Advanced Signal Decomposition Analysis and Anomaly Detection in Photovoltaic Systems

Mahya Qorbani, Daniel Fregosi, Devin Widrick, and Kamran Paynabar

***Abstract*—With the rapid expansion of large-scale photovoltaic (PV) plants, it is paramount for solar stakeholders to understand the reliability and efficiency of their plants to inform maintenance decisions, increase production, and understand the design factors that impact performance. Diagnosing underperformance in PV plants is challenging due to the relatively few monitoring points with respect to the large geographic footprint of the plant. This study introduces a cutting-edge method that transforms the analysis and management of key factors influencing PV plant performance, including performance loss rate, recoverable soiling, and major system changes. Identifying these factors is critical for deriving actionable insights. Leveraging advanced analytical techniques, such as wavelet transformation, robust regression, and extreme point analysis, this approach provides a nuanced understanding of these factors. This method has been tested across two synthetic datasets and one real dataset, consistently surpassing existing benchmarks by achieving a lower median mean absolute error and reduced error variability across all comparable components.**

***Index Terms*—Degradation, performance loss rate (PLR), performance monitoring, photovoltaics (PVs), soiling, statistical learning.**

I. Introduction

In LARGE-SCALE photovoltaic (PV) plants, performance is assessed by monitoring the energy produced over time compared against the expected energy. The performance index (PI) is the ratio of actual to expected energy. The PI is comprised of various factors like major system changes, seasonality, recoverable soiling, and gradual degradation, or performance loss rate (PLR). System changes refers to sudden shifts in plant performance that persist for a meaningful period of time which significantly alter plant behavior. Seasonality refers to yearly patterns in the PI signal due to physical phenomena not included in the expected energy model [1]. Soiling is the accumulation of dust and debris on solar panels, particularly significant in arid, desert-like environments [4]. The impacts of soiling can be mitigated by cleaning or washing modules. An accurate assessment helps optimize cleaning schedules. Recognizing its importance, numerous studies over the decades have focused on quantifying soiling losses to optimize solar panel efficiency [5], [6], [7], [8]. The effect of soiling on efficiency is often studied in isolation [9], without considering its interaction with other critical factors. PLR refers to a gradual reduction of capacity over time. PLR encompasses both module and system level degradation caused by the accumulation of many unfixed faults and failures. It is a crucial factor for the long-term economics of a plant. Quantifying and evaluating PLR is discussed in various studies [1], [9], [10], [11], [12], [13], [14], [15].

This study addresses a critical gap by introducing a comprehensive methodology that decomposes the PV system's PI signal into its constituent components. By sequentially isolating each component, our method eliminates its effect from the analysis, enabling a more accurate estimation of the subsequent component. This streamlined approach yields a deeper insight into the genuine impact of each factor on PV system efficiency. In addition to enhancing our understanding of all component impacts, our methodology offers broader implications for the renewable energy sector. By providing a more granular and accurate analysis of PV system performance, our approach enables operators and engineers to implement more targeted maintenance and optimization strategies. This is especially critical for accurately assessing PLR and soiling. The implementation of such analytical techniques could lead to substantial improvements in solar energy efficiency and reliability, contributing to the global effort to transition to sustainable energy sources. A holistic approach is presented for understanding and quantifying PV system losses through a novel methodology that synergizes wavelet transformation, robust regression, and extreme point analysis, thereby enabling the dissection and quantification of the compounded effects of system changes, seasonality, soiling, and PLR factors. Various hyperparameters are employed to estimate each component, as detailed in the methodology section, making the method adaptable to different sensitivity preferences and noise levels. The analysis reveals the detailed impact of these factors on PV system performance and offers flexibility in tuning for different sensitivity levels.

II. Methods

The main goal of this analysis is to extract the system changes, soiling, seasonality, PLR, and noise signals from an overall, inverter-level, performance signal. We refer to this performance signal as the PI and it represents how each inverter is performing

Received 28 May 2024; revised 9 October 2024; accepted 14 October 2024. Date of publication 30 October 2024; date of current version 25 December 2024. This work was supported by the U.S. Department of Energy Solar Technologies Office under Award DE-EE-0008976. *(Corresponding author: Kamran Paynabar.)*

Mahya Qorbani and Kamran Paynabar are with the School of Industrial, Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: kpaynabar3@gatech.edu).

Daniel Fregosi and Devin Widrick are with the Electric Power Research Institute, Charlotte, NC 20005 USA.

Digital Object Identifier 10.1109/JPHOTOV.2024.3483258

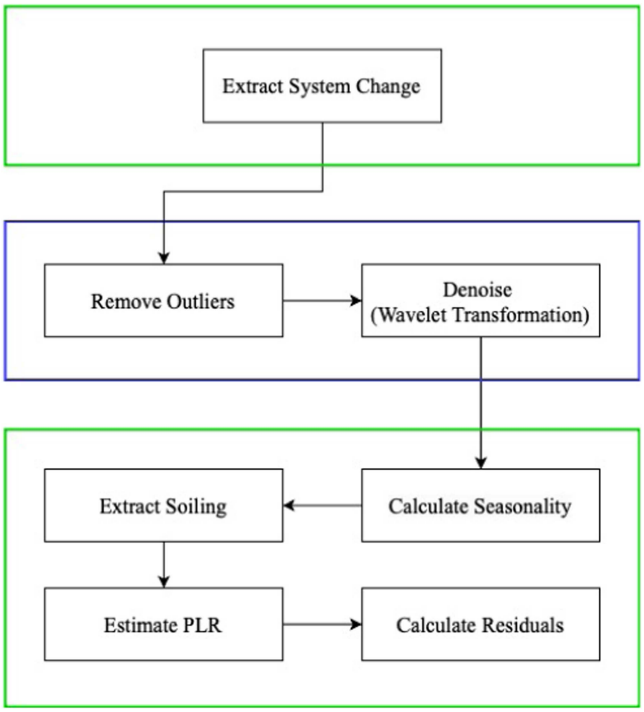

Fig. 1. Illustrates the methodological flowchart of our study, detailing the procedural steps involved. Green squares represent the components (system changes, seasonality, soiling, PLR, noise) extraction phase. The blue square denotes the preprocessing steps, underscoring the initial treatment and preparation of data for subsequent analysis.

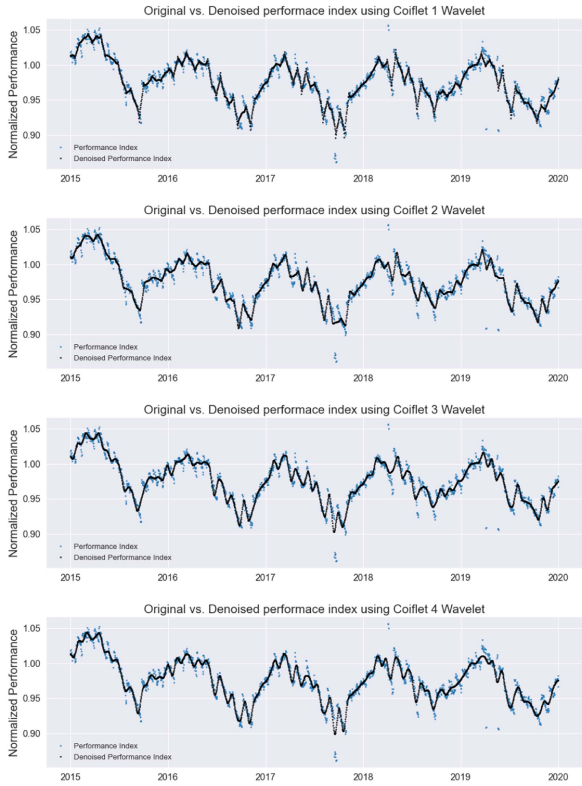

Fig. 2. Illustrates the effect of denoising on the PI using different Coiflet wavelet functions (Coif1 to Coif4).

relative to its initial operating state. It is represented as a ratio and a value less than 1 indicates the system is producing less power under the same conditions compared with the initial state of the plant. The *Synthetic Data* and *Real Data* subsections as follows cover the two different ways we generated the PI in this study.

The analysis covered in this report begins with the extraction of the system changes component from the PI. Following this, preprocessing steps are undertaken to refine the PI. Subsequently, the other key components of the PI-seasonality, soiling, degradation, and noise-are extracted. This structured approach allows us to isolate and understand the impact of each factor on the performance of solar panels comprehensively. For an in-depth visual representation of this methodology, refer to Fig. 1.

A. System Changes

The initial component we extract from the PI is the system changes signal, which substantially impacts plant behavior. Examples of such system changes include groups of trackers no longer functioning properly, damage from weather events, a downed combiner box, or parts of the plant going offline for maintenance, each of which significantly alters the plant's operational dynamics. We identify these changes by looking for sudden shifts in original PI, without any preprocessing steps, that persist for over 20 d. By directly analyzing the raw PI, we ensure that significant changes are not inadvertently smoothed over or eliminated. Preprocessing, such as outlier removal or denoising, might obscure these crucial changes by altering the very fluctuations we seek to identify. The duration threshold of 20 d serves as a tunable hyperparameter, allowing us to fine-tune our detection algorithm for optimal performance across various datasets. This adaptability enables the algorithm to recognize a broad range of change magnitudes, tailored to specific requirements in different contexts. To implement this, fluctuations in the PI exceeding 3% are flagged as a system changes. This approach ensures that our detection method remains both sensitive and specific to meaningful changes in system behavior.

B. Preprocessing

After the system changes is extracted from the PI, the preprocessing stage of the study begins, where signal outliers are replaced with the 25th and 75th quartiles, effectively mitigating the effects of outliers. Subsequently, we employ wavelet transformation for denoising [16], [17], utilizing the Coiflet 3 (coif3) wavelet. It is a powerful mathematical tool that breaks down a signal into its constituent parts at different scales, or resolutions. This method allows us to capture both the fine details and the broader trends within the signal. The choice of coif3, a specific type of wavelet, in our analysis stems from its ability to provide a good balance between the accuracy of representing signal details and the efficiency of computation. Fig. 2 illustrates the effect of denoising on the PI using different Coiflet wavelet functions (Coif1 to Coif4). The original PI is plotted alongside the denoised versions, demonstrating how each wavelet impacts signal smoothing. As shown, Coif3 achieves

a balanced denoising effect, effectively retaining key trends, while minimizing high-frequency noise. Other Coiflets either oversmooth the signal or leave more residual noise. Notably, Coif3 and Coif4 produce comparable results. This comparison supports our choice of Coif3 for the preprocessing step, as it provides an optimal balance between noise reduction and feature preservation. The transformation is executed at a decomposition level of four, segmenting the signal into approximation coefficients, which capture the signal's general shape or trend, and detail coefficients, which capture the finer, rapid changes. By analyzing and modifying these coefficients across multiple scales, we can conduct a nuanced analysis and selectively enhance or reduce specific aspects of the signal. To denoise the signal, we set a threshold value of 0.6, applying it selectively to modify the wavelet coefficients. This thresholding specifically targets the detail coefficients, which represent the high-frequency components typically associated with noise. At the first level, a uniform threshold of 0.6 is employed, whereas at levels two through four, thresholds are dynamically assigned based on the maximum absolute value of coefficients at each level. The threshold value of 0.6 was specifically chosen as it produced lower errors in our analysis compared with other potential values (see Appendix A for detailed comparison plots). Through a technique known as soft thresholding, we reduce the magnitude of the detail coefficients by this threshold value. Coefficients falling below the threshold are effectively considered noise and minimized, allowing us to focus on the more significant, underlying trends in the data. This method offers considerable flexibility, enabling the adjustment of threshold levels to suit different signal characteristics or analysis needs, thereby highlighting the model's adaptability. After denoising, we reconstruct the signal with an inverse wavelet transform, using the modified coefficients. The result is a cleaner signal, with the significant noise components reduced, providing a robust foundation for subsequent analyses.

C. Seasonality

The second component that we aim to extract from the remaining part of the PI, after extracting the system changes and preprocessing, is seasonality. To discern yearly seasonality within PV systems, the initial step involves detrending the data using moving averages. Specifically, for our multiplicative model, we opt for division to remove the trend, distinguishing it from the subtraction method used in additive models. After detrending through division, we compute the seasonal average from the adjusted data to evaluate the seasonal impact accurately. The statsmodels library [18] is employed to facilitate this analysis, offering a systematic approach for examining time-series data, including trends, seasonality, and other fluctuations. Grasping these patterns is vital for the precise forecasting and in-depth understanding of PV system dynamics.

D. Soiling

Following the removal of seasonality, the next component we aim to segregate from the PI is soiling. Soiling has a considerable effect on solar panel performance, making its precise and thorough evaluation essential for understanding its impact. Our contribution to this field involves the development of a novel method that utilizes extreme points and regression analysis to estimate the soiling component. Our algorithm leverages local minima and maxima within the log-transformed signal to identify the onset and resolution of soiling periods, marked by a performance decline that is subsequently reversed through cleaning or natural events like rainfall. To ensure accuracy, we implement a filtering process that distinguishes significant soiling events by eliminating minor variances. This process is governed by criteria, such as a minimum required decline in signal strength relative to the start of a segment. Upon isolating these crucial soiling intervals, our method applies robust linear regression to the data segments both preceding and succeeding each local minimum. By employing the insights derived from these regression analyses, our algorithm effectively corrects the signal by removing the estimated soiling effects. This correction process simulates an ideal scenario, presenting the solar panels' performance as unaffected by soiling. Through this refined methodology, we enhance the precision of soiling effect assessments, thereby allowing for improved maintenance strategies for a specific site.

E. Performance Loss Rate

The last component that is extracted from the PI is PLR. To do this, we employ Huber regression, which is robust to outliers, on the remaining part of the PI. The regression slope serves as a proxy for the PLR. Our methodology is comprehensive, we first isolate and remove the impacts of system changes, seasonality, and soiling components to ensure that what we are measuring is purely the PLR effect. This process allows us to extract a clearer, more precise signal of PLR from the noise of regular operational data. Our approach not only advances our comprehension of the aging process in solar power plants but also establishes a foundation for devising maintenance and optimization strategies aimed at prolonging their service life. A high PLR, serving as an indicator of poor performance, can alert operators to investigate potential causes of diminished efficiency, potentially leading to corrective actions that restore performance. Furthermore, a high PLR may signify that modules are degrading faster than expected, which could justify a warranty claim. Comparing well-performing plants with those underperforming can inform improved operations and maintenance practices, enhancing overall fleet performance.

F. Residuals

Residuals refer to the portion of the PI that remains after accounting for system changes, seasonality, soiling, and PLR. The residuals represent a combination of sensor and model error and appear as random noise. More information about the characteristics of the residuals is given in the Data section.

III. Data

A. Synthetic Data

The synthetic data are generated loosely following the concepts developed by [20] and [21]. Instead of producing the PI

through a normalization step, like with the real data, the synthetic data PI is generated as a product of five synthetically generated components previously discussed: system change, seasonality, soiling, PLR, and the residuals. Synthetic data were used in this study due to the ability to know the real values of the difference components which allowed for tuning and development of the algorithms used.

In the synthetic data generation, system changes is simulated as a pulse with duration between 50 and 150 d and magnitude deviation of between -15% and 5% from unity. Seasonality and soiling implementation is based on the method used in [20] and [21]. Seasonality is applied as a simple sinusoid with a period of 1 a. The soil cleaning events are chosen at random, and the soiling slopes are randomly distributed. The PLR is assumed to be linear and the value for each synthetic inverter is chosen at random from a normal distribution with mean of -0.005 and a standard deviation of ± 0.002. The final component, noise, represents all unknown or unaccounted behaviors of the plant, besides PLR, soiling, seasonality, and system changes. This includes modeling errors in the normalization step. A more accurate normalization model gives a lower noise signal in the PI since it captures more dynamics of the plant. Errors in weather data, weather factors not captured by the data, and errors in sensor data will appear as noise in the PI. The characterization of noise used here is a significant contribution of this work. The noise generator is based on a frequency spectrum which was found empirically by measuring PI signals on actual plants. Rather than simple white noise, where the magnitude is equal at all frequencies, we found PI signals typically contain noise that decreases in magnitude with frequency, like brown or pink noise, which creates a random walk. To generate a random noise signal with the desired frequency spectrum characteristics, a random number is selected at each frequency, according to the defined standard deviation at that frequency which matches the empirical noise profile. The randomly selected frequency spectrum is converted to a time-series signal with an inverse fast Fourier transform. The frequency spectrum profile characteristics can be easily adjusted by two parameters that control the magnitude and slope of the profile. The magnitude can changed to adjust the level of noise.

An example plot is shown in Fig. 3. The parameters are adjustable to qualitatively match the composition of the PI signal from an actual plant. In particular, the noise can be adjusted to mimic actual normalized performance data.

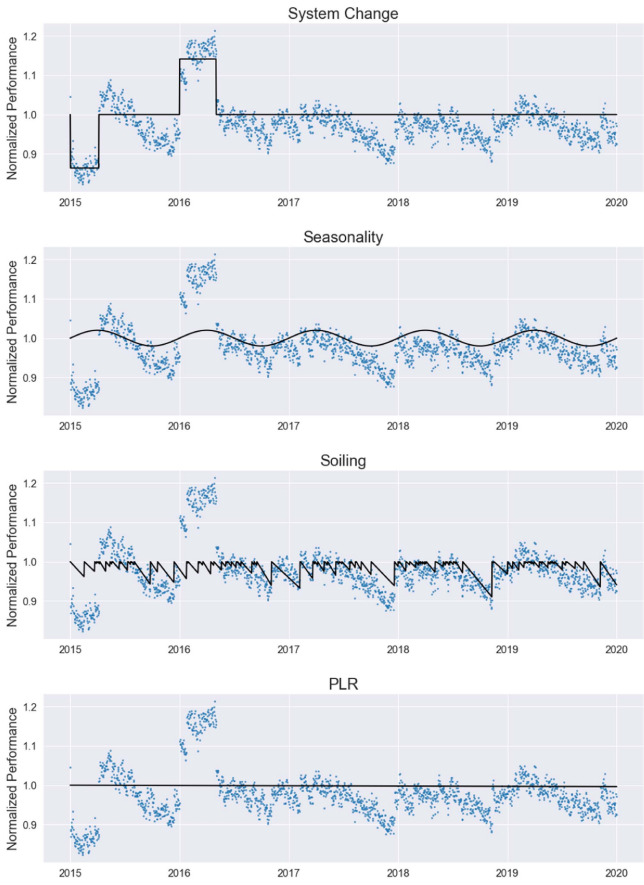

Fig. 3. Example of a synthetically generated, inverter-level, PI. The blue dots represent the synthetic PI signal, while each separate plot has a solid black line that represents the true value for either the system changes, seasonality, soiling, or PLR.

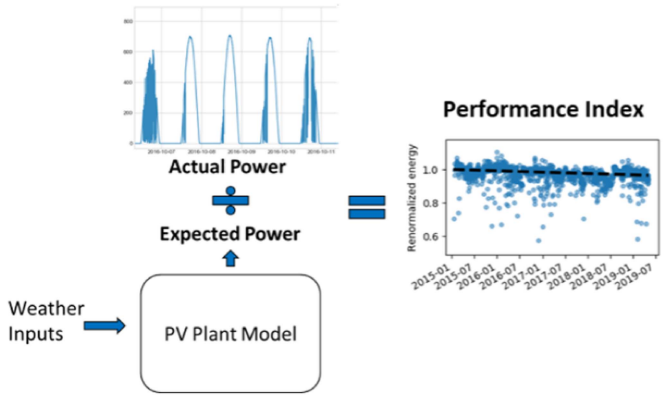

Fig. 4. PI is the ratio between the actual power measured by the inverter and the expected power generated by the data-driven model that uses site specific metrics and local weather data as inputs.

B. Real Data

When using real data, the PI is calculated during a normalization step. This normalization step occurs after training a data-driven model on early-life inverter performance with the measured weather data. The weather data are generated using either the specific site's weather station data or a satellite source's weather data. The PI is then calculated at the inverter-level and is a comparison between the actual energy generated and the model-estimated amount of energy produced under a given set of weather conditions, Fig. 4. A comparison of model accuracy when using site weather stations, free satellite data, or commercial satellite data paired with a physics-based model, simple regression models, or complex ensemble-based models is reviewed in another study [19].

Fig. 5 presents four illustrative plots that exhibit the characteristics of the real dataset, notably its high noise levels and the presence of missing data points. Unlike synthetic datasets,

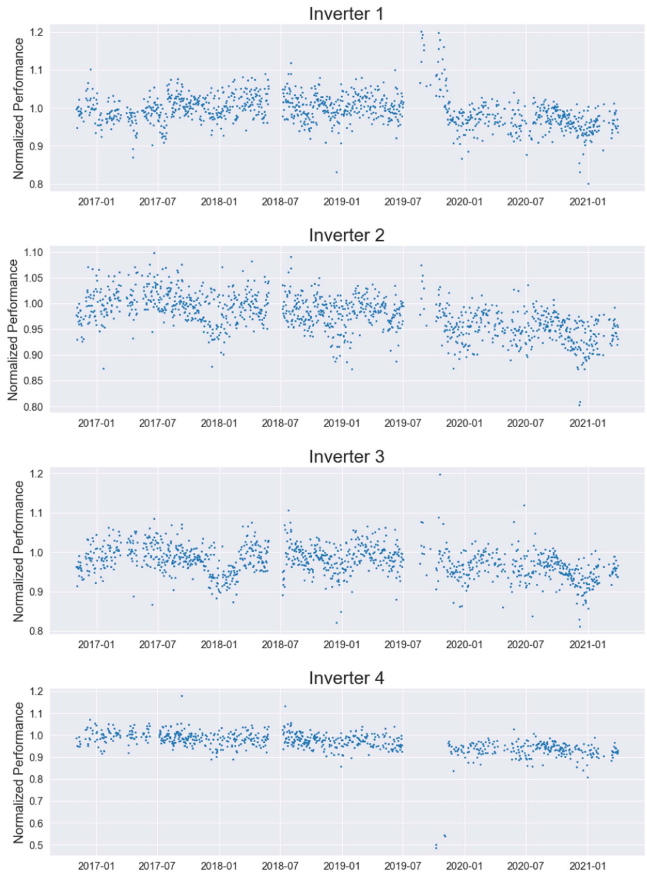

Fig. 5. Four examples of real data, inverter-level PI. The blue dots represent the real PI signal for each inverter.

TABLE I
SYSTEM CHANGES ESTIMATION ERRORS FOR S-LOWNOISE AND S-HIGHNOISE

Inverter	Error for S-LowNoise	Error for S-HighNoise
1	0.1045	0.0624
2	0.0668	0.2359
3	0.0389	0.1144
4	0.0137	0.0022
5	0.0213	0.0033
6	0.0093	0.0230
7	0.0799	0.1357
8	0.0629	0.1850
9	0.0805	0.0208
10	0.0378	0.1292

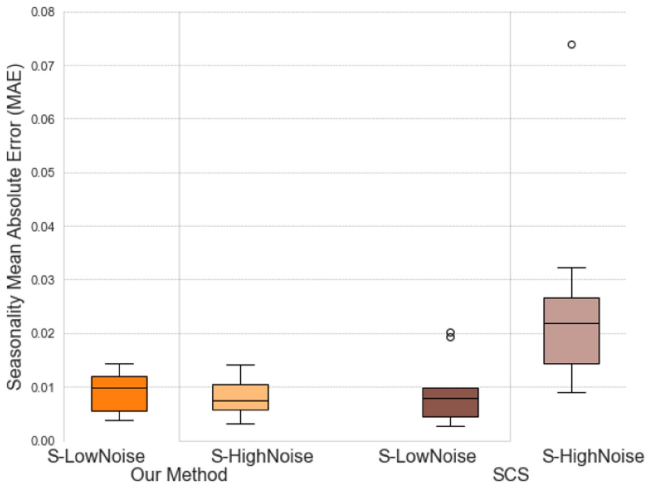

Fig. 6. Comparative Box Plot of MAE values for seasonality estimation: This figure illustrates the distribution of MAE values across our method and SCS.

which are typically cleaner, this real dataset necessitates a more comprehensive preprocessing approach to address these issues effectively and ensure accurate analysis.

IV. RESULTS

In this section, we present the outcomes of applying our method to both synthetic and real-world datasets. Synthetic datasets are particularly valuable as they contain known values for system changes, seasonality, soiling, and PLR, thereby enabling a direct evaluation of our method's performance. In this study, our methodology is evaluated using two distinct synthetic datasets, named S-LowNoise and S-HighNoise, to reflect their differing noise levels. S-HighNoise exhibits greater noise levels than S-LowNoise, potentially offering a more accurate representation of real-world conditions. We employ the mean absolute error (MAE) as the evaluation metric for seasonality, soiling, and PLR. To evaluate our model's ability to detect system changes, we use the proportion of system changes that were not accurately identified as a metric for assessing our model's performance in this regard. Our results are benchmarked against two established models for comparison. The first comparison is made with the stochastic rate and recovery (SRR) method [9]. The second is made with the statistical clear sky (SCS) method [22], [23]. This comparative analysis aims to highlight the efficacy and accuracy of our approach in various scenarios.

It is noteworthy to emphasize our model's superiority in detecting all components, a capability that base models lack. For instance, SRR can only assess soiling and degradation, while SCS fails to account for system changes, which critically influences the estimation of other components when present.

In Table I, we present the performance of our proposed method in identifying system changes across two synthetic datasets, each with ten synthetically generated inverters. Our analysis reveals that the error associated with our method in estimating the system changes remains below 11% and 24% for any inverter on S-LowNoise and S-HighNoise datasets, respectively. Furthermore, the average error recorded for S-LowNoise stands at 5.15%, while for S-HighNoise, it is 9.11%.

Fig. 6 provides a box plot comparison of MAE values between our method and SCS for the estimation of the seasonality component within both synthetic datasets. This comparison aims to highlight the precision of our method in capturing seasonal variations. On the S-LowNoise dataset, both methods demonstrated commendable precision, with our method and the SCS method recording median MAEs of 0.0098 and 0.0017, respectively. The real distinction emerges in the S-HighNoise dataset, where our method significantly outperformed SCS, achieving a median

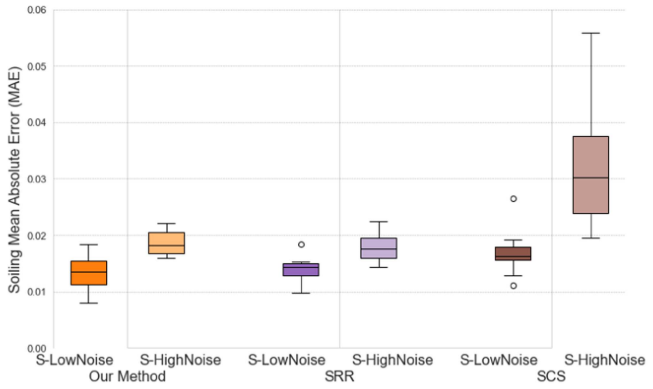

Fig. 7. Comparative Box Plot of MAE values for soiling estimation: This figure illustrates the distribution of MAE values across our method, SRR, and SCS.

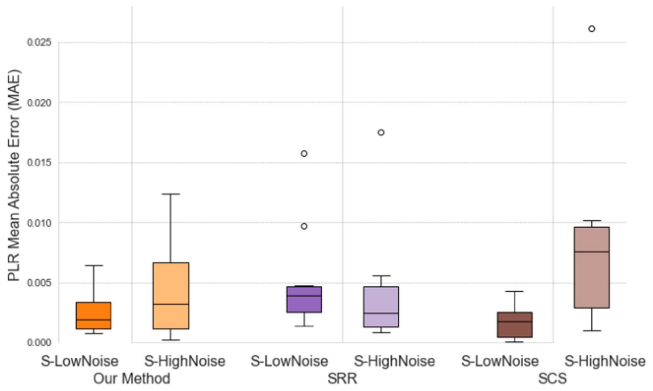

Fig. 8. Comparative Box Plot of MAE Values for PLR Estimation: This figure illustrates the distribution of MAE values across our method, SRR, and SCS on both synthetic datasets.

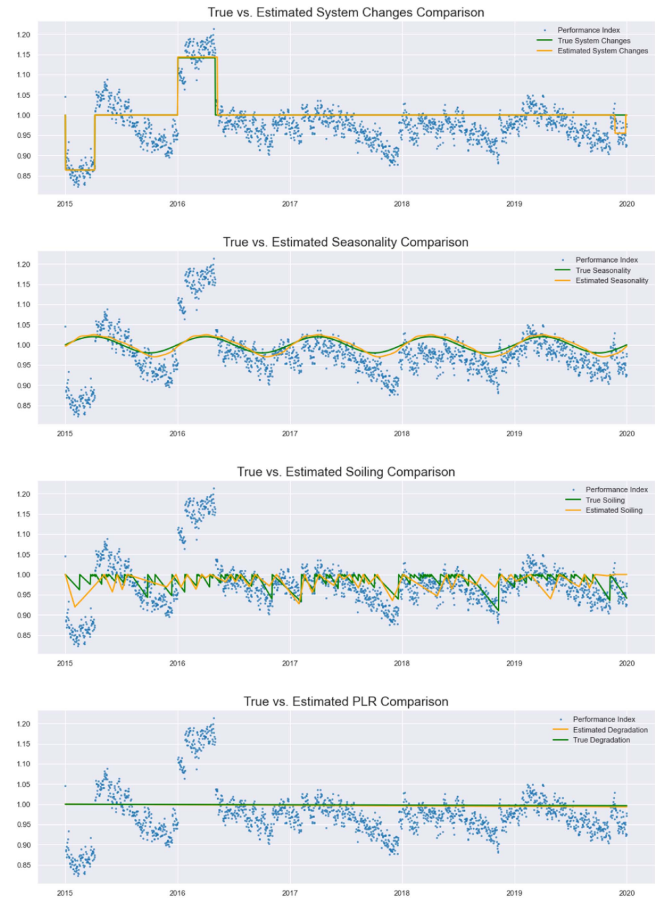

Fig. 9. Comparison of estimated versus true values for system changes, seasonality, soiling, and PLR components in one of the inverters in S-HighNoise dataset.

MAE of 0.0074 compared with SCS's 0.0218. Moreover, our method exhibited greater consistency across both datasets. For instance, the variability (max-min) in the S-LowNoise dataset for our method was 0.0105, compared with SCS's 0.0175, and in the S-HighNoise dataset, it was 0.0117 for our method versus 0.0650 for SCS.

Furthermore, Fig. 7's box plot illustrates the MAE comparisons for soiling estimation using our method, SRR, and SCS on S-LowNoise and S-HighNoise datasets. In low noise conditions, our method's median MAE (0.0134) is closely aligned with SRR (0.0143) and shows a slight improvement over SCS (0.0163). For the high noise scenario, our method (median MAE 0.0179) demonstrates a clear advantage over SCS (0.0302) and is comparable to SRR (0.0174). Regarding variability, our method exhibits less fluctuation with a max-min of 0.0104 in S-LowNoise and 0.0044 in S-HighNoise, outperforming SCS in both cases and showing greater stability than SRR in high noise conditions. This evidence suggests our method maintains consistency and reliability, particularly in environments with significant noise.

Fig. 8 provides a box plot comparison of MAE values for our method, SRR, and SCS when assessing PLR on S-LowNoise and S-HighNoise datasets. Our method demonstrates precision with a median MAE of 0.0018 on the low noise dataset, which is slightly higher than SCS's 0.0017 but lower than SRR's 0.0039. In the high noise environment, our method achieves a median MAE of 0.0036, which is significantly better than SCS's 0.0075 and slightly higher than SRR's 0.0024. Regarding variability in S-LowNoise dataset, our method exhibits a max-min of 0.0056, which indicates a larger variability compared with SCS's 0.0042, but is still substantially lower than SRR's 0.0144. In S-HighNoise dataset, our method's variability, with a max-min of 0.0123, is notably lower than both SRR's 0.0166 and SCS's 0.0251, demonstrating our method's robustness in more challenging conditions.

Fig. 9 demonstrates the performance of our method in estimating the system changes, seasonality, soiling, and degradation components of one of the inverters in S-HighNoise dataset. The first plot reveals that, while the method accurately detects the majority of system changes, it overestimates during periods of sharp transitions. In the second plot, the estimated seasonality closely follows the true seasonality curve, indicating a strong alignment with the actual seasonal patterns. The soiling estimation, shown in the third plot, suggests a precise tracking of the soiling events, albeit with some minor deviations during

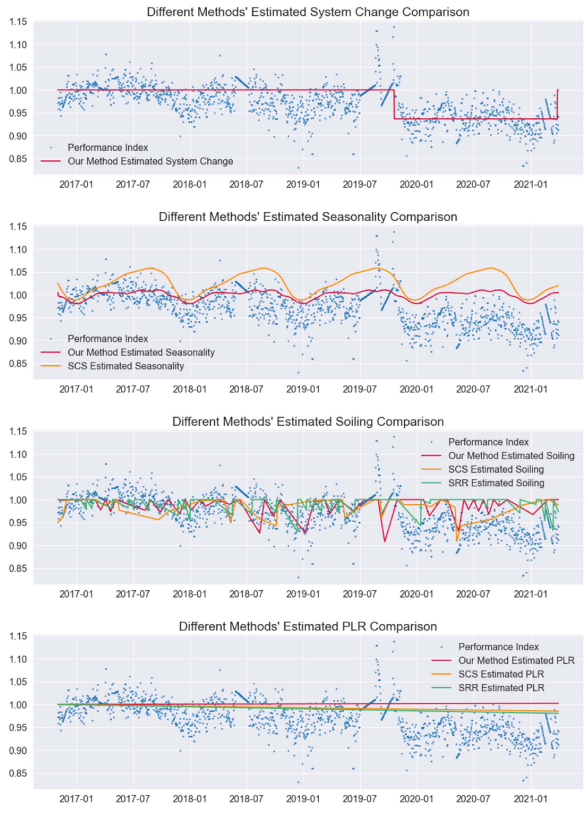

Fig. 10. Comparison of estimated values of our method, SRR, and SCS for system changes, seasonality, soiling, and PLR components in one of the inverters in real dataset.

rapid soiling and cleaning instances. Lastly, the PLR component is estimated with a high degree of consistency, as seen in the fourth plot, where the estimated degradation maintains a near-constant offset from the true degradation line, implying an accurate representation of the gradual performance decline. This assessment confirms the method's efficacy in addressing multiple facets of PV system behavior.

Our methodology was also applied to a real dataset, alongside the two baseline models, SRR and SCS, for comparative evaluation. As highlighted in the Data section, the real dataset poses challenges due to its inherent noise and missing values. To address the missing data, we employed a technique where we established a regression line between the known points at the beginning and end of the gaps. This regression line was then used to predict and fill in the missing values, ensuring continuity in the data. In alignment with our approach for synthetic datasets, we initially refrained from any preprocessing when identifying system changes in the real dataset. After the initial detection of system changes, we proceeded with our preprocessing routine; removing outliers and denoising the signal. Following these steps, we extracted the seasonality, soiling, and PLR components employing the same methodology delineated in the Methods section. Fig. 10 showcases the performance comparison between our method, SRR, and SCS in estimating system changes, seasonality, soiling, and PLR components within one of the inverters in the real dataset. Since the real dataset lacks actual values for these components, we are compelled to rely on a qualitative evaluation to compare the efficacy of the different methods. As previously noted, the baseline models do not possess the capability to identify system changes. However, as depicted in Fig. 10, our method successfully detected a discernible shift in the average data at the onset of the year 2020. Both our method and SCS have demonstrated the ability to estimate seasonality within the real dataset. However, it is evident that SCS exhibits greater sensitivity to outliers, resulting in larger magnitude fluctuations of the seasonality component when compared with our method. This indicates that our approach provides a more tempered and possibly more accurate reflection of seasonal variations, less influenced by extreme data points. When estimating the soiling component, the SCS method struggles to deliver accurate results, largely due to its susceptibility to the outliers and noise inherent in the real data. On the other hand, SRR demonstrates a somewhat improved performance, yet it fails to consistently estimate soiling across the entire signal, erroneously indicating periods without soiling where it is expected. In contrast, our method demonstrates a robust ability to discern the soiling component throughout the entire dataset, delivering a judicious estimate of soiling intervals. It shows resilience against the disruptive influence of noise and outliers, and unlike SRR, it provides a continuous and sensible estimation of soiling across the signal. In the analysis of the PLR components, all models demonstrate the capability to extract PLR effectively. However, it is observed that the SRR and SCS methods appear more susceptible to distortion by outlier values, potentially affecting the accuracy of their PLR estimations.

V. Discussion

In this study, we have developed an comprehensive method capable of isolating various components from the performance signals of PV systems. Our approach surpasses the capabilities of established baseline models, such as SRR and SCS, by concurrently extracting system changes, seasonality, soiling, and PLR. While SRR is limited to isolating soiling and PLR, and SCS lacks the ability to delineate system changes and is negatively impacted by data outliers and noise, our method provides a more nuanced view of each component's impact on PV plant performance. Our comparative analysis reveals superior performance in detecting seasonality within the S-HighNoise dataset and competitive performance within the S-LowNoise dataset relative to SCS. The susceptibility of SCS to outliers further underscores the resilience of our model, especially when applied to real-world datasets. In the realm of soiling detection, our method exhibits comparable effectiveness to SRR and demonstrates clear superiority over SCS across both synthetic datasets. In addition, while all models show competence in estimating PLR, our method distinguishes itself by yielding more consistent error variability across the evaluated synthetic datasets. This enhanced clarity benefits stakeholders in the solar industry by offering a more transparent understanding of the system's behavior, thus facilitating informed maintenance decisions. The versatility of our methodology opens the door for broad applications, ranging

from small residential systems to expansive solar farms. The adoption of such analytical techniques promises significant enhancements in the efficiency and dependability of solar energy systems, contributing vitally to the advancement of sustainable energy solutions worldwide.

VI. Conclusion

As large-scale PV plants grow, understanding their reliability and efficiency becomes crucial for solar stakeholders. This knowledge aids in making informed maintenance decisions, boosting production, and comprehending design factors that affect performance. This study introduces a novel method for analyzing PV system performance signals, addressing a critical gap in the existing methodologies represented by baseline models SRR and SCS. We employed wavelet transformation, robust regression, and extreme point analysis to develop our method. It stands out for its ability to concurrently extract crucial components including system changes, seasonality, soiling, and PLR, offering a holistic analysis toolkit for the solar energy community. Unlike SRR and SCS, which exhibit limitations in isolating specific components and are susceptible to data noise and outliers, our approach offers robustness and clarity in analyzing PV system performance. One of the notable strengths of our methodology is its proficiency in identifying system changes, a task at which the other models falter. Our comparative analysis demonstrates the superior or comparable performance of our method in identifying the aforementioned components against baseline models in terms of median MAE and consistency in error values (variability) across diverse datasets, including those with high noise levels and missing data points.

Appendix A
Threshold Value Analysis for Wavelet Denoising

To validate the selection of a threshold value of 0.6, we conducted an analysis of MAEs across a range of threshold values, from 0.1-1.2. Figs. 11–13 illustrate the impact of these thresholds on the MAEs for estimating the seasonality, soiling, and PLR components within the S-HighNoise dataset. Considering the mean MAE, median MAE, and variability, the threshold of 0.6 emerges as the optimal choice for denoising in our context.

The results reveal that both lower and higher threshold values lead to suboptimal outcomes: lower thresholds tend to under filter the noise, while higher thresholds risk oversmoothing, thus losing essential signal details. While different threshold values produce comparable results, the differences, though not substantial, further underscore that a threshold of 0.6 effectively balances noise reduction and signal preservation.

As discussed in Section II-C, we use a moving average approach to calculate seasonality, which inherently smooths out noise. Consequently, the threshold values of the wavelet do not significantly impact the error in seasonality estimation. Therefore, Fig. 11 shows similar bar charts and mean values across all threshold values.

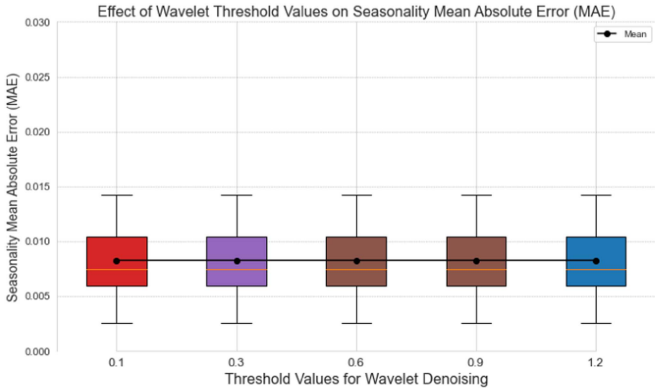

Fig. 11. MAE versus threshold values for wavelet denoising. The plot displays how different threshold values (0.1 to 1.2) affect the MAE values of estimating seasonality component in the S-HighNoise dataset, with all threshold values achieving the same error due to using the moving average approach in estimating the seasonality components.

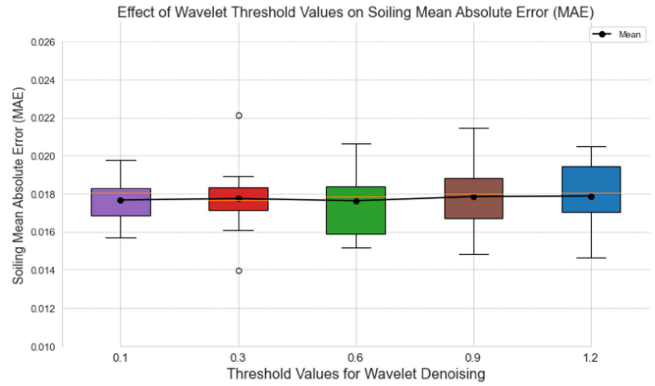

Fig. 12. MAE versus threshold values for wavelet denoising. The plot displays how different threshold values (0.1 to 1.2) affect the MAE values of estimating soiling component in the S-HighNoise dataset, with the threshold of 0.6 achieving the lowest error.

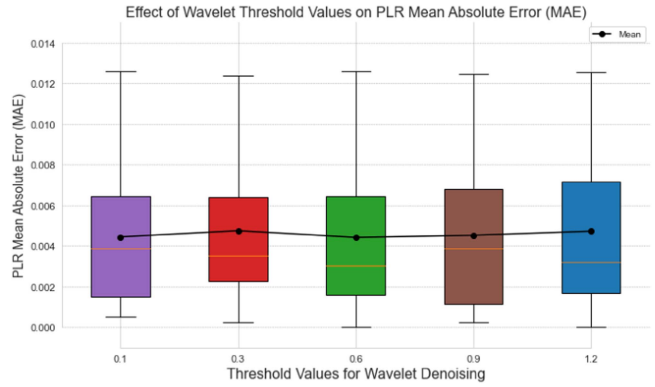

Fig. 13. MAE versus threshold values for wavelet denoising. The plot displays how different threshold values (0.1 to 1.2) affect the MAE values of estimating PLR component in the S-HighNoise dataset, with the threshold of 0.6 achieving the lowest error.

REFERENCES

[1] D. C. Jordan, C. Deline, S. R. Kurtz, G. M. Kimball, and M. Anderson, "Robust PV degradation methodology and application," *IEEE J. Photovolt.*, vol. 8, no. 2, pp. 525–531, Mar. 2018, doi: 10.1109/JPHOTOV.2017.2779779.

[2] V. P. Singh, V. Vijay, S. H. Gaurishankar, D. K. Chaturvedi, and N. Rajkumar, "Analysis of solar power variability due to seasonal variation and its forecasting for Jodhpur region using artificial neural network," *Power Res.- J. CPRI*, pp. 423–430, 2013. [Online]. Available: https://cprijournal.in/index.php/pr/article/view/883

[3] M. Z. Farahmand, M. E. Nazari, S. Shamlou, and S.-k. Miadreza, "The simultaneous impacts of seasonal weather and solar conditions on PV panels electrical characteristics," *Energies*, vol. 14, no. 4, 2021, Art. no. 845, doi: 10.3390/en14040845.

[4] M. Maghami et al., "Power loss due to soiling on solar panel: A review," *Renewable Sustain. Energy Rev.*, vol. 59, no. C., pp. 1307–1316, 2016. [Online]. Available: https://EconPapers.repec.org/RePEc:eee:rensus:v:59:y:2016:i:c:p:1307-1316

[5] C. Araujo et al., "Assessing the impact of soiling on photovoltaic efficiency using supervised learning techniques," *Expert Syst. Appl.*, vol. 231, 2023, Art. no. 120816, doi: 10.1016/j.eswa.2023.120816.

[6] R. R. Cordero et al., "Effects of soiling on photovoltaic (PV) modules in the Atacama desert," *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 13943, doi: 10.1038/s41598-018-32291-8.

[7] J. Diazgranados-Garzón et al., "Analysis of the soiling effect on solar-panel power efficiency in the Colombian Caribbean region," *Revista Facultad de Ingeniería*, vol. 97, pp. 22–29, 2020, doi: 10.17533/udea.redin.20191156.

[8] H. Abuzaid, M. Awad, and A. Shamayleh, "Impact of dust accumulation on photovoltaic panels: A review paper," *Int. J. Sustain. Eng.*, vol. 15, pp. 264–285, 2022.

[9] M. G. Deceglie, M. Leonardo, and M. Muller, "Quantifying soiling loss directly from PV yield," *IEEE J. Photovolt.*, vol. 8, no. 2, pp. 547–551, Mar. 2018, doi: 10.1109/JPHOTOV.2017.2784682.

[10] D. Fregosi, M. Bolen, and B. Paudyal, "Analysis of variability in calculated performance loss rates of large-scale PV plants," in *Proc. 47th IEEE Photovoltaic Specialists Conf.*, 2020, pp. 1742–1748, doi: 10.1109/PVSC45281.2020.9300710.

[11] M. G. Deceglie et al., "Perspective: Performance loss rate in photovoltaic systems," *Sol. RRL*, vol. 7, no. 15, 2023, Art. no. 2300196, doi: 10.1002/solr.202300196.

[12] D. C. Jordan et al., "Reducing interanalyst variability in photovoltaic degradation rate assessments," *IEEE J. Photovolt.*, vol. 10, no. 1, pp. 206–212, Jan. 2020, doi: 10.1109/jphotov.2019.2945191. [Online]. Available: https://escholarship.org/uc/item/448492x2

[13] T. Rahman et al., "Investigation of degradation of solar photovoltaics: A review of aging factors, impacts, and future directions toward sustainable energy management," *Energies*, vol. 16 no. 9, 2023, Art. no. 3706, doi: 10.3390/en16093706.

[14] J. Singh, J. Belmont, and G. TamizhMani, "Degradation analysis of 1900 PV modules in a hot-dry climate: Results after 12 to 18 years of field exposure," in *Proc. IEEE 39th Photovoltaic Specialists Conf.*, 2013, pp. 3270–3275, doi: 10.1109/PVSC.2013.6745149.

[15] M. G. Deceglie, D. C. Jordan, A. Nag, A. Shinn, and C. Deline, "Fleet-scale energy-yield degradation analysis applied to hundreds of residential and nonresidential photovoltaic systems," *IEEE J. Photovolt.*, vol. 9, no. 2, pp. 476–482, Mar. 2019.

[16] Ç P. Dautov and M. S. Özerdem,, "Wavelet transform and signal denoising using wavelet method," in *Proc. 26th Signal Process. Commun. Appl. Conf.*, 2018, pp. 1–4, doi: 10.1109/SIU.2018.8404418.

[17] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bull. Amer. Meteorological Soc.*, vol. 79, no. 1, pp. 61–78, Jan. 1998, doi 10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2.

[18] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with Python," in *Proc. 9th Python Sci. Conf.*, 2010, pp. 92–96.

[19] D. Fregosi, R. Dhakal, and D. Widrick, "Evaluating the use of satellite data and machine learning models for PV performance monitoring," in *Proc. IEEE Photovoltaic Specialists Conf.*, 2023, pp. 1–6, doi: 10.1109/PVSC48320.2023.10359946.

[20] A. Skomedal and M. G. Deceglie, "Combined estimation of degradation and soiling losses in photovoltaic systems," *IEEE J. Photovolt.*, vol. 10, no. 6, pp. 1788–1796, Nov. 2020, doi: 10.1109/JPHOTOV.2020.3018219.

[21] A. Skomedal, "Simulate PV time series," GitHub Repository, [Online]. Available: https://github.com/asmunds/simulate_pv_time_series

[22] B. Meyers, E. Apostolaki-Iosifidou, and L. Schelhas, "Solar data tools: Automatic solar data processing pipeline," in *Proc. IEEE 47th Photovoltaic Specialists Conf.*, 2020, pp. 0655–0656, doi: 10.1109/PVSC45281.2020.9300847.

[23] B. Meyers, "Estimation of soiling losses in unlabeled PV data," 2022, *arXiv:2209.09457*.