Assessing Spatial Representation of Large-Scale Environmental Datasets

A Case Study of Dissolved Oxygen Sampling in Utah's Bonneville Cutthroat Trout Range

Greg Goodrum
Graduate Research Assistant, Department of Watershed Science, Utah State University 8200 Old Main Hill, Logan, Utah 84322 (goodrum.greg@gmail.com)

Submitted: December 7th, 2018
Course: CEE6440 – GIS in Water Resources
Instructor: Dr. David Tarboton

Table of Contents

1. INTRODUCTION

While Utah contains the majority of North America's Bonneville cutthroat trout (*Oncorhynchus clarkii utah*) habitat, human alterations to their native stream environments have led to significant declines in fish populations and their listing as a Species of Special Concern within the state (Utah Department of Wildlife Resources, 2008). To understand and manage these environmental impacts, biologists employ habitat suitability models to quantify the degree to which environmental conditions meet the requirements for a given species (de Kerckhove et al, 2008). For cutthroat trout, environmental conditions include limiting factors such as stream temperature, gravel permeability, and dissolved oxygen (Hickman and Raleigh, 1982). When combined in a habitat suitability model, the results offer a numeric prediction of a given stream segment's ability to support trout.

For managers dealing with range-wide population declines, assessing the impacts of habitat degradation requires applying habitat suitability models at large spatial scales. However, in order for aquatic habitat models to effectively represent aquatic habitats, the input data must capture the full variability of stream conditions within the scope of the project. While large spatial scale datasets are increasingly abundant and available, many are composed of data collected by multiple researchers or agencies with different research objectives. The result is that, while abundant and available, these datasets do not guarantee the capture of stream segment variability required for effective habitat suitability modeling.

In order to determine whether large spatial scale environmental datasets can be used with aquatic habitat models, they must capture the varied attributes of stream segments within the scope of the model. In this report, I address this question by focusing on a single environmental condition important to Bonneville cutthroat trout, dissolved oxygen, and using GIS to examine how currently-available dissolved oxygen data (Figure 1) capture the variability of stream segments within their current range.

2. PROJECT OBJECTIVE

This project addresses whether currently-available dissolved oxygen data can be used to model stream habitat suitability for Bonneville cutthroat trout. To answer this question, I divided this objective into three goals. The first goal is to understand how existing dissolved oxygen data captures the diversity of physical attributes that define stream segments within the habitat range of Bonneville cutthroat trout. The second goal is to determine which stream segment attributes have no representative samples in the data. Finally, the third goal is to identify where new monitoring locations can be added to ensure a complete representation of stream segment attribute variability within the range of Bonneville cutthroat trout in Utah.

3. METHODS

3.1 Study Area

Bonneville cutthroat trout currently occupy 2,728 miles of stream habitat in 21 watersheds ranging across Utah, Idaho, Nevada, and Wyoming (Western Native Trout Initiative 2018). Of this occupied area, 62% lies within Utah alone (Figure 2). In the United States, wildlife management is administered primarily by state governments. This results in management policy that is restricted by administrative boundaries instead of the natural delineation of watersheds. For the purposes of this project, I defined the study area as the current range of Bonneville cutthroat trout within the state of Utah. This allowed me to examine a majority of stream habitat occupied by Bonneville cutthroat trout, while acknowledging the restrictions faced by resource managers implementing similar projects in a professional setting.

Dissolved Oxygen Monitoring Locations in Utah's Bonneville Cutthroat Trout Habitat

