

## RESEARCH ARTICLE OPEN ACCESS

# A Gaussian Process Regression IV Model for PV Outdoor Data

Timon S. Vaas<sup>1,2</sup>  | Bart E. Pieters<sup>1</sup>  | Evgenii Sovetkin<sup>1</sup>  | Andreas Gerber<sup>1</sup>  | Uwe Rau<sup>1,2</sup> 

<sup>1</sup>IMD3-Photovoltaics, Forschungszentrum Jülich GmbH, Jülich, North Rhine-Westphalia, Germany | <sup>2</sup>Jülich Aachen Research Alliance (JARA-Energy) and Faculty of Electrical Engineering and Information Technology, RWTH Aachen University, Aachen, North Rhine-Westphalia, Germany

**Correspondence:** Timon S. Vaas ([t.vaas@fz-juelich.de](mailto:t.vaas@fz-juelich.de))

**Received:** 22 November 2024 | **Revised:** 25 March 2025 | **Accepted:** 12 June 2025

**Funding:** This work has been partially funded by the Federal Ministry of Research, Technology and Space under the Helmholtz LLEC Project and the Federal Ministry of Economic Affairs and Energy, BMWK under the grant FKZ 0325517B (PV-Klima project).

**Keywords:** big data | GPR | IV characteristics | PV outdoor data | statistical analysis

## ABSTRACT

Outdoor data are essential to study the reliability of PV modules and systems. Each electrical performance measure is dependent on the conditions the measurement is conducted at and, therefore, needs to be considered in the context of dynamically changing outdoor conditions. In this paper, we introduce a statistical model designed to analyze PV outdoor data. This model uses a timeseries of current-voltage (IV) characteristics, alongside meteorological data, including plane-of-array irradiance ( $G_{\text{POA}}$ ) and module temperature ( $T_{\text{Mod}}$ ). The model aims to utilize all available information to predict the respective performance measure as well as its uncertainty at arbitrary conditions and times. First, to ensure its quality and relevance, a suitable filtering approach is applied to the IV curves,  $G_{\text{POA}}$  and  $T_{\text{Mod}}$  data from nine modules from five locations (Arizona USA, Germany, India, Italy, and Saudi Arabia) observed for over 2 years. Following this, we utilize the extended solar cell parameters (ESPs), a descriptive model for IV characteristics using 10 parameters. The ESPs, then, undergo a principal component analysis (PCA), which transforms the ESPs into a set of uncorrelated principal components (PCs). Individual Gaussian process regressions (GPRs) are then trained on these principal components (PCs). Once the GPRs are trained, the model is capable of reproducing and predicting the complete IV characteristics at any given time  $t$ , for specified values of  $G_{\text{POA}}$  and  $T_{\text{Mod}}$ . This prediction includes an assessment of its standard deviation, which is derived from data noise and the distance from the observations. This model serves as a versatile tool for various applications, such as analyzing acclimatization effects, degradation trends, seasonal variations, and the performance ratio (PR) of PV modules or systems.

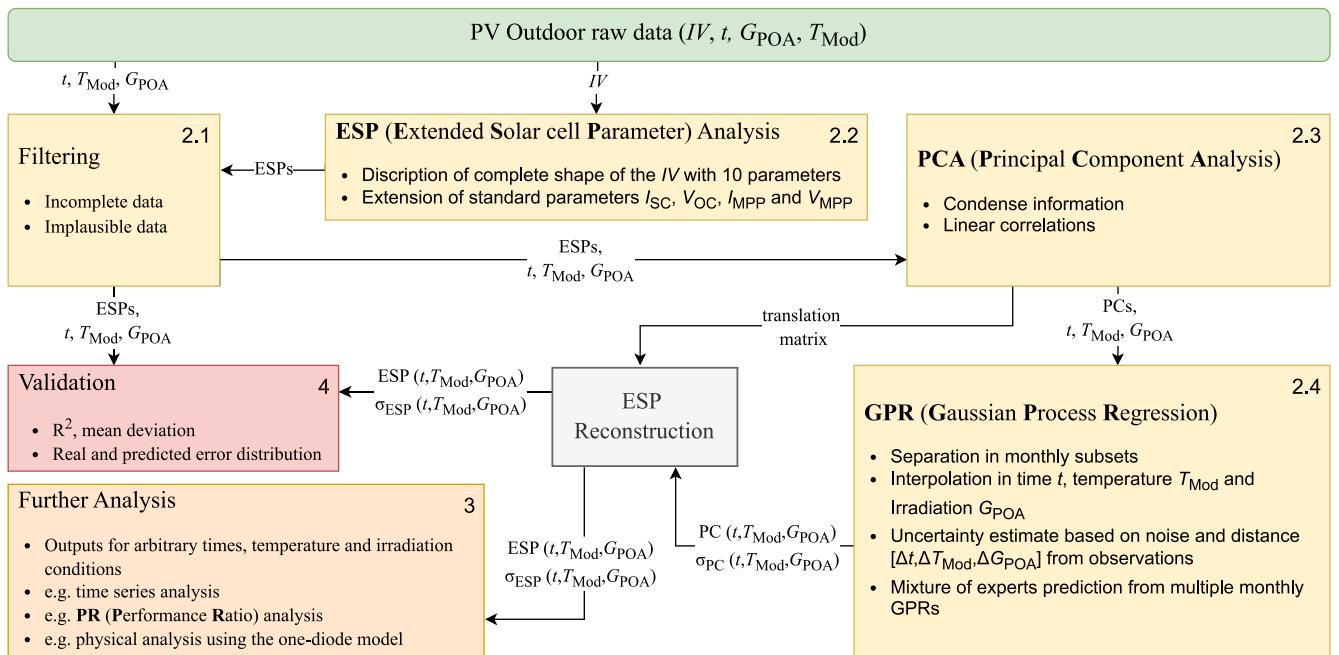
## 1 | Introduction

The reliability of photovoltaic (PV) modules critically affects the amortization of PV systems. Long-term outdoor monitoring of PV modules is essential for analyzing degradation, and, thus, also for assessing the quality and performance of PV products and technology. PV outdoor data generally includes electrical performance measures and additional meteorological and temperature measurements.

Electrical performance is commonly assessed through measurements of the complete current-voltage (IV) characteristic, which provide a reliable means to verify the status of individual modules over time [1–4]. Meteorological and temperature measurements, such as module temperature ( $T_{\text{Mod}}$ ) and plane-of-array irradiation ( $G_{\text{POA}}$ ), are highly correlated with the measured performance. Each electrical performance measure needs, therefore, to be considered in the context of dynamically changing outdoor conditions.

This is an open access article under the terms of the [Creative Commons Attribution](#) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). Progress in Photovoltaics: Research and Applications published by John Wiley & Sons Ltd.



**FIGURE 1** | Illustration of the presented GPR  $IV$  model and its PV outdoor data processing steps. The arrows depict the data flow from one step to another, while the numbers and color code give a reference to the structure of the paper.

While many outdoor datasets provide the complete  $IV$  curve for analysis, the evaluation of outdoor PV data typically focuses on just a few key solar cell parameters: open-circuit voltage ( $V_{OC}$ ), short-circuit current ( $I_{SC}$ ), and the voltage and current at the maximum power point ( $V_{MPP}$  and  $I_{MPP}$ ). This approach simplifies the analysis by reducing the complexity to only four parameters instead of the entire  $IV$  curve. However, this simplification may miss potentially important information. For example, relying solely on  $V_{OC}$ ,  $I_{SC}$ ,  $V_{MPP}$ , and  $I_{MPP}$  makes it impossible to distinguish between an  $IV$  curve with an “S”-shape (see for an overview of this phenomenon [5]) and one with high series resistance [6].

