Pacific Northwest
NATIONAL LABORATORY

PNNL-33102

Improved Data Interpretation through Identification of Time Series Periodicity Changes

July 2022

Sophia Lazcano
Christian Johnson

U.S. DEPARTMENT OF
ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

**Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov**

**Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>
Online ordering: http://www.ntis.gov**

PNNL-33102

Improved Data Interpretation through Identification of Time Series Periodicity Changes

July 2022

Sophia Lazcano
Christian Johnson

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Summary

Analysis and interpretation of time series data is easiest when the data values occur at uniform time intervals, but actual data may have differing data sampling frequencies. When data include significant segments of time with differing sampling frequencies, different treatments may be appropriate for each time block (segment). For example, data collection could occur manually at a monthly or quarterly basis until the installation of a sensor, which could result in much higher data collection frequency (perhaps ranging from seconds to daily).

Applying data analysis techniques, such as smoothing, to a dataset with mixed measurement frequencies may not give a representative result. The ability to automatically determine time segments of differing data collection frequency (data periodicity) would provide a mechanism for applying data analysis independently to each segment, though a suitable blending at segment boundaries would be required. A method for detecting data collection frequency changes was developed and applied to Gaussian kernel smoothing of groundwater hydraulic head (water elevation) data. The process identifies time segments of high-frequency (daily) or low-frequency (greater than daily) data using adjusted-bandwidth Gaussian kernel density estimation and a kernel probability density threshold value. These segments are further refined to address small blocks of low-frequency data within larger blocks of high-frequency data.

User-selectable levels of smoothing (e.g., fine, medium, coarse) can then be applied independently to the time segments prior to combining the segment results for a single smoothed data set. As illustrative examples, the process is applied to groundwater elevation measurements from specific groundwater wells at the U.S. Department of Energy Hanford Site in southeastern Washington state. The examples span a range of time segment patterns to demonstrate the effectiveness of the procedure. This time segment identification approach provides effective low- and high-frequency data separation, which provides a method to apply data analysis independently to each time segment for a more useful overall interpretation of the dataset.

Acknowledgments

I would like to thank my mentor, Christian Johnson, for helping me tremendously through the process of research and this work. His insightful guidance, as well as the support from the PNNL SULI program and Office of STEM Education staff, have been a great help in making this research possible.

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program.

SOCRATES, including GALEN, was developed at the Pacific Northwest National Laboratory under the Deep Vadose Zone − Applied Field Research Initiative, which is funded by the U.S. Department of Energy Richland Operations Office.

Acronyms and Abbreviations

AWLN	Automated water level network
CUSUM	Cumulative sum change-point detection algorithm
GALEN	Groundwater AnaLytics for the ENvironment
HEIS	Hanford Environmental Information System
HFD	High frequency data
IQR	Interquartile range
JS	JavaScript
KDE	Kernel density estimation
LFD	Low frequency data
NGVD88	North American Vertical Datum of 1988
PNNL	Pacific Northwest National Laboratory
SOCRATES	Suite Of Comprehensive Rapid Analysis Tools for Environmental Sites
SULI	Science Undergraduate Laboratory Internships
STEM	Science, Technology, Engineering, and Mathematics
VBA	Visual Basic for Applications

Contents

Figures

Tables

1.0 Introduction and Background

Time series data is ubiquitous, occurring in contexts ranging from electrocardiograms to flow rates to temporal concentration data and beyond. Analysis of time series data is thus important for understanding trends, controlling systems, detecting patterns, or other needs. Mathematical analysis and statistics are used to quantify and understand the behavior of time series data, giving information such as mean value, variance, regressions, etc. Such analysis and interpretation of time series data is easiest when the data values occur at uniform time intervals, but actual data may have non-uniform data sampling (measurement) frequencies. These measurement frequency variations can occur due to a variety of reasons, such as change of measurement method (e.g., installation of a sensor) or a change in objectives of the data collection (e.g., a focused study with a different data collection plan). Changes in measurement frequency (i.e., changes in data periodicity) could, for example, result when shifting data collection from occurring manually at a monthly or quarterly basis to occurring at a higher frequency (e.g., seconds, daily, or weekly) when a sensor is installed.

Applying data analysis techniques, such as smoothing, to a dataset with non-uniform spacing of measurements may not give a representative result because of the differing characteristics of the data over time. Time segments with different measurement frequencies may bias the results in the form of over-smoothing of high-frequency data (HFD) or under-smoothing of low-frequency data (LFD). This nonrepresentative result has the potential to affect other analyses, such as outlier identification using an absolute mean average deviation (error) of data points from the smoothed data curve. Thus, when data include segments of data that have significantly different measurement characteristics, it is worthwhile to detect the change-points to identify these time segments and to analyze the time segments separately.

Change-point detection is a well-studied field where a variety of approaches have been developed (van den Burg and Williams 2022, Truong et al. 2020, Shi 2020, Wang 2008) to identify changes in time series data to facilitate, for example, quality control, process operations, or data analysis. Methods of change-point detection can identify where changes in mean, variance, periodicity, or other patterns occur (Schroth et al. 2021), as illustrated in Figure 1. Mathematical models typically form the foundation for change-point detection methods, setting criteria to identify the change points and sometimes taking advantage of transformation of data to a different domain (Lavielle 2017). The CUSUM (cumulative sum) algorithm (Shi et al. 2022) is an early and commonly applied algorithm based on a sequential sum of weights that, along with a similar sequential procedure using sliding window measurements of sum of squares, has proven useful in finding the change points where the mean changes (e.g., Chen and Kuo 1996). Other methods, such as the generalized likelihood ratio, window-sliding, binary segmentation, or kernel techniques (Truong et al. 2019, 2020, Aminikhanghahi and Cook 2017), can also perform change-point detection. While identifying segments with significantly different means is a common use for change-point detection, existing methods can also detect changes in data periodicity, which could represent changes in the data sampling/measurement frequency.

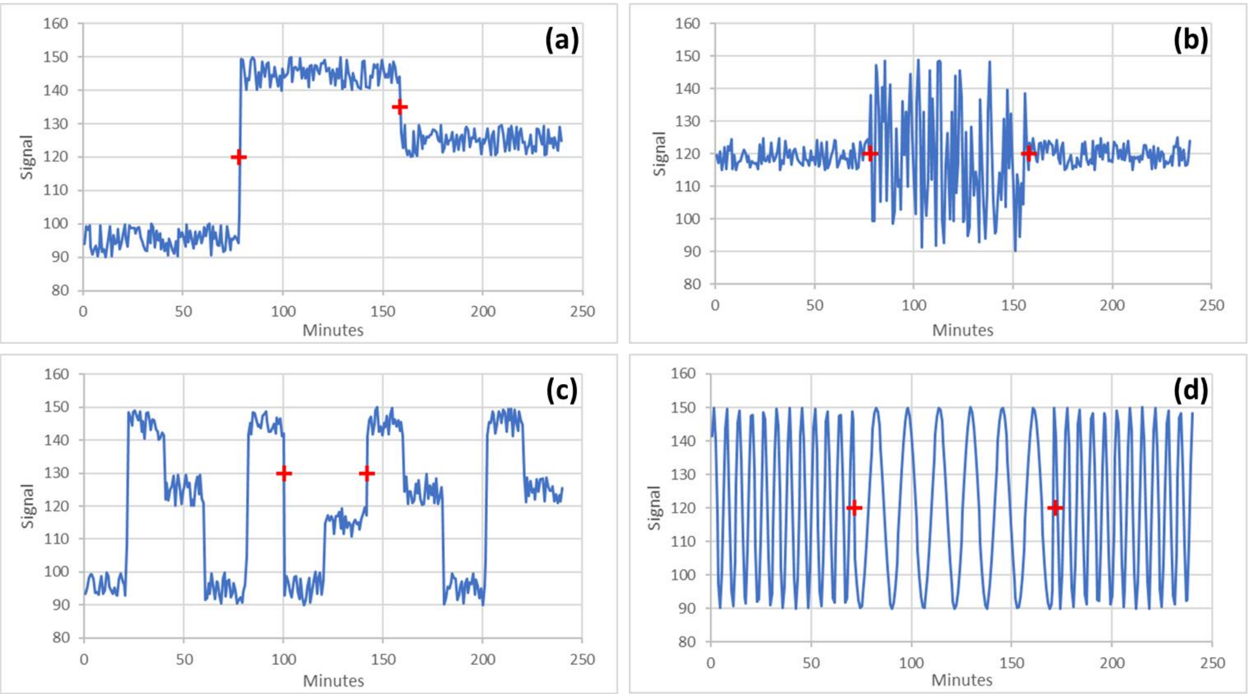


Figure 1. Example datasets with changes in mean (a), variance (b), pattern (c), and periodicity (d), with the change-point between segments indicated by the red pluses.

Fourier analysis is commonly used with time series data (e.g., in signal processing and a range of other applications) to transform data from the time domain to the frequency domain, where trends of periodicity are more readily interpreted and change-point detection can identify time points where there is a significant difference in periodicity (Schroth et al. 2021). Fourier transformed data can reduce noise in the dataset, thereby facilitating application of the CUSUM algorithm to determine change points (Lombard 1988). A drawback to Fourier analysis is that data must be measured at evenly spaced time intervals.

While the abovementioned methods have their uses, they have complexities and sometimes limitations. This work took an empirical/applied approach, investigating a method to automatically determine time segments of different data collection frequency for applications with groundwater level data collected at unequally spaced intervals. The developed method uses an adjusted bandwidth kernel density estimation (KDE) approach to classify LFD and HFD time segments, as described in Section 2.0. Application results are discussed in Section 3.0 and conclusions are given in Section 4.0.

2.0 Procedure for Time Segment Identification

An approach was developed to meet the objective of identifying time segments with differing data collection frequencies and applying smoothing separately to each segment. The data context and exploration of the data is described in this section, followed by an explanation of the mathematics and approach for using a kernel density estimator and additional logic to identify HFD and LFD time segments.