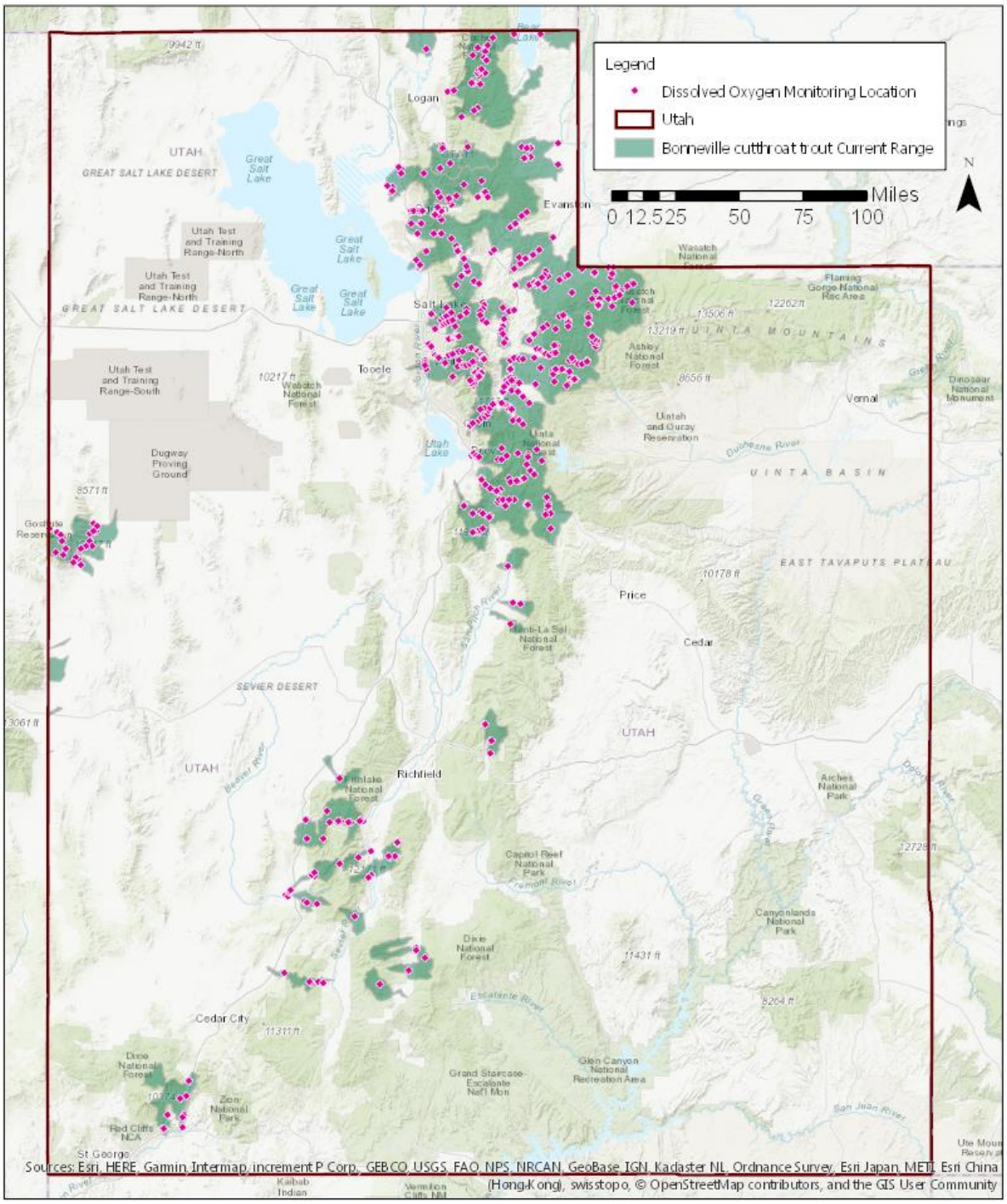

Figure 1. A map illustrating the current distribution of dissolved oxygen data sources in Utah's Bonneville cutthroat trout habitat range.

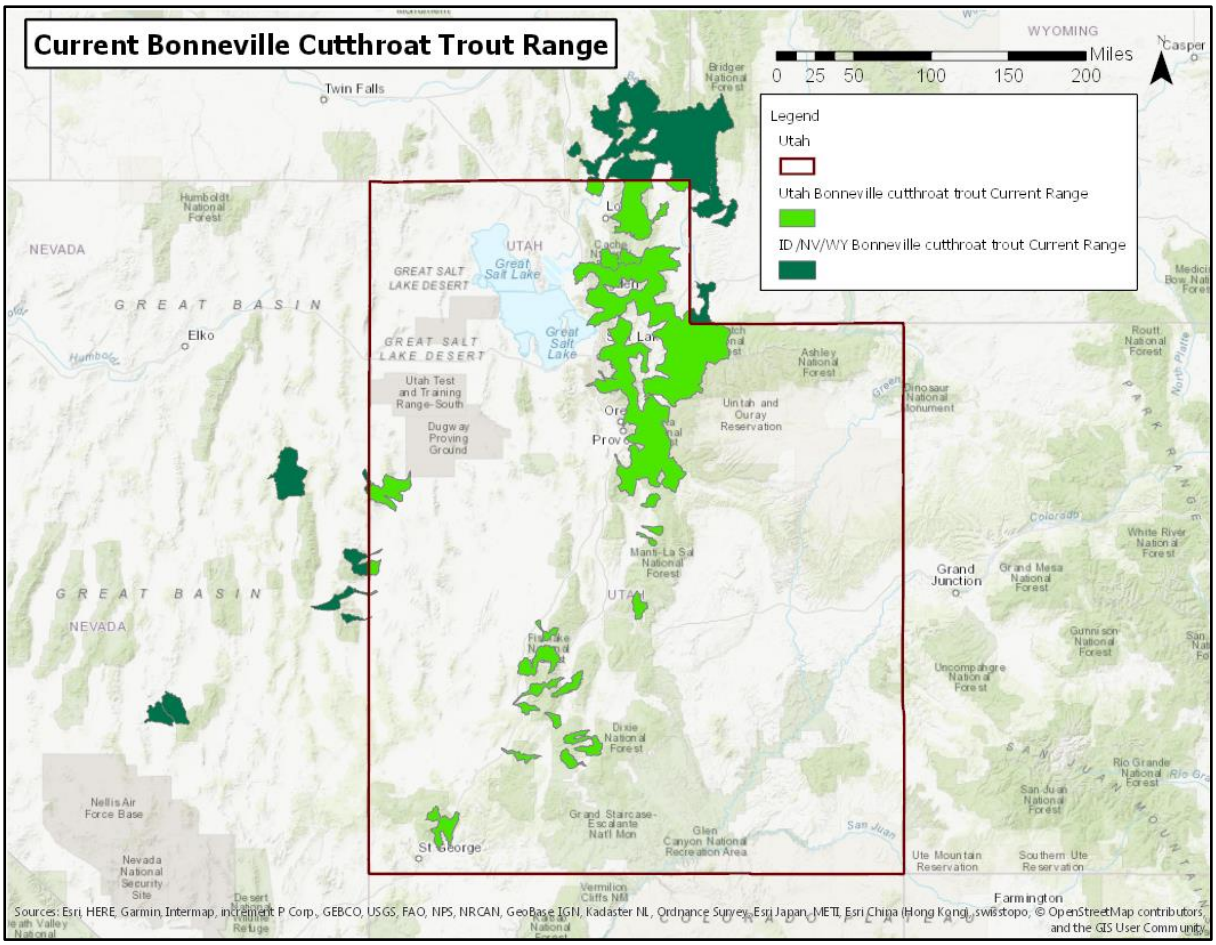

Figure 2. A map illustrating the current range of Bonneville cutthroat trout.

