Technical Memorandum

Date: March 30, 2018

To: Demian Ebert, PacifiCorp

From: Jennifer Vaughn, Watercourse Engineering, Inc.
Mike Deas, Watercourse Engineering, Inc.

Re: Quality Assurance Review of KHSA 2016 dataset

Suggested Citation: Vaughn, J., M. Deas. 2018. Technical Memorandum Re: QA Review of KHSA 2016 dataset. March 30. 4 pp.

1. Introduction

A quality assurance (QA) review of the Klamath Hydroelectric Settlement Agreement (KHSA) 2016 dataset was conducted by Watercourse Engineering, Inc. (Watercourse) by comparing the dataset values of randomly selected Sample ID with the values of those Sample ID results in the original lab reports. Sample IDs were selected from the Baseline Monitoring Program (Baseline) general dataset, the Baseline Algae Species dataset, and the Public Health Monitoring Program (Public Health). The review was accomplished with the cooperation of field crews and stakeholders involved in the KHSA sampling programs, including PacifiCorp, the Karuk Tribe, Yurok Tribe, U.S. Bureau of Reclamation - Klamath Falls office, and E&S Environmental Chemistry, Inc.

2. Review Methods

The 2016 KHSA data review followed the same methods of the review of the 2009-2015 KHSA datasets and the QA review method for each monitoring program was accomplished using the same method. For each program, one percent of the Sample ID values in the 2016 KHSA dataset were compared with values in the original lab reports provided to stakeholders by the respective labs during the course of the sampling programs. To select samples for review, the number of Sample IDs in 2016 was counted and each Sample ID was assigned an integer as a reference number. One percent of the total number of Sample IDs in 2016 was then calculated and rounded to the nearest integer. Using the Random function in MS Excel (RANDBETWEEN), a random value was generated between 1 and the total number of Sample IDs for 2016. A random value was generated again until the total number of random integers equaled one percent of the total number of Sample IDs in 2016. The random values were then used to choose random Sample IDs, selecting those Sample IDs with integer reference numbers matching the randomly generated values.

Once the Baseline general dataset, Baseline algae species dataset, and the Public Health dataset had a sufficient number of randomly selected Sample IDs to total one percent of the total number of Sample IDs for each dataset, the appropriate stakeholders and field crews were contacted and all original lab reports associated with the randomly selected Sample IDs were requested. Original lab reports included pdf versions of paper lab reports, MS Excel files of paper lab reports, or MS Excel files of digital lab data. Sample IDs and their associated lab reports were documented by Watercourse.

When the original lab reports were obtained, a comparison of each possible result for each Sample ID was carried out. In the Baseline general dataset, there were a total of 19 constituents possible for each Sample ID, and in the Public Health dataset, there were a total of 26 possible results (counting each toxic algae species being tracked in the dataset). In the Baseline algae species dataset, the number of possible comparisons was calculated as the number of fields in the pdf original lab reports (seven) multiplied by the total number of species identified for all selected Sample IDs. Each possible result was examined to determine if the KHSA dataset value matched the original lab report value. If a Sample ID did not have a value for a specific constituent within the dataset, Watercourse confirmed that constituent had not been analyzed by any lab associated with that Sample ID using lab reports obtained from the appropriate stakeholders or field crew.

If the result in the KHSA dataset did not match the result in the original lab reports, that result was flagged. If the values did not match, the error was labeled a “true” error. If the value was rounded from the original lab report result, the error was labeled as a “significant figure issue.” Error rates were calculated as the percentage of results that did not match the original lab report values. Error rates were calculated for: (a) all non-matches; (b) only true errors; and, (c) only significant figure issues. The accuracy of the datasets for the Baseline general dataset, the Baseline algae species dataset, and Public Health programs were calculated as the percentage of results that matched the original lab report values, ignoring significant figure issues (i.e., only true errors).