2.1 Data Context

The context for testing and application of the time series segmentation method is groundwater elevation (hydraulic head) measurements from specific groundwater wells at the U.S. Department of Energy Hanford Site in southeastern Washington state. The Hanford Site is the location for former plutonium production, with the site mission having now transitioned to environmental restoration activities related to facilities, stored waste, and soil/groundwater contamination. Thus, Hanford environmental monitoring data includes groundwater elevations, which are used to understand groundwater flow and contaminant transport. The historical waste disposal practices at Hanford included release of significant aqueous streams to cribs and trenches in the 200 West Area of the site, which resulted in a groundwater mound in the underlying water table aquifer. Since cessation of disposal in that manner, the groundwater mound continues to decrease over time. The Hanford Site is also bounded on the north and eastern sides by the Columbia River, whose flow is controlled by at the upstream Priest Rapids Dam, resulting in daily (as well as seasonal) fluctuations in river stage. The fluctuations in Columbia River stage impact groundwater levels in near-river wells.

The data used in this work were obtained from the GALEN module of the SOCRATES software (Royer et al. 2018, Freedman 2021, Brouns and Johnson 2021). GALEN excludes data values flagged as “reject” or “dry.” The groundwater elevation data consist of both manual well measurements (designated as HEIS [Hanford Environmental Information System] data) and daily average sensor data (i.e., averages of automated water level network [AWLN] sensor data collected at minutes to hour intervals). Groundwater elevation data for many Hanford groundwater wells was reviewed, and a subset of wells representing a range of typical and interesting patterns in data measurement frequency were selected for use as test cases for this work. The selected wells are summarized in Table 1. Well 199-K-20 groundwater elevation data is plotted in Figure 2 as a typical example of this type of data.

Table 1. Hanford groundwater wells whose data was used in this work.

Well Name	Date Span	Duration (d)	Number of Values [a]	Data Source [b]	Max. / Min. (m NGVD88)	Std. Dev. (m NGVD88)
199-D5-38	11/15/1999 to 5/19/2022	8,222	6,311	ALWN, HEIS	118.941 / 115.999	0.414
199-F5-43A	1/21/1993 to 6/3/2021	10,361	1,084	AWLN, HEIS	116.727 / 112.544	0.963
199-H4-5	1/22/1985 to 5/18/2022	13,631	6,692	AWLN, HEIS	118.931 / 113.857	0.707
199-K-20	12/29/1957 to 5/20/2022	23,519	5,786	AWLN, HEIS	128.757 / 117.203	0.965
199-N-72	11/27/1991 to 5/22/2022	11,135	4,308	AWLN, HEIS	122.395 / 117.512	0.406
299-W15-763	5/8/2001 to 3/30/2021	7,267	451	AWLN, HEIS	136.879 / 129.859	0.976
299-W23-2	8/17/1955 to 12/4/2000	16,547	31	HEIS	143.096 / 137.317	1.530
699-66-32	7/21/2016 to 5/18/2022	2,128	2,138	AWLN, HEIS	121.082 / 118.150	0.397
1199-33-18B	3/13/2002 to 3/25/2021	6,953	17	HEIS	106.176 / 105.277	0.260

[a] AWLN data is a daily average of the sensor data, which may be collected at a higher frequency (minutes to hourly).

[b] AWLN = Automated Water Level Network data (i.e., in-well sensors), HEIS = Hanford Environmental Information System (manual measurements)

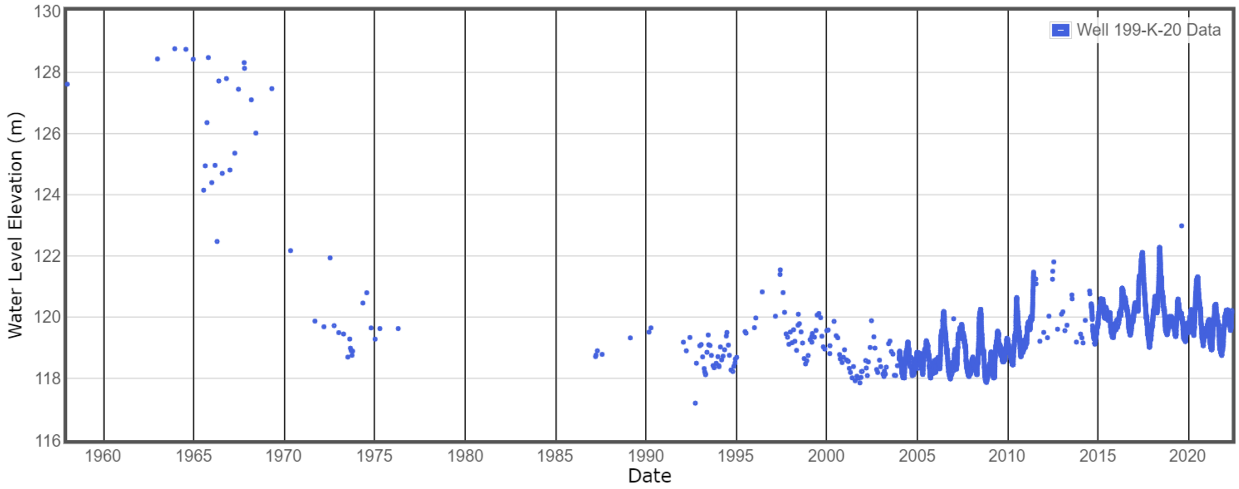


Figure 2. Time series groundwater elevation data for Hanford well 199-K-20 recorded via manual measurements and sensor data. Two segments of daily average sensor data can be seen between (roughly) the years of 2004 to 2012 and 2015 to 2022. All other data points are less frequent manual measurements.

2.2 Data Exploration

Prior to developing a method to distinguish time segments, mixed-frequency groundwater elevation data were explored to investigate two possible approaches. The time difference between successive data points is an intuitive metric for distinguishing the density of data over time. Kernel density estimation is a method to estimate the probability density function of a random variable based on kernels as weights. Both metrics were explored to ascertain how well they provide insight into the data. Figures 3 and 4 show examples of both metrics applied to the groundwater elevation data for wells 199-K-20 and 299-W15-763, respectively.

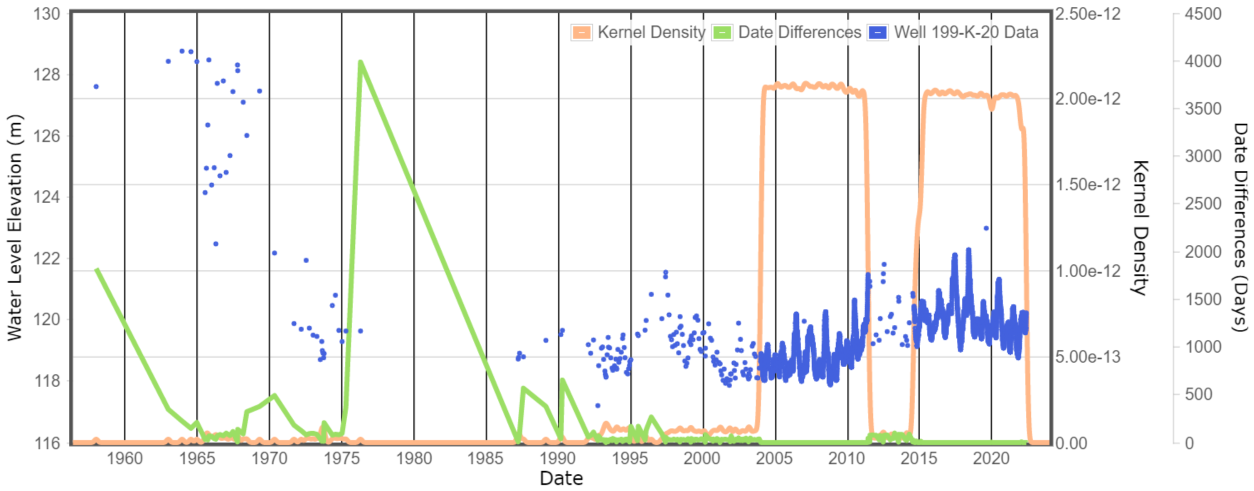


Figure 3. Data for well 199-K-20 (blue) along with the time differences between data points (green) and the KDE with an adjusted bandwidth (orange).

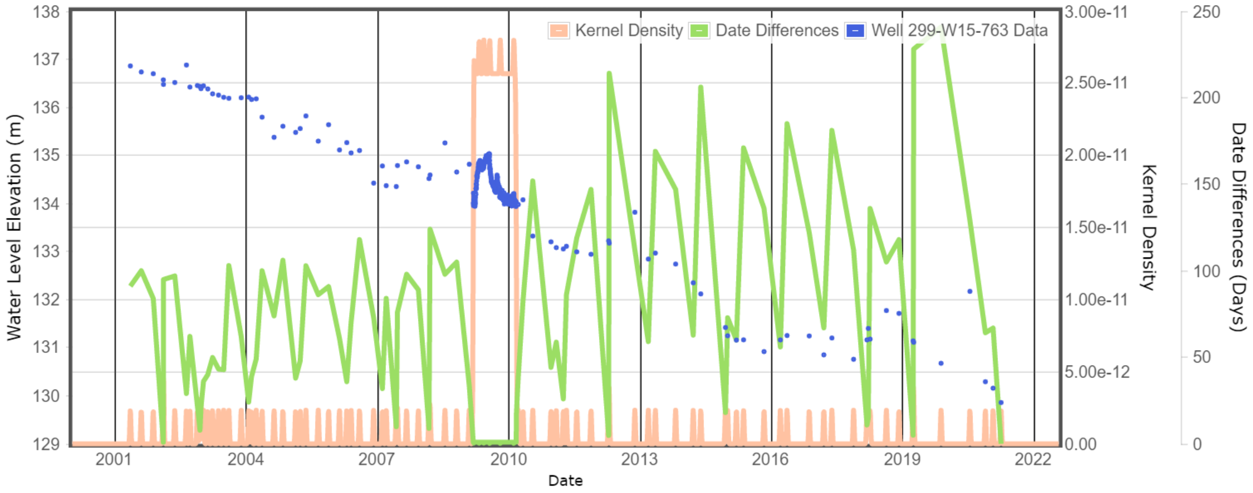


Figure 4. Data for well 299-W15-763 (blue) along with the time differences between data points (green) and the KDE with an adjusted bandwidth (orange).

Comparing time difference versus KDE results, the time difference curves were found to be more variable/erratic than the KDE curves. The KDE probability density representation, while less intuitive than time differences, gave a generally well-behaved result that was a promising basis to identify time segments of data with differing measurement frequencies. Subsequent sections describe the KDE calculation in detail, followed by the overall approach for identifying time segments with differing measurement frequency.

