution. The probabilistic description of the fork lengths $f(FL_{t,1:N})$for juveniles at a specific sampling time t is shown in Equation 2.

Equation 2.

$$f(FL_{t,1:N}) = \sum_{i=1}^{N} \square\, \pi_{t,i} f(FL_{t,i}; \boldsymbol{\theta}_{t,i})$$

Where:

$\tilde{f}(FL_{t,1:N})$ is the fork length distribution given the parameters $\theta_{i,t}$, and

$\pi_{i,t}$is the proportion (or weights) of the mixture distribution for run type i at time t.

The specific probability distribution for fork lengths *f*() used in the PLAD is the lognormal distribution.

Equation 3.

$$FL_{t,i} \sim \text{lognormal}(\mu_{t,i}, \sigma^2_{t,i})$$

Where:

$\mu_{i,t}$ is the log mean of the lognormal distribution, and

$\sigma^2{}_{i,t}$ is the log variance.

A hypothetical example of the finite mixture model with three run types is provided in Figure 2-1.

We have described the PLAD model in terms of its mathematical equations, but to gain some insight into how the model functions under different combinations of run proportions (π), we provide several mixture distributions with different values the mixture proportions $\pi_{i,s,t}$ for a single site and a single time (Figure 2).

If there was a direct correspondence between run and fork lengths such that each run was defined uniquely by a specific fork length or range of fork lengths, then the assignment of fork lengths to run type would be straightforward. Namely, the individual fork length would fit into mutually exclusive categories of fork lengths for each run. Under this scenario, the values of $p_{j,1:N}$ would consist of a vector in which a single element was 1 and all others were 0. Importantly, this was one of the assumptions underlying the original LAD models. While this assumption allowed the run assignment to be calculated easily, it also created misclassifications if the boundaries did not truly define mutually exclusive categories.

2.2.2 Modeling the Parameters of the Fork Length Distribution

As defined to up to this point, the PLAD model requires a value for the mean and variance of the fork lengths ($\mu_{t,i}$ and $\sigma_{t,i}$, respectively) and the component proportions ($\pi_{,t,i}$) of each run type *i*, at each time *t*. To reduce the number of parameters in the PLAD model, we develop models for the following underlying components: 1) the change in the mean fork length over time at a site and 2) the change in the run proportion ($\pi_{t,i}$) over time at a site. We use models with a few coefficients and use time as a covariate to reduce the number of parameters required in the PLAD model. Also, we model the mean fork lengths and proportions of each run type with functional forms that have biological interpretations. Both of these aspects will be important when we estimate the parameters of the PLAD model by fitting to data.

We model the mean fork lengths over time by using a log-linear growth model. This functional form has historical relevance as this was the form under the original LAD work (Harvey et al. 2014). The parameters of the log-linear model also have relatively straightforward biological interpretations.

Equation 4.

$$\log(\mu_{i,t}) = \alpha_i + \beta_i t$$

Where:

α_i is the intercept which is related to the emergence timing, and

βi is the growth rate per unit time (days) *t* for run *i*.

Simpler models can be constructed by collapsing this parameterization. For example, if the growth rates of all runs of juveniles at a site are the same, then $\beta I = \beta$.

While the log-linear growth model has been used previously for modeling juvenile Chinook salmon growth in the Central Valley (Harvey et al. 2014), the PLAD framework can use other growth models as well. For example, growth functions that reflect a change in growth rate with size, such as the sigmoidal function or von Bertalanffy could also be used to model the change in mean fork lengths over time to reflect such patterns in fish growth. Although we begin by using the log-linear growth function due its simplicity, our modeling framework allows for other forms of growth equations may be evaluated and substituted in the future.

The variance of the lognormal distributed fork lengths can also be modeled as a function of site, run type, or time. In the initial implementation of the PLAD model, we simplified the model structure by collapsing the variance terms to be equivalent across site, time, and run type. For example, if the modeled variation in growth rates are assumed to be the same across all time periods and across all runs at a site then $\sigma^2_{i,t} = \sigma^2$, this parameterization assumes that the variance in fork lengths across run types is similar at a given site.

2.2.3 Modeling the Proportion of Each Component

The proportions of each run type ($\pi_{t,i}$) can vary through the season due to several biological processes including the total production of each run type, the survival of the juveniles from the location of spawning to the rotary screw trap site, and the timing of outmigration. In a two-component site (e.g., spring-run and fall-run) the pattern in outmigration is typically dominated by a single run type in the early part of the season and then dominated by the second run type in the later part of the season. In a three-component site (e.g., winter-run, spring-run, and fall-run), the intermediate run type has a unimodal shape in its proportion when that run type is outmigrating. As a result, the structure of the models for the proportions should be capable of reflecting the unimodal shape expected for sites that have three-components in the juvenile Chinook salmon samples (e.g., Figure 3).

The run proportions of the PLAD model can be represented with N-1 underlying parameters. The proportions sum to 1, that is $\Sigma_{i=1:N}\ \pi_{t,i} = 1$, thus, one fewer parameters are needed to model the proportions than run types. For example, if there are three run types at a site, then two parameters can be used to define the proportions $\pi_{t,1:3}$. More formally for the dynamics with three runs as show in Equation 5.

Equation 5.

$$\pi_{t,1} = \rho_{t,1}$$

$$\pi_{t,2} = \rho_{t,2}(1 - \pi_{t,1})$$

$$\pi_{t,3} = 1 - \pi_{t,1} - \pi_{t,2}$$

Where:

$\pi_{t,i}$ is the proportion at site *s* for run *i* = (1,2,3) at time *t*, and

$\rho_{t,i}$ is the underlying proportion parameter, which is restricted to the range (0,1).

This structure can be expanded for sites with more than three runs by repeating the process on line 2 of Equation 5 in a similar fashion.

Using similar logic as in the growth modeling, rather than attempting to estimate a large set of parameters $\rho_{t,i}$ for each run and time, we use models with fewer coefficients and time as a covariate.

The underlying proportion parameters $\rho_{t,i}$ can be modeled using logistic regression as a polynomial function of time, which can reflect a unimodal relationship as shown in Equation 6.

Equation 6.

$$\text{logit}(\rho_{t,i}) = \delta_{0,i} + \delta_{1,i}t + \delta_{2,i}t^2$$

Where:

$\delta_{0,\mathrm{i}}$ is the intercept, and

$\delta_{1,\mathrm{i}}$ and $\delta_{2,\mathrm{i}}$ reflect the effect of time on the underlying proportion parameter $\rho_{t,i\ t}$.

The logit() function is defined as log(x/(1-x)) and keeps the values of $\rho_{i,t}$ in the interval (0,1).

Under the dynamics where there is a site with three runs that outmigrate chronologically, the dynamics may be represented as a decline in the first migrating run, a unimodal shape for the second migrating run, and an increasing proportion over the season for the final run (Figure 3).

The set of parameters in the PLAD model includes the α, β and σ parameters of the log-linear growth functions and the δ parameters that define the proportions of each run type. The full set of parameters is θ = (α, β, σ, δ).

3 Summary of Genetic and Catch Data

There are two sources of data that can be used to estimate the parameters of the PLAD model. The first is the genetic identification of a subset of individuals captured in the RST samples at each site, which provides information on the fork length of fish of a known run assignment. These data are useful for modeling the change in mean fork lengths for each run over time at that site to estimate the α and β and σ parameters of the log-linear growth functions for each site and run type. There is also information in the proportions of each run in the genetic samples that can be used to inform the ρ parameters of the PLAD model. The second is the shape of the catch distribution from the RST over a finite time interval (e.g., week), which may also provide information on the relative proportions of each run (i.e., the ρ parameters) in the overall composition.

3.1 Genetic Data

Samples for genetic analysis were collected from captured juvenile Chinook salmon for JPE years 2022–2024. Two types of samples were collected: mucus swabs (2022–2023) and fin clips from the upper caudal lobe (2022–2024). To collect mucus for genetic analysis, fish were swabbed 2–15 times along the lateral body surface. The cotton swabs were then dipped and swirled into a 1.5-milliliter (mL) microcentrifuge tube containing either phosphate-buffered saline (2022–2023) or low-EDTA Tris-HCl (2023), and the swab was discarded. Mucus DNA samples were either used directly in SHERLOCK genotyping reactions (2022–2023) or were subjected to a rapid DNA extraction (2023) prior to SHERLOCK genotyping. DNA was extracted from fin clips using either a commercial kit (Qiagen DNeasy 96 Blood and Tissue Kit) or a rapid Proteinase K digestion followed by a 95-degree Celsius (°C) heat-kill (2024).

Subsets of mucus DNA samples (2022–2023) and fin-clip extracted DNA samples (2024) were subjected to SHERLOCK genotyping assays described in Baerwald et al. (2023). First, individuals were genotyped at a portion of the *Greb1L* locus on chromosome 28 to determine whether they are early-migrating (spring or winter adult return migration timing) or late-migrating (fall or late-fall adult return migration timing) (Prince et al. 2017; Thompson et al. 2020). Salmon with early-migrating genotypes were subjected to subsequent SHERLOCK assays to differentiate spring-run and winter-run individuals.

Genotyping-in-thousands by sequencing (GT-seq; Campbell et al. 2015) was performed on fin-clip extracted DNA from all sampled individuals (2022–2024). Individuals were genotyped for 208 genetic markers designed to distinguish run types in the Central Valley (Anderson et al. 2025). This marker panel includes seven *Greb1L*-linked markers and 201 other markers distributed across the Chinook salmon genome. Early- versus late-migration phenotypes were inferred by analyzing the set of seven genetic markers associated with the *Greb1L/rock1* locus in the program STRUCTURE (Pritchard et al. 2000). Individuals identified as early-migrating were assigned to either spring- or winter-run populations, and all individuals (regardless of early versus late-migration genotype) were assigned to one of the four Central Valley populations—fall, late-fall, spring, or winter—using the full suite of genetic markers in the R package *RUBIAS* version 0.3.4 (Moran and Anderson 2018).

Final run type assignments were made with consideration to both SHERLOCK and GT-seq genotyping results. Individuals determined to be late-migrating using *Greb1L/rock1* were assigned to a "fall or late-fall" run-grouped category. Individuals determined to be early-migrating were assigned to either spring or winter-runs based on subsequent SHERLOCK assays and/or population assignment in RUBIAS. Individuals that displayed a heterozygous genotype at *Greb1L/rock1* (i.e., displayed genotypes characteristic of both early- and late-migrating populations) were assigned to the run associated with the highest posterior probability reported by RUBIAS. Heterozygous *Greb1L/rock1* individuals with less than an 80% posterior probability of assignment to any of the four runs were designated "unknown."

There were 5,631 genotyped juvenile Chinook salmon that were assigned to run types from collections from the winter of 2022 through spring of 2024 (Table 1). Most of the juveniles were assigned to fall- or late-fall and spring-run (Table 1, Figure 4). Winter-run were identified in small numbers at the Battle Creek, Tisdale, and Delta entry sites. In addition, there were a small proportion of "unknown" fish at most sites. We focus on the fish of known genetic assignment for fitting the PLAD model, although heterozygous *Greb1L/rock1* individuals could be modeled as a distinct run described for Equations 5 and 6.

4 Estimation

4.1 Bayesian Estimation

We use Bayesian methods to estimate the PLAD model parameters. We provide a simple introduction to Bayesian estimation here, but it is beyond the scope of this chapter to provide a full treatment of these topics. A very good reference is Gelman et al. (2013) for an introduction to Bayesian statistical approaches and Kéry and Schaub (2011) for an introduction to Bayesian modeling for ecologists.

