Data Management Strategy for the Spring-Run Chinook Salmon Juvenile Production Estimate

October 1, 2022

Authors

Brett Harvey & Peter Nelson — *California Department of Water Resources*

Sadie Gill, Ashley Vizek, & Erin Cain — *FlowWest*

Contributor

Brook Jacobs — *California Department of Fish and Wildlife*

Introduction

The Incidental Take Permit (ITP) issued to the California Department of Water Resources (DWR) by the California Department of Fish and Wildlife (CDFW) requires the development of a spring-run Chinook salmon juvenile production estimate (SR JPE). The SR JPE is a forecast of the number of juvenile spring-run Chinook salmon expected to enter the Delta each migration season. Efficient access to quality data and metadata is critical to the successful development of a method for calculating an annual SR JPE and then implementing it. The annual SR JPE will depend on over 40 individual data sources, including abundance estimates of multiple life-history stages spread across eight regions and several agencies and involving the cooperation and coordination of at least 20 data stewards and managers.

The Spring-Run Chinook Salmon Juvenile Production Estimate Interim Monitoring Plan (SR JPE Monitoring Plan) identified seven spring-run spawning streams in the Sacramento River valley whose monitoring data would be used to model the SR JPE. DWR and FlowWest formed a Data Management Team and conducted extensive outreach to the monitoring programs for these streams to design a data management system supporting method development and an annual SR JPE. The success of the Data Management Team, and ultimately the SR JPE, will depend on close alignment in data structure and availability and effective coordination among data managers and stewards across all regions.

This report summarizes the availability of historical data, data quality, and metadata documentation for the Sacramento River spring-run. Additionally, this report recommends a data management strategy for future data generated by monitoring efforts and special studies to meet the data requirements for annual production of a SR JPE. These requirements are centered on timely access to machine-readable monitoring data and metadata acquired automatically from third-party data repositories. To meet the annual JPE deadlines, partners involved in producing data for Central Valley Chinook Salmon and the SR JPE will need to provide current monitoring data to public repositories — which for some life-stage monitoring may require near real-time uploading of data. With technical support from the SR JPE Data Management Team, partners are already

engaged in transitioning existing data management practices to achieve these goals.

Recommended SR JPE Data Management Strategy

An annual SR JPE model will require timely and automated access to machine-readable monitoring data and metadata. The current variation across data and data sharing practices in salmon monitoring need to be more closely aligned to meet the needs of the SR JPE. The historic variation in approaches to storing and reporting monitoring data from individual tributaries or projects stems from the fact that each monitoring program was established in response to a specific identified question and data need. These management questions often arose independently from new monitoring requests and questions in neighboring tributaries. Under current data management practices, data is collected and analyzed by multiple personnel within agencies and external contractors, making it difficult at times for external partners not familiar with the history of individual monitoring programs to identify who is responsible for the data. Indeed, one of the most exciting and novel aspects of the SR JPE effort is that it will unify much of the existing monitoring and establish new monitoring under a common goal and purpose. In turn, the new data management system needs to align with this new unified goal and purpose. Outlined below are the elements of a data management strategy that will result in a public repository of high-quality monitoring data that is reported in near real-time in a standard format.

Monitoring Data Management Objectives

- Standardize data collection methodologies, schemas, encodings, and processing protocols.
- Produce machine-readable metadata for all monitoring programs.
- Upload data in near real-time to a shared data management system.
- Make data accessible via application programming interfaces (APIs).

Proposed below are the distinct roles, responsibilities, and technologies needed to achieve these data management objectives.