Another common approach to simplify the information contained in a measured  $IV$  characteristic is to extract the five parameters of the one-diode model [7]. Compared with SSPs, the one-diode model provides more detailed and physically meaningful insights into device performance, as its parameters have a direct physical interpretation. While widely used in the PV community, achieving a robust and efficient parameterization of the one-diode model remains a challenge. A comprehensive overview of existing parameterization methods can be found in [8–10]. However, the one-diode model has inherent limitations: it cannot account for certain physical effects, such as  $IV$  curves with an S-shape, restricting its applicability as a feature vector for  $IV$  characteristics. Additionally, the complex nonlinear optimization methods required to fit the one-diode model to measured  $IV$  curves pose further challenges.

A more descriptive and comprehensive representation of the full  $IV$  curve shape can be achieved using extended solar cell parameters (ESPs) [6]. The ESPs distill the information contained in the  $IV$  curve into 10 key parameters, preserving much detail of the shape of the  $IV$  characteristic. By analyzing 10 parameters instead of the conventional four or five, ESPs offer a more

refined and accurate representation of the original  $IV$  curve. Moreover, these 10 ESPs enable precise reconstruction of the original  $IV$  characteristic [6].

Analyzing PV outdoor data remains challenging due to continuously changing conditions, such as variations in plane-of-array irradiance ( $G_{POA}$ ) and module temperature ( $T_{Mod}$ ). Individual  $IV$  characteristics, as well as their corresponding ESPs, cannot be directly compared without accounting for the influence of irradiance and temperature on the  $IV$  curve shape. Furthermore, additional factors such as (partial) shading and humidity affect data quality and introduce uncertainties. Another significant challenge in PV outdoor data analysis is the lack of standardized filtering methods within the PV community [11]. Different filtering approaches can lead to varying results, for instance, in PV degradation rate estimation [12].

In this paper, we introduce a statistical model for analyzing PV outdoor data. The model is designed to reproduce and predict the  $IV$  characteristic as well as its standard deviation at any given time and temperature and irradiation conditions. Such the model not only outputs the temperature- and irradiation-dependent temporal development of the  $IV$  characteristic, but also sets the output in context of a standard deviation arising through the uncertainty of the measurement itself and the lack of data, that is, the measured  $IV$  characteristics are only available for distinct times at distinct conditions.

The presented model is depicted in Figure 1. The arrows depict data flow between the processing steps, while the color code and numbering reflects the structure of the paper. At the root of the model stands PV outdoor raw data, acquired by TÜV Rheinland [4]. The data consists of up to 3 years of  $IV$  characteristic data as well as  $T_{Mod}$ ,  $G_{POA}$  and the time  $t$  (up to

approximately 60,000 data points each) for various commercial modules operated in different climate zones. Note that the modules were new at the time of installation, that is, the data might be expected to show effects of acclimatization (a performance drop in the beginning of operation). The model builds upon a simple filtering concept (see Section 2.1), excluding incomplete data points as well as data where the effective irradiation on the PV module (i.e.,  $I_{SC}$ ) does not match the measured plane-of-array irradiation  $G_{POA}$ . The filtering approach ensures to keep a representative share of the available data, filtering only approximately 2% to 4% of the data. The time series of *IV* characteristics from the raw data are reduced to a 10-dimensional time series of ESPs (see Section 2.2). After the ESP analysis and the filtering routine, a Principal Component Analysis (PCA) is applied to the ESP time series (see Section 2.3), resulting in a time series of 10 linearly uncorrelated principal components (PCs). These PCs, along with time  $t$ , plane-of-array irradiance ( $G_{POA}$ ), and module temperature ( $T_{Mod}$ ), serve as inputs for training multiple Gaussian Process Regressions (GPRs), where each PC is treated independently (see Section 2.4). To manage the challenges associated with large datasets,<sup>1</sup> the PC time series are segmented into monthly subsets.

Once the individual GPRs are optimized and trained, their outputs can be combined to predict the PCs in dependence of  $t$ ,  $T_{Mod}$  and  $G_{POA}$ . The reconstruction of the ESPs using the PCA translation matrix (gray in Figure 1) results the complete shape of the *IV* characteristic for any given input  $[t, T_{Mod}, G_{POA}]$ . Note that the model is not capable respectively designed to extrapolate (e.g., forecasting). Additionally, the GPRs provide insights into the standard deviation of the predictions, which arises from both data noise and gaps in the input data. Specifically, predictions for outputs far from observed data points will have higher standard deviations compared with those close to observations.

The result of this approach is a compact yet comprehensive description of the temporal, temperature, and irradiation dependencies of the modules ESPs and, thus, *IV* characteristics. The applicability of this concept is demonstrated further analyzing exemplary outdoor datasets (see Section 3) using  $ESP(t, T_{Mod}, G_{POA})$  predictions. Furthermore, the output of the GPRs can be used for validation (see Section 4) using the  $ESP(t, T_{Mod}, G_{POA})$  predictions at the measured input conditions. Such, the from the *IV* characteristics determined ESPs are reconstructed, while test and training data are temporarily separated with a common training to test ratio of 80:20.

The presented model is exemplary applied to PV outdoor data consisting of complete *IV* characteristics alongside  $G_{POA}$  and  $T_{Mod}$ , but is in principle applicable to any kind of PV outdoor data consisting of an electrical performance measure (e.g., AC or DC power) and data capturing the conditions, the performance is measured at.

In Section 2, we provide a detailed explanation of the model concept, highlighting how it integrates the data filtering approach, ESPs, PCA, and multiple GPRs. Following this, the concept is applied to an exemplary outdoor dataset in Section 3. Finally, in Section 4, we evaluate the model's accuracy before summarizing the main results in Section 5.

## 2 | A Gaussian Process Regression IV Model

### 2.1 | Filtering

An essential step in all PV outdoor data analyses is the proper filtering of data to eliminate erroneous measurements (e.g., missing data, physically unreasonable *IV* shapes, or shading-induced errors). To achieve this, ESPs (see Section 2.2) are extracted from an *IV* dataset using the PV-CRAZE library [13]. During ESP extraction, PV-CRAZE automatically flags *IV* data with unusual properties. Specifically, PV-CRAZE flags *IV* characteristics that do not allow for reliable ESP extraction due to issues such as excessive noise or insufficient data. Additionally, nonmonotonic *IV* curves are identified and removed, including cases where the slope at  $V_{OC}$  is negative or where the maximum power point current  $I_{MPP}$  exceeds the short-circuit current  $I_{SC}$ . Furthermore, incomplete data points—those missing *IV* characteristics,  $G_{POA}$ , or  $T_{Mod}$ —are also discarded. This initial filtering step effectively removes the most obvious errors from the dataset, ensuring higher data reliability for further analysis.

Beyond filtering for obvious errors, we also apply the more nuanced filtering approach presented in the Appendix A. This method involves filtering out data points where there is a mismatch in the effective irradiation levels experienced by the PV module and the irradiation sensor, addressing the most common source of systematic errors in the datasets under consideration.

The short-circuit current ( $I_{SC}$ ) is modeled as a function of  $G_{POA}$  and  $T_{Mod}$  using multiple GPRs, randomly splitting the complete dataset into 20 subsets, similar to a bootstrap aggregating approach. We compare the overlaid prediction of the 20 GPRs with the measured  $I_{SC}$ . A high filter threshold is set so that only significant discrepancies between the expected and actual  $I_{SC}$  are filtered out. Furthermore, we use an iterative approach, resulting in an effective filter excluding roughly 2% to 4% of the data points. Details to the filtering approach can be found in the Appendix in Section A.

### 2.2 | Extended Solar Cell Parameters

For completeness and readability, we begin with a brief introduction to the extended solar cell parameters (ESPs); a more detailed discussion can be found in [6]. The ESPs comprise a set of 10 parameters that describe the shape of the *IV* characteristic of a solar cell. As the name suggests, the ESPs extend beyond the standard solar cell parameters (SSPs) of  $V_{OC}$ ,  $I_{SC}$ ,  $V_{MPP}$ , and  $I_{MPP}$ . The ESPs extend the SSPs with the slopes of the *IV* at short and open circuit ( $G_{SC}$  and  $R_{OC}$ ). Savitzki–Golay filters are used to obtain noise robust estimators for the slopes at short and open circuit [14]. Furthermore, the ESPs add two additional key points (“upper quasi maximum power point” [ $I_{qmp+}, V_{qmp+}$ ] and “lower quasi maximum power point” [ $I_{qmp-}, V_{qmp-}$ ]) to comprehensively describe the complete shape of the *IV* characteristic.

