A Comparison of Time Series Gap-Filling Methods to Impute Solar Radiation Data

Alexis Denhard[1], Soutir Bandyopadhyay[1], Aron Habte[2] and Manajit Sengupta[2]

[1] Colorado School of Mines, Golden (United States)

[2] National Renewable Energy Laboratory, Golden (United States)

Abstract

Complete solar resource datasets play a critical role at every stage of solar project phases. However, measured or modeled solar resource data come with significant uncertainties and usually suffer from several issues, including but not limited to, data gaps, data quality issue, etc. In order to mitigate these issues an appropriate data imputation method should be implemented to build a complete and reliable temporal (and spatial) database. Being motivated by this, in this study we compare the performances of eight different gap filling methods extensively by creating random and artificial data gaps in (i) hourly irradiance data for one year using a few locations of the National Solar Radiation Database (NSRDB) and (ii) one-minute ground measurement dataset from Surface Radiation Budget Network (SURFRAD) and the National Renewable Energy Laboratory (NREL) stations.

Keywords: Global horizontal irradiance, clearness index, time series, gap fill

1. Introduction

In recent years, due to significant technological improvements in different cost competitive PV systems, accurate and complete solar resource data have become essential for predicting the solar energy output of these conversion systems and in reducing the expense associated with mitigating performance risks. However, solar resource data are prone to data gaps (missing data) due to instrument malfunctions, discarded data due to data quality problems and/or human error or satellite issue in terms of modeled satellite derived solar resource data. Therefore, from a practical point of view, it is of primary importance to fill those data gaps in order to make the data useable. In this study we apply different existing statistical temporal gap-fill methods to both measured and modeled datasets. In particular, here we consider (a) the ground measurement data from seven National Oceanic and Atmospheric Administration Surface Radiation Budget Network (SURFRAD) and NREL stations and (b) the modeled data from the National Solar Radiation Database (NSRDB V3). Furthermore, in this study, we restrict our attention to the solar irradiance data and other meteorological inputs to gap-fill natural and artificial data gaps.

2. Methodology

One year of data (2017) was obtained from the NSRDB and the ground measurement from SURFRAD and NREL sites (Figure 1) which contain the following three components for solar irradiance - global horizontal irradiance (GHI), direct normal irradiance (DNI), and diffuse horizontal irradiance (DHI) - as well as related meteorological parameters (Augustine et al. 2000, Sengupta et al. 2018). Further, the NSRDB data are available at thirty-minute intervals and the associated GHI time-series are occasionally incomplete, containing varying gap sizes. To reduce the scale of values between 0 and 1 and to remove the effects of low sun angle, in this study we consider clearness index *Kt* (GHI normalized by extraterrestrial radiation). We also consider the clearness index *Kt*, for ground-based measurements at one-minute resolution from NOAA's SURFRAD and NREL locations. In Section 2.1 we briefly describe the different imputation methods considered in this study to fill missing *Kt* values followed by descriptions of different metrics we use to compare their relative performances. The statistical metrics used are described in Habte et al. (2017).