Bayesian methods are based in probability. The approach is to fit a probability model to data and to summarize the inference by probability distributions on the parameters of the model. Bayesian methods use Bayes' theorem to integrate both prior knowledge about the parameters of the model and information in the data about the parameters. Bayes theorem is:

Equation 7.

$$P(a) \propto p(a) \cdot L(data|a)$$

The posterior probability of parameter a *P*(*a*) is proportional to the prior probability of taking a specific value *p*(*a*) times the likelihood of the data given the value of *a* *L*(*data*|*a*). Thus, Bayesian statistics requires two sources of information for a parameter *a*: the prior probability distribution and the likelihood. The prior summarizes the expected value of parameter *a* prior to evaluating the data. Information for the prior can come from ecological theory or from studies that were performed before analyzing the specific data set. In some cases, the prior distribution is structured so that it is uninformative, which allows the information in the data to "speak for themselves." The likelihood summarizes the probability of observing the data given the value of *a.* Bayes theorem is applied for many possible values of *a* to develop the posterior distribution for that range of values. Importantly, by combining the prior information and the information in the data, the posterior contains all the information on parameter *a* up to, and including, the current analysis of the *data*.

One important aspect of Bayesian methods is that the posterior information is influenced by both the prior and the posterior. The prior has an important role in defining the range of values that can be included in the posterior. The adage "not in the prior, not in the posterior" can be seen by looking at Bayes' theorem. When the prior probability of *p*(*a*) = 0 then the posterior of a *P*(*a*) = 0 as well. While both contribute to the posterior, the component that has more information will contribute more to the posterior. Thus, if there is little prior information on a parameter prior to analyzing the data, the likelihood will dominate the prior. If, on the other hand there is good information on a parameter prior to analyzing the data and the data are weakly informative, the prior will dominate the posterior. We have situations where we employ both conditions in the development of the PLAD model, and we will point them out in the workflow for parameter estimation below.

Bayesian updating refers to a cyclical use of Bayes theorem to update the prior to the posterior as new data become available. Because the posterior distribution contains the current information, Bayes theorem can be applied iteratively. In iteration *i* the prior of the current iteration equals the posterior from the previous iteration, that is $p_i(a) = P_{i\text{-}1}(a)$. The likelihood in iteration is calculated by using previously unanalyzed $L(data_i|a)$. Bayesian updating can be applied to data collected over time, for example from annual survey results.

Finally, once we have a model in which the parameters have been estimated using Bayesian methods, we often want to make predictions from the model. The advantage to using Bayesian methods, which are based in probability, is that predictions from the model are also probability distributions. This is advantageous when an ecologist would like to express uncertainty in the predictions from a model. In the PLAD model, we have multiple outputs from the models where expressing uncertainty in probabilistic terms is an advantage. For example, if we collect a juvenile Chinook salmon at a site in which we have estimated the parameters of the PLAD model, we can make a prediction on the run assignment for that fish along with a probabilistic statement of the certainty of that assignment. For example, the probability that the fish is a spring-run is 0.8, but the 95% posterior confidence interval is 0.73 to 0.92.

4.2 Data and Likelihoods

The source of information for estimating the θ parameters of the PLAD model are fork lengths of the genetically identified fish from the RSTs. To utilize the information in the data, we must define likelihood functions that allow us to compare predictions from the PLAD model to observations.

Fish that were genetically identified to run type were used to develop the models for fork length. The fork lengths of an individual fish *j* genetically identified to run *i* sampled at time *t*, $FL_{i,i,t}$ has a lognormal likelihood with mean $\mu_{s,i,t}$ and variance $\sigma^2_{i,t}$ for run type *i*.

Equation 8.

$$FL_{j,i,t} \sim \text{lognormal}(\mu_{i,t}, \sigma^2_{i,t})$$

Note that this equation is very similar to Equation 4. Here we are using the fork lengths of juveniles with known run assignments to provide information on the mean and variance of the size distribution. These observations are similar to the samples or “rugs” in Figures 1 and 2 that are associated with each of the run types in the mixture distribution.

4.3 Priors

The models were developed in a Bayesian framework to provide estimates of uncertainty on the model parameters θ. Under the Bayesian modeling paradigm, priors for all model coefficients are required. In general, we used non-informative or vague priors to allow the data to provide information on the underlying processes.

There is one parameter in which we provided an informative prior due to information being available on it from previous studies. We used a prior on β, the parameter that describes the daily growth rate to reflect growth rate estimates from LAD models developed for Central Valley juvenile Chinook salmon runs (6.57×10^{-3} $\log_e$(mmFL)/d) (Harvey et al. 2014).

4.4 Software

Posterior samples of the model parameters were obtained by using Markov Chain Monte Carlo (MCMC) sampling, which is an efficient method for Bayesian estimation (Gelman et al. 2013). The model was fit using Just Another Gibbs Sampler (JAGS) (Plummer 2003) in the R programming language (R Core Team 2022) using the packages RCT (Isidoro 2020) and runjags (Denwood 2016). Each of the models was run using three chains with 15,000 iterations each thinned to every 15th iteration with a 3,000-iteration burn in and 2,000-iteration adaptive phase. Model convergence was checked using the Gelman-Rubin statistic in which values near 1.0 indicate that the chains are sampling from the same posterior distribution, and thus the model has converged.

5 Application of the Probabilistic Length-at-Date Model to Sacramento River Tributaries

We fit site-specific PLAD models to genetic data from RST sites on spring-run producing tributaries of the Sacramento River, and for several sites on the mainstem Sacramento River. PLAD model results for all currently modeled RST sites are shown in Appendix A. In addition, biweekly spring-run assignment probabilities from completed site-specific PLAD models are shown in Chapter 4 Appendix B for tributary RST sites and Chapter 5 Appendix B for mainstem sites. However, to demonstrate the PLAD model, we used the Battle Creek site as a case study. We chose Battle Creek for this purpose because it has a mixture of spring-run and fall-run types along with a few winter-run in the system. Below we talk through the steps to fit the PLAD model in general and for Battle Creek.

5.1 Definitions and Filtering Data

The time construction that we used was based on the juvenile outmigration schedule. Because historical records show winter-run can start their outmigration in late August or early September, time is defined as the number of days starting from July 1 of the brood year. Under this convention, the beginning of January is approximately day 180 of the outmigration season, and the beginning of March is approximately day 245.

Although yearling juvenile salmon are captured in the RSTs, our current modeling focus is on young-of-year juveniles due to limited monitoring data for estimating yearling abundance and outmigration; however, we return to how yearling fish may be included in the modeling framework in the discussion. To remove yearlings from the data sets prior to fitting PLAD models, we first plotted historical catch data to visually identify a LAD threshold separating yearlings from young-of-year (Figure 5). All spring-run, fall-run or late-fall-run juveniles that were greater than the size threshold and captured prior to day 275 were identified as too large and were excluded from the data set used to estimate PLAD model coefficients.

Some fish were identified as unknown due to their genetic assignments being heterozygous for the *Greb-1L* run-timing genotype and their population-based assignment based on other genetic markers having indeterminate results (e.g., no population assignment with a value greater than 0.8). These samples were also excluded from the data set used to estimate the PLAD model coefficients.

5.2 Fitting to Genetic Data

The Battle Creek site collected a few juveniles that were identified as winter-run (Figure 4, Table 1), so this site provided an opportunity to evaluate how the PLAD parameter estimates respond to a low number of observations for a run. We also note that overall the number of samples were at the lower end relative to other sites (Table 1).

In Battle Creek, winter- and spring-run were the first captured in December (Day 150) and collections of both runs continued into May (Figure 6). Fall-run captures were later in the season beginning in early March (Day 245). The PLAD model described the trends in the size distributions of all three runs. The shaded region indicated the 95% credible region of parameter values, i.e., the range of parameter uncertainty, whereas the dotted lines indicated the range of observations with lognormally distributed fork lengths. The spread in the parametric uncertainty differed among runs, reflecting the amount of information in the genetically identified fish across time at the site. Winter-run had the least information, and uncertainty in the size of fish on a specific day was attributable to parametric uncertainty. In contrast, spring and fall-run 95% credible intervals for parametric uncertainty were substantially smaller (Figure 6).

The PLAD model fits for Battle Creek indicated that spring-run had an earlier emergence (larger α parameter) than fall-run (Table 2), while fall-run had faster growth rates (larger log(β) parameter; Table 2). Winter-run had the earliest emergence and the slowest growth rates, although the estimates were not well informed from the small numbers of winter-run captured on Battle Creek.

The run proportions (π) of the PLAD model showed variable pattens among the run types in Battle Creek (Figure 7). Spring-run had the highest proportions early in the outmigration season and declined as the season progressed. Fall-run were not present early in the outmigration season and therefore had a low proportion in the mixture distribution, which increased toward the season end. Finally winter-run were in low proportions throughout the season with a slight increase toward the end of the outmigration season (Figure 7).

6 Application of the Probabilistic LAD Models

6.1 Predicting Probability of Run Type

One application of the PLAD model is to make a prediction of run type given a juvenile Chinook salmon's site of capture, date of capture and fork length.

The posterior predictive proportion of each run type ($\boldsymbol{\pi}$) on a given date was predicted from the posterior distributions of the δ parameters. The fork length distributions described by the mean ($\boldsymbol{\mu}$) and standard deviation (σ) of the lognormal distributions was predicted from the posterior distributions of the growth parameters ($\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma$).

The probability of a given sized individual FLj from any run at site s and time t in the population can be obtained from the mixture distribution (e.g., Equation 8), which is the weighted sum of a given fork length arising from the size distributions of each of the component distributions. If we define the probability density function for the lognormal distribution as $\Phi_{LN}(\mu, \sigma)$ then the probability of fork length k (*FL*$_k$) is:

Equation 9.

$$\mathrm{Pr}\ (FL_{k,t}) \sim \sum_{i=1}^{N} \widetilde{\pi_{i,t}} \Phi_{LN}(FL_k | \widetilde{\mu}_{i,t}, \widetilde{\sigma_{i,t}})$$

Where:

> the tilde (e.g., $\tilde{\mu}$) indicates that these are posterior predictive distributions of the PLAD model parameters.
>
> Note: in Equations 9–13, we suppressed the subscript s for site in the equation for clarity, although the predictions are unique for each site, as in $\mathrm{Pr}(FL_{k,s,t}) \sim \ldots)$

To generate the complete size distribution for a given site and time, the probability density in Equation 9 is computed across the full range of sizes.

We may also be interested in the probability of run type assignment for an individual k with fork length *FL*$_{k,t}$ that is captured at time *t*. The probabilities of each run type are computed for the fork length and then normalized to calculate the probability of each run type. The probability that individual *k* with fork length *FL*$_{k,t,}$ given that it is run type *i* is:

Equation 10.

$$\mathrm{Pr}(FL_{k,i,t}) = \widetilde{\pi_{i,t}} \Phi_{LN}(FL_k | \widetilde{\mu}_{i,t}, \widetilde{\sigma_{i,t}})$$

To calculate the probability that the individual *k* is run type *i*, the fork length probabilities from each run type are normalized as

Equation 11.

$$\Pr(k \in i|FL_k) = \frac{\Pr\ (FL_{k,i,t})}{\sum_{i=1}^{N} \Pr\ (FL_{k,i,t})}$$

To incorporate uncertainty into the predictions of run type, samples from the posterior distributions of the PLAD parameters (θ) are used to repeat the above calculations on the order of thousands of times. Estimates of the probability of run assignment for an individual with fork length $FL_{k,t}$ captured at time *t* are generated from the repeated outcomes and characterized by median and 95% credible intervals on the probability of run type. The prediction of run type for a new individual can be calculated in this way. The utility of this approach is that the run type can be predicted (with uncertainty) for individuals captured during any point in the outmigration season for a given site and day of the outmigration season.

