NREL
NATIONAL RENEWABLE ENERGY LABORATORY

Evaluation of Methods for Comparison of Spatiotemporal and Time Series Datasets

Dan Getman, Brian Bush, Danny Inman, and Ryan Elmore
National Renewable Energy Laboratory

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Technical Report
NREL/TP-6A20-62647
April 2015

Contract No. DE-AC36-08GO28308

NREL
NATIONAL RENEWABLE ENERGY LABORATORY

Evaluation of Methods for Comparison of Spatiotemporal and Time Series Datasets

Dan Getman, Brian Bush, Danny Inman, and Ryan Elmore
National Renewable Energy Laboratory

Prepared under Task No. 0664.1401

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Technical Report
NREL/TP-6A20-62647
April 2015

Contract No. DE-AC36-08GO28308

NOTICE

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at http://www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
phone: 865.576.8401
fax: 865.576.5728
email: mailto:reports@adonis.osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
phone: 800.553.6847
fax: 703.605.6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/help/ordermethods.aspx

Cover Photos: (left to right) photo by Pat Corkery, NREL 16416, photo from SunEdison, NREL 17423, photo by Pat Corkery, NREL 16560, photo by Dennis Schroeder, NREL 17613, photo by Dean Armstrong, NREL 17436, photo by Pat Corkery, NREL 17721.

NREL prints on paper that contains recycled content.

Nomenclature or List of Acronyms

DIR	direct normal irradiance
NREL	National Renewable Energy Laboratory
NSRDB	National Solar Radiation Data Base
PCA	principal component analysis
RDA	redundancy analysis

Executive Summary

Data used by the National Renewable Energy Laboratory (NREL) in energy analysis are often produced by industry and licensed or purchased for analysis. While this practice provides needed flexibility in selecting data for analysis it presents challenges in understanding the differences among multiple, ostensibly similar, datasets. As options for source data become more varied, it is important to be able to articulate why certain datasets were chosen and to ensure those include the data that best meet the boundaries and/or limitations of a particular analysis.

This report represents the first of three phases of research intended to develop methods to quantitatively assess and compare both input datasets and the results of analyses performed at NREL. This capability is critical to identifying tipping points in the costs or benefits of achieving high spatial and temporal resolution of input data.

This report describes the first phase of a longer-term research effort in which several methodologies are evaluated to determine their capability in describing differences between parameters in large spatiotemporal datasets. The report also presents an overall methodology for performing this type of analysis. The work described in this report will be followed by a second phase of research involving the application of these methods to evaluate three large spatiotemporal renewable energy resource datasets. Selection and preparation of these datasets are being completed at this time.

Table of Contents

List of Figures

1 Introduction

Performing analysis in any domain of research can involve the problem of selecting data from a collection of seemingly similar spatial or temporal data. In some cases, these data can be highly variable in terms of their resolution (both in space and in time), quality, geographic coverage, or availability. In others, the differences are much less obvious and may present challenges when interpreting analysis results. This is particularly true in studies in which resource assessment is critical. For example, different analytical results can be derived from numerous solar resource datasets that are currently available at varying temporal and spatial resolutions. By acknowledging, understanding, and then leveraging the variations and patterns that exist across space and time within and across these datasets, researchers can ensure their analysis take advantage of the strengths, and avoids the weaknesses, of the available data. The methodology described in this report can be used to compare the results of different analysis methods to determine the impact of using fewer or different datasets. Such comparisons address the question of minimally sufficient data, which is critical to expanding research from data-rich (domestic) to data-poor (international) analysis domains. Of particular interest is reducing the cost of unnecessarily high temporal or spatial resolution in data collection and acquisition.

The objective of the research described in this report is to develop statistical methods for the cross-comparison and relative-quality evaluation of large spatiotemporal datasets. This research outlines several methods that can be used to facilitate interpretation of analytical results and perform validation of modeled data through comparison to known or source datasets.

2 Methodologies Evaluated

This report describes the first phase of a longer-term research effort. In the first phase, several methodologies were evaluated to determine their capability in describing differences between parameters in large spatiotemporal datasets. Additionally, an overall methodology for performing this type of analysis was organized, and it is proposed here for further work in this area. The most promising experiments are described, each in a separate sub-section outlining the methodology and specific problem that is being addressed, along with the overall methodological approach.

This analysis was performed using the 1991–2010 update to the National Solar Radiation Database (Wilcox 2012). These data were sourced from 1,454 weather stations across the United States, and they are comprised of hourly values representing 41 weather and atmospheric parameters. The original format of these data was CSV files representing each year for each station. These data were processed to create zoo (Zeileis and Grothendieck 2013) time series data files appropriate for processing in R. These data are often used as a point of reference or as validation data for other solar resource datasets; their size and format provided a convenient starting point from which to develop and evaluate methods that are intended to work on much larger but similarly formatted data.

2.1 Evaluation of Data Formats for Statistical Analysis of Large Matrices on High Performance Computing

As part of our preliminary experimentation, we ran benchmarks to see how standard analyses might scale on Peregrine, NREL's flagship HPC providing 1.19 PetaFLOPS of computing capability. We timed a principal component analysis (PCA) on datasets of varying size (7.6 megabytes [MB], 76 MB, 763 MB, and 7.45 gigabytes [GB]), differing numbers of cores (96, 144, 192, 240, and 288), and numbers of cores per node (16 and 24). We chose PCA as the basis of this comparison because (1) we plan to use PCA in our future analysis and (2) the matrix operations involved in a PCA are commonly used in other statistical procedures. Our results are based on the average timing over ten independent trials.

The results of this experimentation are presented in Figure 1. There are several interesting observations of note. First, it may make sense to undersubscribe nodes (i.e., request 16 cores rather than 24) for datasets less than 100 MB in size (see the left panel in Figure 1). This effect seems to disappear as we approach the 1 GB size. Our second observation is related to requesting the number of compute nodes for a given job; more nodes do not always yield better outcomes (right panel). Again, for small datasets (< 100 MB), requesting fewer nodes yields better results. However, a substantial speedup is attained when using 288 cores on the 7+ GB datasets relative to just using 96 cores.