doi:10.18086/swc.2021.38.03 Available at http://proceedings.ises.org

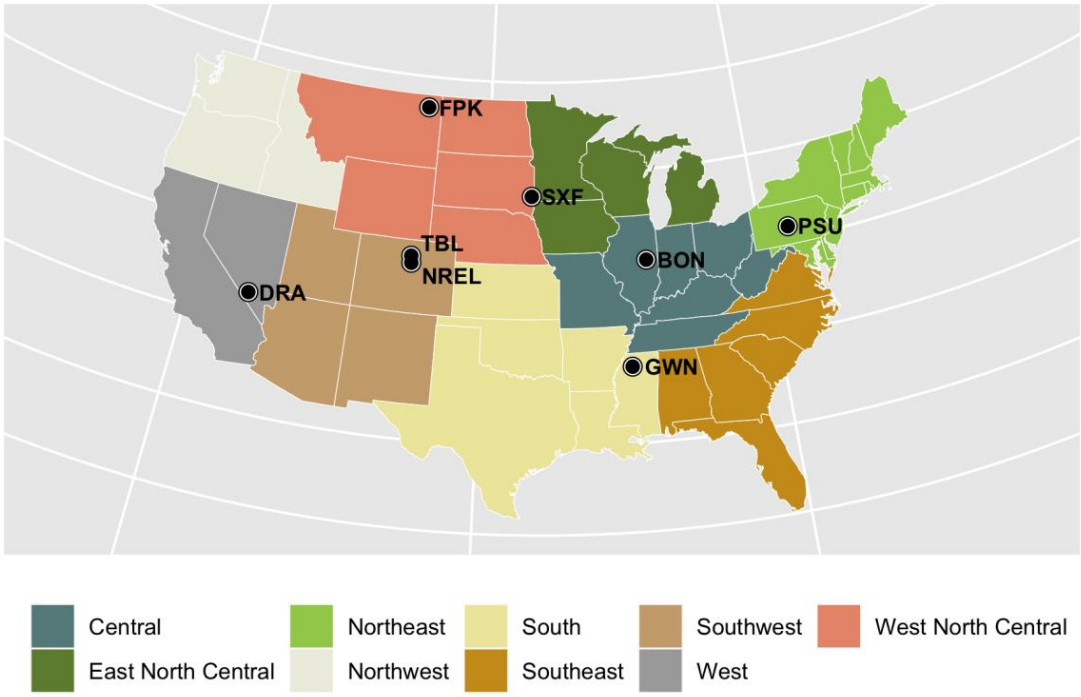

Fig. 1: NOAA SURFRAD and NREL locations over United States climate regions (modified from Habte et al., 2017)

2.1 Different gap-filling methods

a) *Kalman filter and smoothing for structural times series* – Kalman filtering, also known as linear quadratic estimation (LQE), is a recursive algorithm that produces missing values of a univariate time series by estimating a joint probability distribution over the variables for each timeframe. The structural time series models can be viewed as regression models in which the predictors are a function of time and the parameters varying with time.

b) *Kalman filter and smoothing for state space representation of an autoregressive integrated moving average (ARIMA) model* – In this method, Kalman filtering is applied to the most popular parametric class of nonstationary time series, namely, Autoregressive Integrated Moving Average (ARIMA) model, which primarily uses several lagged observations of time series to forecast observations.

c) *Linear interpolation* - Linear interpolation, the simplest and probably the most widely used curve-fitting method, uses a straight line between two points for interpolation.

d) *Spline interpolation* - Spline interpolation uses a similar approach as in the linear interpolation, except the interpolant is now constructed using a piecewise polynomial function, commonly known as splines.

e) *Stine interpolation* - Stine interpolation can be considered as an extension of the linear and spline interpolation methods, where the interpolant uses piecewise rational interpolation to construct data points within a curve.

f) *Simple moving average* – In time series, simple moving average (SMA) is defined as the unweighted mean of an equal number of data on either side of a central value.

g) *Linear weighted moving average* – In this method, instead of unweighted mean as described in SMA we use a weighted mean of the observations, where weights decrease in arithmetical progression. For example, the observations directly next to a central value, have weights 1/2, the observations one further away have weights 1/3, and so on.

h) *Exponential weighted moving average* - An exponential weighted moving average calculates the weighted

mean where the weights decrease exponentially. For example, the observations directly next to a central value, have weights 1/2, the observations one further away have weights 1/4, the next have weights 1/8, and so on.

Additional methods examined include: last observation carried forward, next observation carried backward, mean value, and random sample. These methods were excluded during preliminary trials due to applicability issues for time series and high deviations of the resulting statistical metrics. Furthermore, gap-filling methods that adjust for a seasonal component were also examined, such as seasonally decomposed and seasonally split missing value imputation. However, these methods were also excluded due to no presence of seasonality within the NSRDB or ground-measurement derived *Kt* series.

2.2 Data Pre-Processing

Before imputing on the NSRDB data for each location, two data pre-processing steps were implemented as follows:

- Use data when solar zenith angle is less than 80º. Remove the first four rows of missing values (denoted by NA) from each data set.
- The NSRDB is a serially complete dataset with missing data being filled through interpolation. Fill flags with non-zero values identify the filled points. Where the fill flag variable does not equal zero, we set the corresponding *Kt* observation to missing. Note that, the fill flag variable indicates missing data for values other than zero due to following conditions: 0: no fill; 1: missing cloud type; 2: full time series missing cloud type; 3: missing cloud property; 4: full time series missing cloud property; 5: GHI exceeds clear sky; 6: missing irradiance.