Predictions of run assignment from the PLAD model on Battle Creek show how the growth dynamics and the mixture proportions interact to create unique patterns in run assignment across time (Figure 8). Early in the outmigration season on October 6, almost all sizes of juvenile Chinook salmon would be classified as winter-run, with 40-millimeter (mm) fork length fish having a small probability of being spring-run (Figure 8, top left). On December 15, the predicted run type varies across the size of the individual; juveniles less than approximately 50 mm in fork length have a higher probability of being spring-run, whereas individuals with fork lengths greater than 50 mm would have a higher probability of being winter-run (Figure 8, top right). On February 23, individuals with a fork length of 40 mm would be predicted to be a mixture of fall-run and spring-run, whereas individuals with a fork length of 60 mm could be any of the three run types with the highest probability of being spring-run, and individuals with a fork length greater than 100 mm would be predicted to be a mixture of spring-run and winter-run (Figure 8, bottom left). Finally, toward the end of the outmigration season on May 3, individuals with fork lengths less than 80 mm are more likely to be fall-run, with the probabilities becoming approximately equal for individuals with fork lengths of 90 mm; larger individuals are predicted to be mostly spring-run with some chance of being winter-run (Figure 8, bottom right).

The uncertainty in the predictions reflects both the amount of information that was available for estimating the parameters of the PLAD model and the likelihood of fish of that size being present. For example, in the first two example dates in Figure 8 (Days 98 and 168), winter-run are predicted to be the most likely run type for larger fish with very high certainty, despite small sample sizes in the Battle Creek data set. Later in the season, the run assignment becomes much less certain as can be seen by the widths of the 95% credible intervals for the second two dates (Figure 8, Days 238 and 308).

6.2 Out-of-Sample Prediction

To evaluate the predictive ability of the PLAD model, we used the PLAD model that was developed on genetic run identifications from the 2022, 2023, and 2024 outmigration seasons to predict individuals that were captured during the 2025 outmigration season, and then compared these predictions to individual genetic run assignments from the 2025 season. As an example, we continue to work with the PLAD model for Battle Creek and evaluate the predictions for that site. We used the individual's date of capture and fork length at the time of capture and the PLAD model structure (e.g., Equations 1–6) to calculate the probability of run assignment, working backward from probability of proportion and probability of size to the categorical assignment in Equation 1. The model coefficients (θ) in these equations were the posterior draws from fitting to the genetic data from the 2022–2024 outmigration seasons. For each fish, the probability of run type was calculated from 3,000 draws from the posterior distribution of the PLAD model parameters (θ). As a result, the probability of run type for each run in the mixture distribution was calculated 3,000 times for each individual. From the 3,000 run type predictions, several point and interval estimates can be calculated for evaluating the predictive ability and its uncertainty of the PLAD model. For example, the mean (or median) probability across all of the 3,000 run type predictions can be used to assign a run type that would summarize the central tendency of the distribution of probabilities of run type. In contrast, the prediction intervals (e.g., 95% interval on run-type probability) can provide information on the certainty in the run type assignment by the PLAD model.

Using the mean probability of run type, the most likely run can be assigned to each individual. A useful way to display the errors in prediction versus genetic assignment is a confusion matrix. The assignments that agree are on the diagonal and the mis-assignments are on the off diagonals. This display provides information both on the correct assignments and on the patterns in mis-assignments, which can be helpful for understanding the bias in the classifications.

In addition, there has been a significant development in metrics to evaluate the predictive ability of algorithms for classification. The expansion of machine-learning methods for classification has been one of the main drivers of metric development. One of the more common metrics for two classes is the receiver operating characteristic curve (ROC), in which true positive assignments are plotted against false positive assignments (Fawcett 2006). The area under the ROC curve, or area-under-curve metric indicates the classification ability of the algorithm. A value of 1.0 is a perfect classifier (i.e., all true positives and no false positives). A model with a 50% chance of correct classification will have a metric value of 0.5, which represents the minimum classification score. The binary classification was extended to a multi-class metric by Hand and Till (2001), and we use it to evaluate the predictive ability of the PLAD model for assigning run type to juvenile Chinook.

For Battle Creek, 60 of the salmon captured during the 2025 outmigration season were genetically tested (Figure 8). The assignments for the three run types were accurate for spring-run and fall-run, but winter-run were misassigned (Table 3). There were 12 winter-run and they were predominantly assigned to fall-run, leading to an M-metric value of 0.77.

We also calculated a confusion matrix for the Fisher (1992) LAD criteria assignment. Under the LAD assignments, almost all genetically tested juvenile Chinook salmon captured at Battle Creek during the 2025 outmigration season were assigned to fall-run with one genetic winter-run being assigned to spring-run (Table 3). As a result, the four genetic fall-run were classified correctly by LAD, but the remaining 56 genetically tested salmon were misclassified by LAD. It is worth noting that the version of the LAD criteria used by field crews on Battle Creek do not have a category for predicting winter-run, so genetic winter-run are guaranteed to be misclassified by LAD; however, all but one of the 49 genetic spring-run were also misclassified.

Under the PLAD model, the misclassification of winter-run to fall-run was likely due to the higher numbers of winter-run captured during the 2025 outmigration season relative to the other years used to fit the PLAD model coefficients. This was because of a recently implemented winter-run jump start program on Battle Creek, which continues to increase numbers of winter-run produced in Battle Creek. During the 2025 outmigration season, 12 winter-run and four fall-run were captured in Battle Creek, compared to only seven winter-run and 36 fall-run over the previous three years combined (Table 1). Estimates of the probability of each run type for the 12 genetically identified winter-run from 2025 are presented in Table 5. Because of the change in winter-run to fall-run ratios in 2025 relative to the previous three years, the PLAD model for Battle Creek expected fish in the 70–90 mm fork length range to have a higher probability of being fall-run or spring-run than winter-run (Figure 7, Figure 8), while the genetic assignments showed the opposite pattern (Figure 9). Furthermore, the 95% intervals for the probability of run assignment for fall-run and spring-run were broad, demonstrating a general lack of precision in the PLAD estimates late in the season (Table 5).

In contrast, juveniles that were captured early in the season had high precision in the run assignment, as demonstrated for Day 98 in Figure 8. The PLAD model assigned all individuals during this period to spring-run with a median assignment probability of approximately 0.99 (95% intervals of 0.90, 1.0), and genetic tests also assigned these individuals to spring-run. This suggests there may be periods in the sampling season at any given site when the PLAD models can precisely assign juveniles to the correct run-type, and other periods when the PLAD model assignments are imprecise and incorrect. The uncertainty interval of PLAD predictions can provide insight into when the PLAD model is likely to be less precise, and continuing to perform out-of-sample predictions for additional years will help identify these periods.

Battle Creek presents an interesting case study in the application of the PLAD model to out-of-sample prediction. First, Battle Creek is one of the few tributary RST sites where winter-run are present, creating a three-component mixture model. Second, the proportions of winter-run are changing in Battle Creek because of active reintroduction efforts (Lipscomb et al. 2025). For cases like this, we may need to put more consideration into which years of genetic data are used to estimate PLAD model coefficients. For example, perhaps only the initial years of genetic tests should be used to estimate Battle Creek PLAD coefficients when the application is to assign run to years prior to genetic testing. Conversely, as the reintroduction of winter-run to Battle Creek proceeds, it will be important to continue collecting genetic data and to track the relative proportions of winter-run to improve PLAD run assignment accuracy for future years.

6.3 Spring-run Proportion in Historical Catch Data

One of the primary purposes of the PLAD model is to estimate the proportion of spring-run in the historical mixed stock catch prior to the recent advent of genetic testing. Estimation of run proportions were needed to estimate run-specific abundance from these historical catch records (refer to the BT SPAS-X models described in Chapters 4 and 5), which could then be applied in models to predict spring-run juvenile production (refer to the stock-recruit model described in Chapter 7, and the and inseason outmigrant model described in Chapter 8). The approach we employed was to use the fork lengths of individuals captured during biweekly periods at each RST site to predict the proportions of each run expected during those biweekly periods. We selected biweekly periods as a balance between temporal resolution in run-specific abundance estimates while ensuring adequate sample sizes to fit model parameters. Given the Bayesian approach in developing the PLAD model, the estimates of the spring-run proportion in the historical catch allows the incorporation of run assignment uncertainty into biweekly abundance estimates.

Although the posterior predictions from the PLAD model for the proportions of each run type (π) would be sufficient for predicting the total abundance of each run for biweekly periods, application of survival and travel time models (Chapter 9) required predictions of run proportion for specific size categories: fry (less than a 45-mm fork length) and smolts (greater than a 45-mm fork length), because outmigrant size is a covariate in these models. We accomplished this by integrating over the size distribution on a given date and site using the upper and lower bounds of the life stage categories. Uncertainty in the proportion of spring-run was again incorporated into the multi-model framework using draws from the posterior distributions of the PLAD model.

This approach is similar to the approach described above for estimating the probability of run type for an individual with a specific fork length, although for this application we were interested in a range of fork lengths. To generate the proportion of the run from a specific size range on a given day and location, the probability of fork lengths in each size range from the lower size to upper size (*l*,*u*) and from run type *i* is calculated by integrating across the size range:

Equation 12.

$$\Pr\left(FL_{k=l:u,i,t}\right) = \int_{k=l}^{\widetilde{u}} \square\, \pi_{\iota,t} \Phi_{LN}(FL_k | \widetilde{\mu}_{i,t}, \widetilde{\sigma_{s,\iota,t}})$$

Where:

we are integrating across the lower and upper fork lengths, and

all other symbols of the equation are the same as in Equation 6-2.

To calculate the probability that individuals in the size range (*l*,*u*) are a specific run type, the probability of a fork length from each run type in the size range (*l*,*u*) are normalized as:

Equation 13.

$$\Pr(k \in i | FL_{k=l:u}) = \frac{\Pr\,(FL_{k=l:u,i,t})}{\sum_{i=1}^{N} \Pr\,(FL_{k=l:u,i,t})}$$

Again using Battle Creek as the example, predictions of the proportion of spring-run juveniles were computed for each Julian week beginning with Julian week 45 and ending with Julian week 22. This range of Julian weeks coincide with the historical RST catch dates and facilitate using the PLAD model outputs for analyzing the historical catch data. Across all fork lengths, spring-run had the highest proportion among juveniles captured early in the season and decreasing proportions over the outmigration season (Figure 8, top left). Fry-sized fish (i.e., with a fork length less than or equal to 45 mm) were predominantly spring-run early in the season (Figure 8, top right). This pattern was due to fall-run fry entering the samples around Julian week 5. Smolt-sized fish (i.e., with a fork length greater than 45 mm) were predominantly spring-run in the middle part of the outmigration season (Figure 8, bottom left), which reflected the small size of spring-run in the early portion of the season and the addition of fall-run smolt-sized fish later in the outmigration season.

It is important to note that the proportions of spring-run generated in Figure 8 include the influence of winter-run on the mixture of juvenile Chinook salmon in Battle Creek. Winter-run juveniles were first reintroduced into Battle Creek in 2018 with adult winter-run returning to spawn beginning in 2020. To evaluate the historical RST catch data in Battle Creek prior to 2018 (i.e., before winter-run were reintroduced), the winter-run component could be removed from the PLAD data set such that a pre-reintroduction estimate of spring-run proportion would be estimated. However, given the relatively small number of winter-run in the dataset on Battle Creek (7 of 187; Table 1), the overall impact on the predicted spring-run proportion for years prior to winter-run reintroduction should be minimal.