2.3 Gaussian Kernel Density Estimation

KDE is a nonparametric estimation of a dataset's unknown probability density function (Feigelson 2019). A plot of the KDE curve provides insight into the dataset's probability distribution. The shape of the KDE curve can be used in different applications, such as determining whether a sample dataset follows a normal distribution and thus satisfies assumptions for inference testing (Ahmad and Mugdadi 2003). The procedure developed here uses an adjusted bandwidth KDE applied to a groundwater elevation dataset's time values (timestamps).

The KDE of a dataset is calculated by defining a set of uniform evaluation points and then calculating the probability density for each evaluation point. Some applications choose to use 512 evaluation points that are uniformly spread across the total duration of the data. However, with time series data, it makes sense to select the number of evaluation points based on the time span of the dataset while making sure that sufficient points exist to adequately determine the change points. For this work, 120 evaluation points per year, with a minimum of 120 points for a dataset, were used. The KDE probability density can be calculated for each evaluation point using Equation 1 (Wikipedia 2022a). Computation of the KDE for all the evaluation points, *m*, and a dataset with *n* values is done in O(*mn*) time.

$$KDE(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - \mathrm{x}) \qquad (1)$$

Here, *n* represents the number of actual dataset points, K_h $(x_i - x)$ is the kernel function, *h* is the kernel bandwidth, *x* is a given evaluation point, and x_i is the time value for the i^th^ measured data point.

The KDE is based on a kernel function that applies weights to data in the calculation based on temporal distance from the evaluation point. For this work, a Gaussian kernel function, representing a normal distribution, was selected, though other kernel function options exist, such as an Epanechnikov kernel (Wikipedia 2022b, Soh et al. 2013). The Gaussian kernel function is calculated as shown in Equation 2.

$$K_h(x_i - \mathrm{x}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - x)^2}{2\sigma^2}} \qquad (2)$$

Here, σ is the standard deviation of the dataset, and the evaluation point x, would correspond to the dataset's mean, μ. All other symbols have their standard mathematical values. For our application with time series data, Equation 2 is better described as shown in Equation 3, where the standard deviation is equated to an adjusted bandwidth, 0.1h.

$$K_h(x_i - \mathrm{x}) = \frac{1}{0.1h\sqrt{2\pi}} e^{-\frac{(x_i - x)^2}{2(0.1h)^2}} \qquad (3)$$

The bandwidth, h, of the kernel is computed by using Silverman's "rule of thumb" (Feigelson 2019), which minimizes the mean integrated square error risk function and is calculated using Equation 4.

$$h = 0.9n^{-0.2}(\min{(sd, IQR)}) \qquad (4)$$

Here, *sd* is the sample standard deviation of the groundwater elevation dataset time values, and *IQR* is the interquartile range of the groundwater elevation dataset time values. The minimum of the standard deviation and IQR is chosen to compute h.

The Gaussian KDE using Silverman's "rule of thumb" for the bandwidth gave reasonable results, but lacked a sharp front and flat plateau, which makes sense because the bandwidth determines the "smoothness" of the KDE curve (Moreno 2019). Refinement by narrowing the bandwidth to 10% of the default value increased the sharpness of the KDE curve, which was deemed better for distinguishing change points. The 10% adjustment was selected based on evaluation using the selected well data.

To illustrate kernel density curves, as well as the data they reflect, Figures 5 and 6 show KDE results using the equations above for two groundwater wells.

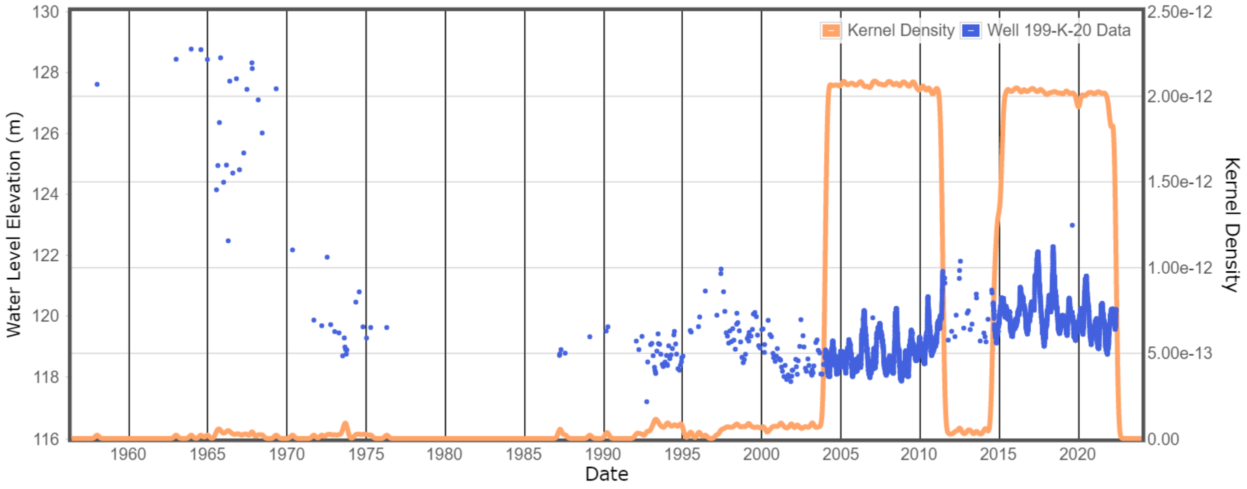


Figure 5. Time series data (blue) for well 199-K-20 in comparison to its KDE curve (orange). This well is comprised of mixed sampling frequencies, where high frequency data is indicated by the higher density of data points.

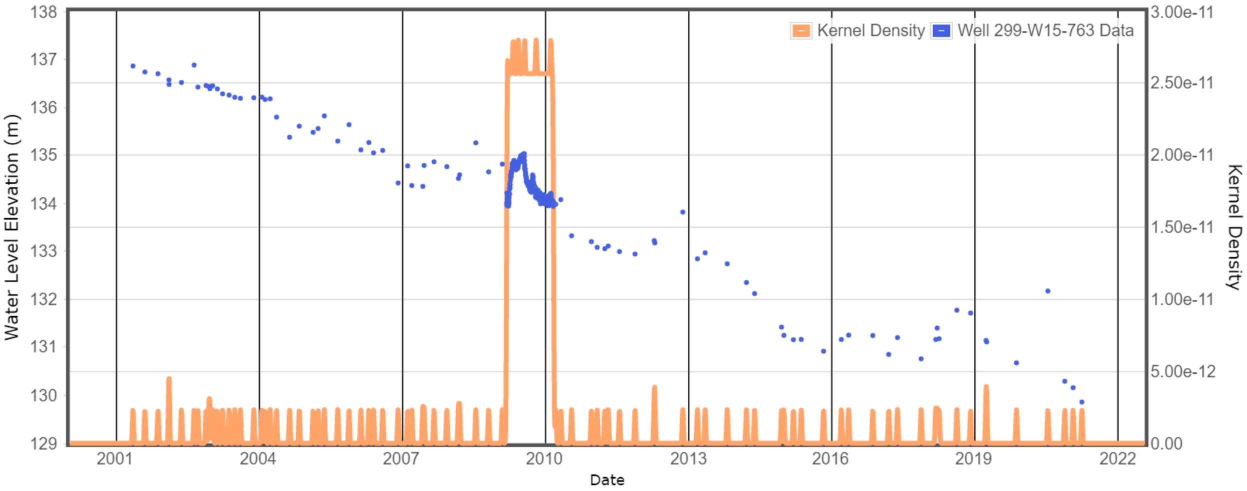


Figure 6. Time series data (blue) for well 299-W15-763 in comparison to its KDE curve (orange). This well is comprised of mixed sampling frequencies, where the low frequency data is sparse on both sides of the high frequency data centered around 2009 and 2010.

2.4 Time Segment Identification Based on Data Sampling Frequency

Given the sharp-front curve shape, a kernel density probability threshold of 1.2×10^{-12} was observed to correlate with known changes in measurement frequency and was selected as the metric to determine the preliminary set of change points. Figure 7 illustrates how this threshold and KDE curve form these segmentations, where segments whose kernel density curve lies above the threshold will be classified as HFD and segments whose kernel density curve lies below will be classified as LFD.

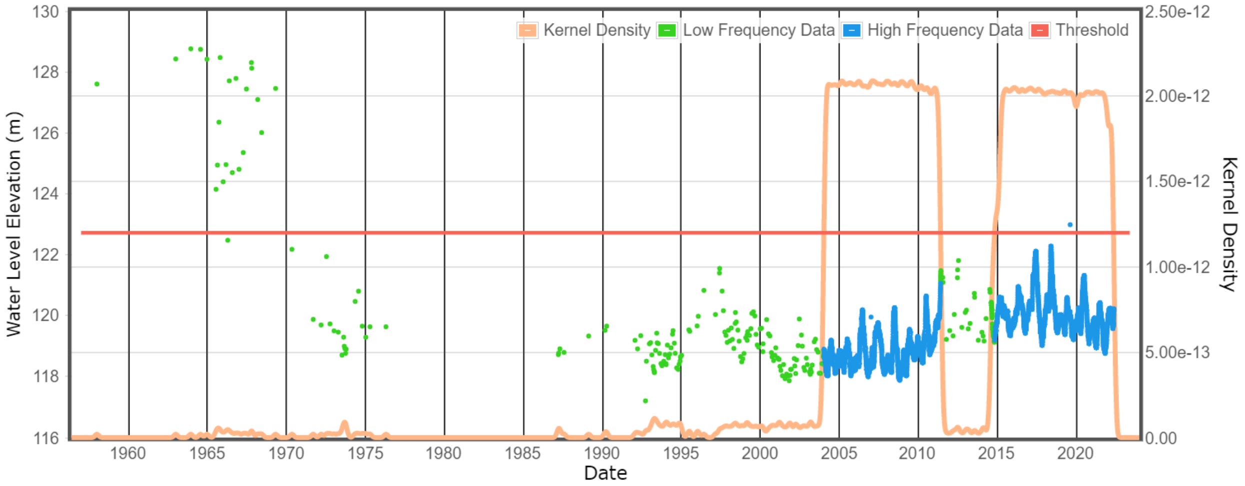


Figure 7. Time series data for well 199-K-20, separated by sampling frequency type (blue = HFD, green = LFD). This segmentation was done using the kernel density threshold (red) as a change-point detection criterion for the KDE curve (orange).