For the ground-based measurements, the following pre-processing step was applied:

- Remove nighttime Kt observations corresponding to a solar zenith angle measurement greater than 89.5º.

2.3 Creating Artificial Gaps

In order to measure the performance of the eight imputation methods, artificial gaps were created in both the NSRDB and ground measurement data. The NSRDB data contains missing data points due to missing satellite measurements whereas the ground measurements are complete and without data gaps.

2.3.1 NSRDB Data

To create synthetic gaps in the NSRDB data, the occurrence of gap sizes was examined for each location.

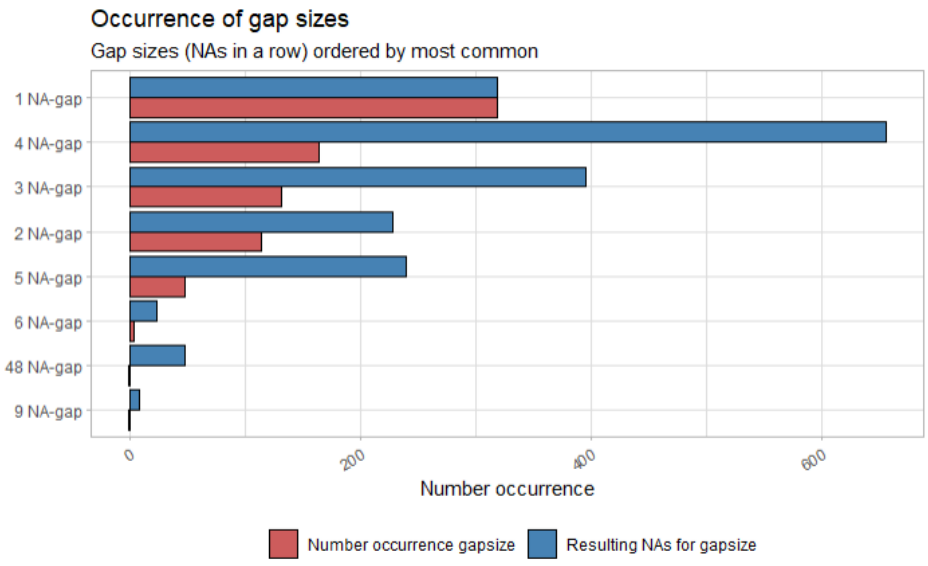

Fig. 2: Distribution of consecutive NA gap sizes and number of occurrences for the Bondville, IL NSRDB data

Figure 2 shows both the number of occurring gap sizes (red) as well as the total resulting NAs for the gap size (blue) for the Bondville, IL NSRDB location. For example, a consecutive NA gap size of 4 occurs approximately 180 times resulting in a total of over 600 NA values. Additionally, a gap size of 48 consecutive NAs occurs once, totaling 48 NA values. Artificial gaps for the NSRDB data were created using a random sampling of the corresponding location's NA distribution.

2.3.2 Ground-measurement Data

All missing values in the measured data were synthetically removed. The same distributions of the original missing values in the NSRDB data were used; however, as the ground measurements are in one-minute intervals, compared to the NSRDB measurements in 30-minute intervals, the missing data lengths were multiplied by a factor of thirty. For example, the creation of one consecutive NA gap size in the NSRDB data corresponds to the addition of a thirty consecutive NA gap size in the ground measurement data. Note that the distributions used for the ground measurement data locations align with the corresponding NSRDB location.

2.3 Statistical Reporting Metrics

As mentioned earlier, in order to measure the performance of different imputation methods listed above, we create synthetic gaps in the Kt series using the distribution of missing (NA) values for NSRDB locations over the SURFRAD and NREL sites. Note that, in order to apply the same distribution to the ground measurements which are obtained in one-minute intervals, each distribution value was multiplied by thirty as previously mentioned. After applying different gap-filling methods to the same datasets, the following criteria, namely, Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and MAE percent of reading (%), and Mean Bias Error (MBE) are calculated for each site as described in Habte et al. (2017). The resulting equations are listed below , where N corresponds to the number of data, ytrue is the actual value of Kt at time t, and yi corresponds to the imputed Kt value at time t.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_{true}))^2} \qquad \text{(eq. 1)}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_{true} - y_i| \qquad \text{(eq. 2)}$$

$$MBE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_{true}) \qquad \text{(eq. 3)}$$