3.2 Data Sources and Initial Data Pre-Processing

My first step in this analysis was to collect and pre-process the data required for the project. I compiled stream segment data from National Hydrography Dataset Plus Version 2 (NHDPlus). Trout unlimited provided current Bonneville cutthroat trout current range data. I acquired road and Utah land cover data through the Utah Automated Geographic Reference Center's Mapping Portal. I downloaded dissolved oxygen monitoring location data in tabular format through the National Water Quality Monitoring Council. Copies of all raw data were compiled inside an ArcGIS Feature Dataset with spatial references set to the GCS North American 1983 geographic coordinate system and the USA Contiguous Albers Equal Area Conic projected coordinate system. Next, I clipped stream segment and dissolved oxygen monitoring location data to the extent of the Bonneville cutthroat trout range to produce a population of stream segments within the current range of Bonneville cutthroat trout.

3.3 Defining Stream Segment Attributes

I defined stream segments for this study based on several criteria. The NHDPlus decimates streams into flowlines divided by their confluence with adjacent segments, thus representing stream segments as the reaches between locations where flows are joined (Mckay et al, 2012). I elected to use this spatial definition of stream segment extent for this study as it was not only inherited from my source data, but also accounts for the influence that conjoined flows can have upon water quality measurements.

Flowing water bodies are defined by a wide array of both physical and biological forms and processes (Thomson et al, 2003). However, the geographic scale of this study required stream segment definitions based on metrics with data available at large spatial scales. This introduced significant restrictions to the attributes available for stream segment definition. I defined variability among stream segments for this study by three physical attributes: total drainage area in square kilometres, stream velocity in feet

per second, and discharge in cubic feet per second. I selected these physical attributes both because they capture important aspects of stream character (size, flow, and water volume, respectively), as well as their near-continuous coverage across all stream segments in the study area as part of the NHDPlus.

3.4 Identifying Stream Segments with Dissolved Oxygen Data

In order to determine how many stream segments had associated dissolved oxygen data available, I intersected points documenting dissolved oxygen data collection locations with each stream segment. To accomplish this, I used the ArcGIS snapping tool to move points within 100m of a stream segment to the nearest stream segment (Figure 3). As many points had multiple stream segments within a 100m radius, I validated the results of my snapping by selecting 5% of dissolved oxygen sampling location points. I validated the data by generating a random list of 5% (31) of the total number of dissolved oxygen locations (611) using the statistical programming language R, then manually assessing whether the stream segment the point was snapped to matched the point's location description by matching my R-generated list to each point's ObjectID. The resulting accuracy was that 29 of 31 points snapped to the correct location, representing 94% accuracy, which I concluded was sufficient to support the analysis.

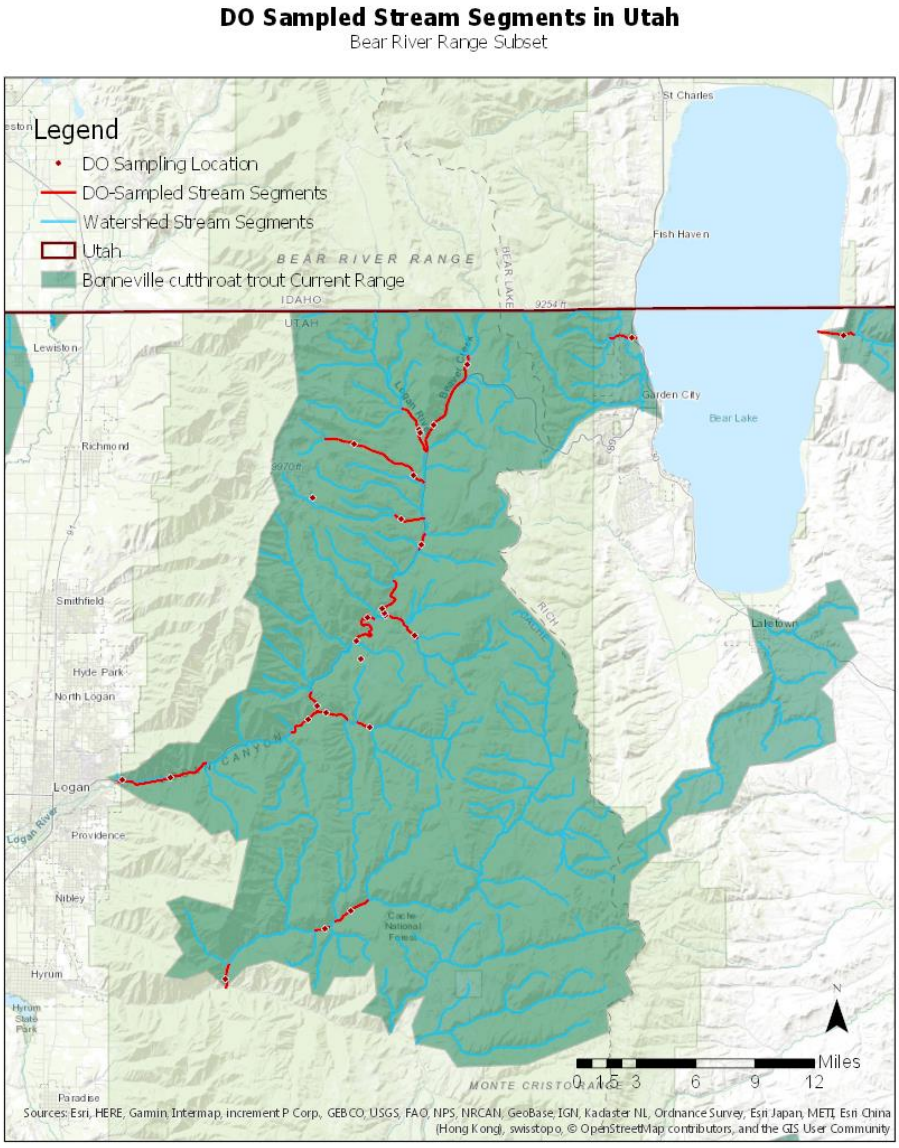

Figure 3. A map identifying stream segments sampled with dissolved oxygen monitoring in the Bear River Range. A subset of the full Bonneville cutthroat trout range is presented here for clarity. Note that some segments were unsampled due to the point location being beyond 100m of a stream segment.

