A Local Correlation Score to Monitor Sensor Drift of Real-Time Environmental Data

Ian Taylor, Dr. Julia L. Sharp, Dr. David L. White, Dr. Jason O. Hallstrom, Dr. Gene W. Eidson

Clemson University, Clemson, SC

Introduction

- Quality control (QC) of environmental streaming data requires multiple levels of automated checks in order to produce quality data in a timely manner.
- The Intelligent River® project's automated QC applies the Local Correlation Score to pairs of sensors in real time to detect sensor drift and changes in overall sensor performance [1].
- The Local Correlation (LoCo) Score can be computed using either of two methods: exponential window or sliding window [2]. Each method depends upon three parameters:
  - Window size (w).
  - The number of eigenvectors to use from the autocovariance matrix decomposition (k).
  - The sliding window method relies on the number of previous windows taken into account (m).
  - The exponential window method relies on the exponential decay factor of the weight of each previous window in consideration (β, $0 < \beta < 1$).

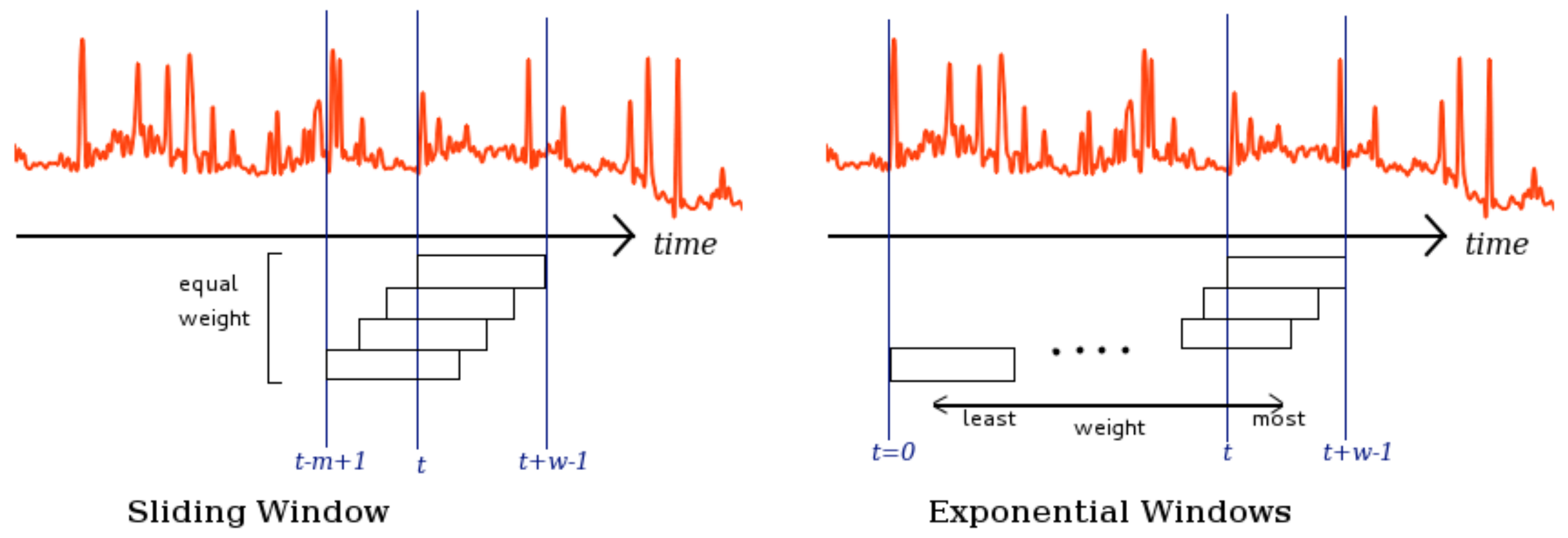

Figure: Illustration of the sliding and exponential window methods. (Adapted from similar figure in [2].)

- Our goal is to determine which combination of values for these parameters produces a sensitive correlation score to detect anomalies in pairs of sensors that may be flagged as potentially erroneous.

Methods

- The **Intelligent River Project®** provides real-time monitoring, analysis and management of water resources in South Carolina through the deployment of many sensors placed in watersheds throughout the state [1].
- The **Local Correlation (LoCo) Score** is a measure designed to capture the local correlations among pairs of time-evolving time series.
  - The exponential window method weights each previous window based on powers of the parameter β, while the sliding window method uses m previous windows all of equal weight.
  - The LoCo score is based on the computation of the autocovariance matrix for each time series, and then comparing these matrices via their matrix decomposition.
  - A generalized notion of linear cross-correlation is produced [2].