Roles and Responsibilities

Identifying the key personnel responsible for data for each spring-run monitoring program is essential for enabling efficient communication and sharing of monitoring datasets. This data management strategy outlines specific responsibilities for individuals and teams across the SR JPE science endeavor:

- **The Data Management Team:** consists of a Data Coordinator and the Stream Team Leads. The Data Management Team will compare Modeling Team requests with existing and proposed monitoring data sources and metadata resources to identify the tools needed to support Data Stewards in meeting the requirements of the SR JPE effort.
    - The Data Coordinator:
        - Acts as a liaison between the Stream Team leads, the Modeling Team, and the Data Management Team, and will be the individual most knowledgeable about the data management needs to support the JPE modeling effort.
        - Provides guidance and templates to produce machine-readable data and metadata.
        - Provides technical assistance establishing data upload workflows to public data repositories.
        - Establishes data pipelines ingesting data from public repositories into a database used to parameterize the forecast model.
        - The Data Coordinator will be a data management system specialist at or under contract to DWR.
    - Stream Team Leads:
        - Ensure that all monitoring datasets within their tributary are delivered in a timely manner and are appropriately formatted and well documented as specified by the Data Management Team and Modeling Team.
        - Identify a Data Steward for each monitoring dataset.

        - The broader role of Stream Team Leads is described in the SR JPE Science Plan.

- **The Modeling Team:** specifies to the Data Coordinator the desired data sources for model development as well as ongoing data inputs to the SR JPE forecasting model. The Modeling Team members will also specify the metadata requirements to inform model development, such as measurement methodology and measurement uncertainty. Additionally, the Modeling Team will describe a schedule for data acquisition that will allow sufficient lead time for SR JPE forecasting. The broader role of the Modeling Team is described in the SR JPE Science Plan.
- **The Data Steward:** should be the person in each monitoring program most knowledgeable about an individual dataset, including source, collection methodology, and data limitations. In most cases, there will be an individual assigned to this task for each spring-run tributary. The Data Steward will be responsible for:
    - Preparing metadata documentation.
    - Cleaning, editing, and preparing data into a machine-readable format.
    - Uploading data to a public repository at regular intervals.
    - Maintain interannual consistency of data publishing format.

Demonstration Data Management Systems

To enable timely updating and reporting of annual SR JPEs, all data management systems developed for the SR JPE will need to use cloud computing for data storage, access, and modeling. Use of cloud computing will also bring all programs contributing data for the SR JPE into compliance with the California State "Cloud First" policy that requires all agencies to shift new technology to cloud computing solutions (SAM 4983 and 4983.1; see references for California State Cloud Computing Policy web address).

Three data systems will be developed for SR JPE data management:

1. Run Identification Data Management System.
2. SR JPE Model Data Management System.

3. Pilot Centralized Data Management System for Chinook salmon juvenile outmigration data.

Each system will be composed of a database, data pipelines, servers, APIs, microservices, and secure websites with automated QA/QC for data entry. The systems will have standard data schemas, encodings, and processing protocols to efficiently deliver quality data. The data management applications will also demonstrate the value of centralized cloud-based data management of monitoring data for science-based decision-making processes in the Central Valley.

Public Data Repository

Currently, data requests to support salmon research and management are fulfilled via email, which requires slow and labor-intensive integration every time the system is updated with new data. This current system for reporting and sharing data, which was developed incrementally in response to individual needs and management questions on each tributary or monitoring location, will not support timely annual production of the SR JPE. In contrast, use of a public repository will allow entities to provide open and accessible data, while also allowing automated data pipelines which integrate and align new data into the existing database. These data will then allow automatic integration into the SR JPE model to allow near real-time updates to model forecasts as new data are collected and uploaded.

Table 1 Data, metadata, and repository features for the four common online locations hosting fisheries data in California

Features	CalFish	SacPAS	CNRA	EDI
Data public view access	YES	YES	YES	YES
Data public download access	YES	PARTIAL	YES	YES
Data machine readable	NO	PARTIAL	YES	YES
Metadata public view access	PARTIAL	PARTIAL	YES	YES
Metadata public download access	NO	NO	YES	YES

Features	CalFish	SacPAS	CNRA	EDI
Metadata machine readable	NO	NO	YES	YES
Repository robust search and discovery	NO	NO	PARTIAL	YES
Repository REST API or queryable urls	NO	PARTIAL	YES	YES
Repository data access documentation	NO	NO	PARTIAL	YES
Repository intuitive user interface	NO	NO	PARTIAL	YES

Note: Websites either fully, partially, or do not meet criteria for each feature. See references for repository web addresses.