3.5 Establishing Stream Character Attribute Classes

In order to develop classes for each stream character attribute, I exported stream segment data to Microsoft Excel, and then plotted the raw attribute values in a histogram with auto-generated bins. The initial results revealed a strong positive skew to the for all three stream segment attributes, which limited my ability to generate useful attribute classes or draw meaningful conclusions from the data. To solve this, I elected to apply a log10 transformation to the data owing to its widespread use in biology for transforming positively-skewed data (McDonald 2014). I applied a log10 transformation to each stream segment within the study area, then plotted the results in a histogram that resulted in normal data distributions for each stream segment attribute. I then applied a Power10 transformation to each attribute class to convert each back to its native unit of measure. For both total drainage area and discharge attributes, I chose to combine all attribute classes with an upper bound smaller than one into the largest attribute class less than one due to the large range of data values in each attribute type and predominance of such classes containing no stream segments. Next I used MS Excel's Data Analysis tool to generate counts from the sampled and full population of stream segments for each attribute class. Each attribute class count was then divided by the total stream segment count for sampled and full study area populations to determine the percentage of total stream segments represented in each attribute class.

3.6 Assessing the Extent of Unsampled Attribute Classes

In this project, I chose to focus on the extent to which stream segment attribute classes were completely unrepresented in current dissolved oxygen datasets as a measure of effective representation. I defined unsampled attribute classes as classes that had full study area stream segment counts of one or greater, but had zero sampled stream segment counts. In order to quantify the extent to which these attribute classes define the total study area stream segments, I selected all stream segments that contained at least one attribute value that fit into one of the unsampled classes. I then divided the result by the total number of stream segments to determine to the total percentage of stream segments in the study area that have at least one unsampled physical attribute. I also calculated the extent to which stream segments were unsampled in each physical attribute by taking the sum of the study area stream segments in all unsampled classes for each physical attribute, then dividing that by the total number of study area stream segments. This provided a percentage of stream segments in each physical attribute type which contained an attribute value in an unsampled attribute class.

3.7 Determining Optimal Locations for Additional Monitoring

The ability to add monitoring locations is largely restricted by the time and monetary cost required to install equipment and collect measurements. In order to address this, I chose to limit my suite of potential locations by establishing a location criteria that favoured sites with low access restrictions, required sampling the fewest number of stream segments required to gain full coverage of all stream segment attribute classes, and selecting segments that are geographically nearest one another. I defined sites with a low access restrictions as stream segments within one kilometer of a road, and having some portion of their length crossing public land (Figure 4). The intent with these criteria was to reduce both the overland travel time to a potential monitoring site by selecting segments that were close to an access point (road), as well as to reduce the difficultly of determining land ownership and complexity of gaining access approval. In order to accomplish this, I created a stream segment population that included all segments with at least one unsampled stream character attribute class, then reduced the population to only sites that met the location criteria using ArcGIS Pro's Select by Location tool, which allowed me to identify both segments that cross land with public ownership, as well as segments that intersect a road within a one kilometer buffer. In order to make sure that my location criteria did not exclude any of the unsampled attribute classes, I individually queried and validated that each unsampled attribute class was associated with at least one stream segment in the population meeting the location criteria.

The final task involved selecting stream segments to monitor. For the first step I used ArcGIS Pro's Select by Attributes tool to identify how many stream segments contained unsampled attribute classes

in multiple attribute types, and identify as many stream segments as possible that could add multiple non-repeating unsampled attributed classes. Once this group of points was identified, I queried the potential sampling stream segments for the remaining unsampled attribute classes, and selected segments that nearest to the multiple-unsampled attribute segments (when multiple adjacent segments with similar distances were returned, I used personal discretion to select from the segments available). The result provided a list of the recommended stream segments to include in future monitoring in order to capture the full range of stream segment variability while maximizing cost and time reductions required for the data collection.