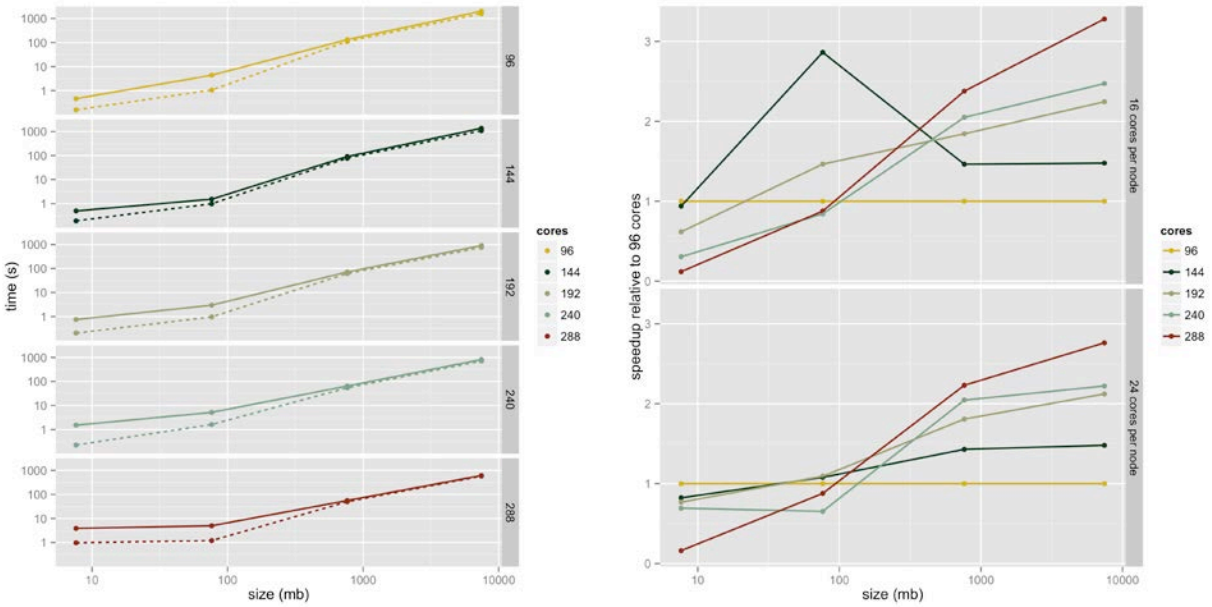

Figure 1. Timings for all combinations and speedups gained in using 144, 192, 240, and 288 cores (left panel) relative to using 96 cores (right panel)

In addition to running these benchmarking experiments, we investigated how to store and access "big data" on Peregrine using R. The majority of Peregrine's compute nodes have 32 GB of memory, with the exceptions having 64 GB. Therefore, we have to distribute data across nodes when analyzing data on the scale that we anticipate in this project (e.g., on the order of terabytes). In our preliminary investigations, we found that the best tool for distributing large matrices within R comes from the "Programming with Big Data in R" project (Ostrouchov et al. 2012) in the form of a ddmatrix object.

Our future work in this domain will focus on a more comprehensive set of benchmarking experiments. Specifically, we will investigate several commonly used statistical procedures to see how they scale with respect to data size and compute (i.e., number of cores). A significant part of our future study will assess memory per node issues for analyzing extremely large datasets on Peregrine. One result of this work will be a list of best practices for various combinations of data analysis tool and dataset size.

2.2 Comparison of Datasets using Map-Reduce

The process of comparing datasets often follows a common pattern (see Figure 2). Analysts are typically faced with deciding which of N datasets is most appropriate for their application. They rarely use datasets in their raw form. Rather, analysts typically aggregate, transform, or summarize datasets into a set of features for their application. In some cases, this might be the output of a complex model, such as the System Advisor Model (SAM) (SAM, 2015) or Plexos (PLEXOS); in other cases, this might simply be a thematic map at a particular resolution.

Rigorous statistical comparison of raw datasets usually indicates that they are statistically different (e.g., have biases or other anomalies), which is not particularly informative to an analyst. Applying statistical tests to the user-defined features of interest can help determine whether the datasets differ for the analyst's application. Statistical tests for comparing datasets

typically rely on transforming, binning, or summarizing the data before applying the test. The result of the test is the identification of the anomalous features, which can then be visualized and used to assess the consequences of using each dataset. Furthermore, statistical inversion techniques can allow one to trace the anomalies identified in the features of interest back to the characteristics of the raw datasets. Subject-matter experts can then focus on determining the fundamental cause of the anomalies and assessing the severity of their impact on applications.

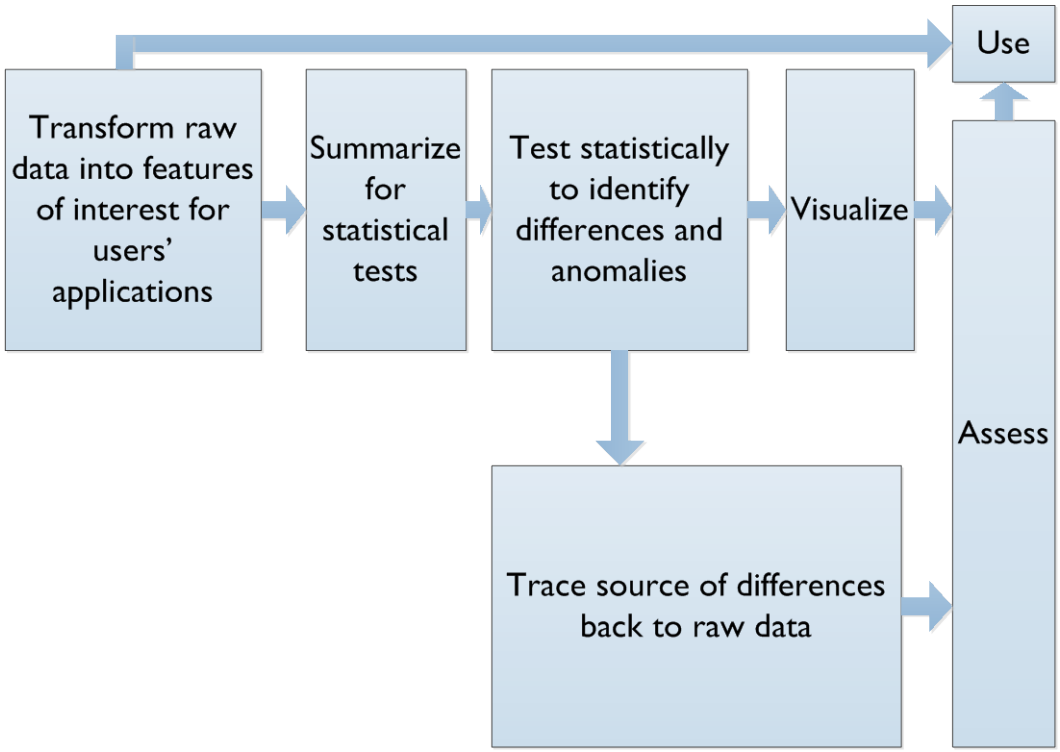

Figure 2. Pattern common in the process of comparing datasets

Abstractly, the preparation of the raw data focuses on presenting information in the form of features meaningful to users, whereas the assessment of the data imposes a model (heuristic or statistical) that highlights the significance of the features and enables the efficient navigation of visualizations of the features. Summarization is often used to organize the data as they are assessed in batches that should be identically distributed under the null hypothesis. Depending upon the dataset and its intended end use, summarization might take several forms:

- Aggressive (i.e., boiling data down to a few numbers)
- Geospatial/temporal slices and windows
- Distributional
- Scale/frequency
- Un-summarized (e.g., individual observations).

Furthermore, these common data-processing patterns can be implemented in high performance computing environments as a map-reduce operation (see Figure 3). We envision that standard map-reduce frameworks, such as Hadoop, MongoDB, and BigTable, can be specialized the

processing and comparative statistical analysis of large renewable-energy datasets. This involves the following generic operations for which an analyst or user provides dataset- and analysis-specific functions:

1. Presentation of features
    A. *Input reader:* extract raw data from its native format
    B. *Map function:* group raw data into the unit of resolution required by the user (e.g., spatial, temporal, frequency, or scale) and transform the raw data into its individual contribution to user-defined features
    C. *Reduce function:* summarize transformed data in the aforementioned units of resolution
    D. *Output writer:* transform summarized data into user-defined features
2. Statistical testing
    A. *Map function:* group feature data (see 1.D above) into the unit of resolution required by the statistical test
    B. *Reduce function:* compute components (e.g., partial sums) of test statistics
    C. *Output writer:* compute test statistic

Thus, the computation of the features can be embodied in one map-reduce operation, and the statistical testing can be embodied in a second map-reduce operation built on the output of the first.