Not all public data repositories are equal in their current level of accessibility and quality. Table 1 details four popular web resources, and their various features, for posting fisheries-related data in California. At minimum, the repository selected for hosting data for the SR JPE must provide machine-readable data and metadata and be accessible through a representational state transfer (REST) architectural style API or queryable urls.

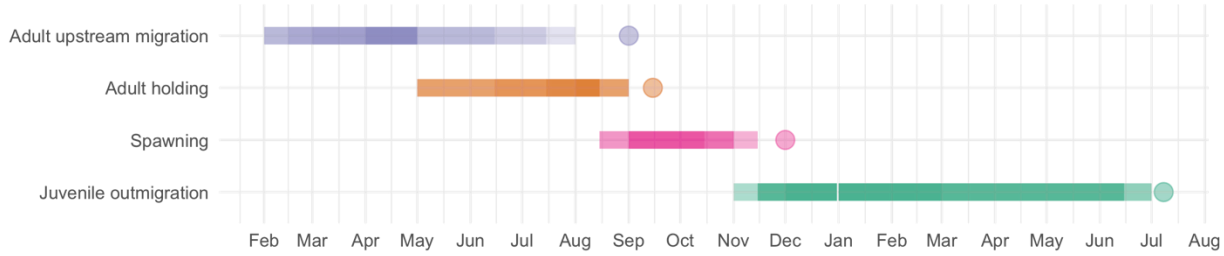
If a monitoring program has a compelling reason for why they are unable to post data to a public repository, the monitoring program can submit their dataset and metadata to the SR JPE Data Coordinator in a mutually agreed upon machine-readable format. The SR JPE Data Coordinator will then store this data within the database that feeds the forecast model.

Current Status of Monitoring Data Collection and Data Management

Salmon monitoring data that will power the SR JPE are collected in the Lower Sacramento River and seven Sacramento River tributaries: Battle Creek, Butte Creek, Clear Creek, Deer Creek, Mill Creek, Feather River, and Yuba River. There are six types of monitoring data: Adult Upstream Passage, Adult Holding Surveys, Redd Surveys, Carcass Surveys, Juvenile Emigration (RST), and Survival Studies. Data from each spring-run producing tributary are processed, analyzed, and stored by the monitoring program that collected the data, rather than by centralized hubs within each agency. In some tributaries, multiple agencies and entities collect data of different types and store that information individually. The development and implementation of the SR JPE will require the collaboration of these multiple monitoring efforts and the alignment of data availability, quality, and formatting across monitoring programs.

Figure 1 and *Table 2* show the relationship between data collection, data availability, and the date when the ITP requires DWR to produce an annual SR JPE. Currently, salmon monitoring programs operate independently from each other, and the data life cycle from collection to dissemination is not coordinated across watersheds, resulting in different monitoring programs having different data elements, encodings, and quality assurance/quality control practices and standards. The sections below summarize the timing of the monitoring of data collection and dissemination and the state of the metadata across the tributaries and monitoring programs. Integrating the data management objectives discussed above into existing monitoring programs would generally strengthen their contribution to the SR JPE and Central Valley salmonid management.

Figure 1 Timing of data collection (bars) and data availability (dots) for different spring-run life stages.



Note: Color gradation in bars indicates variability across watersheds.