7 Future Work

7.1 Expansion of the Modeling Framework with Hierarchical Modeling

Hierarchical modeling is a useful approach for pooling data across related spatial or temporal sampling units that allows the information from these units to be shared (Gelman et al. 2013). For example, we are analyzing the dynamics underlying the growth of juveniles at a site over several years. Thus far, we have implicitly assumed that the growth dynamics at a site are the same across each year and thus we are operating under the hypothesis that the parameters to define the growth of juveniles should be the same values in all years. As a result, any consistent variation in juvenile growth rates is reflected in the posterior distributions of the (α,β,σ) parameters. An alternative hypothesis is that each year is completely different from the other years, and thus the parameters to define the growth of juveniles are unique in each year. Under this formulation, information on the growth dynamics would be assumed to be independent in each year. Hierarchical modeling is intermediate to these two hypotheses about the variation in dynamics among years. The advantage is that the hierarchy allows for individual years to be unique from each other, but to still come from a similar underlying process.

The approach is to allow the parameters to arise from a hyper-distribution. That is, instead of each parameter coming from its own distinct prior distribution, they are drawn from a common hyper-distribution with the same mean and variance. For example, in the simple Bayes' theorem example above (Equation 7), let us assume that the parameter *a* is being estimated in three different years *y*. We define the a_y as coming from the same hyper-distribution, such as a normal distribution with a common mean μ_a and variance σ_a^2:

Equation 14.

$$a_y \sim N(\mu_a, \sigma_a^2)$$

There are some important issues in the development of a PLAD hierarchical model that are worth identifying. The first issue is that the use of a hyper-prior for a parameter assumes that the underlying process this parameter describes is the same across the different units grouped under the hyper-prior. In our example above, we would assume that the years are "exchangeable;" that is, we would assume that the underlying dynamics were the same among years. Also, the number of units where the data could be combined is an important consideration in whether to apply hierarchical modeling. It is generally not worth attempting hierarchical modeling with just a couple of exchangeable units. The hyper-variance determines the degree of sharing of information among the units. Small numbers of units (i.e., three to four units) may not have much information on the posterior of the hyper-variance (σ_a^2). As a result, when the number of exchangeable units is small, additional work is required to develop vaguely informative priors on the hyper-variance to ensure that information from the units is pooled into the hyper-distribution (Gelman et al. 2013 in Section 5.7).

Currently the level of genetic data span four seasons (three of which were analyzed here); thus, we are temporally operating in the design space of a few exchangeable units until more years of genetic data have accumulated. Still, there is value in developing the PLAD models in a hierarchical fashion to understand the degree of inter-annual variability and the degree of shared patterns among years. Also, as additional years of data are collected, the vaguely informative priors required for pooling in the early stages of the analysis can be relaxed to allow the data to define the hyper-variance among years. The use of hierarchical modeling can also be used to group RST sites spatially. The most likely candidates for site grouping are the mainstem sites of Tisdale, Knights Landing, and Delta entry due to the relatively short travel times of juvenile Chinook salmon between these sites, which would result in relatively similar LAD relationships. The use of spatial units follows a similar development to the discussion of hierarchical modeling among years.

7.2 Modeling Feather River

Historically, Feather River spring-run accessed high-elevation streams for spawning, while fall-run spawned closer to the valley floor. Following construction of Oroville Dam, both populations spawn in the several miles of suitable habitat downstream of the dam, which led to interbreeding of spring- and fall-run. Hatchery practices also result in interbreeding of spring- and fall-run. The interbreeding produces juveniles that are heterozygote for run-timing genotypes and can exhibit intermediate juvenile outmigration phenotypes. Currently heterozygotes are assigned to fall-run or spring-run using a suite of genetic markers that are not directly linked to adult run-timing and which may not coincide with their outmigration timing. From the PLAD perspective, the decoupling of assignment from outmigration phenotype introduces potential bias in the estimation of the PLAD model parameters. Because the Feather River is typically one of the most productive tributaries for naturally spawned spring-run, the bias could be substantial when PLAD models are applied to predict a Sacramento Valley spring-run JPE.

One of the next steps in PLAD model development is to explicitly model juveniles with heterozygote run-timing genotypes as a separate “run.” In addition, forms of the PLAD model can incorporate uncertainty in the genetic run assignment, particularly for those heterozygote fish that have been identified to run using population-based markers. The PLAD model can be updated to allow uncertainty in genetic assignment to vary by individual and thus allow fish with high accuracy to have a stronger influence on the growth parameters than fish with lower accuracy. This approach can be implemented through use of a genetic assignment error matrix, such as those used for aging errors (Punt et al. 2008), or by using individual-level genetic typing errors returned from the genetic assignment algorithms (e.g., Moran and Anderson 2018).

To incorporate the release of hatchery juveniles into the Feather River, we can include a hatchery component into the PLAD model structure. Marked individuals of known hatchery origin can be used to develop a unique growth equation and proportion in the population to help identify unmarked hatchery fish. Information on the size and timing at release can be used as priors to help inform the parameters of PLAD models for hatchery fish.

7.3 Use of Catch Data

For the historical catch data that were obtained prior to initiation of genetic sampling, there is information in the catch composition from each year that can theoretically inform the PLAD model for that site and year. For example, discrepancies between the predicted and observed size distribution for a specific site and time can be used to update the PLAD prediction for that site and time (Figure 11).

Attempts to estimate the relationship between the size distribution in catches and the age structure of the underlying population have been a goal of fisheries scientists for some time. Fournier et al. (1990) developed an approach called MULTIFAN that uses a von Bertalanffy growth model to predict the mean size of fishes of different ages. A distribution of sizes for each age is constructed from a probabilistic distribution of sizes given the mean and variance. The catch data at each time step is a mixture distribution composed of different age classes with each age class having its own proportion in the mixture distribution. The MULTIFAN structure is similar in many respects to the PLAD model. The mean size of different components are tied to a growth model through time and the distribution of fish sizes in the catch is a mixture, with the proportions of the components changing through time. The authors of the MULTIFAN framework developed a robust likelihood function for estimation and remark that the fitting of size-structured data is a difficult and sometimes subjective effort. Nonetheless, they were successful in estimating the size structure of southern bluefin tuna (*Thunnus maccoyii*) using this approach. This provides a roadmap for using the historical RST data to understand variability in growth rates and proportions of run types in the Central Valley.

8 References

Anderson EC, AJ Clemento, MA Campbell, DE Pearse, AK Beulke, C Columbus, E Campbell, NF Thompson, JC Garza. 2025. “A Multipurpose Microhaplotype Panel for Genetic Analysis of California Chinook Salmon.” *Evolutionary Applications*. https://doi.org/10.1111/eva.70110

Baerwald, MR, EC Funk, AM Goodbla, MA Campbel, T Thompson, MH Meek, and AD Schreier. 2023. “Rapid CRISPR-Cas13a genetic identification enables new opportunities for listed Chinook salmon management.” *Molecular Ecology Resources* 25(5). https://doi.org/10.1111/1755-0998.13777

Brandes, PL, B Pyper, M Banks, D Jacobson, T Garrison, S Cramer. 2021. “Comparison of Length-at-Date Criteria and Genetic Run Assignments for Juvenile Chinook Salmon Caught at Sacramento and Chipps Island in the Sacramento-San Jaquin Delta of California.” *San Francisco Estuary and Watershed Science* 19(3). https://escholarship.org/uc/item/4dw946ww

Campbell, NR, SA Harmon, and SR Narum. 2015. “Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing.” *Molecular Ecology Resources* 15(4):855–867. https://doi.org/10.1111/1755-0998.12357

Canfield, S, and M Baerwald. 2025. Spring-Run Chinook Salmon JPE Run Identification Program Research and Initial Monitoring Plan Update: March 24, 2025. Prepared for the California Department of Water Resources. March. https://water.ca.gov/-/media/DWR-Website/Web-Pages/Programs/State-Water-Project/Endangered-Species-Protection/SR-JPE-Run-ID-Program-Development-Plan-2023-08-Update.pdf

Denwood, MJ. 2016. “runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS.” *Journal of Statistical Software* 71(9):1–25. https://doi.org/10.18637/jss.v071.i09

Fawcett, T. 2006. “An introduction to ROC analysis.” *Pattern Recognition Letters* 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fisher, FW. 1992. “Chinook Salmon (*Oncorhynchus tshawytscha*) growth and occurrence in the Sacramento–San Joaquin River system.” Draft Office Report. Prepared by the California Department of Fish and Game Inland Fisheries Division. Redding, California.

Gelman, A, JB Carlin, HS Stern, DB Dunson, A Vehtari, and DB Rubin. 2013. *Bayesian Data Analysis*. Third Edition. Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Hand, DJ, and RJ Till. 2001. "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems." *Machine Learning* 45(2):171–186. https://doi.org/10.1023/A:1010920819831

Harvey, BN, and C Stroble. 2011. "Length-at-date criteria to classify juvenile Chinook salmon in the California Central Valley: development and implementation history." *Interagency Ecological Program Newsletter.* 24(3). Available upon request from: https://iep.ca.gov/Publications/Library

Harvey, BN, DP Jacobson, and MA Banks. 2014. "Quantifying the uncertainty of a juvenile Chinook Salmon race identification method for a mixed-race stock." *North American Journal of Fisheries Management* 34(6):1177–1186.

Fournier, DA, JR Sibert, J Majkowski, and J Hampton. 1990. "MULTIFAN a likelihood-based method for estimating growth parameters and age composition from multiple length frequency data sets illustrated using data for southern bluefin tuna (*Thunnus maccoyii*)." *Canadian Journal of Fisheries and Aquatic Sciences* 47(2):301–317. https://doi.org/10.1139/f90-032

Isidoro, GU. 2020. "Design and evaluation of RCTs with RCT." https://github.com/isidorogu/RCT.

Kéry, M and M Schaub. 2011. *Bayesian Population Analysis Using WinBUGS: A Hierarchical Perspective*. Academic Press, an imprint of Elsevier, Inc. https://doi.org/10.1016/C2010-0-68368-4

Lipscomb, TN, Z Siders, S Austing, J Von Bargen, and LA Earley. 2025. "Accelerating the reintroduction of endangered Sacramento River winter-run Chinook Salmon to Battle Creek, California using captive broodstock." *North American Journal of Fisheries Management* 45(2):236–250. https://doi.org/10.1093/najfmt/vqaf009

Moran BM, and EC Anderson. 2018. "Bayesian inference from the conditional genetic stock identification model." *Canadian Journal of Fisheries and Aquatic Sciences* 76(4):551–560. https://doi.org/10.1139/cjfas-2018-0016

National Marine Fisheries Service (NMFS). 2005. "Endangered and Threatened Species: Final Listing Determinations for 16 ESUs of West Coast Salmon, and Final 4(d) Protective Regulations for Threatened Salmonid ESUs." *Federal Register* 70:37159–37204.