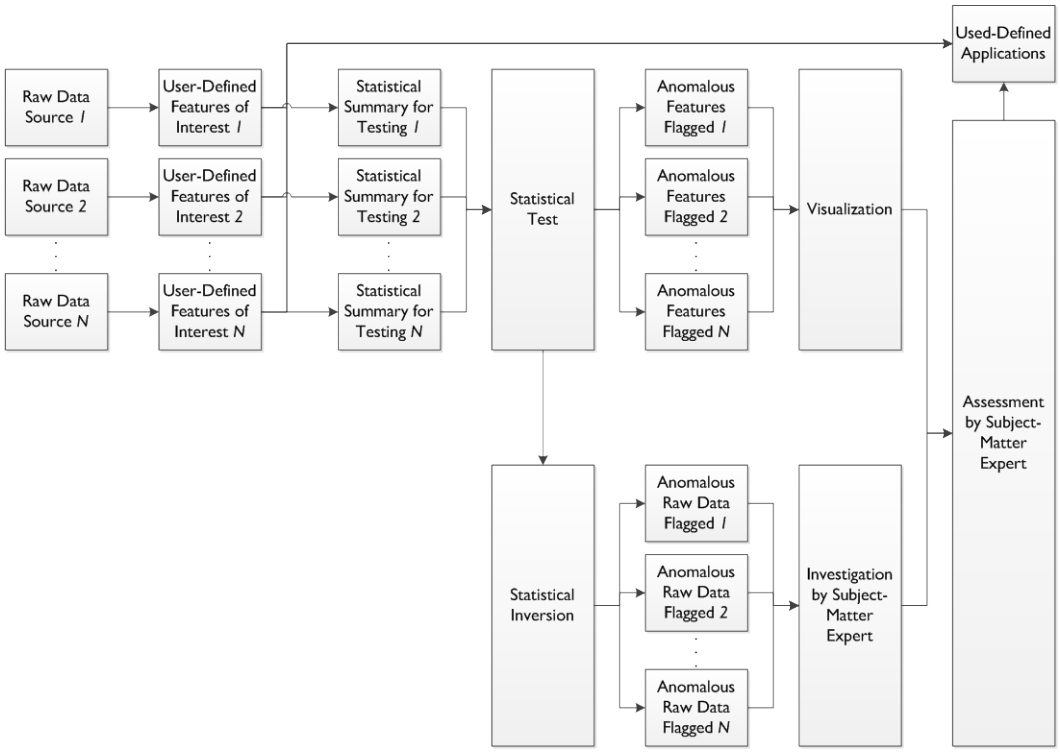

Figure 3. Map-reduce approach to the comparison of spatiotemporal data

2.3 Non-parametric Detection of Bias in Diffuse Solar Irradiance

Non-parametric statistical methods combine power with robustness and have several advantages for routine use in comparing spatiotemporal datasets. These methods are accessible and safe for non-statisticians to use. They apply broadly to many types of energy efficiency and renewable energy data. They do not require strong assumptions about data. They avoid false positives but perform nearly as well as parametric methods. Finally, many non-parametric methods are well suited for rapid calculation in distributed computing environments.

As an example of a study question, we considered how the National Solar Radiation Data Base (NSRDB) station's irradiance "measurements" (Dataset #1) compare with those measurements from the same hour of the same day five years previously (Dataset #2). One of the simplest possible non-parametric statistical tests that can be performed on this pair of datasets to identify a bias in the irradiance is the sign test. This test pairs observations in Datasets #1 and #2 and the same temporal offset (in this case, the day and hour versus the day and hour five years previously). Under the null hypothesis that either of the pair of points is equally likely to be larger, which would be the case if there were no systematic bias between the datasets, the distribution of which point has a larger irradiance would conform to the binomial distribution. Because the sign test is repeated for each spatial point, care must be taken in interpreting the significance of the test; the significance threshold for p-values (say 5% for a single sign test) must be divided by the number of tests performed. (More sophisticated detection techniques, similar to those including multiple comparison corrections or those used in Statistical Analysis of Microarrays [SAM] can also be used to interpret results of repeated tests.) Figure 4 illustrates the performance of this test, which results in the conclusion that a statistically significant bias is detectable between these datasets in many geographical regions. Once the presence of bias has been detected, biased and unbiased spatial locations can be filtered in exploratory visualizations, and the precise nature of the bias can be investigated.

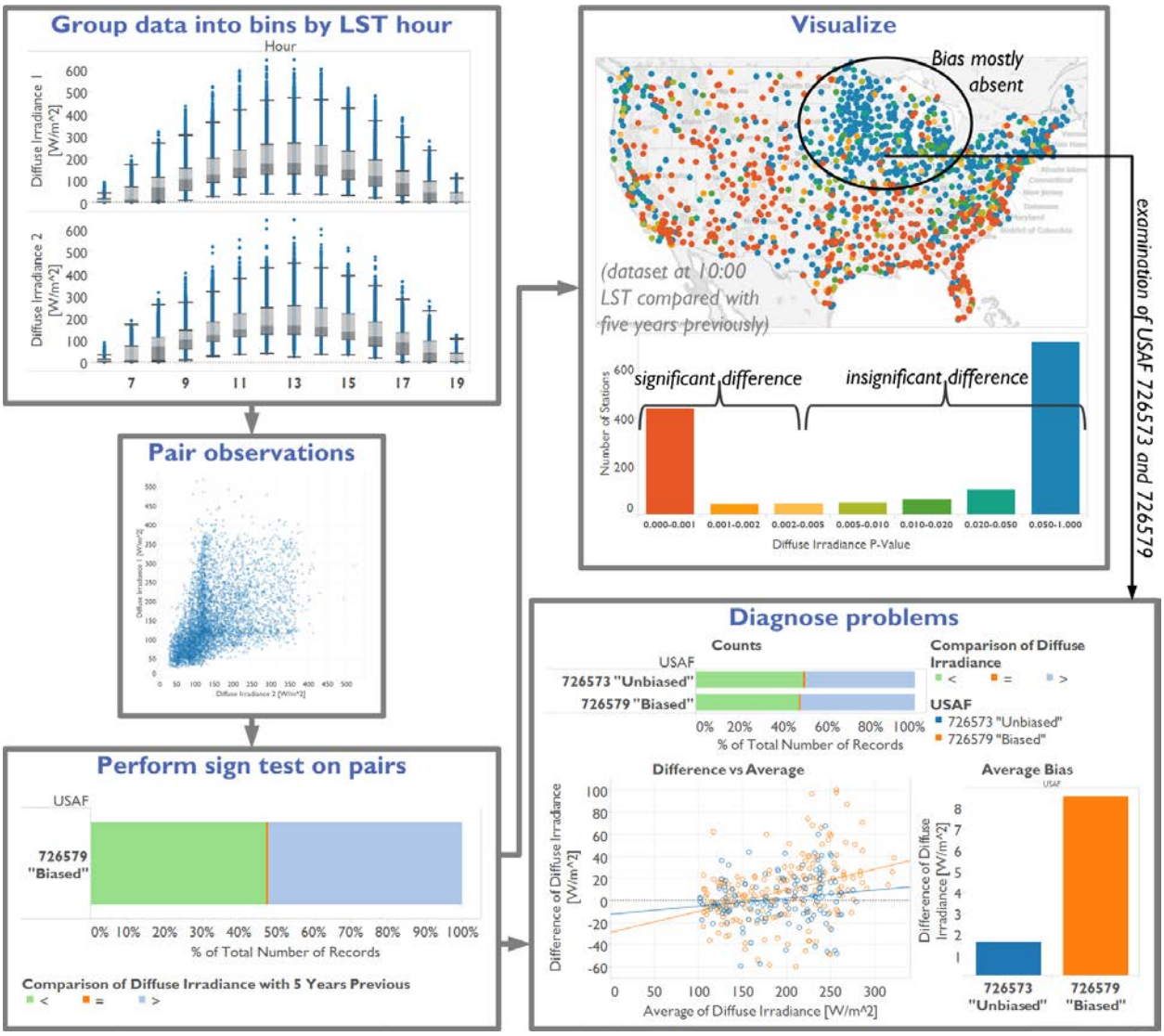

Figure 4. Example of an application of a non-parametric statistical test (the sign test) to detect bias between two datasets