The lower and upper quasi-maximum power points incorporate information on the shape of the *IV* characteristic between short circuit (SC) and maximum power point (MPP) and between MPP

and open circuit (OC), respectively. These points are defined using the power curve ( $P(V)$ ) associated with the  $IV$ .  $V_{qmp+}$  is defined as the voltage between MPP and OC, where the power curve exhibits the largest difference from a straight line drawn between the two coordinates [ $V_{MPP}, P_{MPP}$ ] and [ $V_{OC}, 0$ ].  $I_{qmp+}$  is then the associated current of the  $IV$  at  $V = V_{qmp+}$ . Analogously,  $V_{qmp-}$  is defined as the voltage, where the power curve exhibits the largest difference from a straight line drawn between the two coordinates [0,  $I_{SC}$ ] and [ $V_{MPP}, P_{MPP}$ ] and  $I_{qmp-}$  is the associated current at  $V = V_{qmp-}$ . Thus, the ESPs provide a general-purpose parameterization of  $IV$  characteristics using 10 parameters. Compared with other parameterization methods, ESPs result in less information loss, as they capture more subtle features of the  $IV$  curve. By reconstructing  $IV$  characteristics, it was demonstrated on a large dataset of 2.2 million  $IV$  curves-covering various PV technologies-that ESPs provide an accurate representation of the  $IV$  characteristics. For 90% of the  $IV$  curves, the root mean square (RMS) error relative to the original measurements remains below 0.2% [6]. In comparison, the one-diode model exhibits an RMS error of approximately 1%, while the Karmalkar-Hanefea model [15] yields an RMS error of about 1.2% [6].

### 2.3 | Principal Component Analysis

After fitting the ESPs and applying the filter, we obtain a time series of 10 ESPs that vary along with  $G_{POA}$  and  $T_{Mod}$ . As these time series are generally correlated, we further apply a principal component analysis (PCA) to extract possible linear correlations. To this end, the ESPs are normalized by their respective mean values. A PCA involves a change in the basis of the coordinate system used to describe a set of  $n$ -dimensional data points. Illustratively, this basis change aligns the first principal component (PC) with the direction of the largest variation in the  $m$ -dimensional point cloud of the dataset. Subsequent PCs represent directions with the largest variation in the  $m$ -dimensional point cloud, subject to the constraint of being orthogonal to all previous PCs.

Formally, the  $j$ th PC can be expressed as

$$PC_j = \sum_{k=1}^{10} p_{j,k} \frac{ESP_k - \bar{ESP}_k}{\|ESP_k\|} \quad (1)$$

where the matrix elements  $p_{j,k}$  of the transformation matrix  $\mathbb{P}$  serve as the weights that translate the mean-normalized 10 ESPs into the 10 PCs. The primary advantage of using PCs to describe a dataset is the elimination of redundant information in the form of linear correlations. Additionally, a PCA can reduce the dimensionality of the dataset by revealing whether it can be adequately described with  $d < m$  principal components. However, it is important to note that in this context, the PCA is not necessarily applied to reduce the data's dimensionality but rather to decorrelate the ESP data.

### 2.4 | Mixture of Experts Gaussian Process Regression

After applying the PCA, we obtain 10 PC time series, which are treated separately in the subsequent analysis. To describe the

temporal,  $G_{POA}$ , and  $T_{Mod}$  dependencies of each  $PC_j$ , we use multiple Gaussian Process Regressions (GPRs). A GPR is a probabilistic method that allows for predictions of unsampled inputs, assuming a Gaussian distribution of measurements. Note that this assumption applies to a single measurement, which is assumed to follow a Gaussian distribution if repeated under identical conditions and at the same time. However, we do not assume that the modeled time series itself is Gaussian distributed. We utilize GPRs to interpolate desired outputs, for example, a time series of  $IV$  characteristics under constant  $G_{POA}$  and  $T_{Mod}$  conditions, and to estimate the associated standard deviation based on the available discrete data. That is, the model is designed to predict the  $IV$  characteristic for given inputs of  $t$ ,  $G_{POA}$ , and  $T_{Mod}$  within the parameter space defined by the observations. It is not intended for extrapolation or forecasting.

Each  $PC_j$  time series contains up to  $N \approx 65,000$  data points. To mitigate the computational burden, which scales with  $N^3$  for a single GPR, we adopt a “mixture of experts” approach [16]. We split each  $PC_j$  time series into  $n$  monthly subsets, indexed by  $i$ , and train individual GPRs  $f_{PC_{j,i}}: \mathbb{R}^3 \rightarrow \mathbb{R}$ . Each  $f_{PC_{j,i}}$  is optimized using five hyperparameters within a 3D radial basis function (RBF) kernel, which is multiplied by a constant kernel and supplemented by a white noise kernel. The kernel of each  $f_{PC_{j,i}}$  is defined as

$$K(X, \tilde{X}) = c \exp\left(\frac{\|x_1 - \tilde{x}_1\|^2}{2l_1^2} + \frac{\|x_2 - \tilde{x}_2\|^2}{2l_2^2} + \frac{\|x_3 - \tilde{x}_3\|^2}{2l_3^2}\right) + \sigma^2 \quad (2)$$

where  $X = [x_1, x_2, x_3] = [G_{POA}, T_{Mod}, t]$  represents the 3D input space of irradiation, temperature, and time,  $\mathbb{I}$  is the identity matrix, and  $c$ ,  $l_1$ ,  $l_2$ ,  $l_3$ , and  $\sigma$  are the five hyperparameters optimized during training. The use of an RBF kernel assumes that each  $PC_j$  varies smoothly with time, temperature, and irradiation. The most influential hyperparameters for the accuracy of GPR predictions are the lengthscale parameters  $l_1$ ,  $l_2$ , and  $l_3$ , which define the correlation length in the directions of time  $t$ , irradiation  $G_{POA}$ , and temperature  $T_{Mod}$ . By training separate GPRs for each month, the model allows for temporal variation in these parameters, thereby adjusting the sensitivity to small changes over time. The hyperparameters are optimized using the limited-memory BFGS algorithm with parameter constraints (L-BFGS-B), as implemented in SciPy [17]. To improve reliability, the optimization routine is executed five times with randomly initialized start parameters. The parameter constraints ensure that the hyperparameter values remain within a meaningful range, preventing unrealistic variations.<sup>2</sup>

After training the individual  $10n$  GPRs  $f_{PC_{j,i}}: \mathbb{R}^3 \rightarrow \mathbb{R}$ , the models can predict  $PC_{j,i}(G_{POA}, T_{Mod}, t)$  and the associated standard deviation  $\sigma_{PC_{j,i}}(G_{POA}, T_{Mod}, t)$ . Note that the standard deviation arises from both data noise and the distance from observed data points. In addition to providing predictions and standard deviations, the model outputs can be interpreted as probability densities, with each GPR prediction corresponding to a Gaussian distribution  $\mathcal{N}(PC_{j,i}(G_{POA}, T_{Mod}, t), \sigma_{PC_{j,i}}(G_{POA}, T_{Mod}, t))$ .