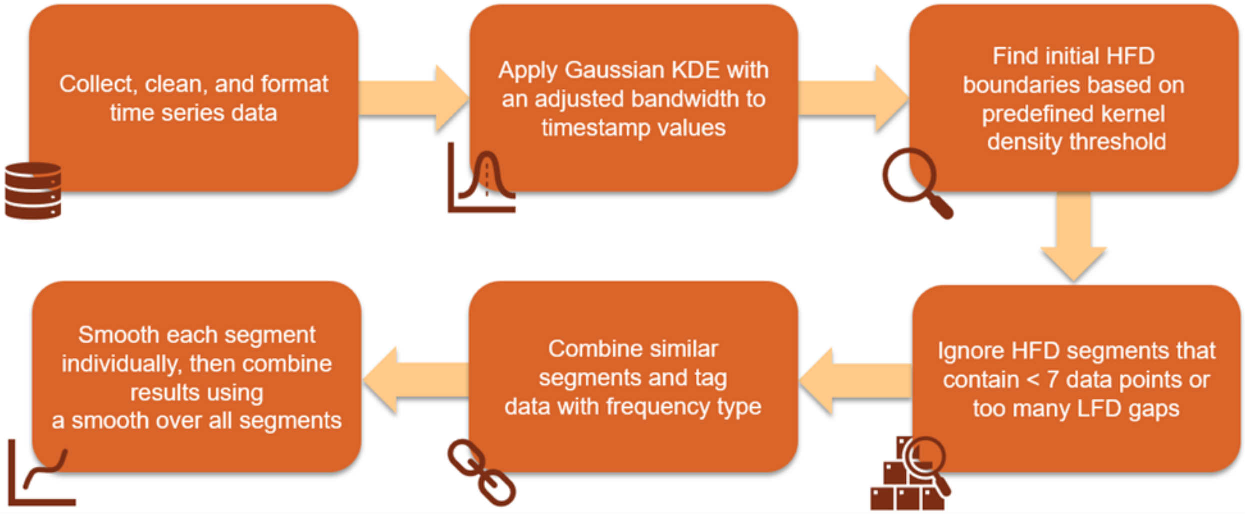
Although the selected threshold captures most change points for mixed measurement frequency data, some refinement was needed because sometimes LFD could be classified as HFD when the dataset's LFD were sparse enough. The procedure detailed in this report addresses incorrect categorization using assumptions and heuristics such as defining HFD as daily data and preferring that HFD segments contain enough data points to be classified as its own segment. Note that wells with only a single-data point are identified as LFD, because there is no basis for categorization as HFD.

Datasets with sparse LFD or those comprised completely of LFD may result in their respective KDE points peaking over the threshold. To verify segments that are truly HFD, these segments are assessed against two more criteria:

1) A HFD segment must contain at least seven datapoints.
2) A HFD segment must not exceed the tolerated percentage of LFD gaps in time between data points that are greater than the high-frequency definition (daily data in this application).

These criteria are defined in the code as parameters that could be adjusted for different kinds of datasets, allowing user definition of what qualifies as HFD (e.g., hourly data, data by the millisecond, etc.). However, the two criteria above suit the needs for groundwater elevation data.

Once the dataset's segments have been identified and filtered with these two criteria, adjacent time segments with the same classification can be combined. For example, if a data set has four segments in the order of low, low, high, and low frequency respectively, then the first two segments of LFD can be combined to form the dataset's new sequence of low, high, and low frequency data segments. Upon combining adjacent segments of similar measurement frequency types, a final set of time segments is obtained. Figure 8 summarizes the procedure described here.


Collect, clean, and format time series data

Apply Gaussian KDE with an adjusted bandwidth to timestamp values

Find initial HFD boundaries based on predefined kernel density threshold

Ignore HFD segments that contain < 7 data points or too many LFD gaps

Combine similar segments and tag data with frequency type

Smooth each segment individually, then combine results using a smooth over all segments

Figure 8. A flowchart summary of the procedure described in this report.

Figure 9 illustrates a case of water elevation data that has sparse LFD segments with kernel density values that exceed the selected threshold, but are classified as LFD based on the dataset assumptions and the defined heuristics/criteria.

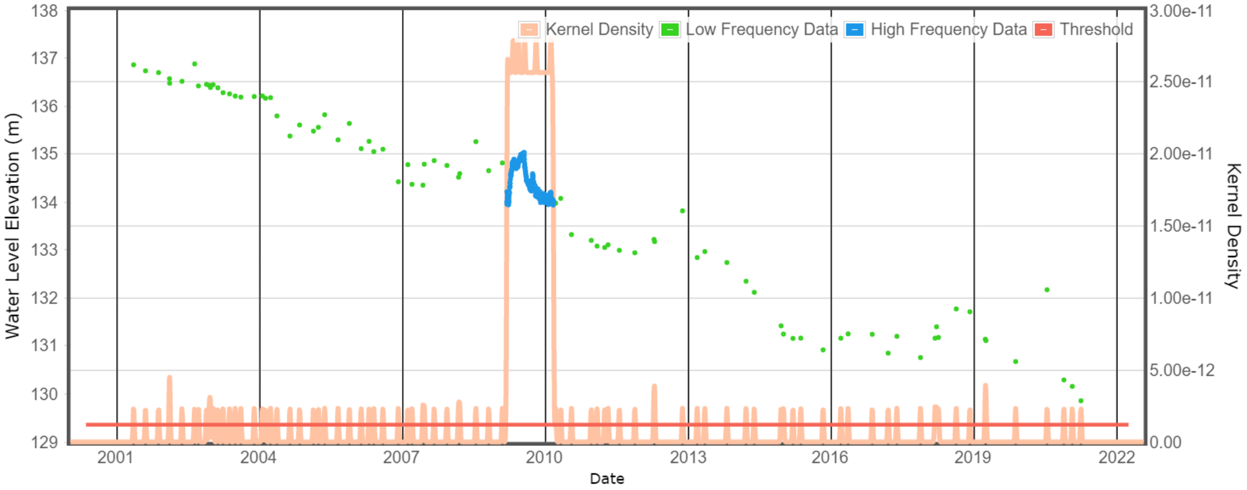


Figure 9. Time series data for well 299-W15-763, separated by sampling frequency type (blue = HFD, green = LFD). This segmentation was done using the kernel density threshold (red) as a change-point detection criterion for the KDE curve (orange), as well as non-qualifying data criteria, and the combination of similar windows.

3.0 Results

This section discussed results of the data frequency segmentation approach and its application to smoothing groundwater elevation data.

3.1 Procedure Results

Specific groundwater well data from the Hanford Site was used with the time segmentation approach developed in this work. Groundwater elevation data for multiple wells (Table 1) were used to test application with a range of different time series distributions, including cases with different patterns of mixed LFD and HFD and datasets of a single frequency type. The data obtained from GALEN were formatted into a two-dimensional JavaScript (JS) array of date-water elevation pairs. The calculations for the developed approach were coded in JS, a language that can handle large amounts of data and which includes robust array methods for data manipulation and analysis. However, this procedure could instead be written in other languages (e.g., Python, Excel/VBA, MATLAB).

The resulting time segments of differing measurement frequencies that were determined by the developed approach match well with the actual low- and high-frequency data segments, as illustrated by the examples in Figures 10 and 11. See Appendix A for additional examples for wells with different sampling frequency patterns.

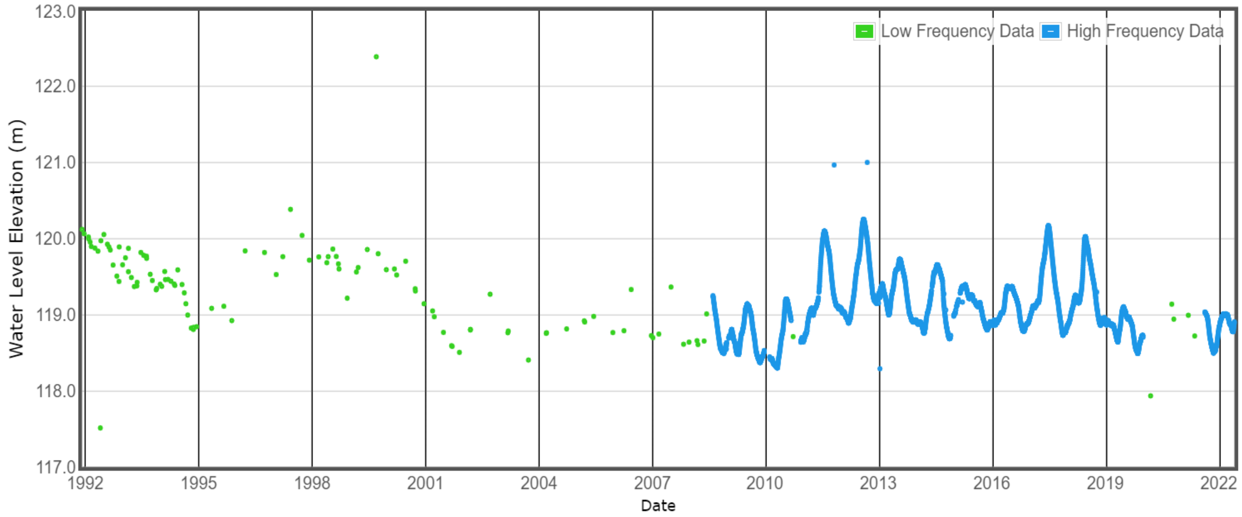


Figure 10. Time series data for well 199-N-72, with sampling frequency type identified by the KDE approach (blue = HFD, green = LFD).