As with the simple sign test, more sophisticated and robust non-parametric tests can be applied to spatiotemporal data, such as renewable-energy resource information. In particular, the following standard tests are applicable (Giddons and Chakraborti 2011), as are their multivariate generalizations (Oja 2010):

1. Median comparison
    A. Sign test
    B. Wilcoxon signed-rank test
2. Distribution comparison
    A. Quantile test
    B. Wald-Wolfowitz runs test

C. Kolmogorov-Smirnov test

D. Mann-Whitney test

3. Location comparison

   A. Wilcoxon rank sum test

   B. Van der Waerden test

4. Scale comparison

   A. Wilcoxon rank sum test

   B. Mood test

   C. Siegel-Tukey test.

If numerous test statistics are produced, multiple-comparison corrections are applied to the test results in order to interpret them accurately. If the test results have spatial or temporal extent, simple classification and clustering methods can be applied to help summarize the spatial and temporal domains and trends where the null hypothesis does not hold.

2.4 Detection of Simultaneous Variability through Analysis of Principal Components

This analysis was performed using the 1991–2010 update to the NSRDB. A subset of the data, representing 122 stations, was created by limiting the data to those stations for which data existed for wind speed and solar irradiance for every daylight hour from 1/1/1992 to 12/31/2010. Additionally, the analysis was focused on the West and Southwest, including stations in Arizona, California, New Mexico, Nevada, and Utah.

From these data, two values were calculated that form the basis of these analyses. First, the variance was calculated for each parameter at each station at daily, weekly, monthly, and yearly aggregation levels. Each calculation was based on the data at the hourly level for the entire temporal period and resulted in 6,940 daily, 991 weekly, 228 monthly, and 19 yearly measures of variability for both solar irradiance and wind speed. At this point, both a correlation and a difference between each measure of variability for wind speed and solar irradiance were calculated. This produced a dataset for which each observation represented either the correlation or the difference in variability of wind speed and solar irradiance at a point in space and in time, at a specific aggregation level (Figure 5). These differences were used in the remainder of the analysis.

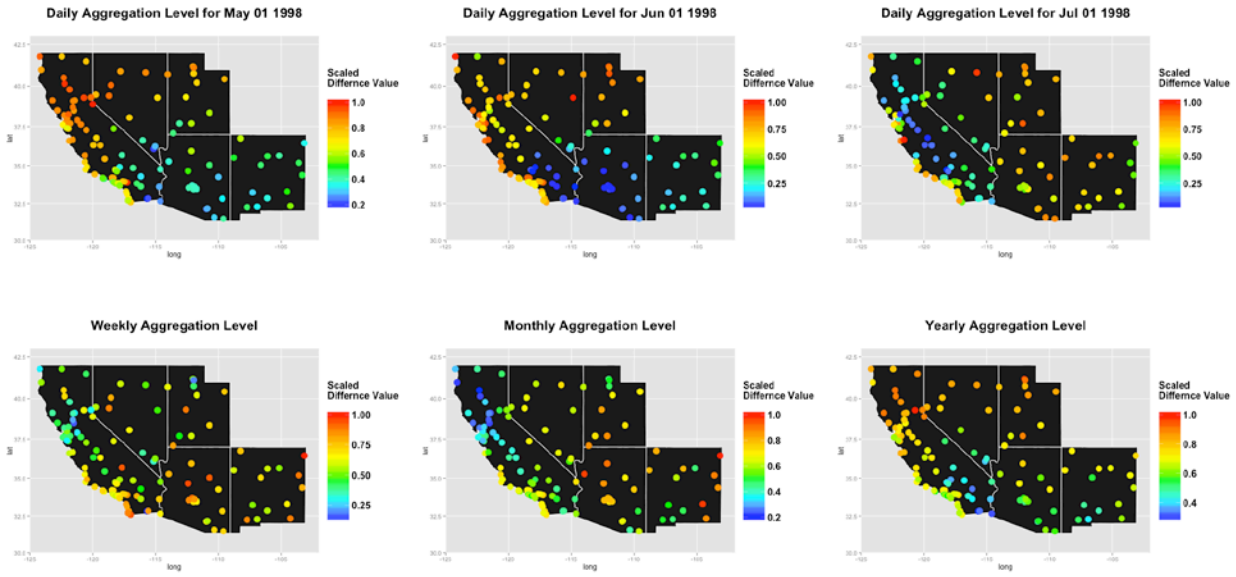

Figure 5. Mapping of the difference between the variability of direct normal irradiance (DIR) and wind speed for several individual days at weekly, monthly, and annual aggregation levels

The pre-processing steps taken in this analysis produced data that could be used directly to evaluate the correlation of variability between the wind speed and DIR. One issue with taking this step has to do with the volume of data and the nature of the variability of the relationship between wind and solar variability. Given a correlation of the values at any aggregation level or a series of charts of the differences at each aggregation level, it is difficult to make statement about the relationships (Figure 6).

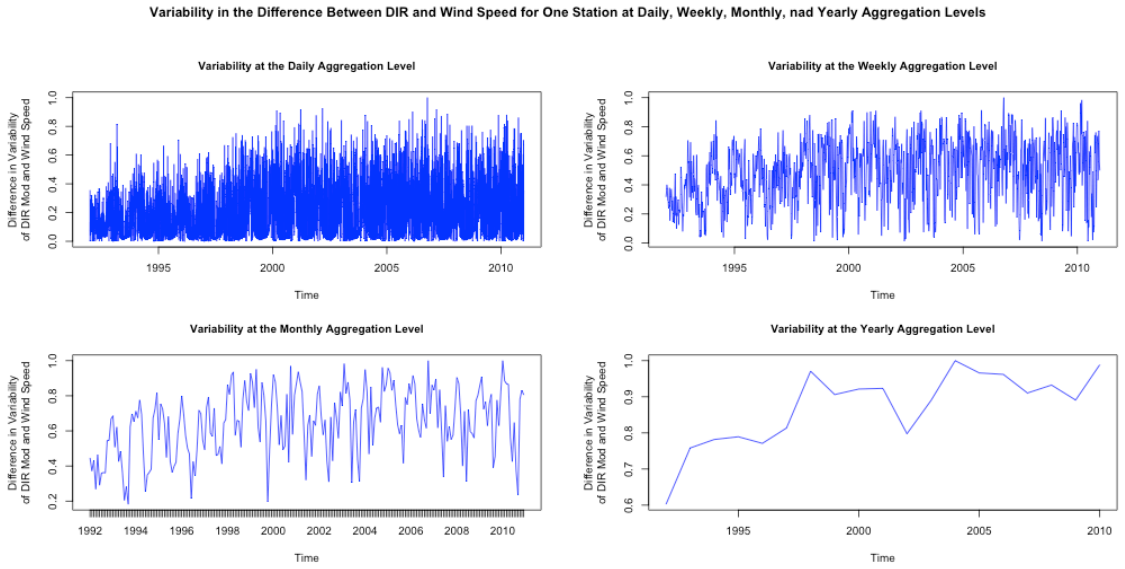

Figure 6: Difference between variability of wind speed and variability of solar irradiance for a single station at four aggregation levels