3. Types of Errors and Issues

There are several types of true errors that can occur within a large dataset, but the most common two are transcription errors and omission errors. Transcription errors occur when there was a value entered into a dataset that clearly did not match the value in the original lab report. Omission errors occur when a value was not included in the dataset that should have been. Other errors in the dataset can include a result added to the dataset that should have been omitted, or a value assigned to a wrong Sample ID.

The issue of significant figures was also included in the QA review of the KHSA dataset. At this time, there has been no decision on the number of significant figures that should be included in the KHSA dataset for each constituent. For this review, if the dataset value was rounded from the original lab-reported value, that value was flagged.

4. Possible Sources of Errors

As a further investigation into the dataset accuracy, Watercourse investigated the possible source of error or significant figure issue for each flagged value. This was done by using the compilation spreadsheets and other files provided to Watercourse by stakeholders that had been used to create the annual KHSA datasets and reports. Because of this, original lab reports could be compared to the KHSA dataset files to determine when an error or significant figure issue most likely had been introduced into the dataset. The number and percentages of each of those possible sources of each error was calculated and documented.

5. Review Results

For the Baseline general dataset, there were a total of 76 possible results reviewed, consisting of 19 possible constituents for each of the 4 randomly selected Sample IDs from 2016 (Table 1). Though the 231 sample IDs included in the Baseline algae species provided only 2 randomly selected Sample IDs for review, 2 of the Baseline general Sample IDs that were randomly selected also had algae species data that was included in the review. These 4 randomly selected Sample IDs resulted in review of 7 fields for each of the 91 total species identifications generating a total of 637 reviewed results for the Baseline algae species dataset (Table 1). A total of 84 possible results were reviewed for the Public Health dataset, consisting of 28 fields for each of the 3 randomly selected Sample IDs from 2016 (Table 1).

The QA review results found that the Baseline general dataset, the Baseline algae species dataset, and the Public Health dataset all had an estimated accuracy of 100 percent. If significant figure issues were included, the accuracy of the Baseline general dataset dropped to 97 percent and the Public Health accuracy dropped to 96 percent (Table 2). The Baseline algae species dataset did not have significant figure issues and its accuracy remained 100 percent.

The source of significant figure issues for the Baseline general dataset was the sampling entity, which had rounded two values in the data that was submitted to Watercourse for use in creating the 2016 KHSA dataset (Table 3). The Public Health significant figure issues were generated by the sampling entity using a spreadsheet-style lab report issued to the sampling entity for data entry but not provided to Watercourse as part of this review. The original lab reports that were provided to Watercourse for the purpose of this QA review were pdf reports that did not contain decimal places for toxic algae counts, while the sampling entity had submitted data to Watercourse for the creation of the 2016 KHSA dataset that did contain decimals and had been taken from the spreadsheet-style lab report (Table 3).

Table 1. KHSA 2016 Possible Results Reviewed by Program.

Program Part	Baseline – General	Baseline – Algae Species	Public Health
Number of Constituents	19	7	28
Number of Sample IDs	395	231	278
Number of Reviewed Sample IDs	4	4	3
Number of possible results examined during QA review	76	637	84

Table 2. KHSA 2016 Accuracy Estimates.

	Baseline – General		Baseline - Algae Species		Public Health	
Number of possible results	76		637		84	
Number of results that were not exact matches	2	(3%)	0	(0%)	3	(4%)
Number of result non-matches that were significant figure issues	2	(3%)	0	(0%)	3	(4%)
Number of results that were true errors	0	(0%)	0	(0%)	0	(0%)
Estimated Accuracy of dataset	**100%**		**100%**		**100%**	

Table 3. KHSA 2016 Sources of Significant Figure Issues.

	Baseline - General		Baseline - Algae Species		Public Health	
Number of significant figure issues	2		0		3	
Significant figure issue, unknown source	0	(0%)	0	(0%)	0	(0%)
Significant figure issue, introduced by sampling entity	2	(100%)	0	(0%)	0	(0%)
Significant figure issues, introduced by Watercourse	0	(0%)	0	(0%)	0	(0%)
Significant figure issues, other sources	0	(0%)	0	(0%)	3	(100%)