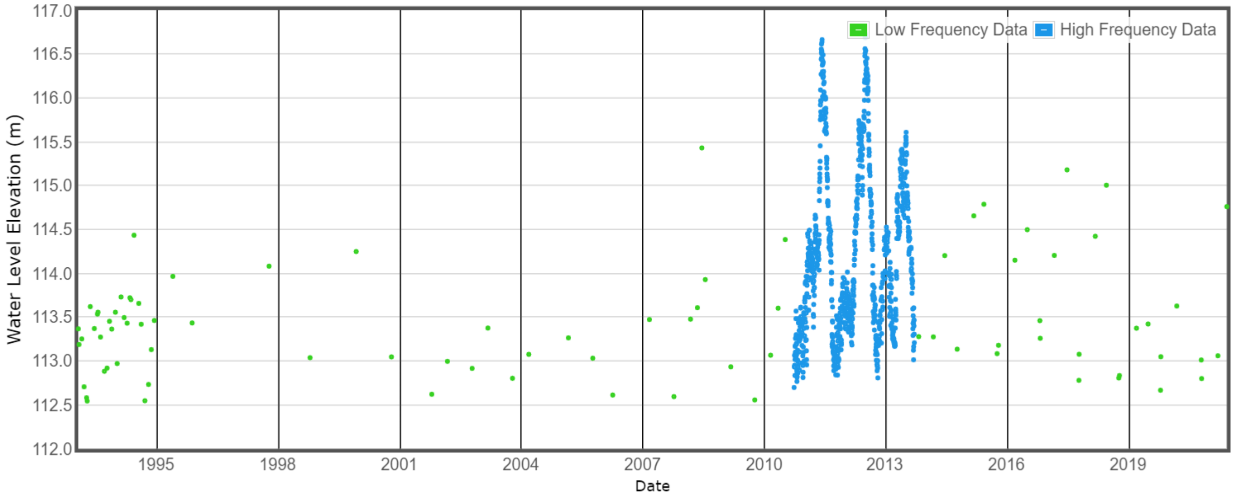


Figure 11. Time series data for well 199-F5-43A, with sampling frequency type identified by the KDE approach (blue = HFD, green = LFD).

3.2 Application

As discussed in Section 1.0, applying data analysis techniques, such as smoothing, to a dataset with time segments of differing measurement frequencies may not give a representative result because of the differing characteristics of the data over time. Smoothing of data to get a representative result regardless of the differences in measurement frequency is a key motivation for this work. Kernel smoothing uses the Nadaraya-Watson estimator (e.g., Jones et al., 1994) to apply a locally weighted average across a sliding time window, using a so-called kernel function as the weighting function. The GALEN module of the SOCRATES software (Royer et al. 2018, Freedman 2021, Brouns and Johnson 2021) currently offers two Gaussian kernel smoothing options (31-day and 181-day time windows), which are applied across the entire groundwater elevation dataset, regardless of the measurement frequency. Figure 12 shows an example of Gaussian kernel smoothing (using a 181-day sliding window) that is applied to the entire groundwater elevation dataset for well 199-N-72.

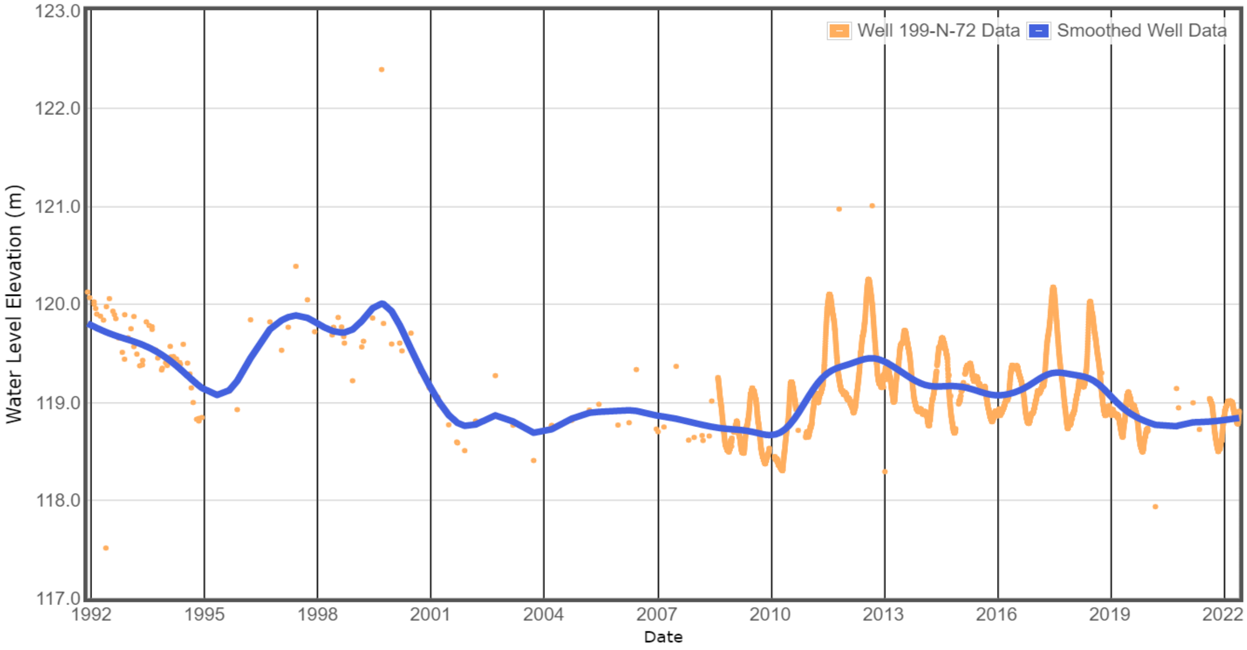


Figure 12. Application of a single Gaussian kernel smooth (blue) for the entire well 199-N-72 groundwater elevation dataset (orange) using a 181-day window size.

The approach developed in this work (Section 2.4) offers the potential for an improved smoothing representation of Hanford groundwater elevation data. To facilitate the use of smoothing across mixed measurement frequency data, qualitative smoothing levels (e.g., fine, medium, coarse) are defined with specific kernel smoothing window sizes for LFD and HFD segments as listed in Table 2. The window sizes were established for the smoothing levels based on an understanding of the type of data generally observed for groundwater elevations (with manual LFD measurements occurring at a monthly to annual rate and sensor HFD measurements generally occurring daily).

Table 2. Smoothing window sizes of LFD and HFD at each smoothing level.

Smoothing Level	LFD Window Size	HFD Window Size
Super Coarse	361	181
Coarse	271	91
Medium	181	61
Fine	91	31
Super Fine	31	11

Kernel smoothing with a Gaussian kernel function was applied to the groundwater elevation data for the selected wells (Table 1) independently for each time segment of different measurement frequency. The smoothing results for individual time segments were then merged by applying a kernel smooth across all of those results using a small smoothing window size (10 days). Figure 13 shows the result of this process using “medium” smoothing on groundwater elevation data for well 199-N-72. The result is more detail in the HFD segments and less detail in the LFD segments, which aligns with the associated density of the data. Additional examples of application of the time segmentation and smoothing are shown in Appendix A for a range of data frequency patterns.

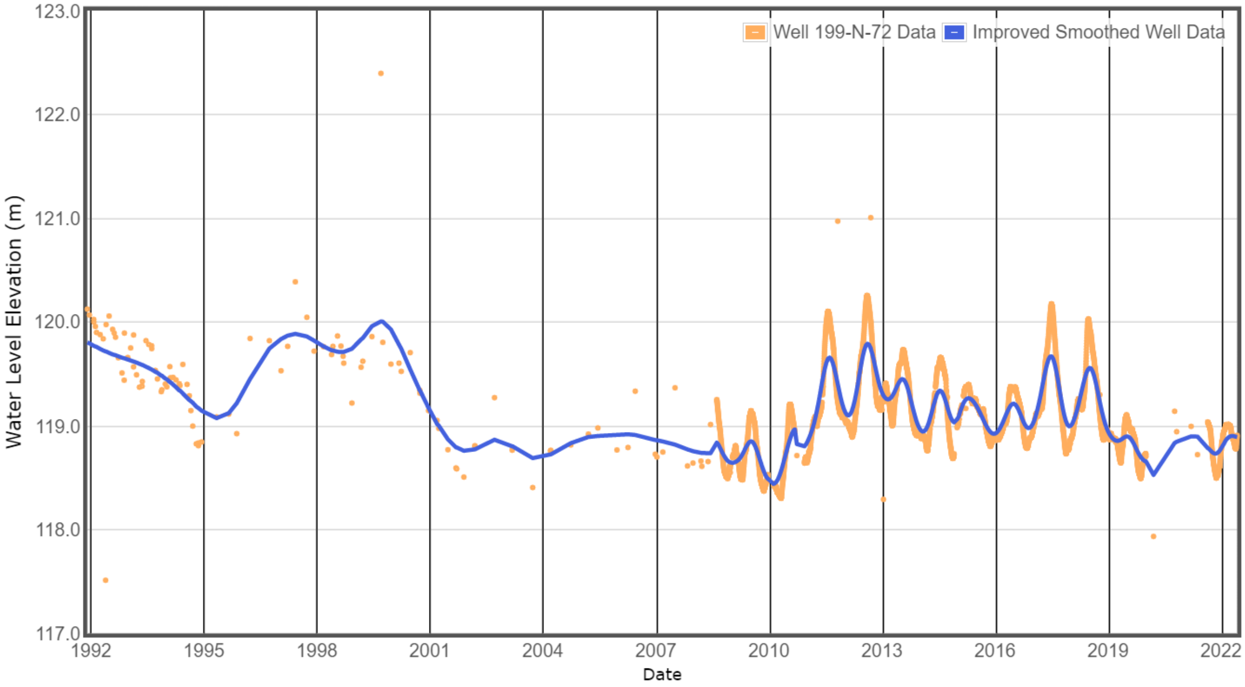


Figure 13. Application of Gaussian kernel smoothing (blue) at a medium level (window size of 181-days for LFD and 61-days for HFD) separately for the identified time segments using groundwater elevation data (orange) for well 199-N-72.

Application of the time segmentation approach to Hanford well data demonstrates that the developed method can improve the smoothing results for better representation and interpretation of the data to better support evaluations and decision making. However, future work could include investigation and further optimization of the approach. Different KDE kernels, standardization (normalization) of parameters and results, testing with different kinds of datasets, and application of other data analyses to the time segments are all items that could be further studied to improve the effectiveness and applicability of this approach. Note that, in Appendix A, the smoothing is extended from Gaussian kernel smoothing to running median smoothing in an example.