The ESPs can be derived from the PC predictions using the matrix elements  $\hat{p}_{k,j}$  of the inverse transformation matrix  $\mathbb{P}^{-1}$  as follows:

$$\text{ESP}k_i(G_{\text{POA}}, T_{\text{Mod}}, t) = \overline{\text{ESP}k} \left( 1 + \sum_{j=1}^{10} \hat{p}_{k,j} \text{PC}j_i(G_{\text{POA}}, T_{\text{Mod}}, t) \right) \quad (3)$$

This allows the model to predict the complete *IV* characteristic at any given  $t$ ,  $G_{\text{POA}}$ , and  $T_{\text{Mod}}$ . The standard deviation of the ESP predictions can be calculated from  $\sigma_{\text{PC}j_i}(G_{\text{POA}}, T_{\text{Mod}}, t)$  using the following equation:

$$\sigma_{\text{ESP}k_i}(G_{\text{POA}}, T_{\text{Mod}}, t) = \overline{\text{ESP}k} \sqrt{\sum_{j=1}^{10} \hat{p}_{k,j}^2 \sigma_{\text{PC}j_i}^2(G_{\text{POA}}, T_{\text{Mod}}, t)} \quad (4)$$

where we assume that the  $j = 1, 2, \dots, 10$  predictions  $\text{PC}j_i(G_{\text{POA}}, T_{\text{Mod}}, t)$  are statistically independent, that is, the covariance matrix elements  $\sigma_{\text{PC}j_i, \text{PC}l_i}$  are zero for  $j \neq l$ . Given that the PCs are linearly uncorrelated (by definition of the PCA) and that each GPR uses individual hyperparameters for predictions, this assumption is a reasonable approximation. Furthermore, because linear transformations preserve the Gaussian property, the ESP predictions can also be interpreted as Gaussian probability densities  $\mathcal{N}(\text{ESP}k_i(G_{\text{POA}}, T_{\text{Mod}}, t), \sigma_{\text{ESP}k_i}(G_{\text{POA}}, T_{\text{Mod}}, t))$ .

The model's accuracy is highest for test points  $[G_{\text{POA}}, T_{\text{Mod}}, t]$  close to observed data, and the model is not intended for extrapolation. Within these limitations, the model is capable of extracting, for example, temperature or irradiation dependencies at a specific time, or the temporal development of the *IV* characteristic under constant conditions. By splitting the dataset into  $i = 1, 2, \dots, n$  monthly subsets, each  $f_{\text{PC}j_i}$  becomes an "expert" for the respective month. As the temporal distance from observations increases, the prediction  $\text{PC}j_i(G_{\text{POA}}, T_{\text{Mod}}, t)$  and, consequently,  $\text{ESP}k_i(G_{\text{POA}}, T_{\text{Mod}}, t)$  exhibit increased standard deviations.

### 3 | Results

The presented concept is applied to all 45 datasets provided by TÜV Rheinland. Each dataset consists of up to 3 years of *IV* characteristic data, along with corresponding measurements of  $T_{\text{Mod}}$ ,  $G_{\text{POA}}$ , and time  $t$ , with up to approximately 60,000 data points per parameter for various commercial PV modules. The studied modules include an amorphous/microcrystalline silicon tandem module, a conventional PERC silicon module, three CdTe modules, and four CIGS modules from different manufacturers.

The same module types from all nine technologies and manufacturers are deployed in Ancona (Italy), Phoenix (USA), Cologne (Germany), Chennai (India), and Thuwal (Saudi Arabia). Depending on the module type and location, the operational period varies between 17 and 35 months, resulting in dataset sizes ranging from approximately 29,000 to 65,000 *IV* characteristics. The *IV* characteristics are measured using programmable loads at 10-min intervals.

The module temperature  $T_{\text{Mod}}$  is determined as the average of two measurements taken with Pt100 temperature sensors mounted on the back of each PV module. The plane-of-array

irradiance  $G_{\text{POA}}$  is measured using a ventilated pyranometer. More details on the datasets can be found in [4].

The data was collected approximately 10 years ago and therefore reflects the PV technologies that were commercially available at that time. The focus of this paper is on the presented model rather than on the analysis of degradation features in the specifically used PV modules. For simplicity, we use an exemplary dataset of a CIGS module (CIGS4) operated for approximately 2.8 years (approximately 60,000 data points), beginning on the 1st of November 2013, in Ancona, Italy, to provide a compact overview of possible applications of the presented GPR *IV* model concept. We provide in the following two examples of how to utilize the presented concept to analyze PV outdoor data, concentrating on the temporal development of the shape of the complete *IV* characteristic and the determination of the performance ratio (PR) in comparison to a classical temperature-corrected performance ratio  $\text{PR}_T$ . A third example of utilizing the models output for a physical analysis using the one-diode model can be found in the Appendix in Section B.

#### 3.1 | Timeseries Analysis

Figure 2 shows one possible output of the GPR prediction, a timeseries of the 10 ESPs (and with that a representation of the complete *IV* characteristic) over the complete operation time of approximately 2.8 years at constant  $T_{\text{Mod}} = 40^\circ\text{C}$  and  $G_{\text{POA}} = 750 \text{ Wm}^{-2}$  for the exemplary chosen Italy CIGS4 dataset. Note that the constant conditions are chosen to represent medium-high irradiation with a realistic module temperature at such irradiation levels. Compared with typical nominal operating cell temperature (NOCT) at  $G_{\text{POA}} = 800 \text{ Wm}^{-2}$  and  $T_{\text{amb}} = 20^\circ\text{C}$  being in the range of 40 to 50°C [18].

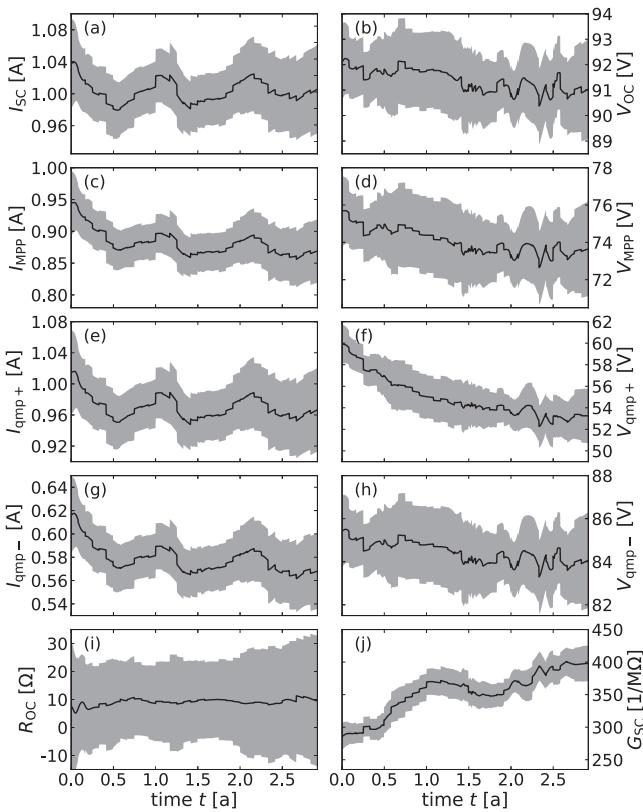
The output shown in Figure 2 is produced with a 5-month sliding window approach, that is, we use the output  $\text{PC}j_i(t)$  (constant  $T_{\text{Mod}}$  and  $G_{\text{POA}}$ ) of  $f_{\text{PC}j_i}$  for  $i = m-2, m-1, m, m+1, m+2$  to predict  $\text{PC}j(t)$  for times  $t$  in month  $m$  using

$$\text{PC}j(t) = \sigma_{\text{PC}j}^2(t) \sum_{i=m-2}^{m+2} \frac{\text{PC}j_i(t)}{\sigma_{\text{PC}j_i}^2(t)} \quad (5)$$

where the standard deviation of the prediction  $\sigma_{\text{PC}j(t)}$  is determined from the standard deviation outputs of the five GPR predictions according to

$$\sigma_{\text{PC}j}(t) = \sqrt{\frac{1}{\sum_{i=m-2}^{m+2} \frac{1}{\sigma_{\text{PC}j_i}^2(t)}}} \quad (6)$$

and the prediction of the five GPRs are weighted with the respective standard deviation output for the predictions.<sup>3</sup> The window size of five months is chosen to suppress the influence of observations with a temporal distance of more than three months, that is, 3 months are used as a threshold for the temporal correlation. The time series of ESPs, including the standard deviation (i.e.,  $\text{ESP}k(t)$  and  $\sigma_{\text{ESP}k}(t)$ ), are determined from  $\text{PC}j(t)$  and  $\sigma_{\text{PC}j}(t)$  according to



**FIGURE 2** | The output of the GPR prediction for a timeseries of the 10 reconstructed ESPs over the complete operation time of approximately 2.8 years at constant  $T_{\text{Mod}} = 40^\circ\text{C}$  and  $G_{\text{POA}} = 750 \text{ Wm}^{-2}$  for the exemplary chosen Italy CIGS4 dataset. For the prediction, a 5-month sliding window approach is chosen, where  $\text{ESPk}(t)$  and  $\sigma_{\text{ESPk}}(t)$  are determined for times  $t$  in the  $m$ th month according to (3) and (4) from  $\text{PC}_j(t)$  and  $\sigma_{\text{PC}_j}(t)$ . The 95 % confidence interval for every single prediction is denoted with the gray area.