Table 2 Range of dates across monitoring programs when data is collected for each life stage and the data availability lag times shown as the most common lag time with the range in parentheses

Life Stage	Monitoring Period	Data Availability Lag
Adult upstream Migration	February–August	1 month (1 month–1 year)
Adult holding	May–September	2 weeks
Spawning	August–November	2 weeks
Juvenile outmigration	November–June	1 week (1 week–6 months)

Table 3 Status of monitoring by data type and watershed

Data Type	Battle Creek	Clear Creek	Mill Creek	Deer Creek	Butte Creek	Feather River	Yuba River	Lower Sacramento River
Adult Upstream Passage	Ongoing	Ongoing	Ongoing	Ongoing	Ongoing	Historical or Proposed	Ongoing	NA
Adult Holding Survey	Ongoing	Ongoing	None	Ongoing	Ongoing	None	None	NA
Redd Survey	Ongoing	Ongoing	Ongoing	None	None	Ongoing	See Note	NA
Carcass Survey	Ongoing	Ongoing	None	None	Ongoing	Ongoing	See Note	NA
Rotary Screw Trap	Ongoing	Ongoing	Historical or Proposed	Historical or Proposed	Ongoing	Ongoing	Historical or Proposed	Ongoing
Survival Study	None	None	None	None	Historical or Proposed	Historical or Proposed	Historical or Proposed	Historical or Proposed

Note: Unable to obtain approval from Army Corps to access Yuba River redd and carcass data.

Adult Monitoring

Adult Passage Monitoring

Adult upstream passage data are currently collected in six tributaries within the SR JPE, with plans to expand to all seven tributaries included in the SR JPE Monitoring Plan by 2022 (*Table 3*). Adult passage data are primarily collected using video monitoring systems, and some tributaries supplement this information with additional trapping data. The timing between when data are collected and when they are made available varies from one month to one year depending on data processing resources, staffing constraints, and program priorities.

Stream Teams from the seven spring-run tributaries monitored for the SR JPE manage and store their data in Microsoft Access Databases or Excel workbooks on individual computers. Many programs currently share passage data via pdf files, which are not machine readable. Several different methodologies are used across tributaries to calculate passage estimates from recording video data. Data attributes, the naming of attributes, and attribute encodings are not consistent across the adult passage monitoring programs.

Adult Holding, Redd, & Carcass Surveys

Data collection methodologies for adult holding, redd, and carcass surveys vary among tributaries in response to the specific historic management questions and staff resource availability in each location. For example, the frequency of sampling varies; some tributaries conduct a single survey annually while others survey consistently throughout the monitoring season. The percentage of river area sampled also varies by monitoring program, and in some cases, these data are collected only as part of post-project monitoring efforts. Additionally, some adult surveys are not collected systematically, only opportunistically while crews are conducting other monitoring activities.

Data from holding, redd, and carcass surveys are typically available one week to one month after survey completion (*Table 2*), and like other data types, are tailored to tributary-specific questions and staff resource availability. As a result, there are inconsistencies among tributaries in which data attributes are collected, the naming of similar attributes, and their

encodings. Some tributaries record only count data for these datasets, while other tributaries collect in-depth measurements and environmental data. The metadata provided are not complete across each monitoring program, and descriptions of the data attributes and encodings are unclear or missing.

There is not a shared data management system for adult holding, redd, and carcass surveys across tributaries. Stream Teams store and share these data in Microsoft Access Databases and Excel workbooks. These databases are stored primarily on individual computers, many using agency specific cloud storage backup systems that are difficult to access by most potential data users. The SR JPE Data Management Strategy will ensure that all data is backed up and secure while also providing public accessibility that will save regional programs the effort often required to make databases available for each request.

Juvenile Monitoring

Rotary Screw Trap

In 1997, the Comprehensive Assessment and Monitoring Program (CAMP) was implemented. CAMP was designed to provide a unified protocol for rotary screw trap (RST) data collection and reporting (U.S. Fish and Wildlife Service 2008). The CAMP platform consists of a Microsoft Access database, a desktop user interface to support data entry, and statistical modules for data analysis (i.e., passage estimates). The platform is a template that can be used and maintained locally by each tributary monitoring program for RST data management and analysis. The CAMP platform has provided an important first step in streamlining salmon monitoring data across watersheds. Though, to better support the unified purpose of the SR JPE program, a database is needed that can ingest standardized data from watersheds. To do this, the new SR JPE database will require consistent data collection, naming conventions, encodings, and quality assurance/quality control practices and standards across programs.

There are historic RST data within all watersheds selected for SR JPE monitoring (Table 3). RST monitoring efforts on Deer Creek and Mill Creek were discontinued in 2010 and then resumed in 2022 for the SR JPE Science Program. Since 2009, there has been no RST monitoring on the Yuba River, though resumption is planned in 2023 for the SR JPE Science Program.

Although most RST monitoring programs use some elements of the CAMP platform, there are large variations in the resulting data structures and encodings as well as metadata availability and completeness. There are also differences in how trap efficiency is being evaluated and juvenile passage estimates are being calculated as well as the environmental covariates being collected. The lag between when data are collected and when they become available also varies from one week to six months across watersheds (Table 2).

Survival Studies

Historical survival studies were conducted on the Feather River and Butte Creek as part of the Central Valley Enhanced Acoustic Tagging Project. This survival data was collected using JSAT acoustic tagging. These data are available on Environmental Research Division's Data Access Program (ERDDAP), a public National Oceanic and Atmospheric Administration (NOAA) data repository (see references for repository web address). After data acquisition, all survival data will require further QA/QC before use in the SR JPE to remove false detections and address other data quality issues. This time-intensive process involves tracing each individual tag to eliminate erroneous data points.

There are two survival studies from 2021 on the Yuba River and Butte Creek, though data from these studies is not yet available for review or use. There will be one survival study on the Sacramento River conducted in 2022. The SR JPE Science Plan outlines additional survival studies, but the dates and locations have not been planned as of the writing of this document. Data produced by new studies is unlikely to be consistent with historical data in structure and will need preparation before use for SR JPE modeling.

Conclusion

Sacramento River valley salmonid monitoring programs are managed by many agencies, involving at least 20 individual data managers and stewards across the valley. These programs were designed and continue to fulfill the goals and objectives for their tributary- or region-specific management questions and resource and staffing availabilities. As such, there is variability in data structure, alignment, and availability, requiring labor-intensive data processing for every new analysis effort. To best integrate into broader regional analysis and synthesis, such as the SR JPE program, consistency is needed in the quality and availability of data collected and the data schema used to record and report that data. The strategy outlined in this plan provides a roadmap to achieving this objective.

The CAMP platform was implemented with the goal of solving many of these issues for RST monitoring data to allow for effective and coordinated regional analysis. The data management strategy recommended in this report builds upon the foundation for data streamlining that CAMP was designed to provide and aims to develop a public repository of high-quality monitoring data in a standard format. An integral and fundamental component of the recommended data management strategy is to develop a coordinated system for data management and reporting across the multiple entities and individuals involved in collecting and managing data. Clear and effective communication and collaboration spanning spatial and institutional boundaries will be critical to producing an annual SR JPE. While this framework will be developed for the purpose of the SR JPE, its utility will support numerous other existing and future analysis and modeling needs at both the regional and valley-wide scale for all Central Valley salmonids.

Table 4 List of addresses for data repositories discussed in this document

Website	Web Address
CalFISH public data repository website	https://www.calfish.org
CNRA public data repository website	https://data.cnra.ca.gov/
EDI website	https://environmentaldatainitiative.org/
NOAA public data repository website	https://oceanview.pfeg.noaa.gov/erddap/tabledap/FED_JSATS_detects.html
SacPAS website	http://www.cbr.washington.edu/sacramento/
California State Cloud Computing Policy	https://www.dgs.ca.gov/PD/Resources/Page-Content/Procurement-Division-Resources-List-Folder/Cloud-Computing-Policy

Reference

U.S. Fish and Wildlife Service (USFWS). 2008. Draft rotary screw trap protocol for estimating production of juvenile Chinook salmon. Document prepared by the U.S. Fish and Wildlife Service, Comprehensive Assessment and Monitoring Program. Sacramento, California. 44 pp.