To examine spatial and temporal trends in this relationship, we performed a PCA on the difference values between wind speed and direct normal irradiance at each aggregation level. PCA is a widely used multivariate data analysis technique due to the simplicity with which it can be used to both interpret and calculate (Wackernagel 2003). PCA can be used to reduce, or transpose, a dataset from some number of correlated parameters to a more limited number of principal components. These components are orthogonal, meaning they are uncorrelated themselves and each component will successively account for the maximum variance, as estimated through the creation of eigenvectors in the original data (Wilks 2011) that will be assessed in this analysis. In this case, PCA was run in two distinct processes. Both processes started from a dataset in which the columns (or parameters) represented the stations and the rows (or observations) represented the time series at each aggregation level.

First, PCA was run using all of the data at each aggregation level to see how the contribution of a particular station to the variance of the difference in variability changes at each aggregation level. Second, PCA was run at the monthly aggregation level separately for each month. The first run took into account all the observations that occurred in January of all years, the second all of the values that were observed in February of all years, et cetera. The intention was to see how the contribution of each station changes monthly or seasonally. Figure 7 illustrates this process.

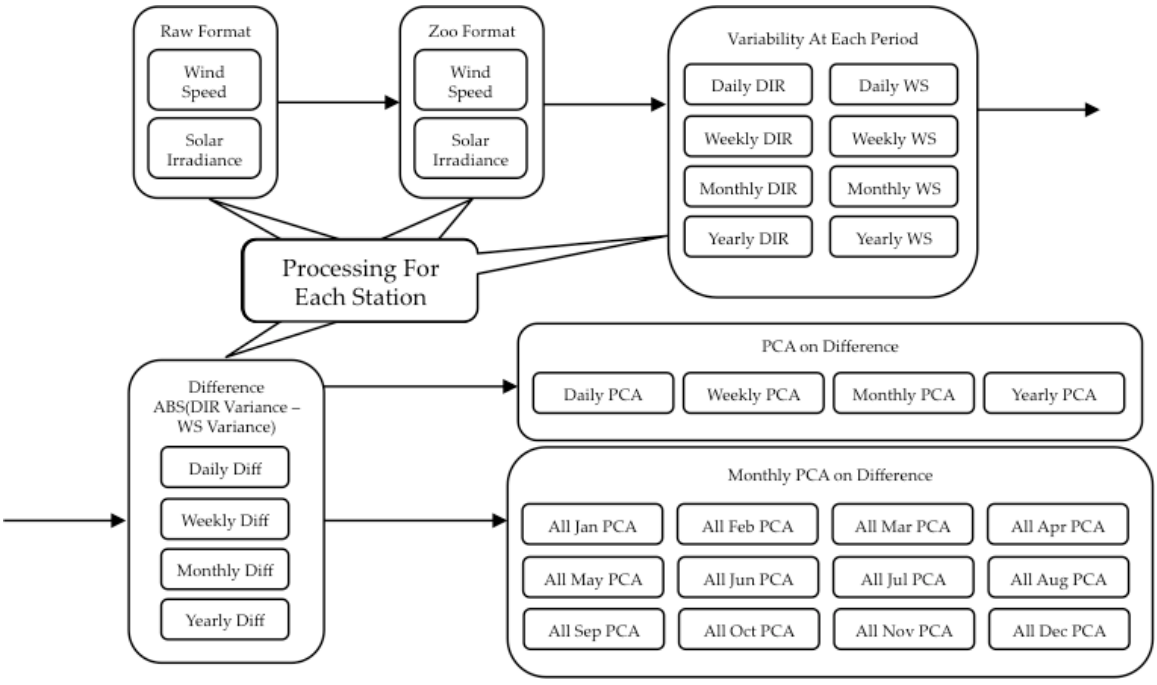

Figure 7. Diagram of data processing and analysis

By comparing stations that represent low and high contributions to the first ten eigenvectors, which represent the majority of the variance in these data, we can evaluate whether this contribution is related to simultaneous variability in wind speed and DIR. In the initial comparison, the contribution to the first ten eigenvectors for each month was calculated and visualized in map form (Figure 8). Visually inspecting these data demonstrates that the resulting pattern for this contribution is more consistent than the pattern that results from simply looking at the difference between the variability of DIR and wind speed (Figure 5).

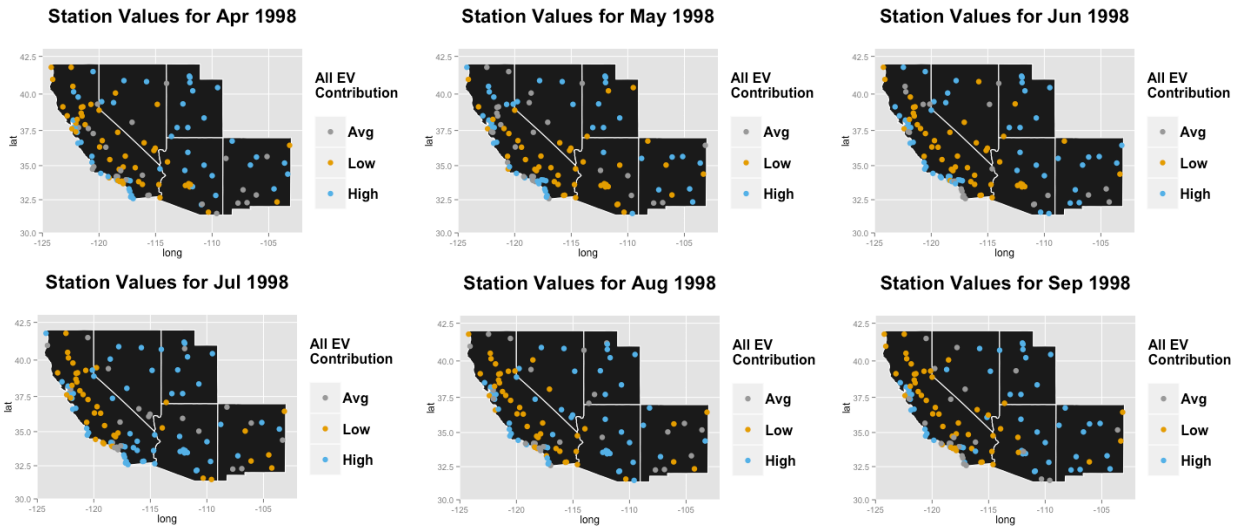

Figure 8. Mapping the contribution of each station to the first ten eigenvectors

A visual inspection of the variability of the difference between the variability of wind speed and DIR for two stations—one a consistent low contributor (blue) and one a consistent high contributor (red) to the first ten eigenvectors—demonstrates more variability at both the daily and weekly aggregation levels (Figure 9).