4.0 Conclusion

The change-point detection and time segmentation approach developed in this work can identify time segments with low or high data measurement frequencies. For groundwater elevation datasets with mixed measurement frequency data, the KDE probability density curve versus time is a good indicator of change points between measurement frequency regimes. Segments of HFD can be determined from the identified change points, which are the intersections of the KDE curve and the established kernel density threshold. The segments delineated by these change-points need to be refined by applying a criterion to check whether the segments contain sufficient data points to warrant being its own segment. Another criterion is that HFD segments cannot contain a surplus of low-frequency measurements (i.e., with large gaps in time within the segment).

Application of the change-point detection and time segment identification is beneficial for time series data analysis because time segmentation can reveal characteristics of the dataset that are a function of different regimes of measurement frequency. The approach developed here using a Gaussian KDE-based segmentation analysis is useful for applications such as data smoothing. With segments classified by their measurement frequency type, data analysis techniques can be applied in a segment-specific manner, using parameters consistent with the time segment data collection frequency. A good example of how identification of measurement frequency segments can lead to a more representative result is the application of smoothing to a time series dataset. A single smooth of mixed-frequency data may suit one data frequency type and may not be as representative for the other type. To remedy this, the identified high and low measurement frequency time segments can be analyzed (e.g., smoothed) individually with appropriate levels of smoothing for each segment type.

While the method developed here is useful, there are opportunities to refine and expand the approach. Future work could consider alternate KDE kernels and a method to normalize the KDE probability density curve (i.e., to span from 0 to 1). Expansion to other data analyses, as well as applications with different types of datasets, would also be useful to extend the applicability of this approach.

5.0 References

Ahmad, I., and A. Mugdadi. 2003. "Testing Normality Using Kernel Methods." *J. Nonparametric Statistics*, 15:273-288. https://doi.org/10.1080/1048525021000049649.

Aminikhanghahi, S., and D.J. Cook. 2017. "A Survey of Methods for Time Series Change Point Detection." *Knowledge and Information Systems*, 51:339-367. https://doi.org/10.1007/s10115-016-0987-z.

Brouns, T., and C.D. Johnson. 2021. "SOCRATES: Suite Of Comprehensive Rapid Analysis Tools for Environmental Sites" (website). Pacific Northwest National Laboratory, Richland, Washington. Available at: https://www.pnnl.gov/projects/socrates (accessed July 6, 2022).

Chen, W., and C.-C.J. Kuo. 1996. "Change-point Detection Using Wavelets." *SPIE Proceedings Vol. 2750: Digital Signal Processing Technology*, Society of Photo-Optical Instrumentation Engineers, Bellingham, Washington. https://doi.org/10.1117/12.241984.

Feigelson, E. 2019. "Nonparametric density estimation or Smoothing the data." *Summer School in Statistics for Astronomers*. Available at: https://astrostatistics.psu.edu/su19/19Lectures/R_IIRdens.pdf (accessed on July 12, 2022).

Freedman, V.L. 2021. "SOCRATES Suite of Comprehensive Rapid Analysis Tools for Environmental Sites." Pacific Northwest National Laboratory, Richland, Washington. Available at: https://www.pnnl.gov/sites/default/files/media/file/PNNL_SOCRATES_brochure_2021.pdf (accessed July 6, 2022).

Jones, M.C., S.J. Davies, and B.U. Park. 1994. "Versions of Kernel-type Regression Estimators." *J. Am. Stat. Assn.*, 89(427):825-832.

Lavielle, M. 2017. "Detection of Change Points in a Time Series." Statistics in Action with R (website). Available at: http://sia.webpopix.org/changePoints.html (accessed on July 5, 2022).

Lombard, F. 1988. "Detecting Change Points by Fourier Analysis." *Technometrics*, 30(3):305-310. https://doi.org/10.2307/1270084.

Moreno, M.C. 2019. "Understanding Gaussian Kernel Density: A 'by (R)Hand' Introduction." (website) RPubs, RStudio, Boston, Massachusetts. Available at: https://rpubs.com/mcocam12/KDF_byHand (accessed on June 16, 2022).

Royer, P.D., C.D. Johnson, and M.J. Truex. 2018. "Suite of Comprehensive Rapid Analysis Tools for Estimation (SOCRATES)." In: *WM Symposium 2019*. WM Symposia, Inc., Tempe, Arizona. Available at: http://amz.xcdsystem.com/A464D2CF-E476-F46B-841E415B85C431CC_finalpapers_2019/FinalPaper_19076_0123012553.pdf (accessed July 6, 2022).

Schroth, C., J. Siebert, and J. Groß. 2021. "Time Traveling with Data Science: Focusing on Change Point Detection in Time Series Analysis (Part 2)." (website) Fraunhofer IESE, Kaiserslautern, Germany. Available at: https://www.iese.fraunhofer.de/blog/change-point-detection/ (accessed on July 6, 2022).

Shi, X. 2020. "A Survey of Changepoint Techniques for Time Series Data." (thesis) All Dissertations #2697, Clemson University, Clemson, South Carolina. Available at: https://tigerprints.clemson.edu/all_dissertations/2697 (accessed July 6, 2022).

Shi, X., C. Gallagher, R. Lund, and R. Killickc. 2022. "A Comparison of Single and Multiple Changepoint Techniques for Time Series Data." *Computational Statistics & Data Analysis*, 170:107433. https://doi.org/10.1016/j.csda.2022.107433.

Soh, Y., Y. Hae, A. Mehmood, R.H. Ashraf, and I. Kim. 2013. "Performance Evaluation of Various Functions for Kernel Density Estimation." *Open Journal of Applied Sciences*, 3:58-64.

Truong, C., L. Oudre, and N. Vayatis. 2019. "Greedy Kernel Change-Point Detection." *IEEE Transactions on Signal Processing*, 67(24):6204-6214. https://doi.org/10.1109/TSP.2019.2953670.

Truong, C., L. Oudre, and N. Vayatis. 2020. "Selective Review of Offline Change Point Detection Methods." *Signal Processing*, 167:107299. https://doi.org/10.1016/j.sigpro.2019.107299.

van den Burg, G.J.J., and C.K.I. Williams. 2022. "An Evaluation of Change Point Detection Algorithms." arXiv:2003.06222v3. Available at: https://arxiv.org/pdf/2003.06222.pdf (accessed July 6, 2022).

Wang, G., S.B. Hariz, J.J. Wylie, Q. Zhang. 2008. "Change-Point Detection for Continuous Processes with High-Frequency Sampling." *Comptes Rendus Mathematique*, 346(7-8): 467-470. https://doi.org/10.1016/j.crma.2008.02.005.

Wikipedia. 2022a. "Kernel Density Estimation." (website) Wikimedia Foundation, Inc., San Francisco, California. Available at: https://en.wikipedia.org/wiki/Kernel_density_estimation (accessed on July 3, 2022).

Wikipedia. 2022b. “Kernel (statistics)” (website). Wikimedia Foundation, Inc., San Francisco, California. Available at: https://en.wikipedia.org/wiki/Kernel_(statistics) (accessed July 6, 2022)

Appendix A – Additional Well Data Examples

This appendix includes examples of the change-point detection and time segment identification approach developed in this work using groundwater elevation data from additional Hanford Site wells. These additional examples represent other data measurement frequency patterns.

Figures A.1 to A.4 show the time segment identification for different patterns of HFD and LFD.

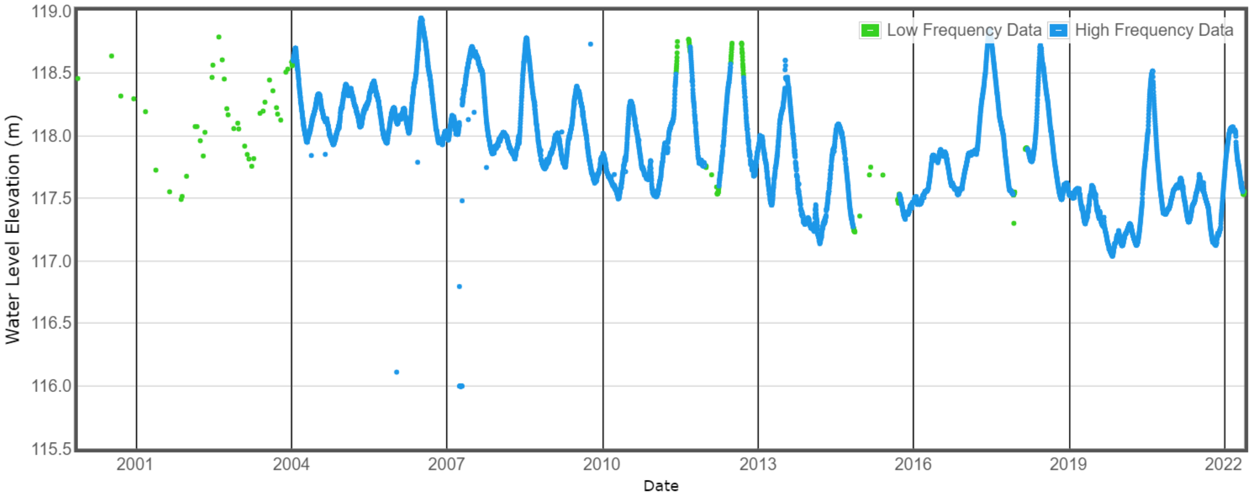


Figure A.1. Time series data for well 199-D5-38, with time segments of both HFD (blue) and LFD (green).