3. Results

3.1 NSRDB Data

To further measure method performance, the strings of consecutive NA values in the NRSDB *Kt* series, both originally missing and synthetically created gaps, were partitioned into varying bin sizes. Each method's imputation performance is measured for all six NA bin sizes across three separate groups, namely, (1) Bin 1: 1 consecutive NA's, and (2) Bin 2: >1 consecutive NA's, (3) Bin 3: 1-2 consecutive NA's, (4) Bin 4: >2 consecutive NA's, (5) Bin 5: 1-3 consecutive NA's, (6) Bin 6: >3 consecutive NA's. Dividing the bin sizes into three separate groupings allowed for testing the sensitivity of the imputation methods to gap size. For example, in the figures discussed below, it is concluded that spline interpolation's imputation performance significantly suffers when implemented on large gap sizes.

Figure 3 shows the RMSE results for each method and bin size for the Bondville, IL NSRDB site. Spline interpolation performs significantly worse in imputing the missing values for bins 2, 4, and 6 compared to the other methods. Kalman filter methods, linear and stine interpolation, and the moving average methods perform similarly for all bin sizes except bin 2 in which stine interpolation has a low RMSE value comparatively.

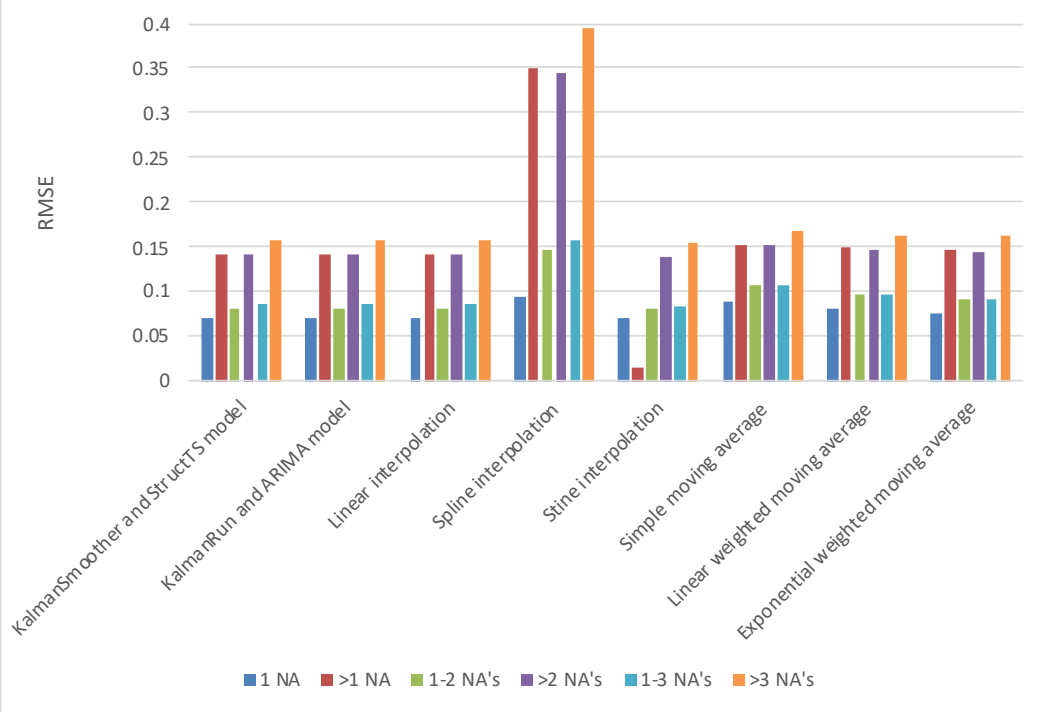

Fig. 3: Resulting RMSE values by imputation method across six NA bin sizes for Bondville, IL NSRDB data

Figure 4 shows the resulting MAE for each method and bin size for the same NSRDB location. Like with RMSE, spline interpolation is the worst performing imputation method. In the case of MAE, there is more difference between the other imputation methods, with the Kalman filter and moving average methods performing most similarly.

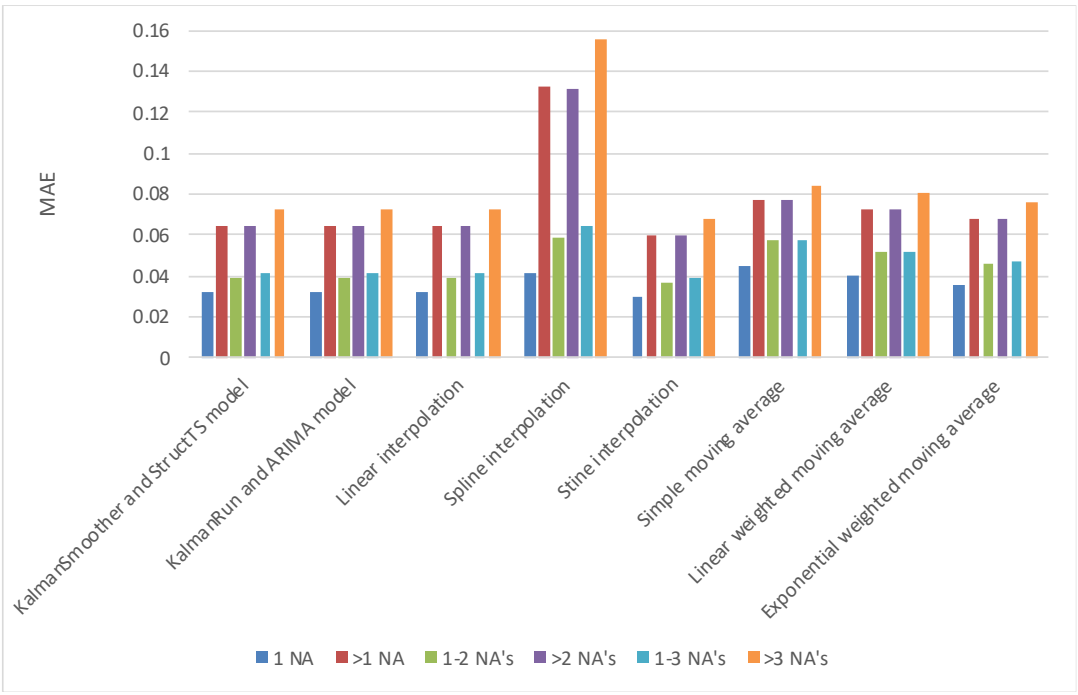

Fig. 4: Resulting MAE values by imputation method across six NA bin sizes for Bondville, IL NSRDB data

Both linear and stine interpolation has the smallest MAE values across all six bin sizes. The values between these two interpolation methods are not statistically different.

Figure 5 shows the resulting absolute values of the MBE results for each imputation method and across the six bin sizes. For graphical purposes, the absolute value was used to visually compare the bias across the methods. Like with RMSE and MAE, spline interpolation has the highest bias across bins 2, 4, and 6. The method also has noticeable higher bias than the other for bin 1. Stine interpolation and the moving average methods exhibit the lowest bias; however, the Kalman filter methods and linear interpolation do not have significant higher bias comparatively. From the statistical metrics shown above and those for the other NSRDB locations, we conclude linear interpolation and stine interpolation are the best performing methods with low RMSE, MAE, and MBE values across all bin sizes. We have also shown that all methods except for spline interpolation are relatively insensitive to change in consecutive NA gap size. As expected, the resulting statistical metrics increase as gap size increases.

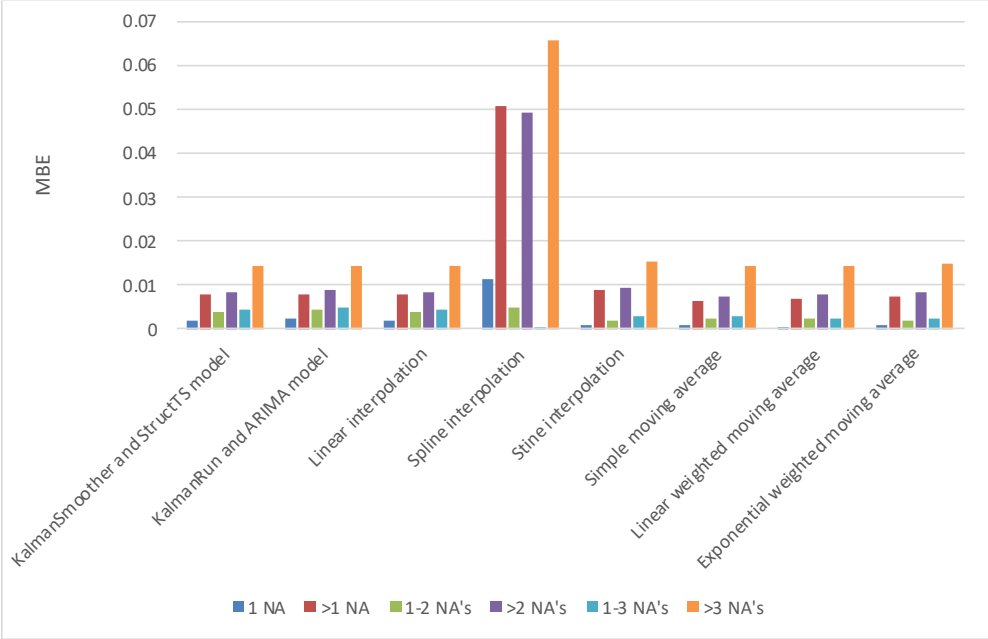

Fig. 5: Resulting absolute MBE values by imputation method across six NA bin sizes for Bondville, IL NSRDB data

3.2 Ground measurement Data

While gap filling the NSRDB data sets, the imputation methods were tested for both prediction performace and sensitivty to numerous NA gap sizes. For the SUFRAD data, six bin sizes were also used to calculate the resulting metrics according to consectuive NA gap sizes. Note that only one trial was run with all siz bin sizes compared to three separate groupings in the NSRDB data. Additionally, only linear and stine interpolation methods were applied to the ground measurement data as follows the NSRDB results. To account for the difference in measurement intervals between the NSRDB and SURFRAD data, the NSRDB NA distributions were multiplied by thirty for their corresponding SURFRAD sites. The converted bins are (1) Bin 1: 1-30 consecutive NA's, (2) Bin 2: 31-60 consecutive NA's, (3) Bin 3: 61-90 consecutive NA's, (4) Bin 4: 91-120 consecutive NA's, (5) Bin 5: 121-150 consecutive NA's, and (6) Bin 6: >150 consecutive NA's.

Figure 6 shows the resulting RMSE for the ground measurement data by location and interpolation method across all six bin sizes. For each location and bin size, linear interpolation has the smaller resulting RMSE, though, these values do not significantly differ for stine interpolation. The methods perform most similarly for the DRA location, which has more clear sky days and, therefore, less missing observations compared to the other locations. We expect the locations with larger NA distributions to result in higher RMSE values.

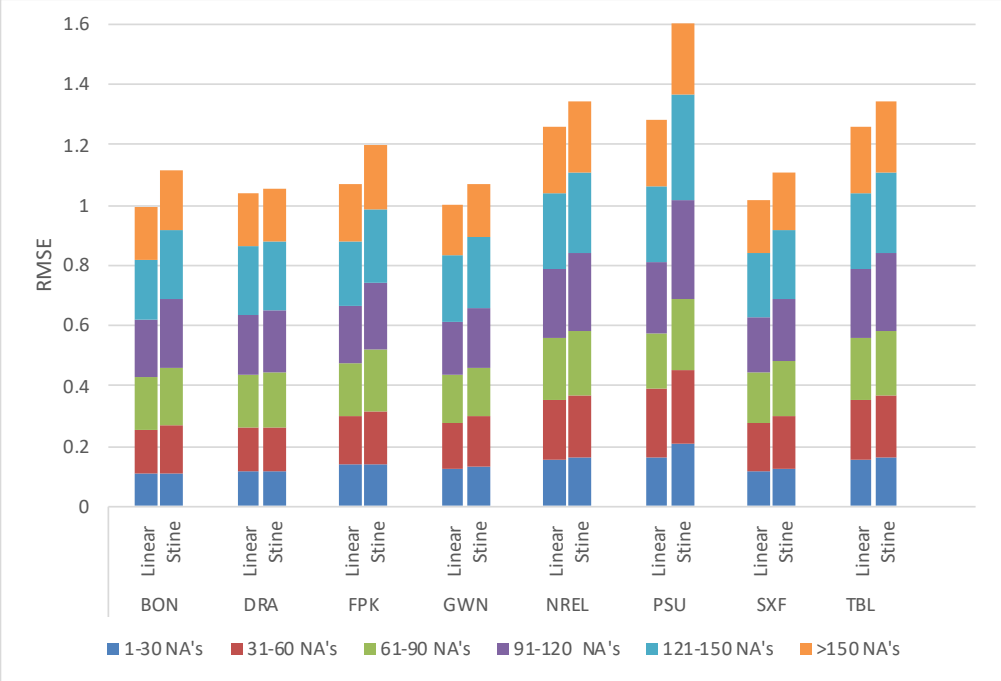

Fig. 6: Resulting RMSE by location and interpolation method for ground measurement data

Figure 7 shows the resulting MAE across all six bin sizes by location and interpolation method for the ground measurement data. Linear interpolation has smaller resulting MAE for locations BON, FPK, and PSU. The methods perform similarly across bin sizes for locations DRA, GWN, and SXF. Stine interpolation has smaller resulting MAE for location NREL. We see that the methods are performing closely in terms of MAE and vary in performance based on location, though, the values are not significantly different.

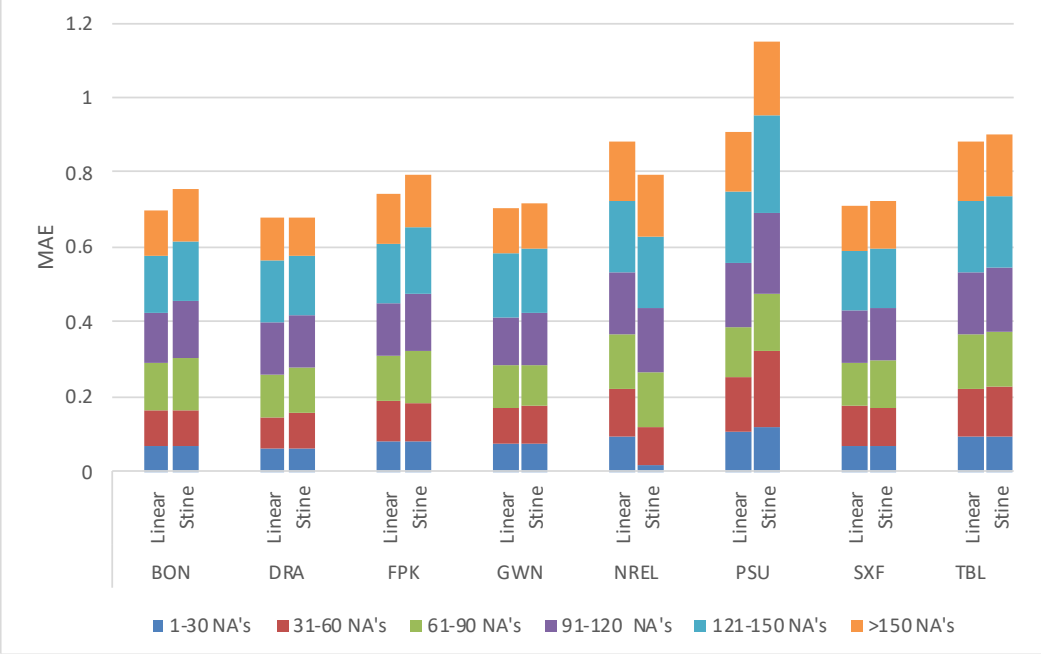

Fig. 7: Resulting MAE by location and interpolation method for ground measurement data

Figure 8 above shows the absolute MBE values per bin size across interpolation methods and location. As

mentioned for the NSRDB results, the absolute value of MBE was taken for graphical comparison purposes since we are focused on observing how far the value is from zero. Unlike the RMSE and MAE results, the resulting MBE greatly varies across bin sizes, interpolation method, and location. As expected, a higher bin size (larger length of consecutive NA's) results in a higher bias among the imputed values. There is not consistently low bias across bin sizes per location. For example, linear interpolation has a higher bias for bin 2 at the DRA location compared to stine interpolation but has a lower bias for the same location for bin 3. The two methods exhibit an array of biases.

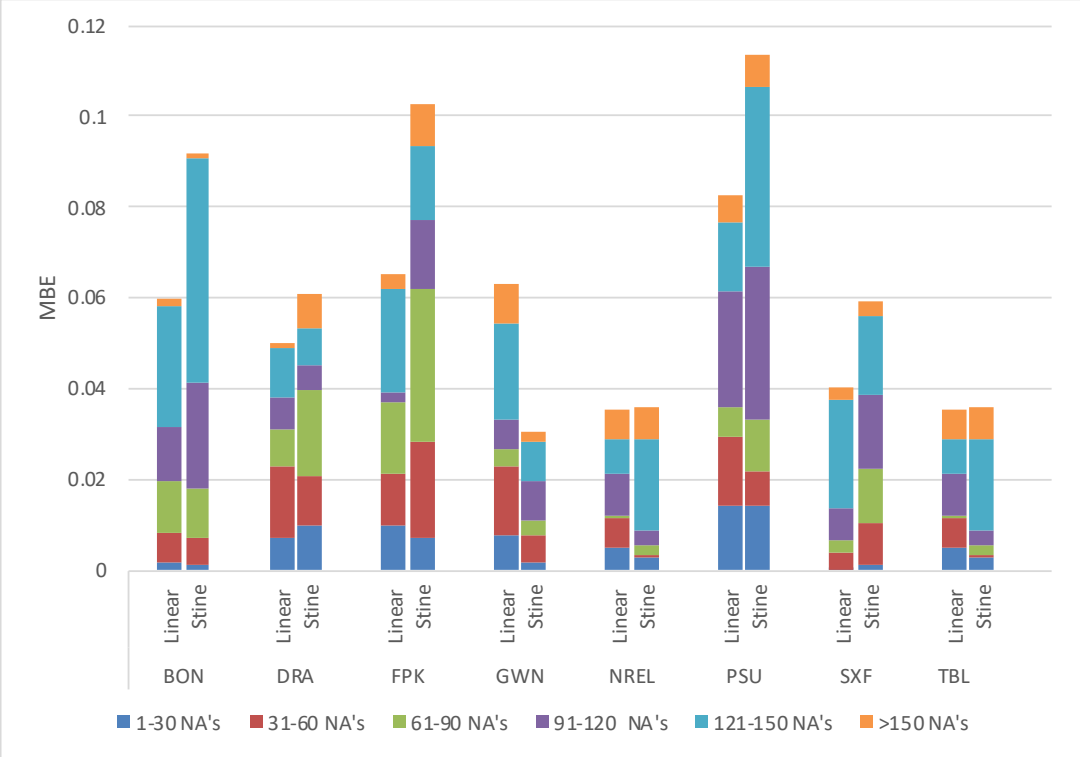

Fig. 8: Resulting MBE by location and interpolation method for ground measurement data. Note: for plotting simplicity, absolute values of the MBEs are considered in this figure.

4. Conclusion

Appropriate usage of solar resource dataset for solar energy project phases depends largely on the completeness of the dataset. This study evaluated eight statistical imputation techniques and selected and implemented two of them to fill NSRDB and ground measurement random and artificial data-gaps.

As described in the results, linear and stine interpolation outperformed the other six imputation methods in terms of RMSE, MAE, and MBE for each NSRDB location across three groupings of two differing NA gap sizes. Hence, when imputing the ground measurements, only linear and stine interpolation were applied. The two perform similarly in terms of RMSE and MAE for the ground measurements with linear interpolation performing slightly better than stine interpolation in RMSE. Both methods exhibited varying bias across location and bin size for the ground measurement.

Our results indicate that some of the *simpler methods* such as stine and linear interpolation methods perform best compared to others for imputing NSRDB and ground measurements, respectively. Linear interpolation is the most appropriate choice of imputation method for the data in question due to its simplicity compared to stine interpolation. However, it is important to apply all methods when imputing new data sets to see which is most appropriate.

5. Acknowledgments

This work is supported in part by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Solar Energy Technologies Office. The National Renewable Energy Laboratory is operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy under Contract No. DE-AC36-08GO28308. The views expressed in the article do not necessarily represent the views of the Department of Energy or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

6. References

Augustine, J. A., DeLuisi, J. J., & Long, C. N. (2000) SURFRAD—A National Surface Radiation Budget Network for Atmospheric Research. *Bulletin of the American Meteorological Society*, vol. 81, no. 10, 2341–2357.

Habte A, Sengupta M, Lopez A. (2017). Evaluation of the National Solar Radiation Database (NSRDB Version 2): 1998-2015. NREL/TP-5D00-67722. https://www.nrel.gov/docs/fy17osti/67722.pdf.

Sengupta, M., Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby (2018), The National Solar Radiation Data Base (NSRDB), Renew. Sustain. Energy Rev., 89, 51-60. https://doi.org/10.1016/j.rser.2018.03.003