(3) and (4). The 95 % confidence interval in Figure 2 is defined for every  $t$  as  $[\text{ESPk}(t) - 1.96\sigma_{\text{ESPk}}(t), \text{ESPk}(t) + 1.96\sigma_{\text{ESPk}}(t)]$  and denoted with the gray area. Note that the 95 % confidence interval is defined for each individual prediction and not for the complete timeseries. Because for each prediction of the ESPs the predicted probability distribution is Gaussian, where the 95 % confidence is equal to a  $\pm 1.96\sigma$  environment.

A clear seasonality superimposed with a slight linear decrease is visible for the four current parameters  $I_{\text{SC}}$ ,  $I_{\text{MPP}}$ ,  $I_{\text{qmp}_+}$  and  $I_{\text{qmp}_-}$  (Figure 2a,c,e,g). For the four voltage parameters  $V_{\text{OC}}$ ,  $V_{\text{MPP}}$ ,  $V_{\text{qmp}_+}$  and  $V_{\text{qmp}_-}$  (Figure 2b,d,f,h) the temporal development shows a more or less linear decrease, where the slope of  $V_{\text{qmp}_+}$  is observed to flatten out. The slope of the IV characteristic at open circuit is described by the parameter  $R_{\text{OC}}$  (Figure 2i), which is observed to be constant over the operation period with a high standard deviation of the prediction.<sup>4</sup> For the slope of the IV characteristic at short circuit, depicted by the parameter  $G_{\text{SC}}$  (Figure 2j), one can observe a steady increase superimposed with a periodic change over the course of the operation time.

The representation of the complete shape of the IV characteristic in terms of the ESPs enables to quantify the temporal development of the PV modules performance. First of all, it is clear that a strong

seasonality affects the current level of the IV characteristic, while a slight degradation is visible in both the current and voltage parameters. The difference in the development of  $I_{\text{SC}}$  and  $I_{\text{MPP}}$  already indicates a loss in the FF over time, as the degradation appears to be more pronounced in  $I_{\text{MPP}}$ . Considering the temporal development of  $G_{\text{SC}}$  as well as of  $V_{\text{qmp}_+}$  one can follow that the slope at SC as well as the complete shape of the IV between SC and MPP changes, while the shape of the IV between MPP and OC remains quantitatively the same (rather a constant  $R_{\text{OC}}$  as well as similar temporal development of  $V_{\text{OC}}$ ,  $V_{\text{MPP}}$ , and  $V_{\text{qmp}_-}$ ).

Regarding the uncertainty estimate of the ESPs we find the highest confidence (lowest relative standard deviation) for the temporal development of the parameter  $G_{\text{SC}}$  (Figure 2j). The uncertainty on the temporal development of the four current parameters (Figure 2a,c,e,g) as well as on the temporal development of  $V_{\text{qmp}_+}$  (Figure 2f) is comparably higher (higher relative standard deviations). Nonetheless, one can clearly observe high confidence for the qualitative temporal development, that is, the amplitude of seasonality and linear degradation is comparable to the amplitude of the 95 % confidence interval. For the three voltage parameters  $V_{\text{OC}}$ ,  $V_{\text{MPP}}$  and  $V_{\text{qmp}_-}$  (Figure 2b,d,h) and especially for the parameter  $R_{\text{OC}}$  (Figure 2i), we find a comparably high standard deviations on the temporal development, where the amplitude of the signal (temporal variation of the prediction) is lower than the respective confidence interval.

### 3.2 | Performance Ratio Analysis

A common way to verify a PV module's performance over time is the performance ratio (PR) respectively the temperature-corrected performance ratio ( $\text{PR}_T$ ). For the AC/DC output performance, the PR is given by

$$\text{PR}_{\text{AC/DC}} = \frac{E_{\text{AC/DC}}/P_{\text{nom}}}{H_{\text{POA}}/G_{\text{STC}}} = \frac{E_{\text{AC/DC}}G_{\text{STC}}}{H_{\text{POA}}P_{\text{nom}}} \quad (7)$$

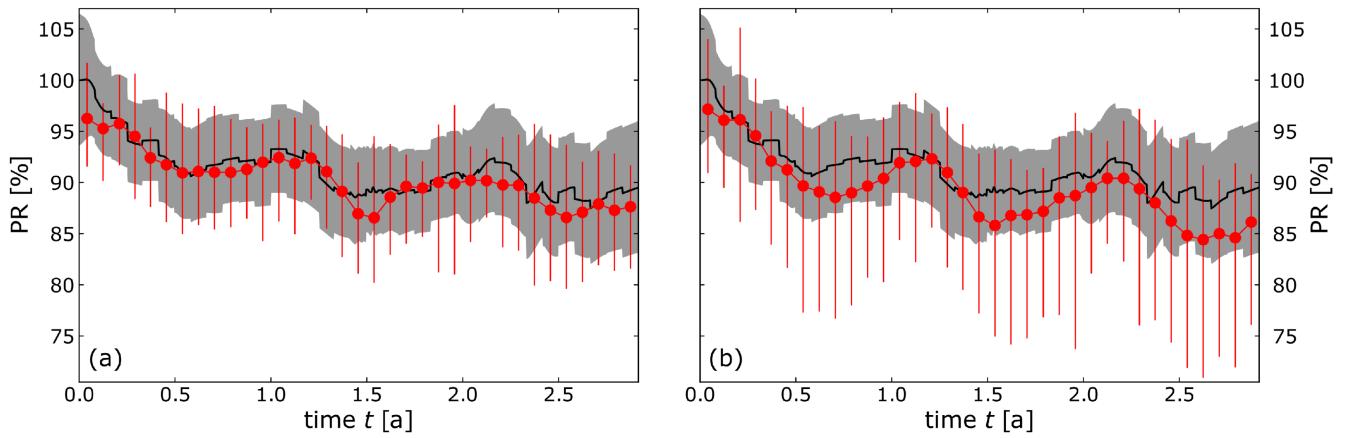
where  $P_{\text{nom}}$  is the nominal AC respectively DC output power at STC conditions,  $E_{\text{AC/DC}}$  is the integrated AC/DC energy yield,  $G_{\text{STC}} = 1000 \text{ Wm}^{-2}$  is the STC irradiation and  $H_{\text{POA}}$  is the normalized integrated in-plane irradiation for the considered time span. The PR is defined on a time interval, but is often generalized for the DC output power to a discrete time using

$$\text{PR} = \frac{P_{\text{DC}}G_{\text{STC}}}{G_{\text{POA}}P_{\text{STC}}} \quad (8)$$

where we substituted  $P_{\text{nom}} = P_{\text{STC}}$  for the DC situation. The performance of a PV module and, thus, the performance of a string of modules or a PV system are dependent on the module temperature  $T_{\text{Mod}}$ . The temperature-corrected performance ratio  $\text{PR}_T$  includes the temperature dependency of the output power using

$$\text{PR}_T = \frac{P_{\text{DC}}G_{\text{STC}}}{G_{\text{POA}}P_{\text{STC}}(1 + \gamma(T_{\text{Mod}} - T_{\text{STC}}))} \quad (9)$$

As can be seen in (8) and (9), PR and  $\text{PR}_T$  are referenced to STC conditions. Because STC conditions are uncommon in real operation scenarios and GPRs are not able to extrapolate