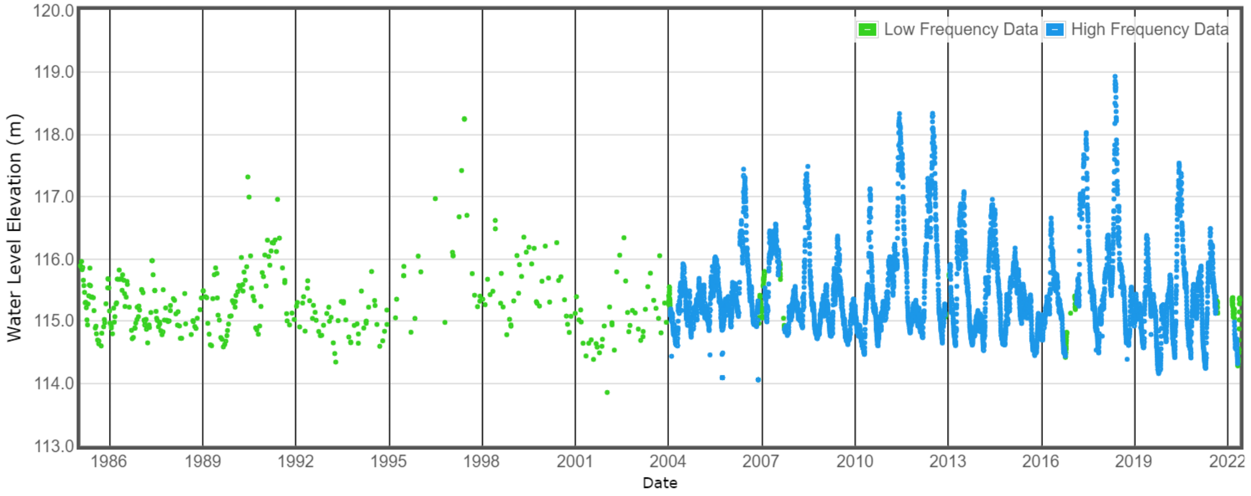


Figure A.2. Time series data for well 199-H4-5, with time segments of both HFD (blue) and LFD (green).

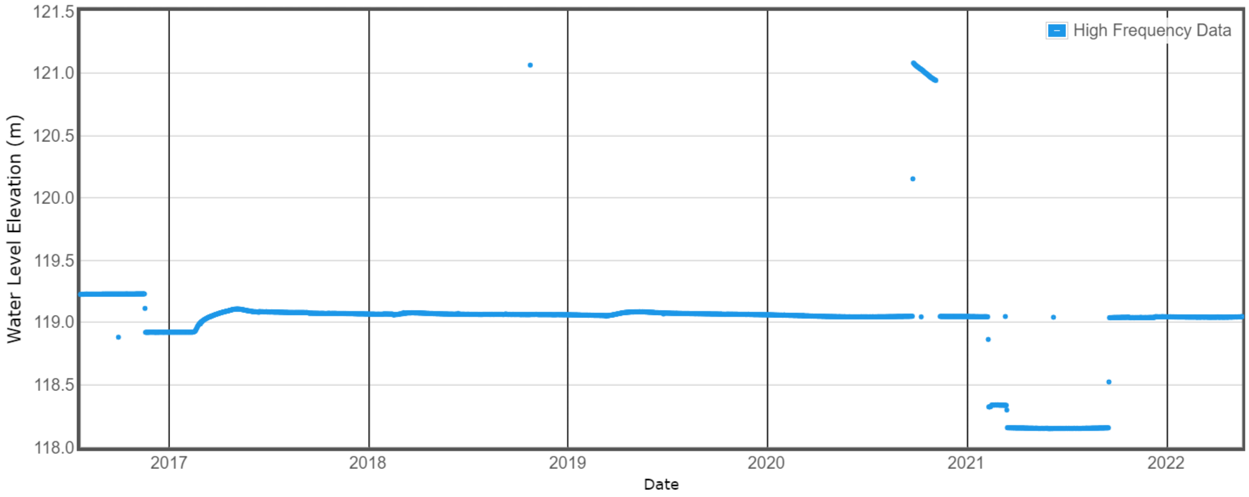


Figure A.3. Time series data for well 699-66-32, which is found to consist of all HFD (blue).

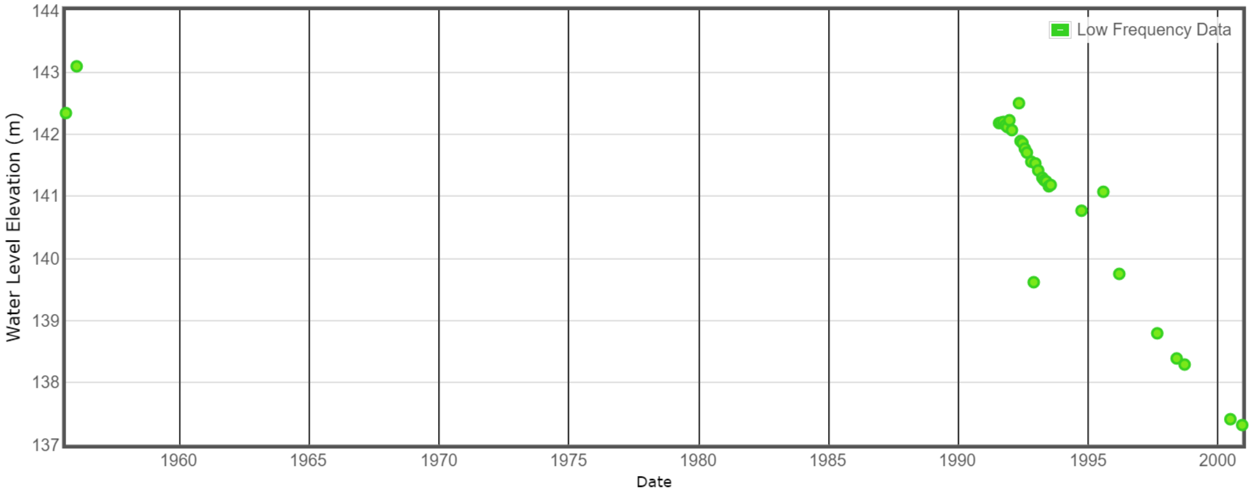


Figure A.4. Time series data for well 299-W23-2, which is found to consist of all LFD (green).

Figures A.5 and A.6 show smoothing results for groundwater elevation data from well 199-N-72. Figure A.5 compares Gaussian kernel smoothing to running median smoothing for the same level of smoothing. Figure A.6 compares five different levels (Table 2) of Gaussian kernel smoothing. Both figures apply smoothing separately to the identified HFD and LFD segments.

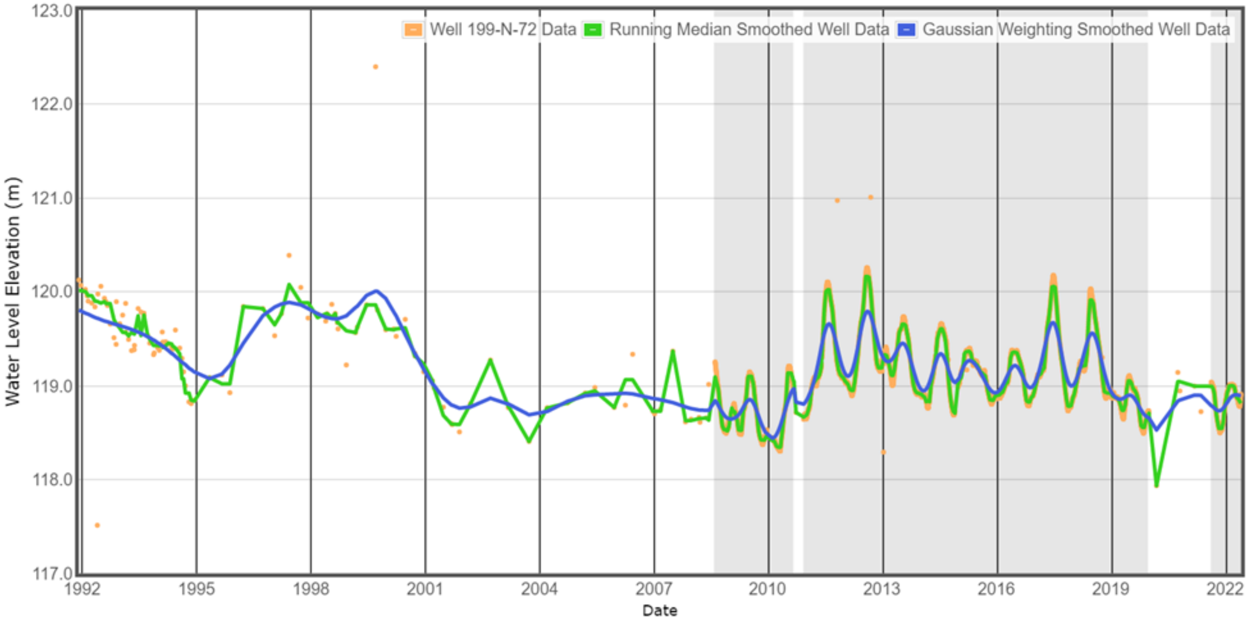


Figure A.5. Time series data (orange) for well 199-N-72 with Gaussian kernel smoothing (blue) and running median smoothing (green) based on the time segmentation approach and a "medium" level of smoothing. The shaded zones are HFD time segments.

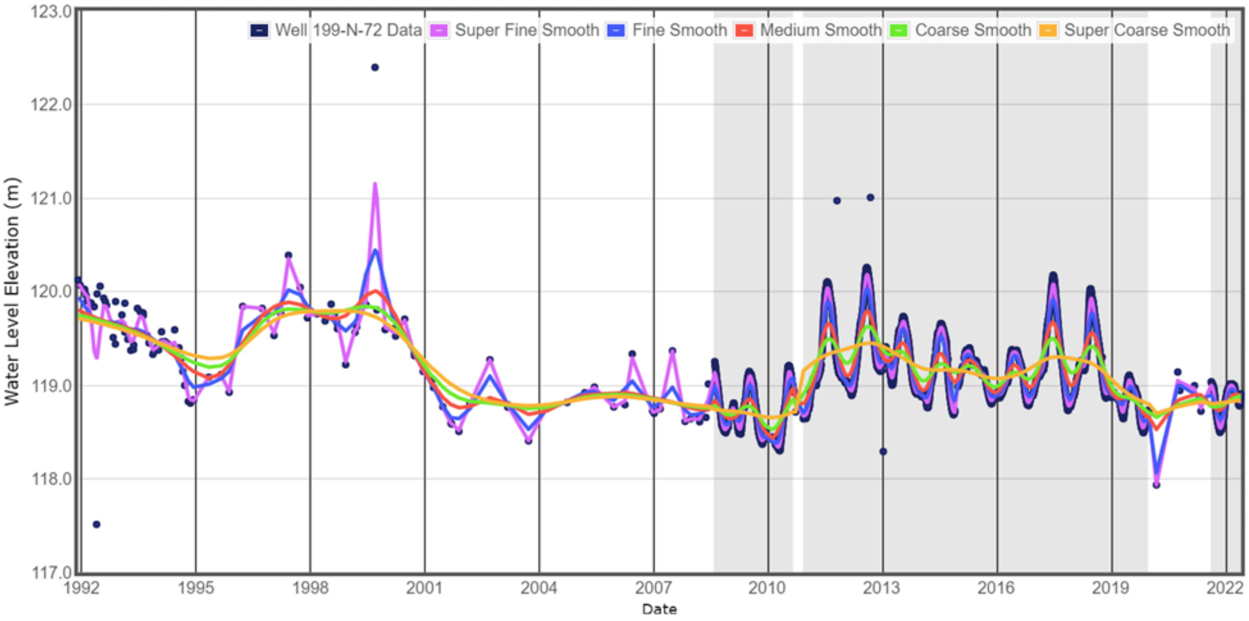


Figure A.6. Time series data (navy blue) for well 199-N-72 with Gaussian kernel smoothing at five different smoothing levels (Table 2), based on the time segmentation approach. The shaded zones are HFD time segments.