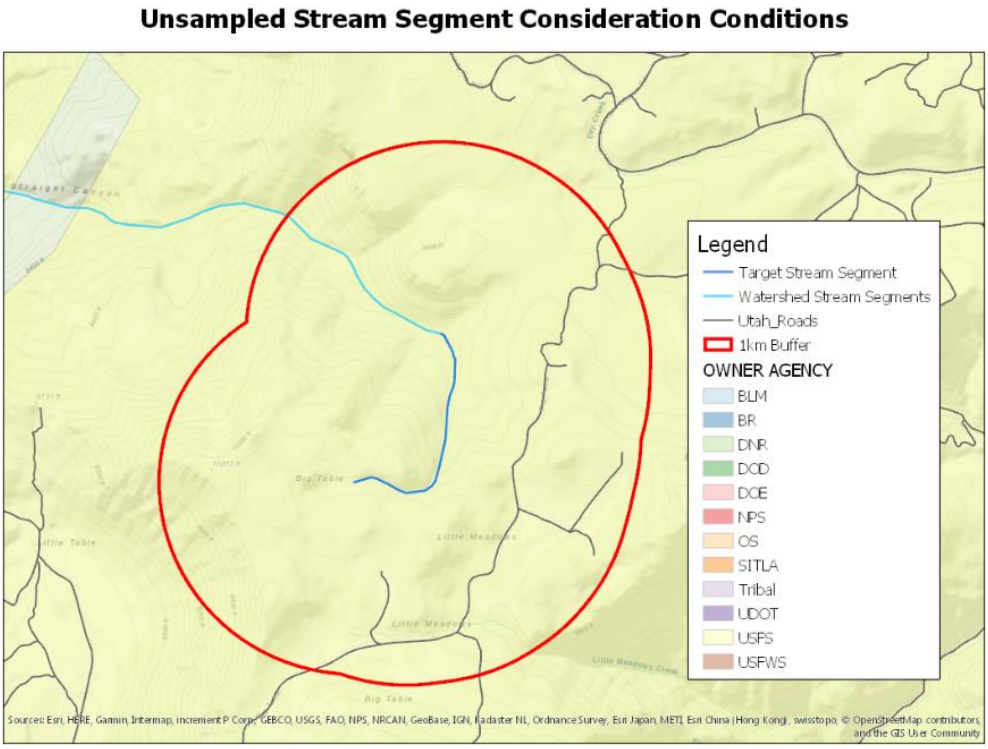

Figure 4. An example stream segment, shown in navy, which meets both location criteria. The red buffer indicates that a road comes within one kilometer of the stream segment, and the stream segment crosses publicly-owned (United States Forest Service) land, displayed here in yellow.

4. RESULTS

4.1 Identifying and Defining Attribute Classes for Stream Segments with Dissolved Oxygen Monitoring

In my initial analysis I identified a total population of 4874 stream segments within the current range of Bonneville cutthroat trout, of which 434 have some associated observation of dissolved oxygen. When exported for statistical analysis, stream segment populations were slightly reduced from total by the removal of segments with no data values (Table 1).

Table 1. Stream Segments by Attribute Type

Attribute Category	Total Stream Segments	Sampled Stream Segments
Total Drainage Area	4483	412
Velocity	4370	409
Discharge	4472	412
All Stream Segments	4874	434

The binning of the raw stream segment population data revealed that all three stream segment attributes had significant positive skews to their data. In order to counteract this, I applied a log10 transformation to each attribute value in all three attribute classes, which resulted in fitting the data to a

normalized distribution. Attribute classes for each stream attribute were then calculated using the log10-transformed values and converted back to native units using the Power10 transformation. This resulted in the set of attribute classes used to group stream segments for analysis. The results of this classification is presented in Appendix A.

4.2 Assessing the Extent of Unsampled Attribute Classes

I initially assessed dissolved oxygen sampling location representation of stream segments by plotting the percentage of stream segments contained in each attribute class for both sampled and total population stream segments (Figure 5). The ideal result would have been a near match in the distribution of population percentages for each attribute class for both the full study area stream segments and sampled stream segments. However, in all three stream attribute types, sampled stream segments showed a tendency to underrepresent attribute classes that define the majority of total stream segments and oversample attribute classes on the upper end of the distribution.