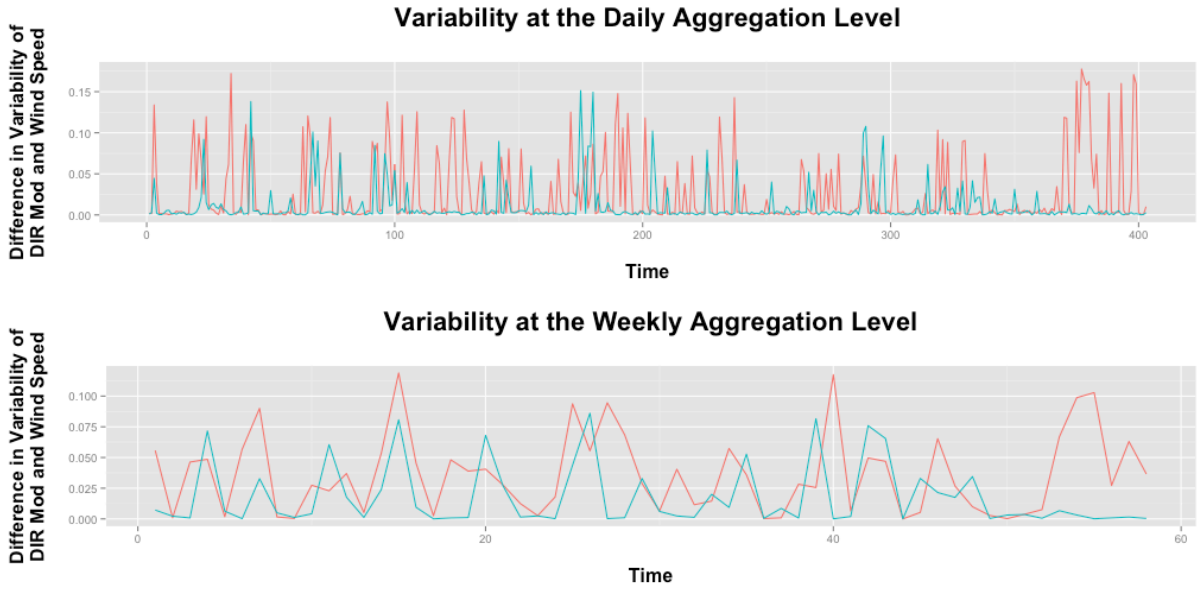

Figure 9. Variability of the difference between the variability of wind speed and DIR for two stations

The station represented in red is a high contributor to the first ten eigenvectors and the station represented in blue is a low contributor. Values represent the month of July across all years at each aggregation level.

Additional statistics were calculated for each station in order to gauge the relationship between the variability of the difference between the variability of wind speed and DIR. These included variability, range, and maximum difference between wind speed and DIR for each station. Quintiles were generated for each of these values, and each station was classified as consistently low (< 3) or high > 3 in these values. By plotting the distribution of the values for the contribution to each eigenvector for both the consistently low and high stations (Figure 10), and by plotting the distribution of the parameters for both the consistently low and high contributors (Figure 11), we obtained an indication of the relationship between these values. Stations with low contributions to the first ten eigenvectors demonstrate lower variability, range, and maximum values of the difference between variability of wind speed and solar irradiance (DIR). Stations with higher contributions demonstrate higher values in these parameters.

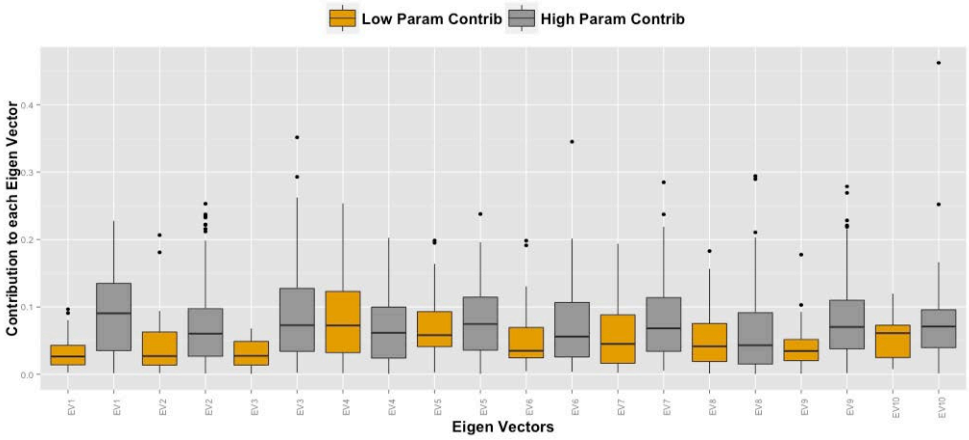

Figure 10. Contribution of all stations to the first ten eigenvectors, grouped by high and low values representing the variability, range, and maximum difference between wind speed and DIR

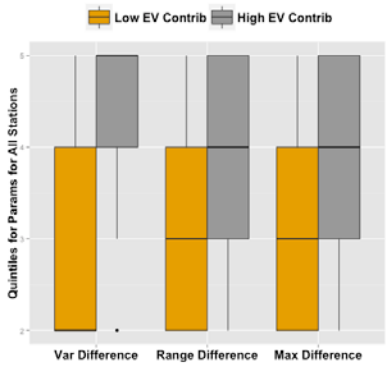

Figure 11. Quintiles for variability, range, and maximum difference between wind speed and DIR for stations, grouped by low and high contribution to the first ten eigenvectors

The next steps in developing this methodology will involve quantifying these classifications in such a way as to automate the process of identifying locations in the dataset, both temporally and spatially, where variability has a high probability of being correlated.

2.5 Non-parametric Distance-based Redundancy Analysis

Understanding the differences between and within large resource datasets is important when analysts decide which datasets to use and whether and how variables may be used within a given dataset. Methods that are quick to implement and require few assumptions are likely to be the most accessible to a wide range of energy efficiency and renewable energy analysts with diverse backgrounds. In this section, we introduce a non-parametric method of performing intra- and inter-dataset comparisons.

Nonparametric analyses are methods in which the analyst does not make *a priori* assumptions regarding the distribution of the data or the distribution of the test statistic used. Sheskin (2011) distinguishes nonparametric from parametric analyses thusly: "inferential tests that evaluate categorical/nominal data and ordinal/rank-order data are nonparametric, while tests that evaluate interval or ratio data are parametric." With regard to time series analysis, Fan and Yao (2005) describe nonparametric methods as those that model a process whose defining parameters lie in infinite dimensional space and/or that have an unspecified (or incompletely specified) probability form. In general, nonparametric methods are considered to have lower statistical power than their parametric counterparts as long as the assumptions required for a given parametric test are not violated (Sheskin 2011). Nonparametric methods are preferred in instances in which one or more parametric assumptions are violated. Nonparametric methods of comparison were well suited for our study for several reasons. Many of the large resource datasets are irregularly distributed, and they violate assumptions of normality, precluding parametric testing. Methods of analysis that do not require distributional assumptions will be more user-friendly and straightforward for analysts to implement.

Redundancy analysis (RDA) is used to assess differences and interactions of factors within large resource datasets. Redundancy analysis is an extension of multiple regression to multivariate response data (Legendre and Legendre 2012). In RDA, Y is constrained and the ordination axes are linear combinations of X. The ordination axes are obtained through principle component analyses (PCA) of matrix $\hat{y}$. RDA consist of two main processes: (1) the regression of matrix Y on matrix X and computation of the fitted values and (2) performance of PCA on the matrix of fitted values to obtain eigenvalues and eigenvectors. Individual canonical axes are tested for significance using a permutation procedure.