Figures A.7 to A.10 show comparisons between a Gaussian kernel smooth for all data at once (with a 181-day window) and Gaussian kernel smooths applied separately to identified HFD and LFD time segments (with 61-day and 181-day time windows, respectively) for groundwater elevation data from four Hanford wells. These comparisons illustrate how the approach developed in this work can improve the smoothed result.

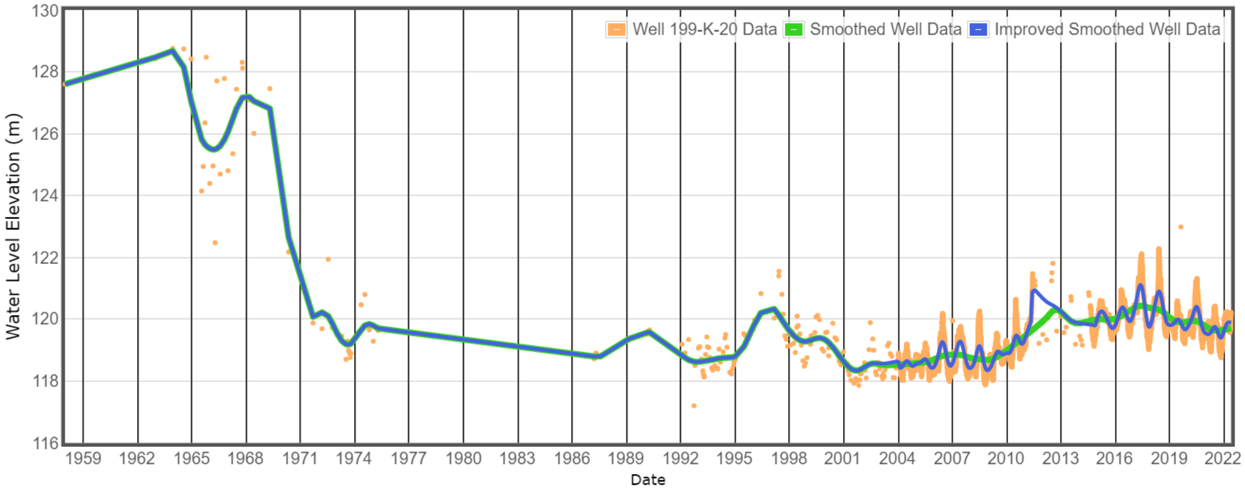


Figure A.7. Time series data (orange) for well 199-K-20 with Gaussian kernel smoothing results for all data (green, 181-day window) and based on the improved time segmentation approach (blue, 61- and 181-day windows for LFD and HFD, respectively).

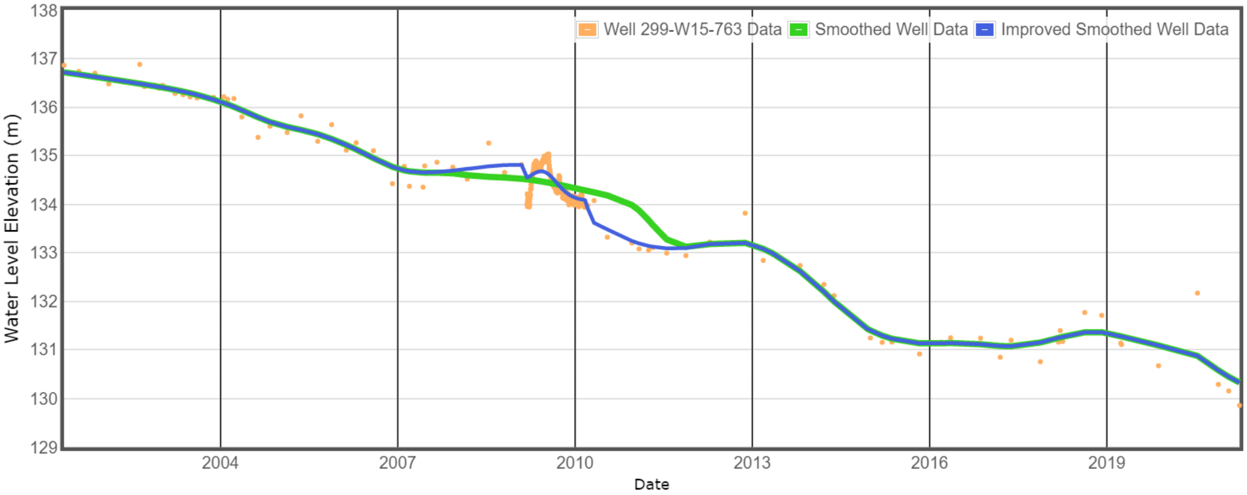


Figure A.8. Time series data (orange) for well 299-W15-763 with Gaussian kernel smoothing results for all data (green, 181-day window) and based on the improved time segmentation approach (blue, 61- and 181-day windows for LFD and HFD, respectively).

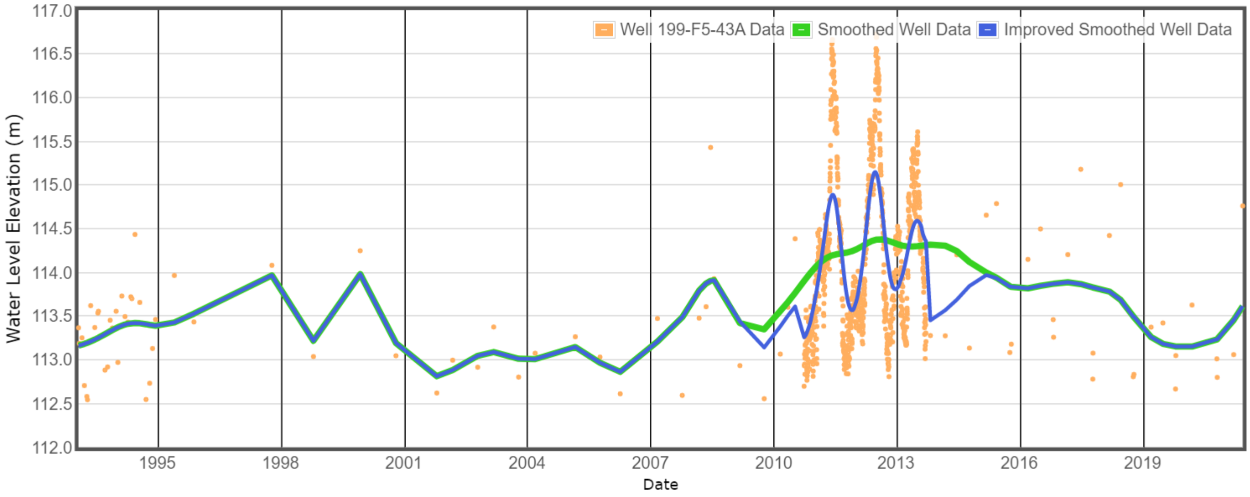


Figure A.9. Time series data (orange) for well 199-F5-43A with Gaussian kernel smoothing results for all data (green, 181-day window) and based on the improved time segmentation approach (blue, 61- and 181-day windows for LFD and HFD, respectively).

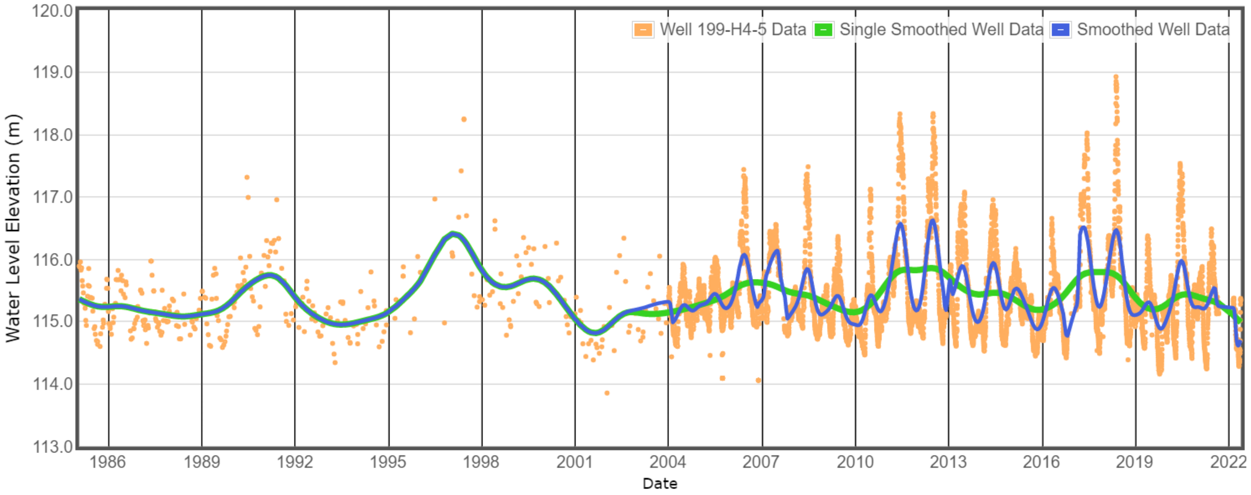


Figure A.10. Time series data (orange) for well 199-H4-5 with Gaussian kernel smoothing results for all data (green, 181-day window) and based on the improved time segmentation approach (blue, 61- and 181-day windows for LFD and HFD, respectively).

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov