STUDENT SUMMER INTERNSHIP TECHNICAL REPORT

Gap-Filling Time Series Using Direct Sampling in the Hanford 100-Areas

DOE-FIU SCIENCE & TECHNOLOGY
WORKFORCE DEVELOPMENT PROGRAM

Date submitted:

December 10, 2021

Principal Investigators:

Christian Gonzalez Lopez (DOE Fellow Student)
Florida International University

Mark Rockhold, Ph.D., Xuehang Song, Ph.D. (Mentor)
Pacific Northwest National Laboratory

Ravi Gudavalli, Ph.D. (Program Manager)
Florida International University

Leonel Lagos, Ph.D., PMP® (Program Director)
Florida International University

Submitted to:

U.S. Department of Energy
Office of Environmental Management
Under Cooperative Agreement # DE-EM0005213

FIU | Applied Research Center
FLORIDA INTERNATIONAL UNIVERSITY

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

This research work has been supported by the DOE-FIU Science & Technology Workforce Initiative, an innovative program developed by the US Department of Energy's Environmental Management (DOE-EM) and Florida International University's Applied Research Center (FIU-ARC). During the summer of 2020, a DOE Fellow intern Christian Gonzalez Lopez spent 10 weeks doing a summer internship at Pacific Northwest National Laboratory under the supervision and guidance of Mark Rockhold Ph.D. and Xuehang Song Ph.D. The intern's project was initiated on June 7, 2021, and continued through August 13, 2021, with the objective of researching more accurate forms of gap-filling time-series.

Cleanup efforts have been ongoing since the late 1990s to remediate contaminated waste sites and groundwater in the 100 Areas at the U.S. Department of Energy (DOE) Hanford Site. One of the primary contaminants of concern is hexavalent chromium (Cr(VI)), which was used as a corrosion inhibitor in cooling water for nuclear reactors that formerly operated along the shoreline of the Columbia River. Cleanup efforts have included 1) removal, treatment (as needed), and disposal of contaminated sediments; 2) in situ redox manipulation as a permeable reactive barrier; 3) pump-and-treat; 4) soil flushing; and 5) monitored natural attenuation. DOE's annual groundwater monitoring reports document the significant reductions in Cr(VI) plume areas that have occurred over the past 10 years or more as a result of these cleanup efforts. The Record of Decision for the 100-HR-3 operable unit specified a cleanup level (CUL) for Cr(VI) in groundwater of 48 µg/L to protect human receptors, and a surface water CUL of 10 µg/L to protect aquatic organisms in the Columbia River. The Record of Decision did not specify point-of compliance locations for the surface water CUL. Data for 2019 from the six groundwater operable units (OUs) in the 100 Areas indicate that the 48 µg/L groundwater CUL has been achieved in 100% of the wells in the 100-BC and 100-NR OUs, and in 89- 97% of the wells in the other OUs (100-KR, 100-HR-D, 100-HR-H, 100-FR). Data for 2019 indicate that 100% of the aquifer tubes monitored for Cr(VI) in the 100 Areas have concentrations below the 48 µg/L groundwater CUL. However, the 10 µg/L standard has not yet been consistently achieved for both inland groundwater monitoring wells and shoreline aquifer tubes. The overarching goal is to identify consistent relationships, if any, between inland well and shoreline Cr(VI) concentrations within the 100 Areas. However, to date attempts to make this correlation have failed. This research was focused on finding another form of gap-filling that would reduce error and increase the probability of performing accurate analysis on the data, strengthening the results. The use of direct sampling for gap-filling proved useful in test cases with artificial gaps.

1. INTRODUCTION

The overarching goal of the research was to assist in the spatiotemporal analyses of groundwater and shoreline hexavalent chromium concentration in the 100 areas at Hanford by researching a better method to fill the extensive and omnipresent gaps in the 100 areas' time series data. The proximity of GW Cr(VI) plumes to the river, and the highly dynamic nature of the river, influence the transport behavior of the plumes and create challenges for monitoring and interpretation of Cr(VI) fate and transport. Understanding the controlling processes, the spatial and temporal relationships between inland GW and SW Cr(VI) concentrations, and the effects of ongoing remediation efforts is critical to establishing technically defensible compliance monitoring for SW and GW CULs. To strengthen the probability of an accurate analysis finding another form of gap-filling that would preserve the original relationship in the time-series became the focus of this internship. The Direct Sampling method (DS) came up as a potentially better alternative.

2. RESEARCH DESCRIPTION

Data is extremely important when it comes to analysis, having enough data at frequent enough intervals to accurately make predictions is desirable. The Hanford 100 areas are very complex with many variables that may affect the concentrations of Hexavalent Chromium in the area. The sampling in the area is not consistent and many have different length time-series both the date that sampling started and concluded varies from well to well. Given, the nature of the Hanford area, it was found beneficial to try and discover a more accurate form for filling those gaps rather than standard interpolation. This research is not only applicable to directly to the Hanford 100 area but with further modification it could be applied to a variety of other time-series datasets. Initially the use of linear interpolation was implemented but quickly realized that it was too rigid for the data that we had. PNNL was already implementing multi-point statistics (MPS) on other tasks however MPS is generally used on special data. After being guided towards potentially modifying the MPS to be used for special-temporal data I eventually found a paper “Gap-filling of daily streamflow time series using Direct Sampling in various hydroclimatic settings” in the Journal of Hydrology about a new implementation of MPS Direct Sampling method (DS) used for special-temporal data.

Before committing to the MPS DS other methods were considered specifically: nearest neighbor, method by data transfer, interpolation techniques, simple and multiple regressions, classification and regression trees, and various forms of artificial neural networks. These methods however, had limitations that led me to pursue the DS method, for example, nearest neighbor brings discontinuity in the time series, interpolation techniques offer limited representation of the space-time structure of the time series, linear regression methods assume linearity between variables, regression trees there is a lack of understanding of the construction of the trees, artificial neural networks complex, high computational cost, results have no physical interpretations.

In this case, Direct Sampling (DS) is used as a non-parametric stochastic method for infilling gaps in streamflow records for data collected in the Volta River basin, West Africa. The main idea behind the direct sampling algorithm is that it uses a provided training dataset (TI). Then new simulated values are generated based on a conditional resampling of TI and these simulated values are called a simulation grid (SG). There were scenarios with artificially produced gaps used to assess the performance as well as real application to the existing gaps in the flow records. In this framework however, the simulated data is sampled from historical values of the same station (TI) rather than stations around it (Dembélé, Moctar, et al., 2019, p. 579).

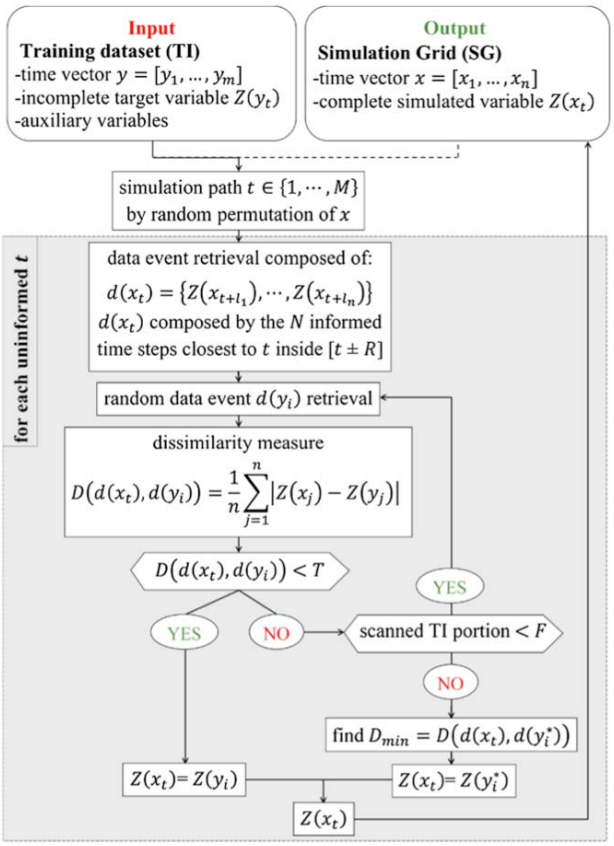

Figure 1. Workflow of the DS algorithm for continuous variables. *R* is the radius of the search neighborhood window ([*t* ± *R*]) composed by a number of neighbors *N* closest to x_t , with lags *L* = {l_1, , l_n}, such that l_i *R* and *n N* . *Dmin* is the minimum dissimilarity found in the TI. *F* is the maximum TI fraction scanned.

2.1 Algorithm Steps

1. Linearly normalize to a range of [0,1]. $Z \rightarrow Z(\max(Z) - \min(Z))^{-1}$
2. A random simulation path $t \in \{1, \dots, M\}$ of the same length as the SG is generated.
3. Following the random path an uninformed time step t is selected.
4. A data event $d(x_t) = \{Z(x_{t+l_1}), \dots, Z(x_{t+l_n})\}$ representing a pattern of neighboring data of t, is retrieved from the SG according to a radius R centered on x_t.
5. A random time step y_i is scanned and the corresponding data event $d(y_i)$ is retrieved to be compared with $d(x_t)$ based on the same time lags.
6. A distance $D(d(x_t), d(y_i))$is calculated as a measure of dissimilarity.
7. If $D(d(x_t), d(y_i))$ is below a defined similarity threshold T ∈ [0,1], the iteration stops and the $Z(y_i)$ is assigned to $Z(x_t)$.
8. Otherwise, the procedure is repeated from step 5 to 7 until a suitable $d(y_i)$ is found or a prescribed TI fraction *F* is scanned.
9. In case no time step corresponding to $D(d(x_t), d(y_i)) < T$ is found, the datum $Z(y_i^*)$ minimizing this distance is assigned to $Z(x_t)$.

10. The procedure from step 3 to 8 is iterated for the simulation at each x_t until the entire SG is completely informed.
11. The variables are linearly back transformed to their original range.

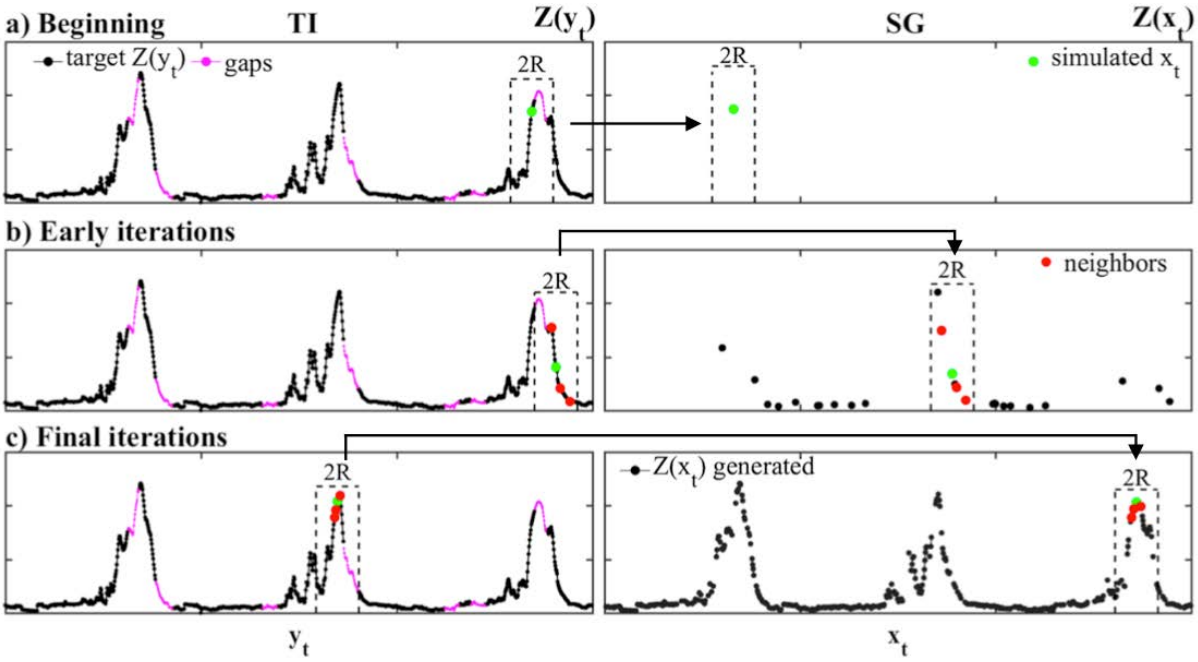

Figure 2. Sequential simulation of streamflow time series with DS. The dashed rectangle represents the search window defined as twice the radius *R*, and contains the data event that is formed by the simulated datum in green and the *N* neighboring data in red.

There are certain conditions that would help get better results from the algorithm, these are, that the target and the predictor stations are in the same sub catchment, the predictor station is well correlated with the target variable, the predictor station is located upstream of the target station, the predictor station has no gaps, and the target station contains relatively short gaps. These are not requirements but mainly steps that would assist in the accuracy of the gap-filling.

3. RESULTS AND ANALYSIS

Table 1 (Dembélé, Moctar, et al., 2019, p. 579) shows the minimum and maximum values of the Pearson and Spearman correlation. The correlation values before the simulation are in the same range as after the simulation with an average error of 2 percent. This shows that the Direct Sampling algorithm can preserve the relationships between the target and predictor stations. This can be used to model missing data in a well with previous years of the same well or wells around it that way it can help model changes that occur due to seasons as well as local changes that can affect every well like a flood for example.

Table 1. Pearson (r (Z, Q)) and the Spearman (rs (Z, Q)) correlation coefficients between the target (Z) and the predictor (Q) variables (SC: same catchment; indicates where the target and the predictor stations are located. SN: scenario number, CN: catchment name, WV: White Volta, BV: Black Volta, LV: Lower Volta). Before gaps, Z and Q are fully informed. After gaps, Z (Daboya station) contains 50% of gaps for all scenarios, while Q contains 30% of gaps for only even-numbered scenarios.

SC	SN	*Q* station	CN	gaps in *Q* (%)	Before simulation				After simulation			
					Before gaps		After gaps					
					r	r_s	*r*	r_s	*r* (min-max)		r_s (min-max)	
yes	1	Yarugu	WV	0	0.593	0.669	0.656	0.680	0.580	0.604	0.611	0.619
	3	Nawuni	WV		0.984	0.913	0.988	0.908	0.988	0.989	0.910	0.916
no	5	Akosombo	LV		-0.309	-0.302	-0.475	-0.356	-0.318	-0.289	-0.256	-0.238
	7	Lawra	BV		0.871	0.809	0.920	0.814	0.840	0.876	0.752	0.781
	9	Saboba	Oti		0.947	0.822	0.951	0.828	0.944	0.950	0.789	0.819
yes	2	Yarugu	WV	30	0.593	0.669	0.689	0.660	0.550	0.606	0.608	0.629
	4	Nawuni	WV		0.984	0.913	0.988	0.904	0.983	0.986	0.881	0.899
no	6	Akosombo	LV		-0.309	-0.302	-0.430	-0.265	-0.301	-0.277	-0.270	-0.260
	8	Lawra	BV		0.871	0.809	0.911	0.824	0.852	0.869	0.720	0.739
	10	Saboba	Oti		0.947	0.822	0.952	0.857	0.939	0.950	0.795	0.816

High

Low

4. CONCLUSION

As time passes, wells in the Hanford 100 areas continue to change. The wells that have and do not have pump-and-treat system and different time-series lengths will persist as different wells may become inactive or active as well as changes in sampling areas. While these issues continue filling those time-series as accurately as possible will become increasingly important. Eventually, with the increased development of the direct-sampling method it may become effective enough to replace traditional interpolation methods in these areas.

The internship provided Mr. Gonzalez with priceless experience in the process that is required in cleanup and monitoring efforts at the Hanford 100 areas. The experience and collaboration of the mentors at (PNNL) taught Mr. Gonzalez about the thinking and communication that goes into solving problems like this.

5. REFERENCES

1. Dembélé, Moctar, et al. “Gap-Filling of Daily Streamflow Time Series Using Direct Sampling in Various Hydroclimatic Settings.” *Journal of Hydrology*, vol. 569, 2019, pp. 573–586., https://doi.org/10.1016/j.jhydrol.2018.11.076.

2. Rockhold, Mark, et al. “Spatiotemporal Analyses of Groundwater and Shoreline Cr(Vi) Concentrations in the 100 Areas at Hanford.” 2020, https://doi.org/10.2172/1734936.