2.6 Illustrative RDA Results and Discussion

For this, the first phase of the study, a subset of a much larger dataset was used as a test set for development and evaluation of statistical methods. Direct solar irradiance data from the NSRDB for four sites in the western United States were used: Abilene, TX; El Torro, CA; Tustin, CA; Twentynine Palms, CA. Autoregression was used to remove the influence of temporal autocorrelation from each time series (Figures 12 and 13). The RDA was performed on the residuals from the autoregressive models. Based on the results of the RDA (Figure 14), salient differences in underlying data structure of the four NSRDB stations exist. Such results could elucidate critical differences in large datasets. This analysis approach could also be implemented to assess the similarity or dissimilarity of datasets that are ostensibly the same. During the next phase of this project, we will focus on applying this analytical approach to much larger datasets, which will require the use of high-performance computing and the parallelization of R and the statistical methods described above.

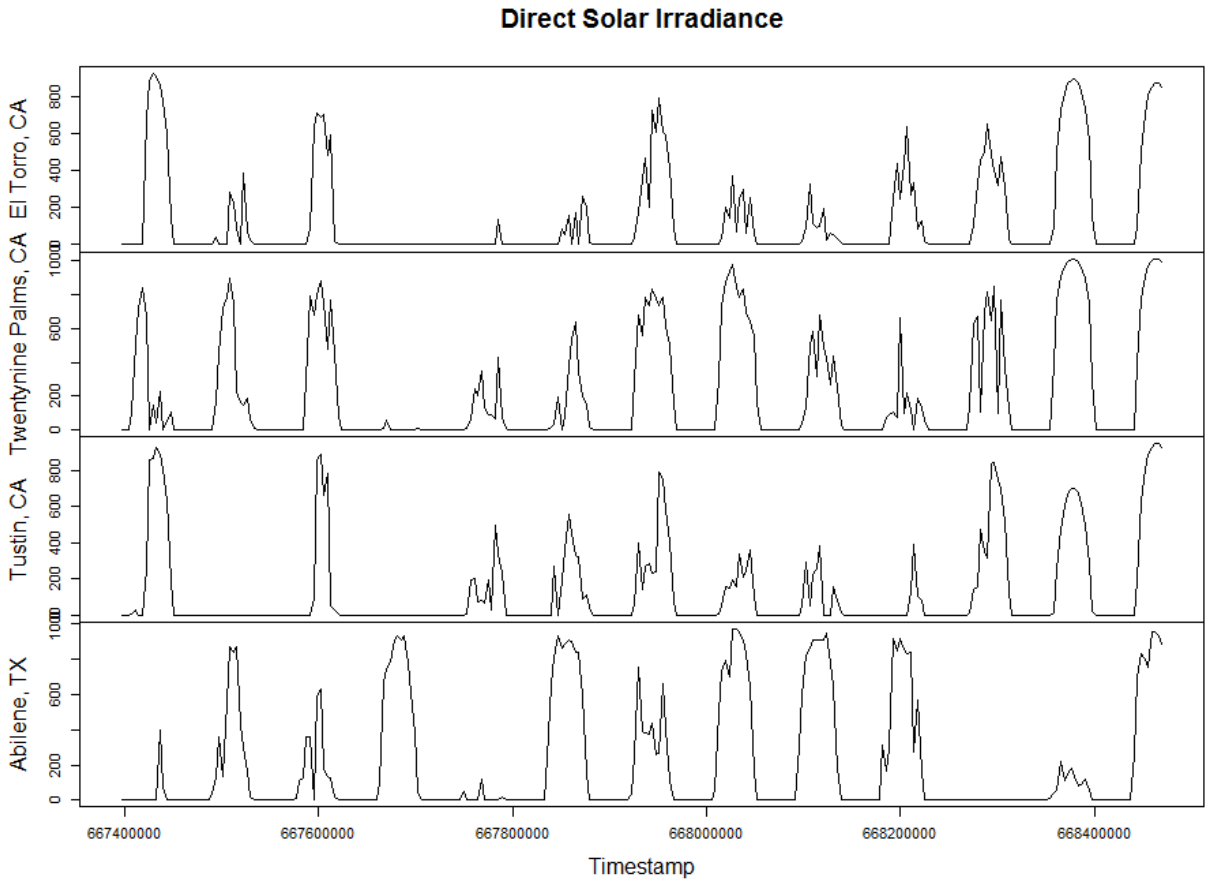

Figure 12. Direct solar irradiance for four NSRDB stations in the western United States

Data presented in this figure have been minimally processed to remove NA values.
The time period presented is 300 hours.

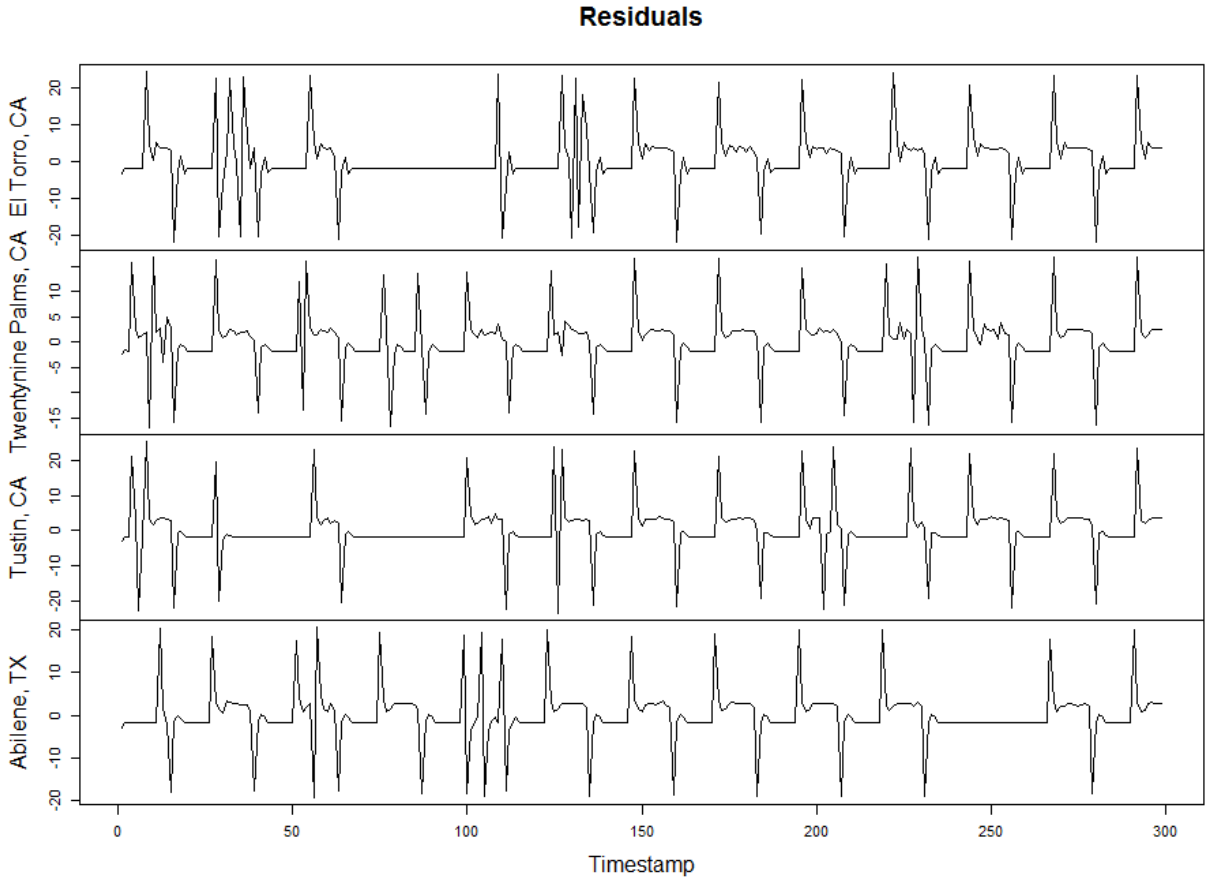

Figure 13. Residual plots of four NSRDB stations in the western United States

An autoregressive model was used to remove the effects of autocorrelation on the data. From this plot, it is salient that a strong underlying structure exists in the data.