Results

- The LoCo score was calculated using turbidity data from second-order streams located in Dunn Hollow and Hembree Hollow in Eastern Tennessee from July 6, 2010 and July 13, 2010. These locations are ideal because of their physical proximity, but have differing levels of anthropogenic activity near each site.
- Observations within the two data sets are typically similar, with occasional anomalies in correlation due to extraneous factors like human interference and environmental conditions.

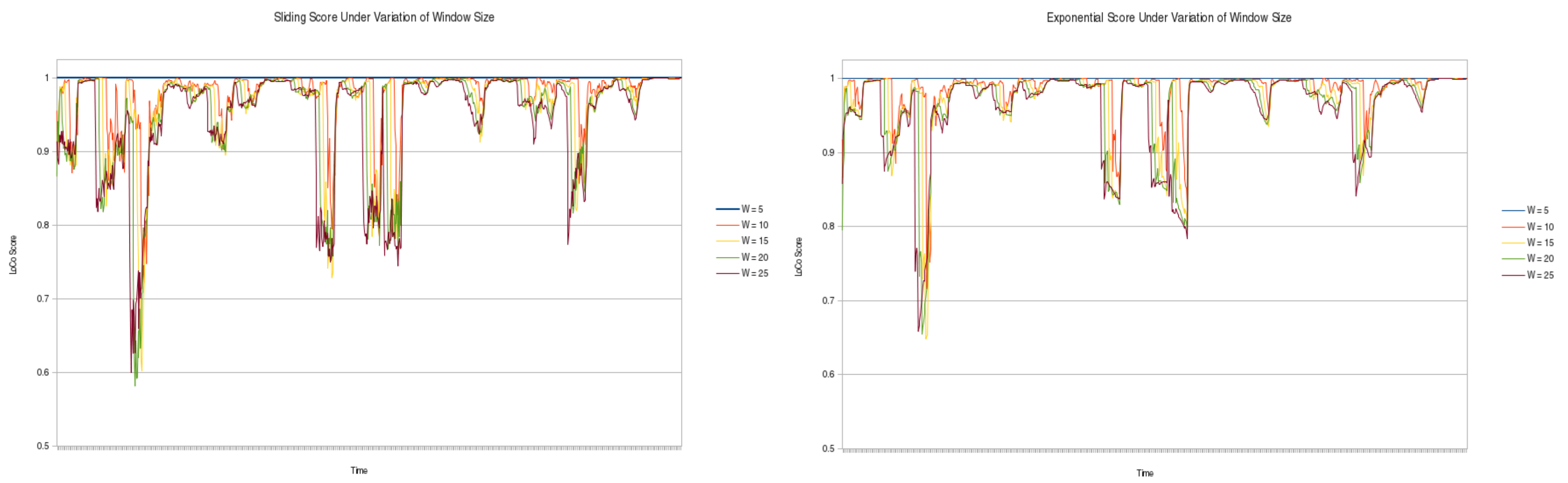

Figure: Results of sliding window method and exponential window method with varying window sizes. Other parameters are held constant at $k = 5$, $m = 2$, and $\beta = 0.5$.

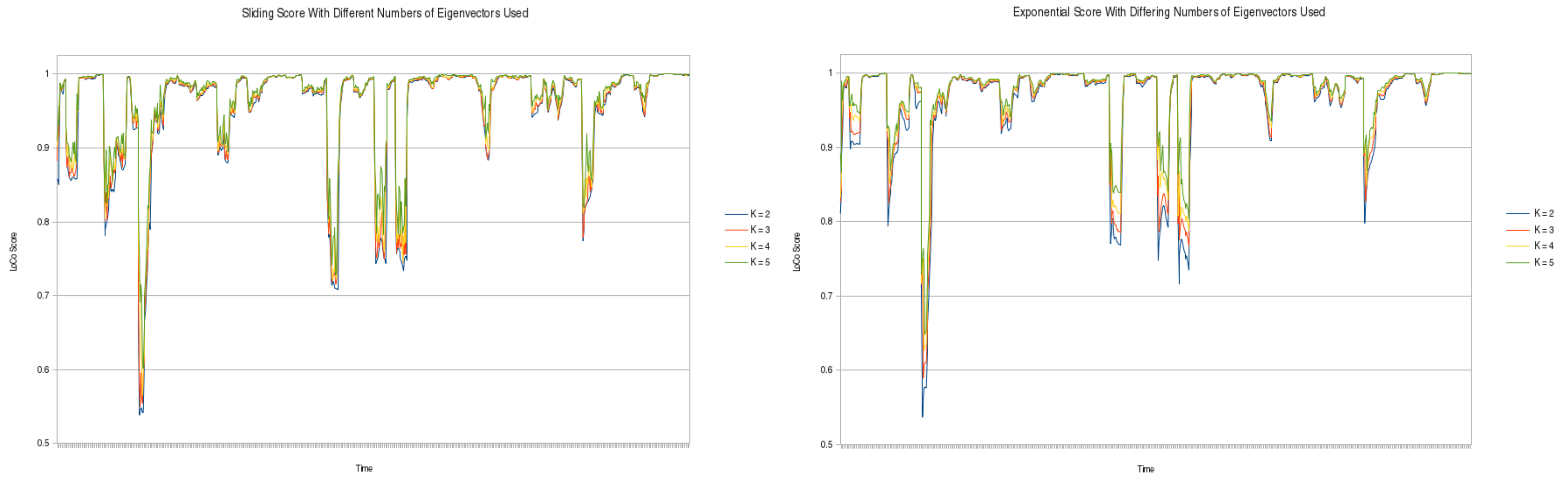

Figure: Results of sliding window method and exponential window method with varying numbers of eigenvectors used in computation. Other parameters are held constant at $w = 15$, $m = 2$, and $\beta = 0.5$.

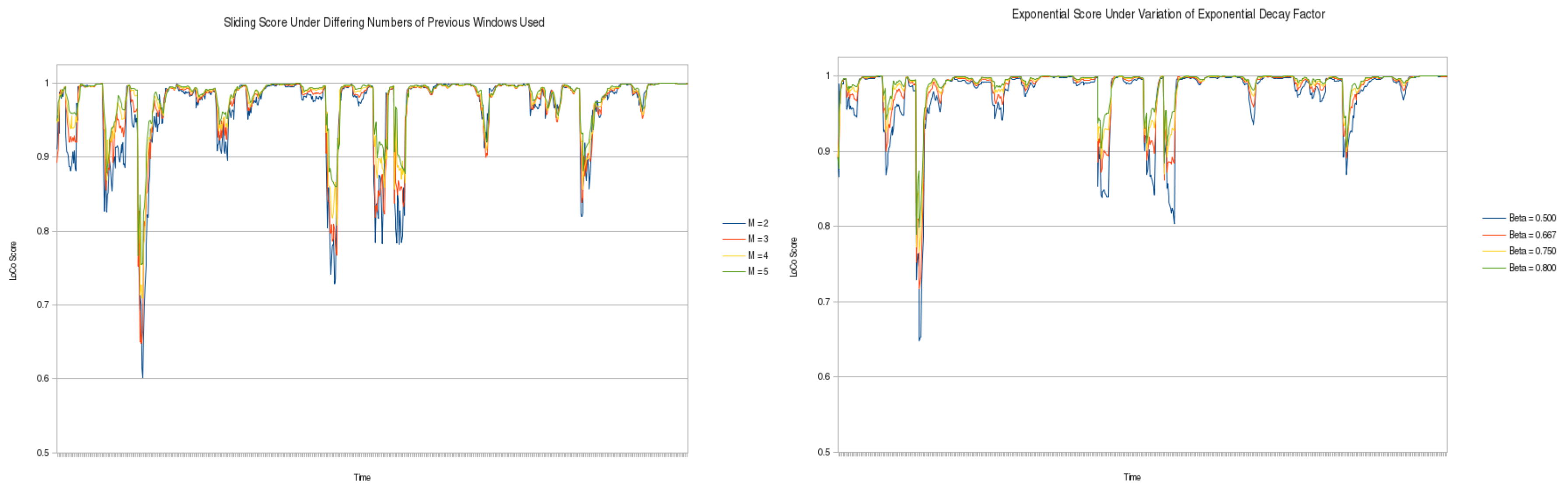

Figure: Results of the sliding window method with varying numbers of previous windows used in computation, and of the exponential window method with varying exponential decay constants. Other parameters are held constant at $w = 15$ and $k = 5$. (Note: the values of m and β were selected because they loosely correspond to each other in function [1].)

Conclusions

- For both the sliding window and exponential window method:
  - Anomalies are detected more reliably as window size increases (i.e., when the window size is large, anomalies are detected more quickly). However, the benefits of a larger window size ($w \geq 15$) are minimal, as the correlation scores do not differ greatly for these window sizes.
  - A fewer number of eigenvectors (k) used in computation seems to also be adequate in detecting anomalies, as well as reducing the LoCo score computation time.
- For the sliding window method, the LoCo score is most sensitive to detect anomalies when a lower value of m is used. However, this may result in the exaggeration of smaller anomalies.
- For the exponential method, the LoCo score is most sensitive to detect anomalies when a lower value of β is used. However, lower values of β do not exaggerate smaller anomalies.
- From our study, we determine that for QC of environmental, streaming data, the exponential LoCo score, with LoCo parameter values $w = 15$, $k = 2$, and $\beta = 0.5$, is most appropriate for identifying anomalies among pairs of sensors.

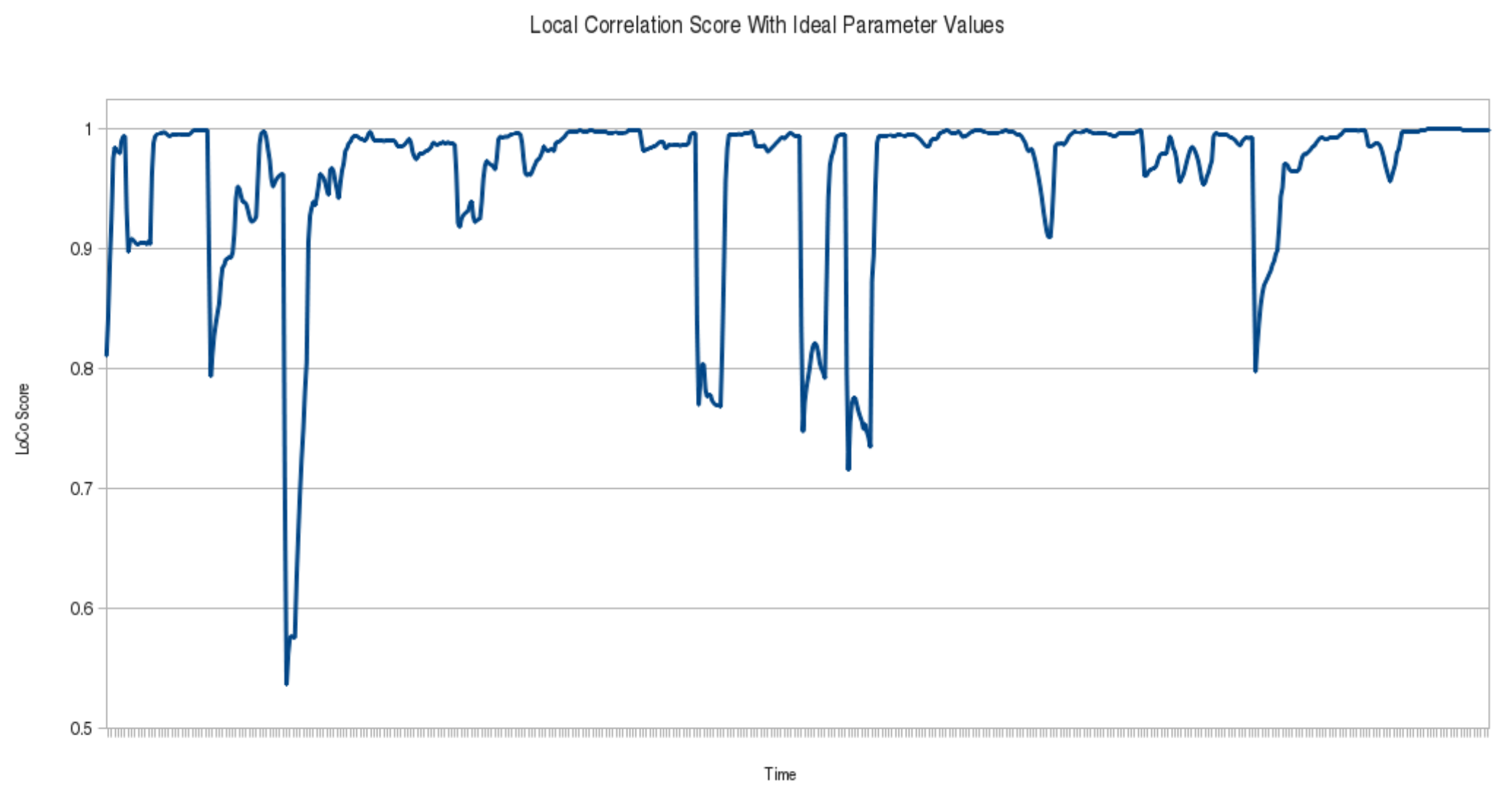

Figure: Exponential LoCo score with LoCo parameter values $w = 15$, $k = 2$, and $\beta = 0.5$

References

1 "Intelligent River®: from observational to operational," website, 2012, http://www.intelligentriver.org.

2 S. Papadimitriou, J. Sun, and P. Yu, "Local correlation tracking in time series," in Data Mining, 2006. ICDM '06. Sixth International Conference on, dec. 2006, pp. 456-465.

Acknowledgments

This work was supported by grants from the National Science Foundation (CNS-1126344), National Park Service - Great Smoky Mountains National Park Grant #2097821, and the City of Aiken, SC.

Mail: itaylor@clemson.edu or jsharp@clemson.edu