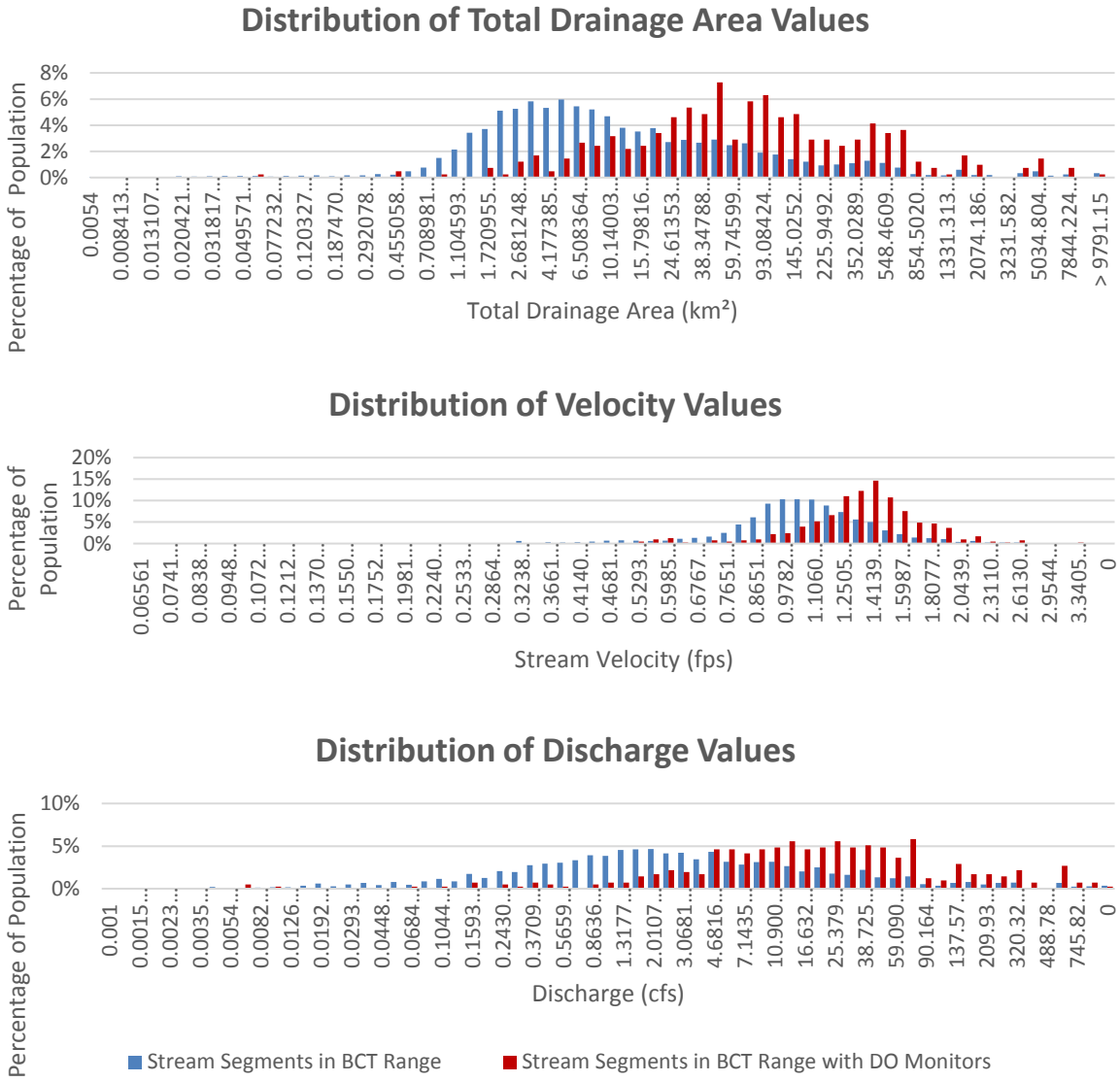

Figure 5. Tables showing the percentage of stream segments for all and sampled stream segments in the study are for each log10-derived attribute class.

However, nearly all attribute classes had at least some representation in the sampled stream segments, and only 13 attribute classes among all 3 stream attribute types lacked any form of representation within the sampled stream segments (Appendix A). I quantified the extent of stream segments characterized by unsampled attribute classes by calculating the percentage of total stream segments for each attribute

type that contained an attribute value that falls into an unsampled attribute class. I also examined the total stream segment population by calculating the percentage of stream segments contained at least one unsampled attribute class among all three stream segment attribute types (Table 2).

Table 2. Percentage of Stream Segments with Unsampled Attribute Classes

Attribute Type	Number of Unsampled Attribute Classes	Total Stream Segments	Stream Segments with Unsampled Attribute Class	Percentage of Population Unsampled
Total Drainage Area	4	4483	266	6%
Velocity	9	4370	213	5%
Discharge	0	4472	0	0%
Total Study Area	13	4874	440	9%

* Total study area unsampled stream segments are less than the sum of total drainage area and velocity unsampled stream segments due to co-occurrence of unsampled attribute classes among both unsampled attribute types.

The results showed that six percent of stream segments by total drainage area and five percent of stream segments by velocity contained attribute values that fell into an unsampled attribute class, while all discharge attribute classes contain a representative sample. As a whole, 9 percent of stream segments in the study area had at least one attribute value that fell into one of the 13 unsampled attribute classes. While the total percentage of stream segments with attributes unrepresented by dissolved oxygen monitoring was relatively high, the small number of attribute classes required to improve dissolved oxygen sensor representation of stream segments made clear that limited effort could result in significant improvements to coverage of stream segment attributes.

4.3 Determining Optimal Locations for Additional Monitoring

The first step in determining how to select additional monitoring locations was to identify how many possible stream segments I could select from. To do this, I queried the total population of stream segments for any segments that contained an attribute value that fell into one of the 13 unsampled attribute classes described above, which returned 440 stream segments. I first implemented the location criteria to return a list of stream segments that prioritized ease of access, which identified 189 possible stream segments that cross public land and are within one kilometre of a road. In order to make sure that I had not excluded any unsampled attribute classes through the location criteria, I queried the 189 location criteria compliant stream segments and validated the presence of all 13 unsampled attribute classes.

From the validated list of 189 location criteria compliant stream segments, I selected all stream segments that contained both an unsampled velocity class and unsampled total drainage area class in order to prioritize sampling sites where multiple unsampled attribute classes could be added at once. I was able to identify 13 different stream segments that contain attribute values that fall into unsampled attribute classes for two different attribute types. From this result, I identified three unsampled velocity classes and two unsampled total drainage area classes represented in the data subset, which revealed that a maximum of two stream segments would allow me to add coverage to non-repeating unsampled velocity and total drainage area classes. Given the small number of stream segments, I hand-selected two stream segments that both provided the maximum four non-repeating unsampled attribute classes and had the smallest distance between them. Using these two points as my base, I queried the 189 location criteria compliant stream segments for each of the nine remaining unsampled attribute classes, and selected each stream segment that was nearest to either of the two multi-attribute class points I had previously identified. The result allowed me to identify a list of 11 stream segments that represented the lowest cost and highest efficiency distribution of new dissolved oxygen monitoring locations required to give complete coverage of all study area stream segment total drainage area, velocity, and discharge attribute classes (Figure 6).

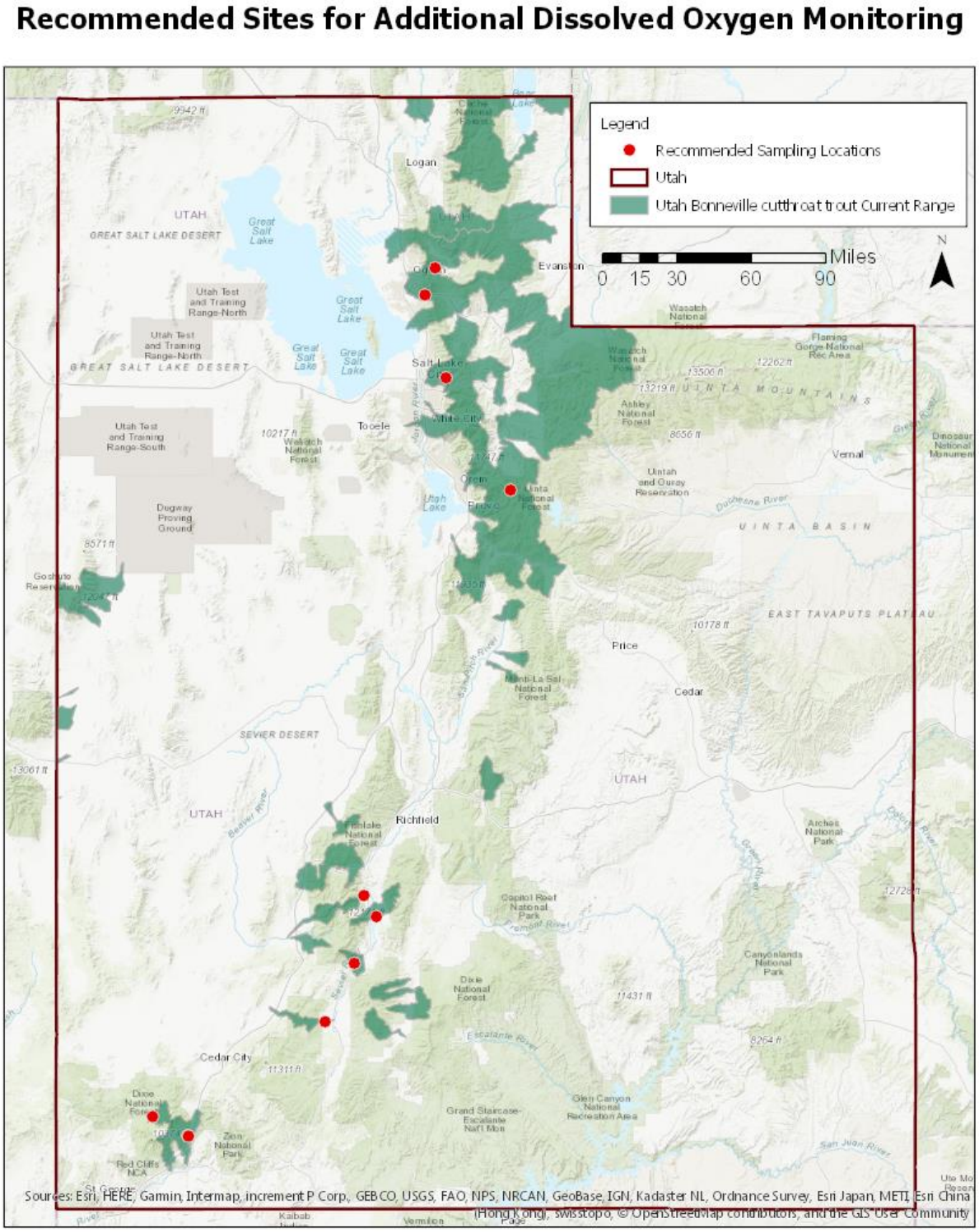

Figure 6. A map showing the 11 recommended sites to add dissolved oxygen monitoring to in order to sample stream segments for each attribute class currently without representation.

5. DISCUSSION

This report presents a methodology for assessing the representation of stream segments captured by current dissolved oxygen monitoring datasets using GIS to facilitate the intersection of both data types.

When considering large-spatial-scale environmental datasets for use in modeling, it is critical that researchers consider the degree to which their data can represent the features they are trying to evaluate. The method presented here offers several advantages when considering representativeness of monitoring locations.

Firstly, this analysis relies on widely available data and attributes contained within the NHDPlus dataset that allow for stream segment characterization that can be repeated anywhere that NHDPlus data is available. In addition, this method relies on a simple intersection of stream segments as line data and sample data represented as points, and can thus be easily expanded to any type of metrics where monitoring locations can be represented by points, such as turbidity, stream temperature, or specific conductivity. These two factors recommend this method, as it can be implemented widely across a variety of aquatic habitat parameters and using an easily understood methodology.

However, the analysis presented here is also subject to several significant limitations. In this analysis, I relied on log10 transformations and histogram binning procedures to determine attribute classes for each stream segment attribute considered. While this did result in classes that captured a normalized distribution of stream segment attribute values, the classes were solely based on statistical methods, and had no direct correlation to real world conditions. This created significant concerns about whether the attribute classes have any substantive relation to real-world stream segment variability, and thus whether they can be used to accurately quantify the degree to which stream segments are represented in a sample population.

Another significant limitation of this methodology concerns the automation of site selection for future monitoring efforts. While the restriction of stream segments based on location criteria and priority of multiple missing attribute classes follows a clearly defined method, the selection of stream segments that optimize efficiency by being grouped by proximity was carried out using my own experience and judgement. While this was successful at generating a list of locations to add dissolved oxygen monitoring, the analysis' reliance on user expertise significantly reduces the repeatability of the method. The method's reliance on user expertise also leaves the method open to the potential for erroneous selections, especially as the size of stream segment populations increase.

The final limitation with this model is the simplistic way in which this analysis characterizes stream segments. The characterization of rivers and streams involves a complex array of physical and biological factors (Thomson et al, 2003). However, this project reduced this complexity to total drainage area, discharge, and velocity, largely based on the availability of the data. While this did facilitate an understanding of how current dissolved oxygen monitoring captures stream segments, it failed to address whether this definition of stream character has the ability to accurately quantify the differences between stream segments. Additionally, the similarity of distribution among sampled stream segments for all three stream segment attribute types revealed a high potential for autocorrelation between the three attribute types chosen for this study. This leaves open the possibility that, while generating usable results, the method could be missing significant gaps in stream segment representation by not considering a wider suite of stream segment attributes.

6. CONCLUSION

Current dissolved oxygen monitoring data captures 91% of stream segments within the current range of Bonneville cutthroat trout in the state of Utah based on stream segments defined by total drainage area, velocity, and discharge. However, only 13 different classes of total drainage area or velocity have no representation among the dissolved oxygen-sample stream segments, and monitoring need only be added to 11 additional stream segments to address this gap in stream segment representation. Thus, through the method presented here researchers can ensure that dissolved oxygen data available within Utah's Bonneville cutthroat trout range provides representation to all levels of stream variability, and validate the use of current dissolved oxygen data for use in aquatic habitat modeling. While the method presented in this report does provide a means to quantifying and correcting representation of stream segments by the spatial distribution of current dissolved oxygen monitoring, it also represents a simplistic approach and relies on a number of assumptions that require further verification.

7. REFERENCES

de Kerckhove, D.T., Smokorowski, K.E., Randall, R.G. (2008). A primer of fish habitat models. Canadian Technical Report of Fisheries and Aquatic Sciences 2817. Fisheries and Oceans Canada, Ontario.

Hickman, T. and Raleigh, R.F. (1982). Habitat suitability index models: Cutthroat trout. U.S.D.I. Fish and Wildlife Service. FWS/OBS-82/10.5.38 pp.

McDonald, J.H. (2014). Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland.

McKay, L., Bondelid, T., Dewald, T., Johnson, J., Moore, R., and Rea, A (2012). NHDPlus Version 2: User Guide. Retrieved from http://www.horizon systems.com/nhdplus/NHDPlusV2_documentation.php#NHDPlusV2%20User%20Guide

Thomson, J.R., Taylor, M.P., Brierley, G.J (2003). Are river styles ecologically meaningful? A test of the ecological significance of a geomorphic river characterization scheme. Aquatic Conservation: Marine and Freshwater Ecosystems. 14: 25-48.

Utah Division of Wildlife Resources (2008). Bonneville cutthroat [Website]. Retrieved from https://wildlife.utah.gov/cutthroat/BCT/index.html

Western Native Trout Initiative (2018). Bonneville Cutthroat Trout Species Status Report [Online Document]. Retrieved from http://westernnativetrout.org/wp-content/uploads/2018/03/BonnevilleCT_WesternNativeTroutStatusReport_UpdatedJanuary2018.pdf

6. Appendices

6.1 Appendix A. Summary Table of All log10-Derived Attribute Classes and Stream Segment Counts
* Unsampled Attribute Classes are shown in red.

Total Drainage Area (km²)

Attribute Class Upper Bound	All Stream Segments Count	Sampled Stream Segments Count
0.88495	219	4
1.104593	96	0
1.378752	154	0
1.720956	166	3
2.148095	229	1
2.681249	236	5
3.346731	261	7
4.177385	239	2
5.214206	268	6
6.508365	244	11
8.123732	233	10
10.14003	210	13
12.65677	171	9
15.79817	158	10
19.71925	169	14
24.61354	122	19
30.72258	129	22
38.34789	119	20
47.86578	130	30
59.746	111	12
74.57487	117	24
93.08425	85	26
116.1876	79	19
145.0252	63	20
181.0203	54	12
225.9493	42	12
282.0296	45	10
352.0289	49	12
439.402	58	17
548.461	50	14
684.5882	34	15
854.502	12	5
1066.588	9	3
1331.314	8	1
1661.744	27	7
2074.187	9	4
2588.997	9	0
3231.583	0	0
4033.657	15	3
5034.805	21	6
6284.436	7	0
7844.225	11	3
9791.15	0	0
> 9791.15	15	1

Velocity (ft/s)

Attribute Class Upper Bound	All Stream Segments Count	Sampled Stream Segments Count
0.3662	45	0
0.389393	12	0
0.414054	15	0
0.440278	19	0
0.468162	28	0
0.497813	32	0
0.529341	29	2
0.562866	25	4
0.598515	31	5
0.636421	50	1
0.676728	58	0
0.719587	69	3
0.765161	109	2
0.813622	195	3
0.865151	266	4
0.919945	407	9
0.978208	450	10
1.040162	450	16
1.106039	446	21
1.176089	386	27
1.250575	319	45
1.329778	244	50
1.413998	218	60
1.503552	135	44
1.598777	95	31
1.700033	60	20
1.807703	54	19
1.922191	45	15
2.043931	17	4
2.17338	25	7
2.311029	11	2
2.457395	10	1
2.613031	9	3
2.778523	3	0
2.954498	0	0
3.141617	0	0
3.340587	2	1
3.552159	0	0
>3.552159	1	0

Discharge (ft^3/s)

Attribute Class Upper Bound	All Stream Segments Count	Sampled Stream Segments Count
0.863609	1399	19
1.066785	172	3
1.317762	203	3
1.627784	207	6
2.010744	208	7
2.4838	186	9
3.06815	189	8
3.789976	155	7
4.681623	193	19
5.783042	141	19
7.143585	126	17
8.824216	139	19
10.90024	142	20
13.46468	118	23
16.63244	90	19
20.54546	112	20
25.37907	79	23
31.34986	72	20
38.72537	99	21
47.83606	59	20
59.09018	55	15
72.992	64	24
90.16441	24	5
111.3769	15	4
137.5799	31	12
169.9475	36	7
209.9301	22	7
259.3192	31	6
320.3278	32	9
395.6895	3	3
488.7812	0	0
603.7739	31	11
745.8204	11	3
921.2854	13	3
>921.2854	15	1