Plummer, M. 2003. "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling." K Hornik, F Leisch and A Zeileis, editors. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* 124(125.10):1–10). https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Punt, AE, DC Smith, K KrusicGolub and S Robertson. 2008. “Quantifying age-reading error for use in fisheries stock assessments, with application to species in Australia’s southern and eastern scalefish and shark fishery.” *Canadian Journal of Fisheries and Aquatic Sciences* 65(9):1991–2005. https://doi.org/10.1139/F08-111

Prince, DJ, SM O’Rourke, TQ Thompson, OA Ali, HS Lyman, IK Saglam, TJ Hotaling, AP Spidle and MR Miller. 2017. “The evolutionary basis of premature migration in Pacific salmon highlights the utility of genomics for informing conservation.” *Science Advances* 3(8):e1603198. https://doi.org/10.1126/sciadv.1603198

Pritchard JK, M Stephens, and P Donnelly. 2000. “Inference of population structure using multilocus genotype data.” *Genetics* 155:945–959. https://doi.org/10.1093/genetics/155.2.945

R Core Team. 2022. R: “A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Thompson, NF, EC Anderson, AJ Clemento, MA Campbell, DE Pearse, JW Hearsey, AP Kinziger, and JC Garza. 2020. “A complex phenotype in salmon controlled by a simple change in migratory timing.” *Science* 370(6516):609–613. https://doi.org/10.1126/science.aba9059

Williams, JG. 2006. “Central Valley salmon: a perspective on Chinook and steelhead in the Central Valley of California.” *San Francisco Estuary Watershed Science* 4(2):1–398.

Yoshiyama, RM, FW Fisher, PB Moyle. 1998. “Historical abundance and decline of Chinook salmon in the Central Valley region of California.” *North American Journal of Fisheries Management* 18(3):487–505.

Tables and Figures

Tables

Table 1. Genetic Data Analyzed for Each Site by Genetic Designation

Site (Abbreviation)	Fall[a]	Spring	Unknown	Winter
Battle Creek (BTC)	36	144	0	7
Butte Creek (BUT)	3	239	0	0
Lower Clear Creek (CLR)	419	88	8	0
Deer Creek (DER)	17	106	1	0
Delta entry (DEL)	213	36	9	10
Feather River RM17 (F17)	510	100	6	3
Feather River RM61 (F61)	1,836	468	38	0
Knights Landing (KNL)	167	31	1	5
Mill Creek (MIL)	138	106	2	1
Tisdale (TIS)	244	49	7	19
Yuba River (YUR)	441	119	3	1

[a] Includes fall and late fall.

RM = river mile

Table 2. Posterior Distributions for Probabilistic Length-at-Date Parameters Fit to Juvenile Chinook Salmon Captured in Battle Creek

Table of posterior distributions for probabilistic length-at-date (PLAD) parameters fit to juvenile Chinook salmon captured in Battle Creek. The effective sample size (SSeff) indicates the number of independent samples in the Markov Chain Monte Carlo (MCMC) (larger is better) and the Gelman-Rubin statistic (psrf) indicates whether chains have failed to converge (values greater than 1.1 indicate lack of convergence).

Parameter	Lower95	Median	Upper95	SSeff	psrf
α_S	1.9592	2.1098	2.1098	6468	1.000
α_F	-0.04830	0.66962	1.4117	1197	1.000
α_W	3.345	3.9223	4.3949	10530	1.000
σ	0.18208	0.20239	0.22429	28402	1.000
$\log(\beta_S)$	-4.9794	-4.8896	-4.8065	6487	1.001
$\log(\beta_F)$	-4.6697	-4.4469	-4.2458	1172	1.001
$\log(\beta_W)$	-7.2830	-6.1645	-5.3445	11567	1.000
$\delta_{0,1}$	0.87205	1.3758	1.8940	30000	1.000
$\delta_{0,2}$	-0.18276	0.61827	1.4195	30000	1.000
$\delta_{1,1}$	-1.6997	-1.2269	-0.76434	29904	1.000
$\delta_{1,2}$	-0.01906	0.74940	1.5644	30000	1.000

Parameter	Lower95	Median	Upper95	SSeff	psrf
$\delta_{2,1}$	-0.39759	0.09008	0.59936	30000	1.000
$\delta_{2,2}$	-0.49628	0.17593	0.86419	30000	1.000

S = spring-run, F = fall-run, and W = winter-run

Table 3. Genetic Run Assignments and Probabilistic Length-at-Date Predictions

Genetic run assignments and PLAD predictions for 60 juvenile Chinook salmon captured in Battle Creek during the 2025 outmigration season.

		PLAD		
		Fall	Spring	Winter
Genetic Assignment	Fall	4	0	0
	Spring	0	44	0
	Winter	11	1	0

Table 4. Genetic Run Assignments and Run Assignments Based on Fisher at Battle Creek in 2025

Genetic run assignments and run assignments based on Fisher (1992) length-at-date (LAD) criteria for 60 juvenile Chinook captured in Battle Creek during the 2025 outmigration season.

		Field ID from LAD			
		Fall	Spring	Winter	Non-race
Genetic Assignment	Fall	4	0	0	
	Spring	38	0	0	1
	Winter	11	1	0	

Table 5. Individuals Genetically Identified to Winter-run in Battle Creek and Probabilistic Length-at-Date Predictions

Individuals genetically identified to winter-run in Battle Creek during the 2025 outmigration season and their PLAD predictions for probability of each run-type median (0.5) and 95% credible intervals (0.025, 0.975) for spring-run, fall-run, and winter-run.

ID	Spring 0.5	Spring 0.025	Spring 0.975	Fall 0.5	Fall 0.025	Fall 0.975	Winter 0.5	Winter 0.025	Winter 0.975
1	0.45	0.17	0.77	0.5	0.2	0.78	0.03	0	0.2
2	0.45	0.17	0.77	0.5	0.2	0.78	0.03	0	0.2
3	0.41	0.15	0.74	0.55	0.23	0.81	0.02	0	0.17
4	0.41	0.15	0.74	0.55	0.23	0.81	0.02	0	0.17

ID	Spring 0.5	Spring 0.025	Spring 0.975	Fall 0.5	Fall 0.025	Fall 0.975	Winter 0.5	Winter 0.025	Winter 0.975
5	0.61	0.23	0.89	0.21	0.05	0.53	0.13	0.01	0.56
6	0.37	0.13	0.71	0.58	0.26	0.83	0.03	0	0.18
7	0.28	0.09	0.61	0.69	0.36	0.89	0.02	0	0.13
8	0.4	0.15	0.74	0.53	0.22	0.8	0.04	0	0.23
9	0.45	0.17	0.78	0.47	0.18	0.76	0.05	0.01	0.28
10	0.4	0.15	0.74	0.53	0.22	0.8	0.04	0	0.23
11	0.44	0.16	0.77	0.48	0.19	0.77	0.05	0.01	0.27
12	0.34	0.12	0.67	0.62	0.3	0.85	0.02	0	0.17

Figures

Figure 1. Hypothetical Finite Mixture Distribution Over Finite Time Interval

A hypothetical finite mixture distribution composed of three run types over a finite time interval (e.g., week). The distribution of fork lengths (in millimeters [mm]) for each component is plotted as a curve and samples from each of the distributions are shown as vertical slashes on the x-axis. The composite distribution is shown in gray.

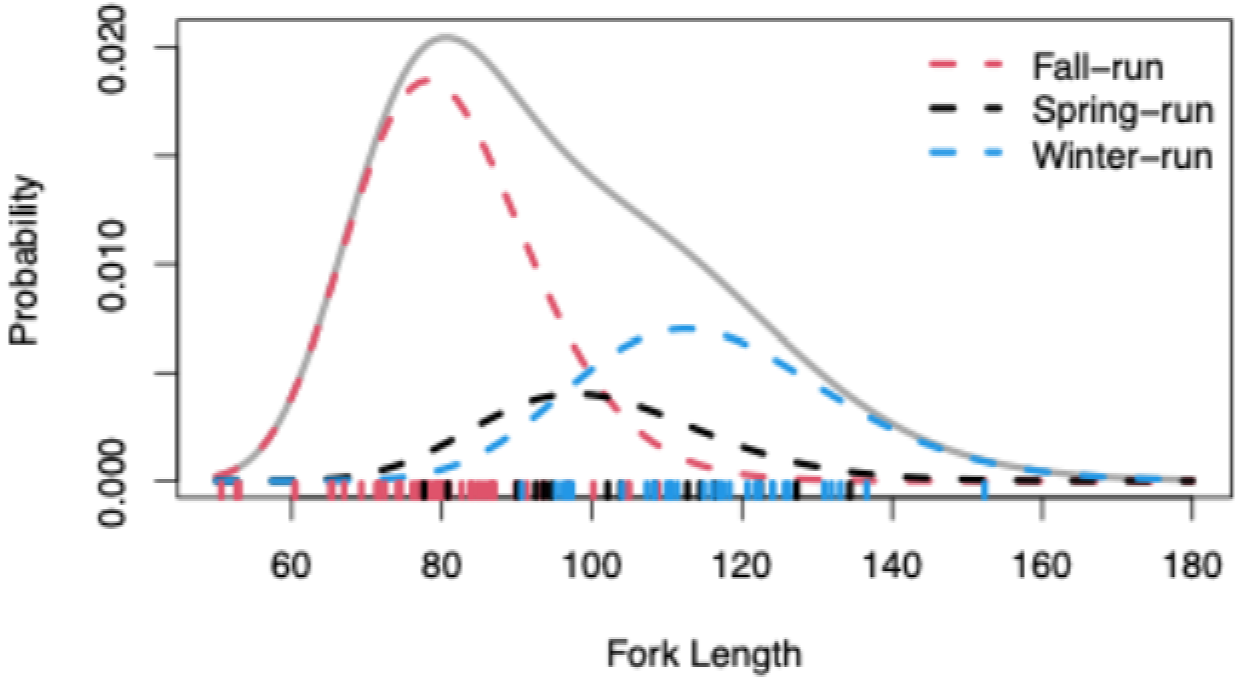


Figure 2. Hypothetical Finite Mixture Distributions with Variable Component Proportions

Hypothetical finite mixture distributions composed of three run types over a finite time interval (e.g., week) with variable component proportions. Equal proportions (top left), fall-run dominated (top right), spring-run dominated (bottom left) and winter-run dominated (bottom right). The distribution of fork lengths (in mm) for each run type is plotted as a curve and samples from each of the distributions are shown as vertical slashes on the x-axis. The composite distribution is shown in gray.

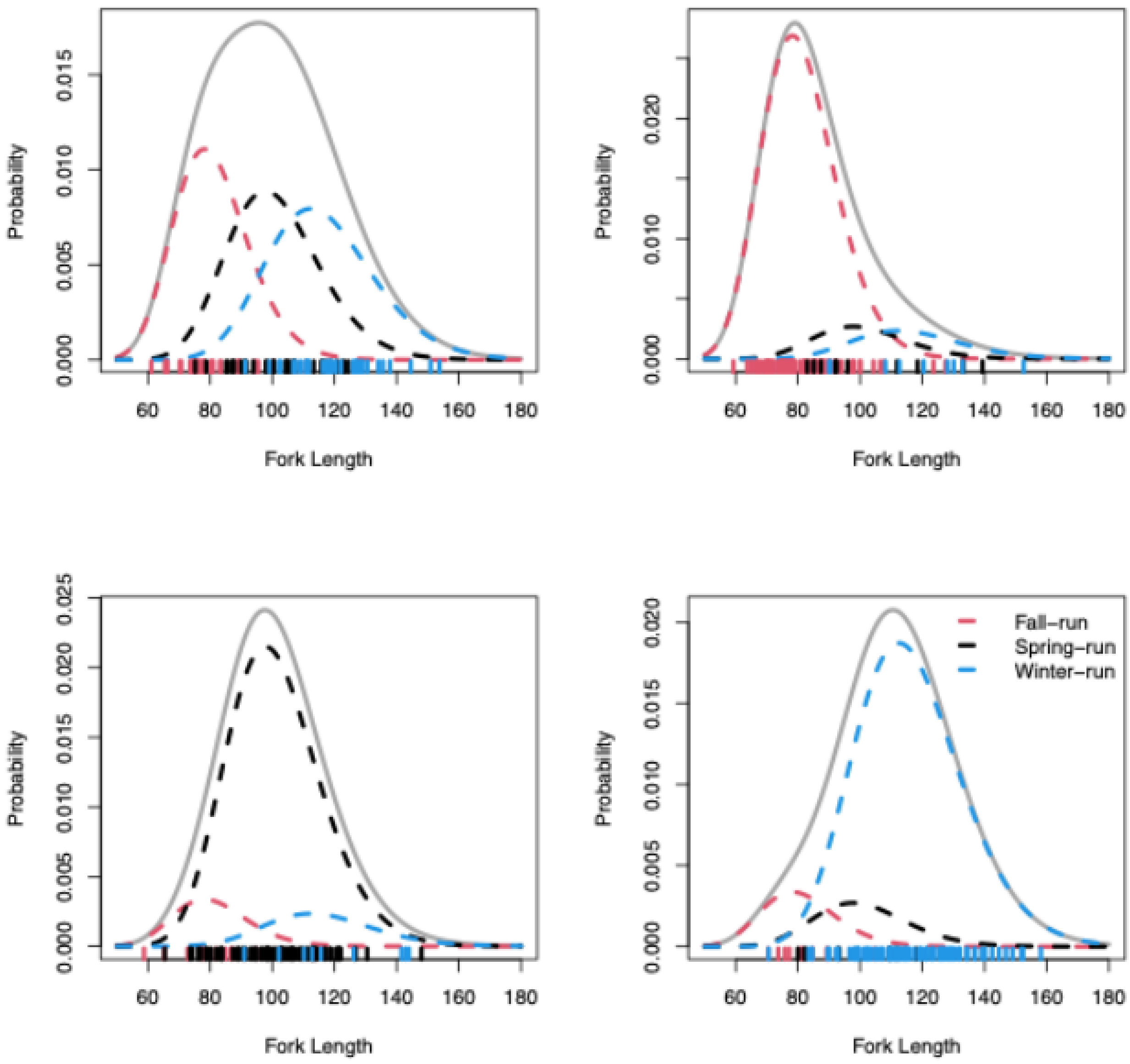


Figure 3. Hypothetical Variation in Proportions of the Three Generic Run Types

Hypothetical variation in the proportions (π) of each of three generic run types over time at a site. The values of π were generated by using Equations 6 and 7 with known values of $\delta_{0,1:2}$ $\delta_{1,1:2}$ and $\delta_{2,1:2}$.

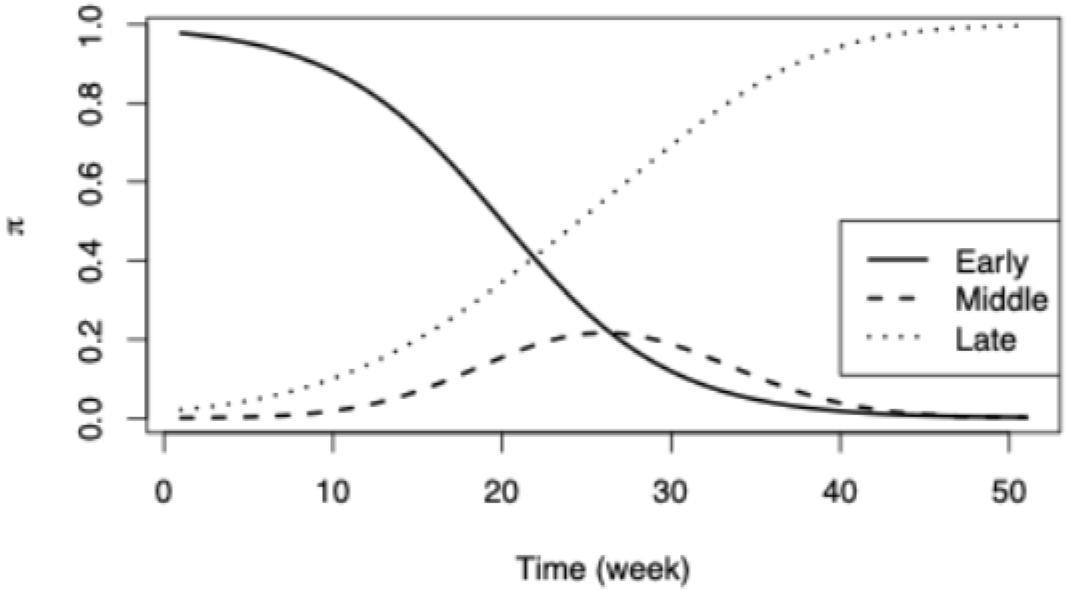


Figure 4. Genetic Samples Categorized by Run Type for Each Site

Genetic samples categorized by run type for each site. Site names are included in Table 1. Days are days since July 1, which was the date used to distinguish between brood years. Samples from multiple sample years are combined for each site.

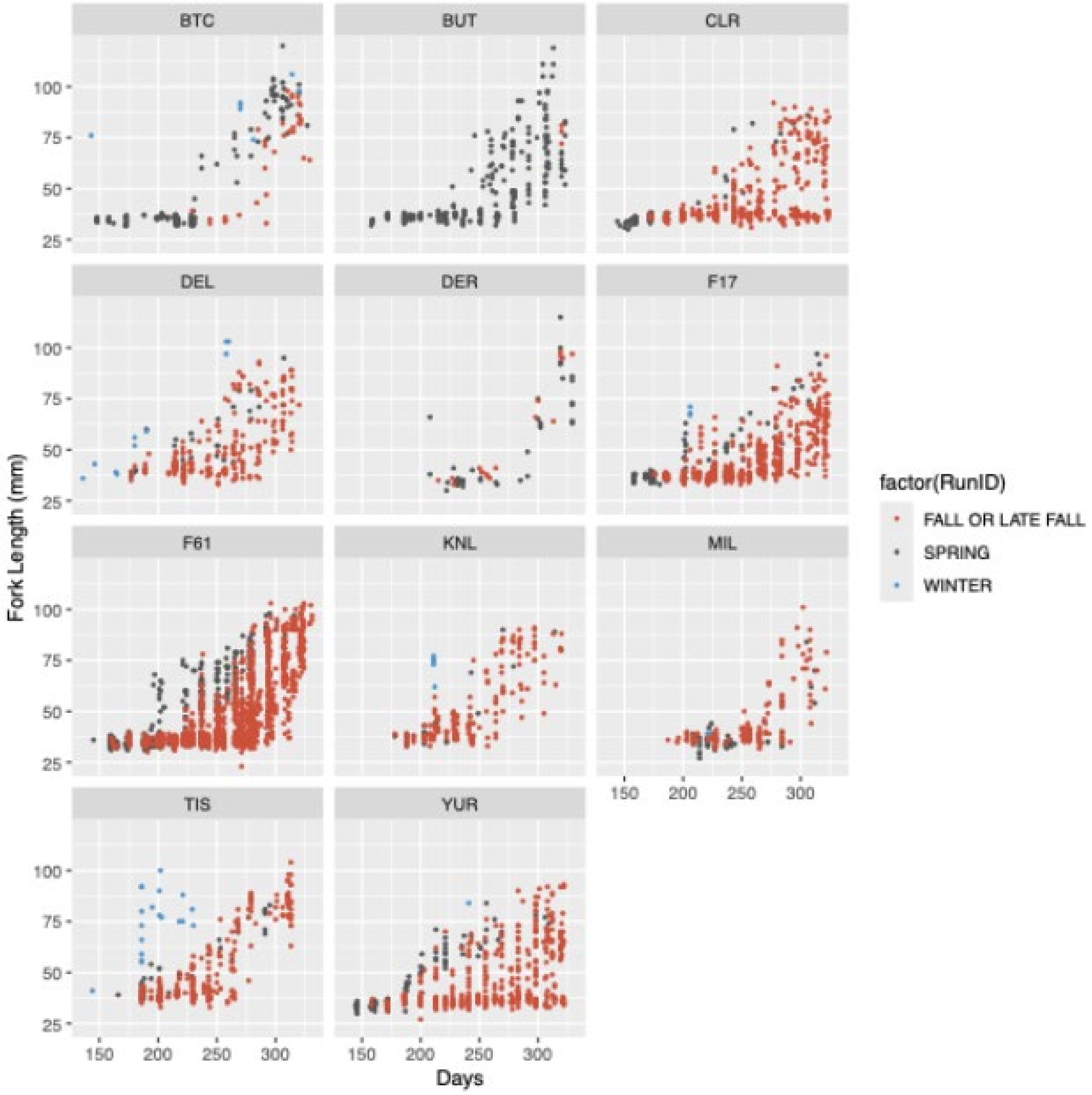


Figure 5. Runs of Juvenile Chinook Captures in Rotary Screw Traps and Identified to Run Type

Winter-run (blue), spring-run (black) and fall-run or late fall-run (red) juvenile Chinook salmon captured in rotary screw traps (RSTs) and genetically identified to run type. Spring-run circled in magenta and fall-run or late fall-run circled in aqua were identified as either yearlings (spring-run) or too large (fall or late fall) and were not included in the data set for estimating PLAD parameters.

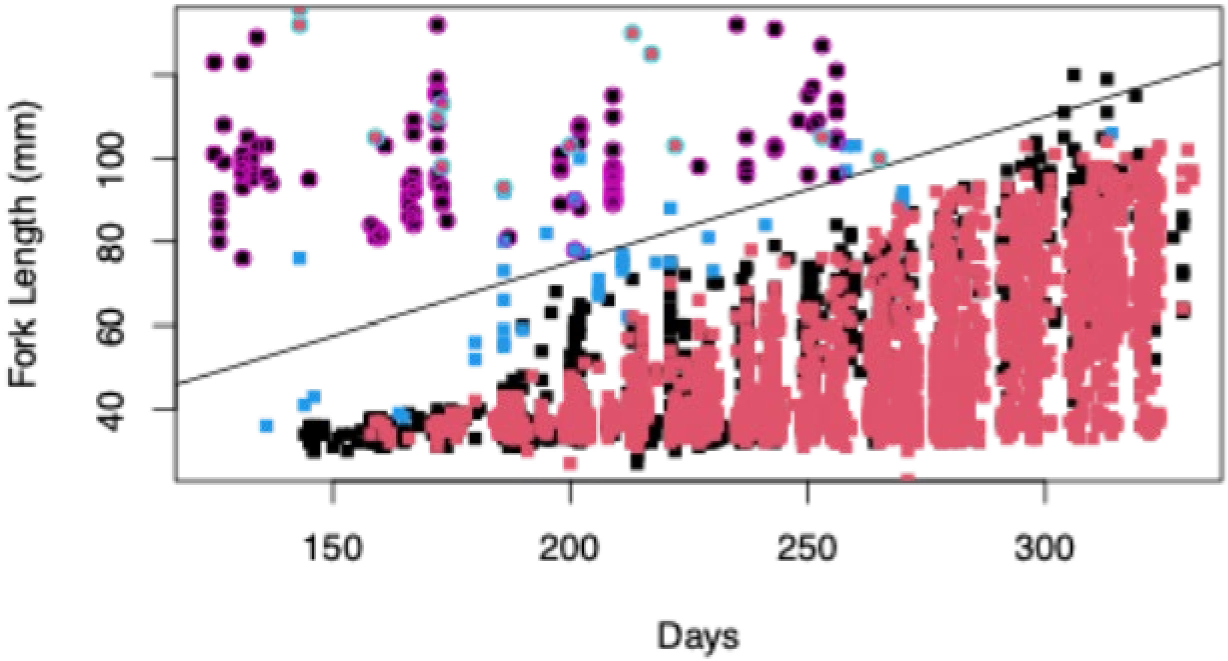


Figure 6. PLAD Model fits to Juvenile Chinook Salmon Captured in Rotary Screw Traps and Identified to Run Type

Fits of the PLAD model to juvenile Chinook salmon captured in RSTs and genetically identified to run type on Battle Creek. Spring-run (top left) in black, fall-run in red (top right), and winter-run in blue (bottom left). For each run type, the median (solid line), interquartile range (darker region), 95% credible intervals on predicted mean size (fork length in mm) (lighter region), 95% credible intervals on observed fish sizes (dotted lines), and observed sizes (squares) are plotted.

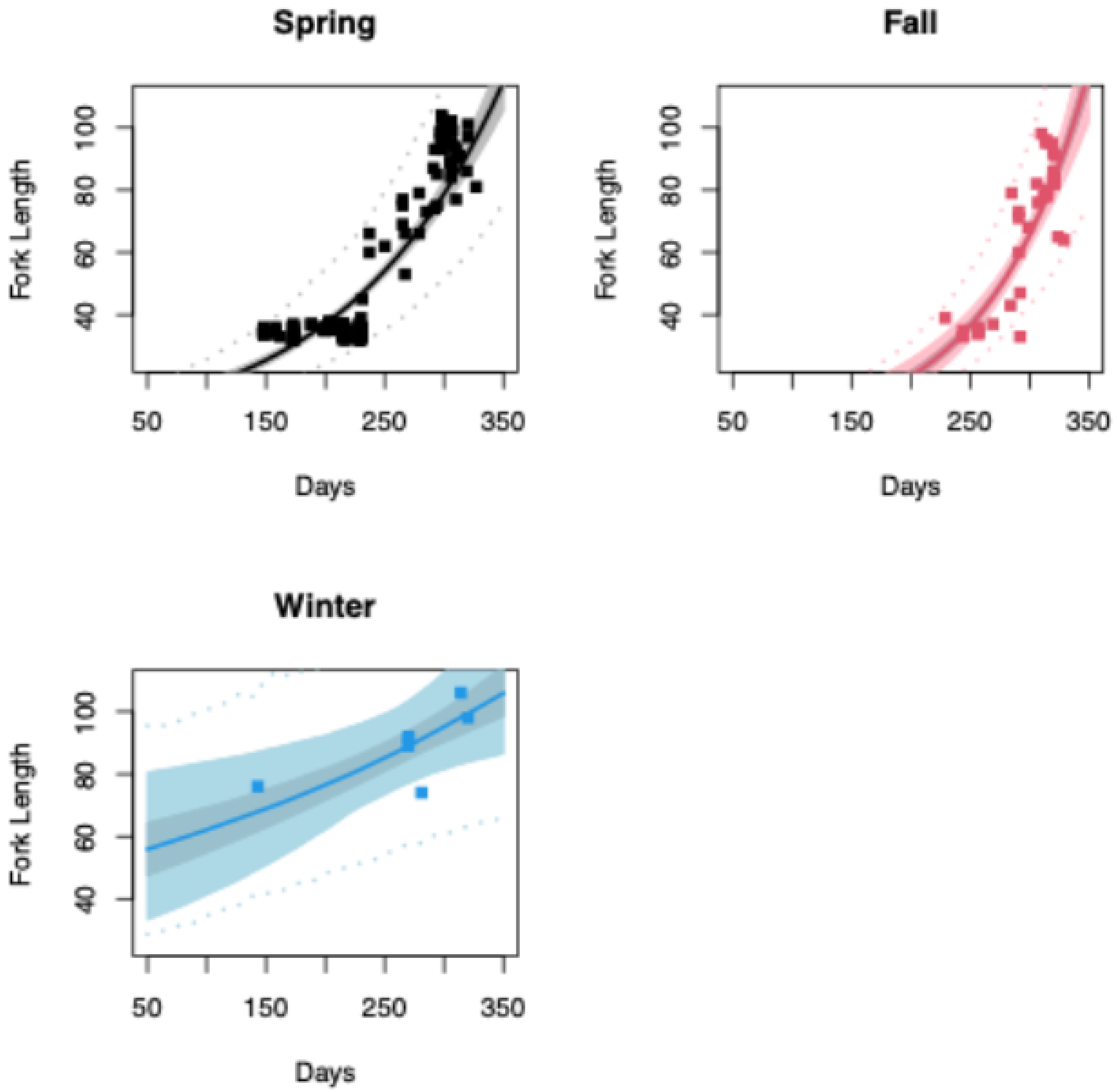


Figure 7. Predicted Proportions in Battle Creek Across the Sampling Season

Predicted proportions in Battle Creek across the sampling season (i.e., the π in the PLAD model). Spring-run (top left) in black, fall-run in red (top right), and winter-run in blue (bottom left). For each run type, the median (solid line), interquartile range (darker region) and 95% credible intervals are plotted.

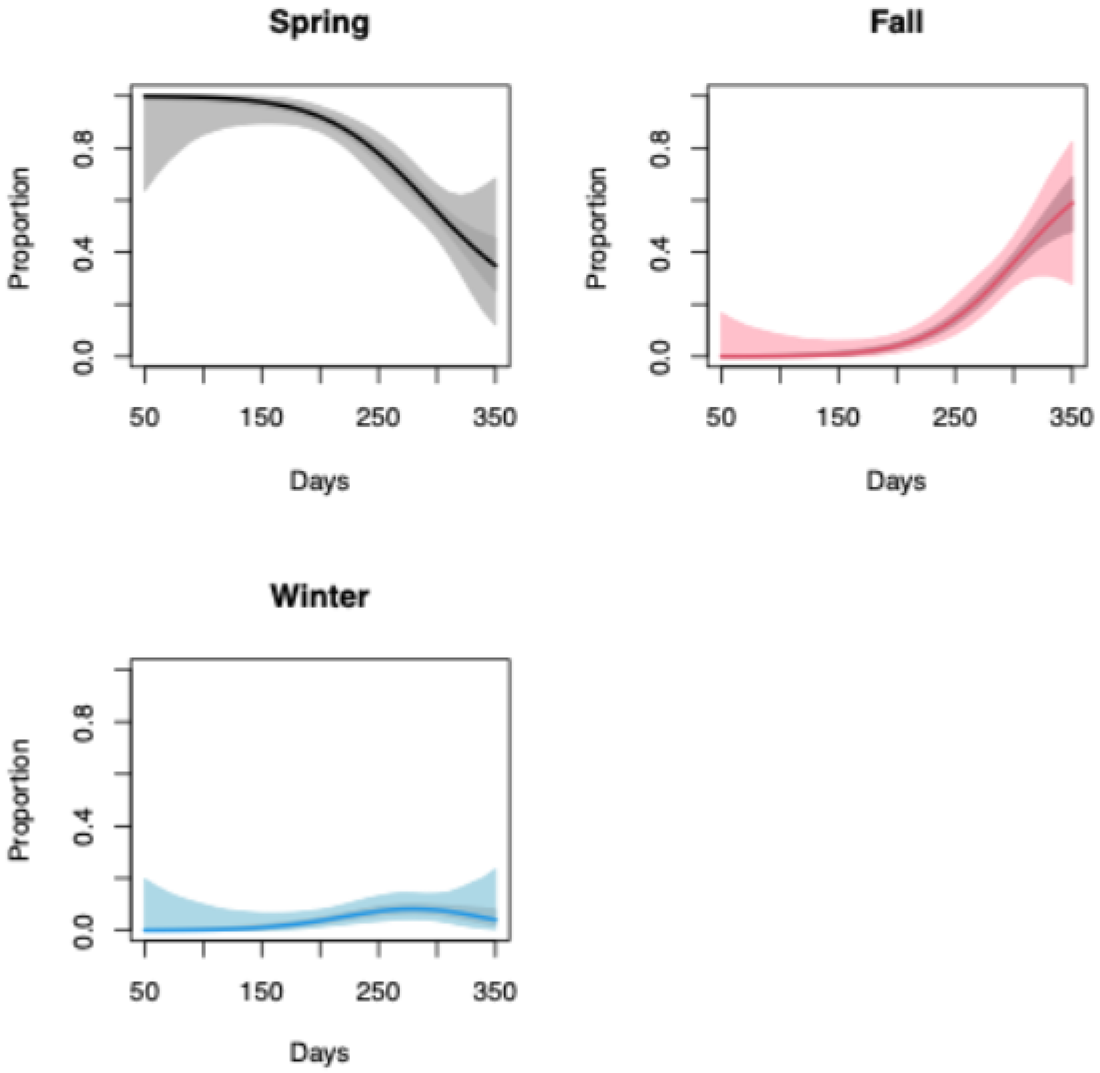


Figure 8. Prediction of Run Assignment by Fork Length at Four Dates in the Sampling Season at Battle Creek

Prediction of run assignment by fork length (mm) at four dates in the sampling season at Battle Creek:

- October 6 (Day 98, top left)
- December 15 (Day 168, top right)
- February 23 (Day 238, bottom left)
- May 3 (Day 308, bottom right)

Spring-run (black), fall-run (red) and winter-run (blue). Median probability of assignment (solid), interquartile range (darker region), and 95%CrI (lighter region) are presented for each run type.

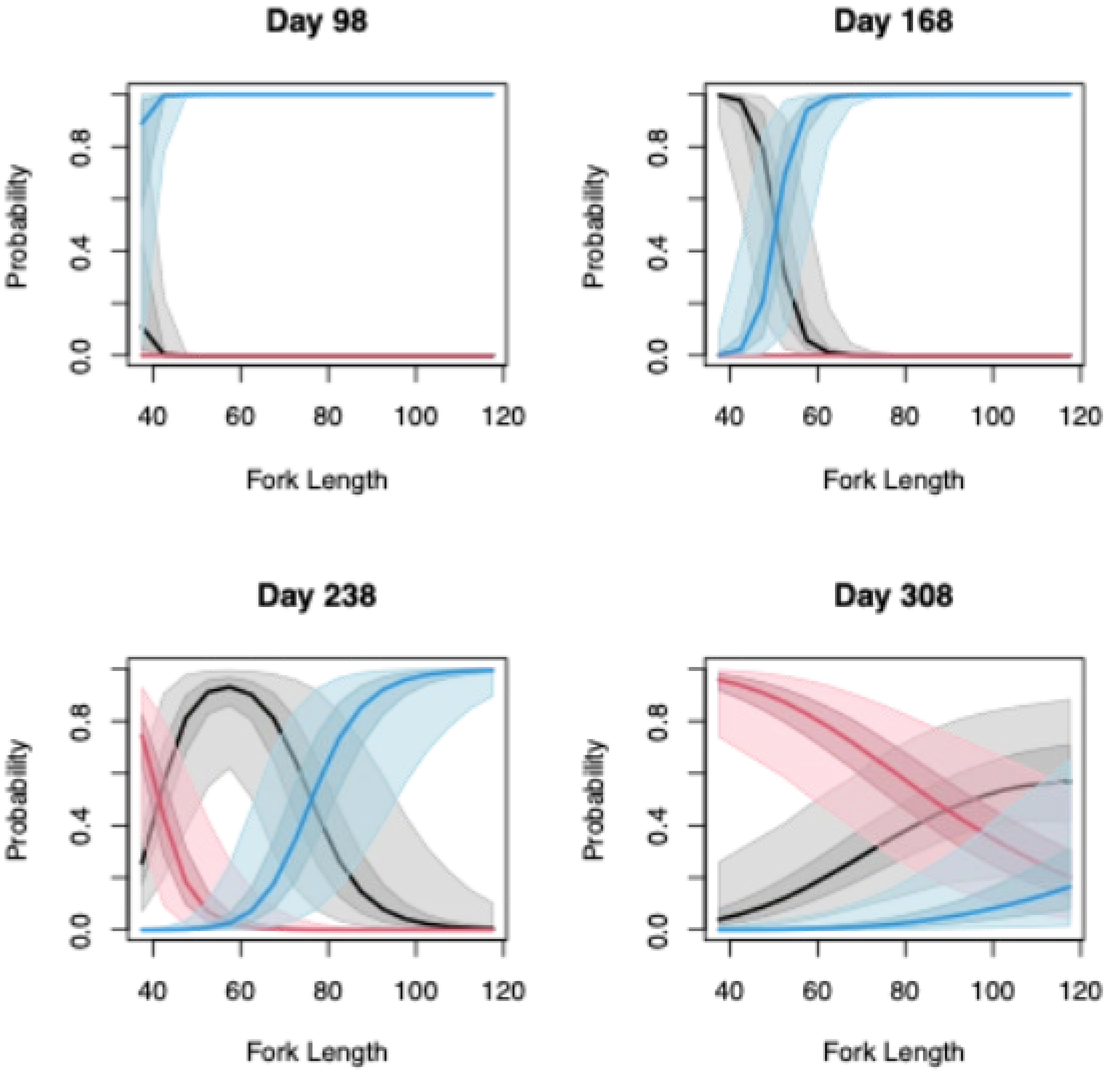


Figure 9. Battle Creek Juveniles Genetically Identified to Run Type during the 2025 Outmigration Season

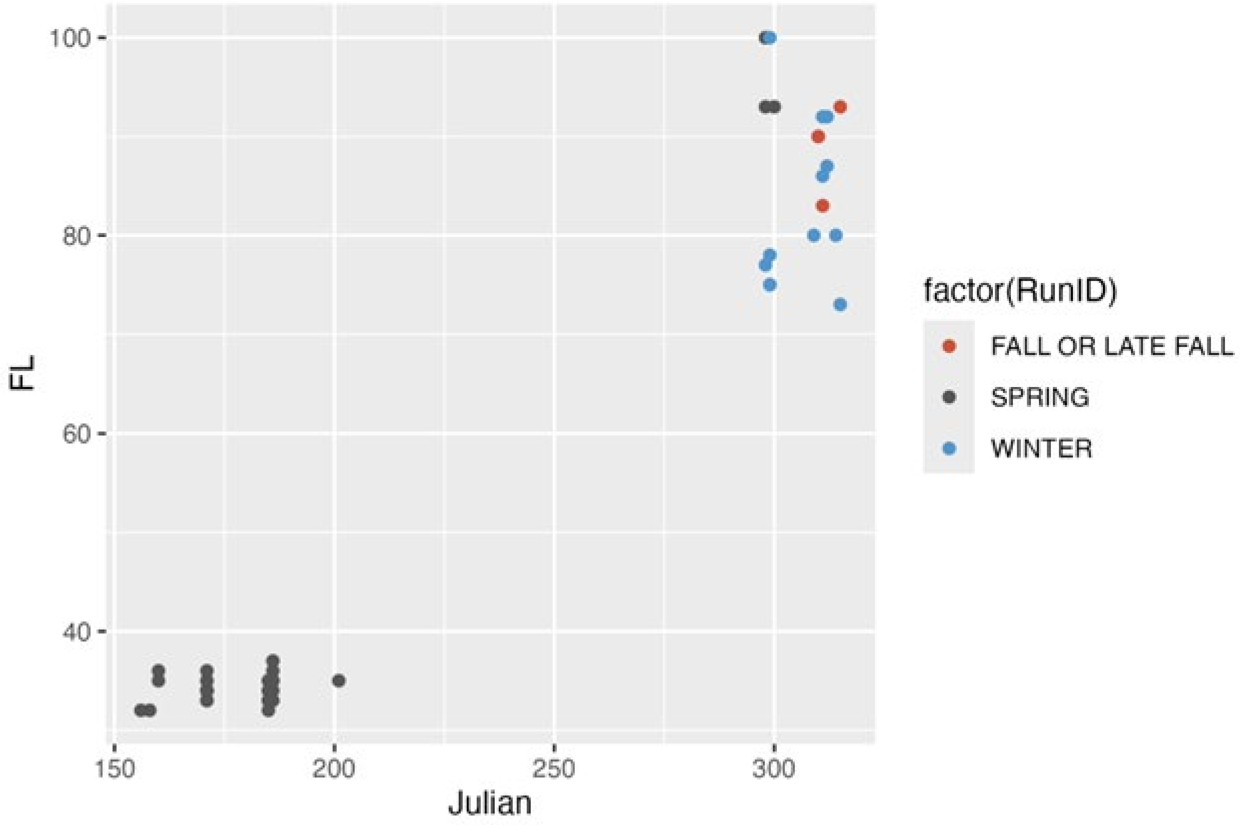


Figure 10. Proportion of Spring-run in Battle Creek

Proportion of spring-run in Battle Creek across all juvenile sizes (top left), fry-sized juveniles (less than or equal to 45mm, top right), and smolt-sized juveniles (greater than 45mm, lower left). Median proportion (line), interquartile range (dark gray), and 95%CrI (light gray) are plotted.

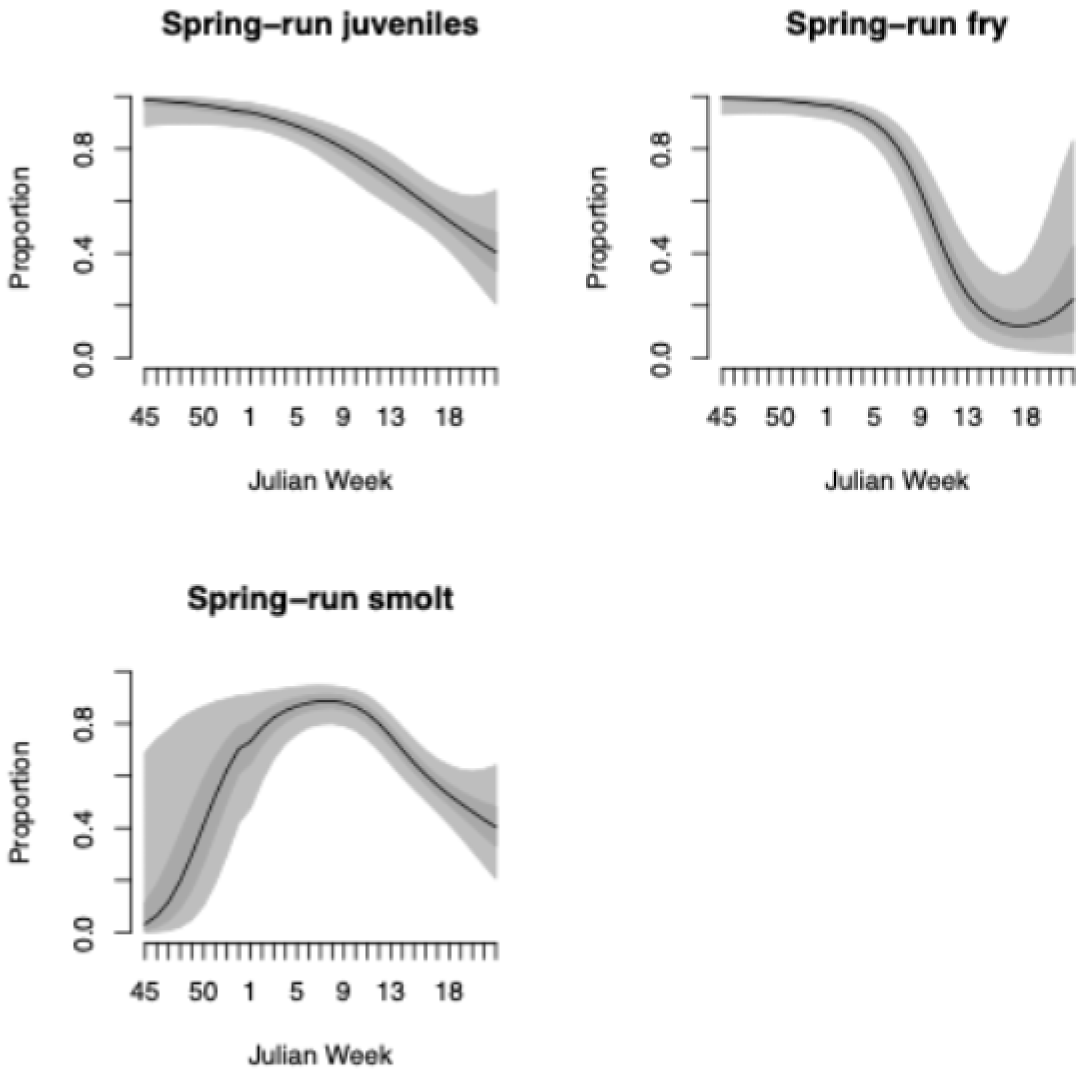
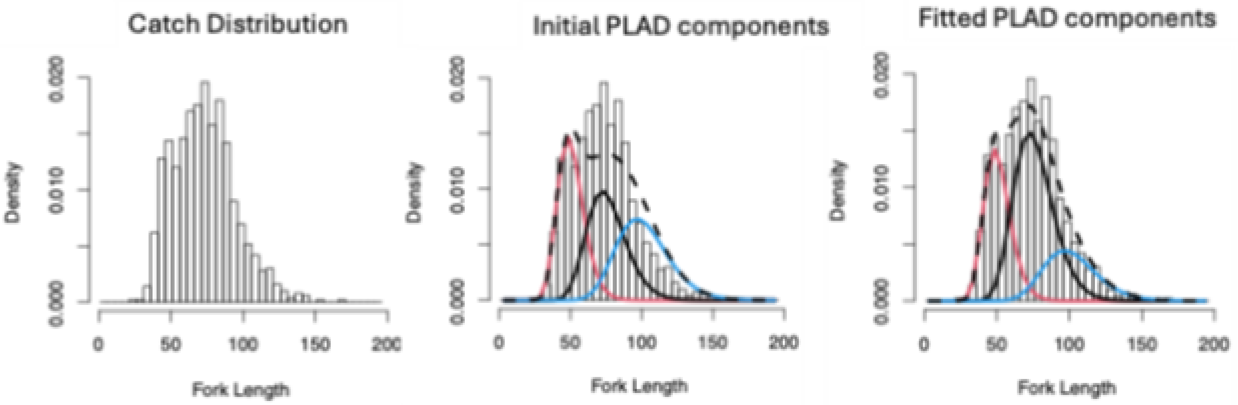


Figure 11. Hypothetical Catch Distribution, Initial PLAD Fit, and Updated PLAD Fit

Hypothetical catch distribution from a site (left). Initial PLAD fit to the fork length distribution (middle) and updated PLAD model fit after re-estimating the PLAD proportions (π)(right). Colors indicate fall-run (red), spring-run (black), and winter-run (blue).



Appendix

A. Probabilistic Length-at-Date Model Run Type Predictions for Sacramento River and Tributary Rotary Screw Trap Sites

Please refer to the file named “Ch 06 App A PLAD Predictions.docx.”