**FIGURE 3** | GPR performance ratio  $PR_{GPR}(t)$  (black) prediction for the constant  $G_{POA} = 750 \text{ Wm}^{-2}$  and  $T_{Mod} = 40^\circ\text{C}$  and classical  $PR_T$  estimate (red) using (a) a narrow irradiation band filter of  $400 \text{ Wm}^{-2}$  and (b) a broad irradiation band filter of  $800 \text{ Wm}^{-2}$  referenced to  $G_{POA} = 750 \text{ Wm}^{-2}$  and  $T_{Mod} = 40^\circ\text{C}$ . The gray area and the red error bars depict a 95 % confidence determined from the numerical approximation to the theoretical probability distribution of  $PR_{GPR}(t)$  for each  $t$  and the distribution of the determined  $PR_T$ , respectively.

to inputs far away from the discrete inputs they are trained on, a model's output like shown in Figure 2 is not suitable to evaluate on classical STC-referenced PR analysis. However, the PR for distinct constant conditions can be computed by comparing the MPP power output  $P_{MPP,GPR}(t) = V_{MPP,GPR}(t)I_{MPP,GPR}(t)$  to the initial  $P_{MPP,GPR}(t=0)$  at the same conditions. As there is no producer information on the nominal output power at the respective conditions available and taking only one single value as reference ( $P_{ref} = P_{MPP,GPR}(t=0)$ ) might result in high deviations when  $P_{MPP,GPR}(t=0)$  is not accurate, we define the initial  $P_{MPP,GPR}$  as the mean of  $P_{MPP,GPR}(t)$  in the first 7 days of operation  $\overline{P_{MPP,GPR}|_{t \leq 7d}}$ .<sup>5</sup> Thus, we define a generalized PR from the GPR IV model at constant  $G_{POA}$  and  $T_{Mod}$  conditions as

$$PR_{GPR}(t) = \frac{P_{MPP,GPR}(t)}{\overline{P_{MPP,GPR}|_{t \leq 7d}}} = \frac{V_{MPP,GPR}(t)I_{MPP,GPR}(t)}{\overline{V_{MPP,GPR}|_{t \leq 7d}}\overline{I_{MPP,GPR}|_{t \leq 7d}}}. \quad (10)$$

Regarding the standard deviation estimate of such defined  $PR_{GPR}$ , one has to consider the standard deviation of  $V_{MPP,GPR}$  and  $I_{MPP,GPR}$ . As already discussed in Section 2.4, the predictions of the ESPs can be interpreted as a Gaussian probability density  $\mathcal{N}(ESP_k(G_{POA}, T_{Mod}, t), \sigma_{ESP_k}(G_{POA}, T_{Mod}, t))$ . For the standard deviation prediction of  $PR_{GPR}(t)$  the correlation between the two probability densities of  $V_{MPP,GPR}(t)$  and  $I_{MPP,GPR}(t)$  needs to be considered.<sup>6</sup> Because all ESPs are a linear combination of the PCs, we can generalize 4 to determine the covariance between two ESPs via the matrix elements  $\hat{p}_{k,j}$  of the inverse translation matrix  $\mathbb{P}^{-1}$  and the matrix elements  $\sigma_{PC_j}^2$  of the diagonal covariance matrix of the PCs. Formally, we compute the covariance between  $ESP_k(t)$  and  $ESP_l(t)$  given by

$$\sigma_{ESP_k,ESP_l}(t) = \overline{ESP_k ESP_l} \sum_{j=1}^{10} \hat{p}_{k,j} \sigma_{PC_j}^2(t) \hat{p}_{l,j}, \quad (11)$$

where  $\hat{p}_{l,j} = \hat{p}_{j,l}^\top$  are the matrix elements of the transposed matrix  $\mathbb{P}^{-1\top}$ . With the covariance of  $V_{MPP,GPR}(t)$  and  $I_{MPP,GPR}(t)$  the standard deviation  $\sigma_{PR_{GPR}}(t)$  can be computed to

$$\begin{aligned} \sigma_{PR_{GPR}}(t) &= \left[ \left( \frac{\sigma_{V_{MPP,GPR}}(t)I_{MPP,GPR}(t)}{\overline{V_{MPP,GPR}|_{t \leq 7d}}\overline{I_{MPP,GPR}|_{t \leq 7d}}} \right)^2 \right. \\ &\quad + \left( \frac{V_{MPP,GPR}(t)\sigma_{I_{MPP,GPR}}(t)}{\overline{V_{MPP,GPR}|_{t \leq 7d}}\overline{I_{MPP,GPR}|_{t \leq 7d}}} \right)^2 \\ &\quad \left. + \frac{2V_{MPP,GPR}(t)I_{MPP,GPR}(t)\sigma_{I_{MPP,GPR},V_{MPP,GPR}}(t)}{(\overline{V_{MPP,GPR}|_{t \leq 7d}}\overline{I_{MPP,GPR}|_{t \leq 7d}})^2} \right]^{\frac{1}{2}}. \end{aligned} \quad (12)$$

While the standard deviation of the probability density for  $PR_{GPR}(t)$  can be computed, defining a confidence interval similar to the confidence interval given in Figure 2 directly from  $\sigma_{PR_{GPR}}(t)$  might not be accurate, as the probability density of a product of two correlated Gaussian distributed random variables is in general not Gaussian. While there is an expression for an exact solution of the probability density (compared with [19]), the computation of the exact solution involves an infinite sum of modified Bessel functions and is not applicable without an approximation (e.g., setting a limit for the infinite sum). To compute a confidence interval for  $PR_{GPR}(t)$ , we use an alternative approach approximating the probability density with a numerical solution. We generate  $n_r = 10^6$  random distributed variables according to the 2D multivariate Gaussian distributed probability density  $(\mathcal{N}(I_{MPP,GPR}(t), \sigma_{I_{MPP,GPR}}(t)), \mathcal{N}(V_{MPP,GPR}(t), \sigma_{V_{MPP,GPR}}(t)))$  with covariance  $\sigma_{I_{MPP,GPR},V_{MPP,GPR}}(t)$  and compute the probability density of  $PR_{GPR}(t)$  for each  $t$ . The probability density computed this way approaches the exact solution with  $n_r \rightarrow \infty$ .

Figure 3a shows the temporal development of the  $PR_{GPR}$  (black, including 95% confidence interval for each  $t$  in gray) for the constant  $G_{POA} = 750 \text{ Wm}^{-2}$  and  $T_{Mod} = 40^\circ\text{C}$ . In Figure 3, the  $PR_{GPR}$  is further compared with a classical determined monthly temperature-corrected performance ratio  $PR_T$  (red, including 95% confidence) referenced to  $G_{POA,\text{ref}} = 750 \text{ Wm}^{-2}$  and  $T_{Mod,\text{ref}} = 40^\circ\text{C}$ . For the determination of the classical  $PR_T$ , we use 9 and substitute the STC with the reference conditions, insert  $P_{DC} = P_{MPP,\text{meas}}$ , and further use  $\overline{P_{MPP}|_{t \leq 7d}}$  from the GPR

prediction as reference for the initial output power at the reference conditions. This results in the expression

$$\text{PR}_T = \frac{750 \text{ Wm}^{-2}}{G_{\text{POA,meas}}} \frac{P_{\text{MPP,meas}}}{\overline{P_{\text{MPP,GPR}}|_{t \leq 7d}}(1 + \gamma(T_{\text{Mod,meas}} - 40^\circ\text{C}))} \quad (13)$$

where the temperature correction coefficient  $\gamma$  is determined by a linear fit to the filtered dataset. The shown temperature-corrected PR is a commonly used monthly  $\text{PR}_T$ , where all measurements in a narrow irradiation band  $550 \text{ Wm}^{-2} < G_{\text{POA}} < 950 \text{ Wm}^{-2}$  around the reference irradiation  $G_{\text{POA}} = 750 \text{ Wm}^{-2}$  are used for the determination of the average  $\text{PR}_T$  and its 95% confidence interval.<sup>7</sup> Thus, the 95% confidence interval is defined as the interval between the boundaries set by the 2.5% and 97.5% quantile of the distribution of determined  $\text{PR}_T$  in the respective month.