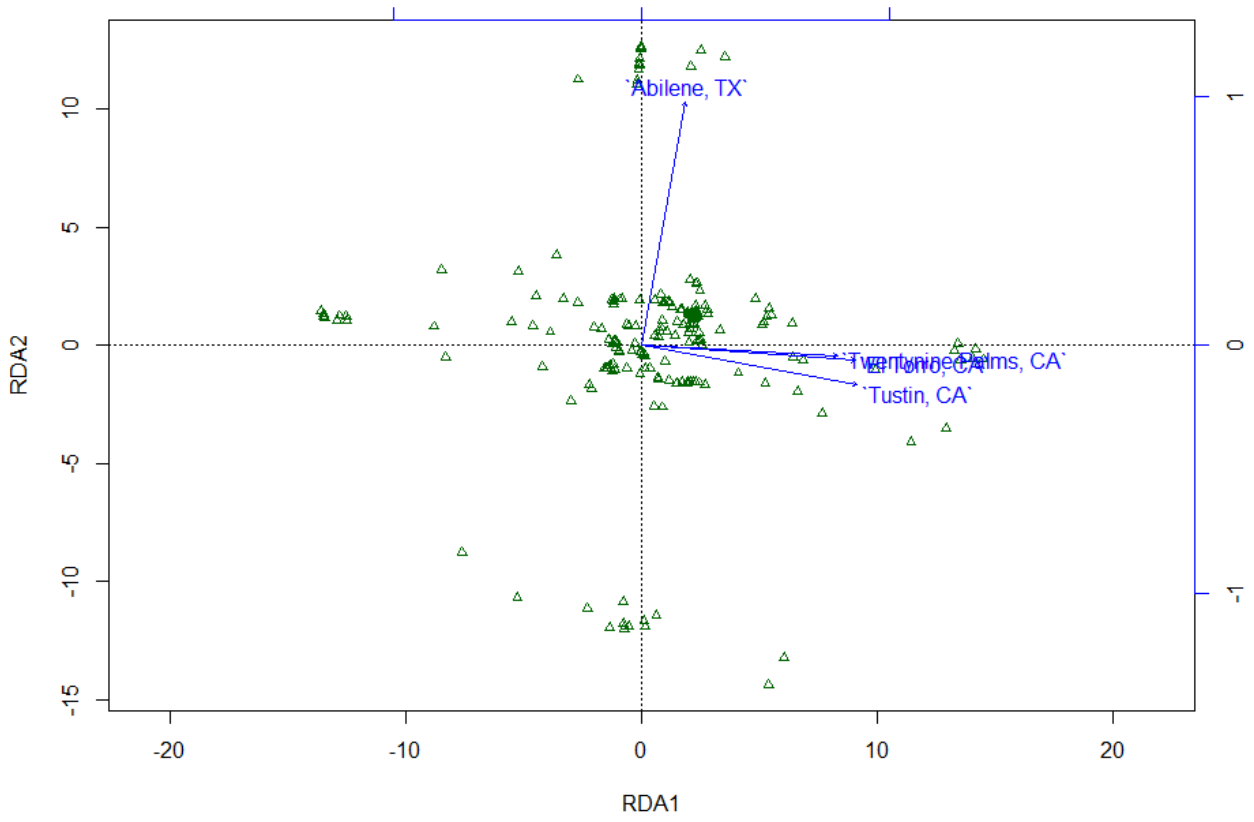

Figure 14. Biplot of the redundancy analysis (RDA)

The green triangles represent the fitted site scores of the dissimilarity measures and the blue lines represent the station data. The two axes are the first (x) and second (y) principle coordinates. The cosine of the angle between any two stations (blue lines) represents the correlation.

3 Discussion

The first phase of our research focused on discovering methods with potential for use in the cross-comparison and relative-quality evaluation of large spatiotemporal datasets. From the initial experimentation, it is apparent that several of these methods can be used for this purpose. In anticipation of the need to perform this analysis on much larger data matrices using high performance computing resources, research into both the most appropriate data format and computing configuration was performed. The next phase of this research will involve applying these methods to evaluate three large spatiotemporal renewable energy resource datasets. Selection and preparation of these datasets is being finalized at this time.

References

Energy Exemplar. (August 2013). PLEXOS for Power Systems. Accessed March 9, 2015: http://energyexemplar.com/software/plexos-desktop-edition/.

Fan, J.; Yao, Q. (2005). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer: New York. ISBN: 10: 0-387-26142-7.

Oja, H. (2010). *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks*. Springer: New York.

Ostrouchov, G.; Chen, W.-C.; Schmidt, D.; Patel, P. (2012). "Programming with Big Data in R." Accessed August 26.2014: http://r-pbd.org/.

Sheskin, D.J. (2011). *Handbook of Parametric and Nonparametric Statistical Procedures*. 5th edition. CRC Press, Taylor and Francis Group: New York. ISBN: 978-1-4398-5801-1.

System Advisor Model Version 2015.1.30 (SAM 2015.1.30) Website. Hourly Output from Multi-Year P50/P90 Simulation. National Renewable Energy Laboratory. Golden, CO. Accessed March 12, 2015. https://sam.nrel.gov/content/hourly-output-multi-year-p50p90-simulation.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Springer: Berlin. ISBN: 3-540-44142-5.

Wilcox, S.M. (2012). *National Solar Radiation Database 1991-2010 Update: User's Manual*. NREL/TP-5500-54824. Golden, CO: National Renewable Energy Laboratory.

Wilks, D.S. (2011). *Statistical Methods in the Atmospheric Sciences*. Academic Press: Burlington, MA. ISBN 13: 978-0-12-751966-1.

Zeileis, A.; Grothendieck, G. (2005). "zoo: S3 Infrastructure for Regular and Irregular Time Series." Journal of Statistical Software (14:6), pp. 1-27. Accessed August 26, 2014: http://www.jstatsoft.org/v14/i06/.

Bibliography

Denholm, P.; Mehos, M. (2011). *Enabling Greater Penetration of Solar Power via the Use of CSP with Thermal Energy Storage*. NREL/TP-6A20-52978. Golden, CO: National Renewable Energy Laboratory.

Gibbons, J. D.; Chakraborti, S. (2011). *Nonparametric Statistical Inference*. 5th edition. CRC Press, Taylor and Francis Group: New York.

Massachusetts Institute of Technology. (2011). *The Future of the Electric Grid an Interdisciplinary MIT Study*. Cambridge, MA: Massachusetts Institute of Technology.