One can observe a strong acclimatization effect within the first few months of operation overlaid with a seasonality and a long-term degradation for both PR estimates. The  $\text{PR}_T$ , determined with the classical monthly approach, matches qualitatively the  $\text{PR}_{\text{GPR}}$ , while a slightly higher standard deviation can be observed for the classical approach. Furthermore, we observe for some months  $\text{PR}_T$  values several percent below the  $\text{PR}_{\text{GPR}}$ , remaining within the 95% confidence interval of the GPR prediction. While the GPR prediction uses all available information of the performance measures at different conditions within one month,<sup>8</sup> the narrow irradiation band filter reduces the dataset to  $\approx 22\%$  and such to a share of the dataset that can not be expected to be representative of the complete dataset.

Despite reducing the dataset to a small share of itself, these narrow irradiation band filters are a common choice while determining the PR respectively  $\text{PR}_T$  of PV modules and systems. Besides the effects of the low-light performance of PV modules gaining more influence on the result, for the determination of the PR a broader irradiation band filter would further increase the influence of the operating temperature, highly correlated with the  $G_{\text{POA}}$ . Such a PR estimate would experience an increased standard deviation. For the  $\text{PR}_T$ , the effect of the operating temperature is reduced due to the linear correction coefficient, but can still affect the result in the presence of non-linear temperature effects. Figure 3b underlines the increasing standard deviation of the  $\text{PR}_T$  with increasing width of the irradiation band filter. Here,  $\text{PR}_{\text{GPR}}$  is compared with the  $\text{PR}_T$  calculated from a filtered dataset with a doubled width of the irradiation band filter, averaging  $P_{\text{MPP}}$  for all measurements for  $G_{\text{POA}}$  with  $350 \text{ Wm}^{-2} < G_{\text{POA}} < 1150 \text{ Wm}^{-2}$ . Note that despite the choice of a broad irradiation band filter, the dataset is reduced to  $\approx 42\%$  in this case.

From the comparison of Figure 3a,b, we find, as expected, an increased standard deviation for the  $\text{PR}_T$  estimate with a broader irradiation band filter. Furthermore, one can observe clear deviations of  $\text{PR}_{\text{GPR}}(t)$  and  $\text{PR}_T$ , where  $\text{PR}_T$  is found to be reduced in the summer months<sup>9</sup> and this way in operation at higher ambient and resulting in higher operation temperatures. This effect of temperature in the time series of  $\text{PR}_T$  shows one major downfall of the classical PR approach. Equation (9) only accounts for temperature and irradiation dependencies

in terms of linear dependencies using a single correction coefficient. In comparison, for example, the IEC60891 norm commonly used for temperature and irradiation corrections of IV characteristics [20] accounts for 6 correction coefficients. Furthermore, while observing degradation mechanisms in PV modules it is expected, that correction coefficients might also change over time, leading to the increased inaccuracy of the classical  $\text{PR}_T$  method.

On the other hand, the PR analysis using GPR prediction allows for arbitrary temperature and irradiation dependencies in the data, allowing the dependencies further to change over time. Furthermore, the GPR PR analysis does not require to exclude a large part of the data, as the influence of each data point is weighted according to the respective distance in temperature, irradiation and time. Therefore, we argue the GPR PR analysis is a more accurate alternative to classical PR analysis techniques.

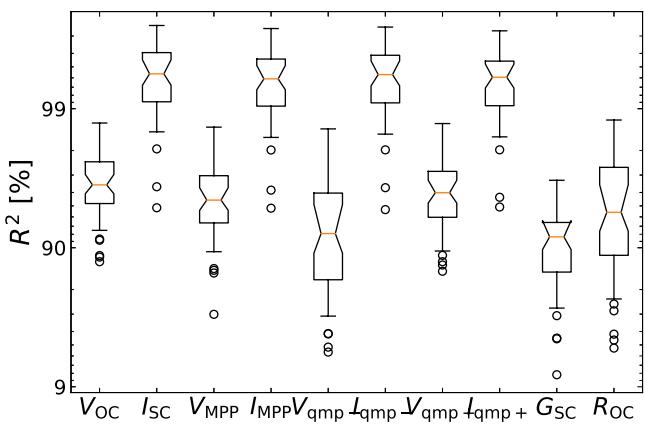
## 4 | Validation

With the two examples shown, we have introduced some possible outputs of the GPR IV model without addressing its accuracy. To this end, we use the trained GPRs  $f_{\text{PC}_j}$  to validate their performance regarding recreating the training data (observations). Because each GPR is inherited to recreate its own input with an accuracy of the estimated noise level, validating a GPR on its own training input has little to no information on the performance. For validation, there needs to be a clear separation of training and test data, where no test data are used for training the model. To further ensure a temporal separation of training and test data, we compare the input of the  $\text{PC}_j(t)$  ( $\text{ESP}_j(t)$ ), that is, the observations in the  $m$ th month, with the overlaid output of  $f_{\text{PC}_j}$  for  $i = m - 2$ ,  $i = m - 1$ ,  $i = m + 1$  and  $i = m + 2$ , that is, the predictions of the four GPRs trained with the data of the two months before and after the month of consideration. Formally, we compute

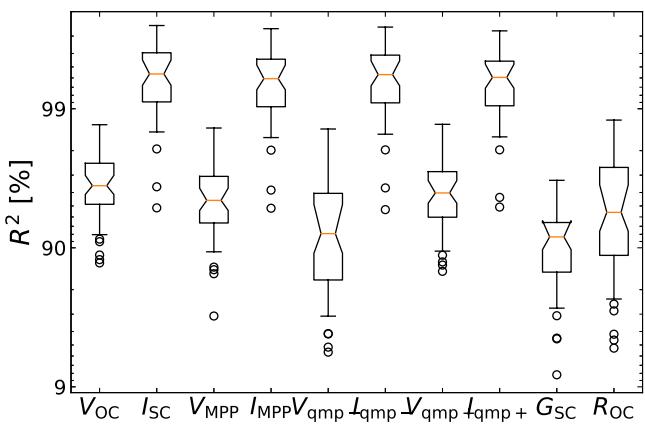
$$\text{PC}_{j,\text{val},m}(G_{\text{POA}}, T_{\text{Mod}}, t) = \sigma_{\text{PC}_j}^2(G_{\text{POA}}, T_{\text{Mod}}, t) \sum_{\substack{i=m-2, m-1, \\ m+1, m+2}} \frac{\text{PC}_j(G_{\text{POA}}, T_{\text{Mod}}, t)}{\sigma_{\text{PC}_j}^2(G_{\text{POA}}, T_{\text{Mod}}, t)}, \quad (14)$$

for all distinct observations in  $[G_{\text{POA}}, T_{\text{Mod}}, t]$  for  $t$  in month  $m$  and determine  $\text{ESP}_{j,\text{val},m}(G_{\text{POA}}, T_{\text{Mod}}, t)$  analogously to 3. This way, we use four instead of five months (w.r.t. the predictions shown in Section 3) for training and the fifth month as test data repeating the procedure for each month of the complete dataset. This results in the commonly used training-to-test ratio of approximately 80:20. Due to the reduced input for validating the prediction for the first and last month (67:33 ratio) as well as for the second and second to last month (75:25 ratio) of each dataset, the actual training to test ratio is a bit lower than the 80:20 ratio. Note that due to the sliding window approach, the validation of the model does not result in a high additional computation time, as the already trained GPRs respectively optimized hyperparameters are used for the predictions.

We compute the difference of prediction output ( $\text{ESP}_{j,\text{val},m}(G_{\text{POA}}, T_{\text{Mod}}, t)$ ) and the test observations



**FIGURE 4** | GPR performance validation in terms of the  $R^2$  values of each ESP prediction, reconstructed from 10 PCs. The PC prediction is computed according to (14).



**FIGURE 5** | GPR performance validation in terms of the  $R^2$  values of each ESP prediction, reconstructed from seven PCs. The PC prediction is computed according to (14).

( $\text{ESPk}_{test}(G_{\text{POA}}, T_{\text{Mod}}, t)$ ) for each month  $m$  of the dataset and set the performance of the models prediction in comparison to the difference of observations to their mean using

$$R^2_{\text{ESPk}} = 1 - \frac{\sum (\text{ESPk}_{test} - \overline{\text{ESPk}}_{test})^2}{\sum (\text{ESPk}_{test} - \text{ESPk}_{val})^2} \quad (15)$$

Illustratively, a  $R^2 = 0$  translates to the prediction having the same RMSE of a model that would predict a constant ESP for all measurements independent of  $G_{\text{POA}}$ ,  $T_{\text{Mod}}$  and  $t$ . The maximum possible value for  $R^2$  is one (100%), translating to a perfect prediction with no deviation for all test data points.

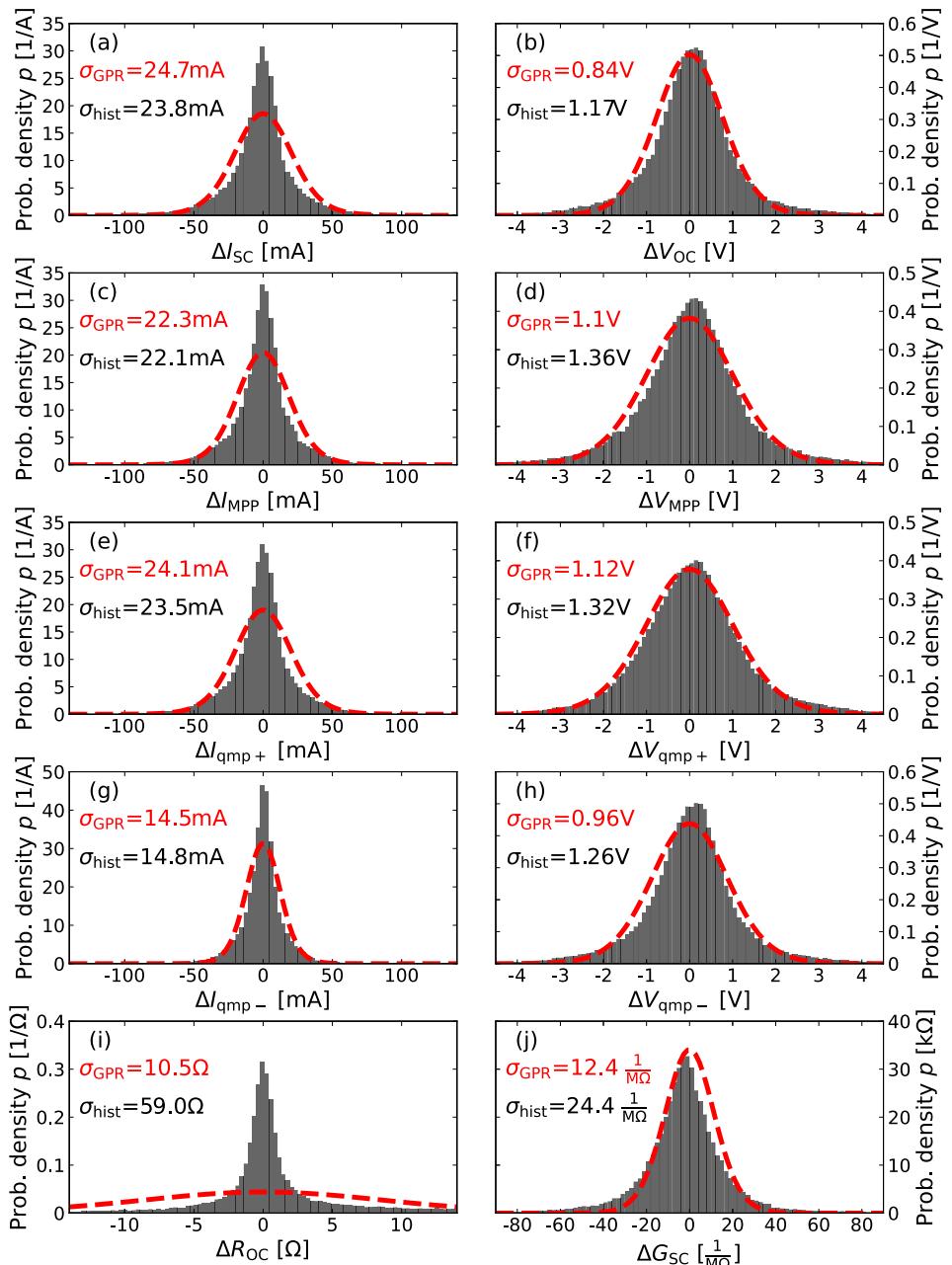
For the 10 ESPs, we find median  $R^2$  values between 91.7% and 99.4% among all 45 datasets (five locations, nine modules). Figure 4 shows a boxplot of the  $R^2$  values of each ESP, reconstructed from 10 PCs. The orange line denotes the median  $R^2$ , the box denotes the range of the first to third quartile, the whiskers extend from the box to the farthest data point lying within 1.5 times the inter-quartile range (IQR, the distance between the first and third quartile) and the single points denote individual  $R^2$  values

outside the described ranges. Note that the logarithmic scale of the y-axis is chosen such that it captures the relevant range of  $R^2$  values with increasing resolution toward most often observed high  $R^2$  values close to 100 %. For the four current parameters, we find high  $R^2$  values among all 45 datasets, with only a few outliers below 98% and the majority above 99%. The three voltage parameters  $V_{OC}$ ,  $V_{MPP}$  and  $V_{qmp+}$  are predicted with median  $R^2$  values around 96%. The median accuracy of  $V_{qmp}$ ,  $R_{OC}$  and  $G_{SC}$  is between 91.5% and 94.5%. In total, we find only very few outliers of predictions with  $R^2$  values below 70 %. Note that the  $R^2$  values evaluate the model's performance relative to the variance of the observations around their mean. As the measured voltage parameters are more accurately represented by their mean compared with the current parameters, it is not surprising that significantly higher  $R^2$  values are observed for the current parameters.

Due to the decreasing amplitude of the PCs, the validation can further be used to evaluate, if the PCA reduces the data to a lower dimensional representation. We find no improvement regarding the  $R^2$  values using more than seven and eight PCs, respectively, depending on the dataset. Thus, the ESPs can be reduced to a seven to eight-dimensional representation, and the computation time can be reduced to 70%–80%. To visualize this high performance already observed using seven PCs, Figure 5 shows the boxplot of the  $R^2$  values for the ESPs reconstructed from seven PCs (computed with limits of the sum in (3) of [1,7]). Comparing Figures 4 and 5, we find only slight deviations for single parameters in some datasets.

The  $R^2$  values give a reasonable indication of how well a model performs, setting the performance in comparison to the deviation of the respective data from its mean. For PV outdoor data analysis, it might be of greater interest to evaluate on (mean) absolute deviations of the prediction, as they determine the accuracy regarding a PV modules performance (and yield). Figure 6 shows the histogram (gray) of the absolute deviations  $\Delta\text{ESPk} = \text{ESPk}_{\text{GPR}} - \text{ESPk}_{\text{test}}$  of the from seven GPR predicted PCs reconstructed ESPs ( $\text{ESPk}_{\text{GPR}}$ ) from the real test data of ESPs ( $\text{ESPk}_{\text{test}}$ ) for the exemplary chosen Italy CIGS4 dataset. The standard deviation of the histogram  $\sigma_{\text{hist}}$  is given in each subplot in black.

Furthermore, Figure 6 depicts the from the GPRs predicted probability density of deviations (red), that is, the prediction of the histogram according to the GPR output. Each GPR prediction has an individual predicted standard deviation (compare 6), that is, every prediction  $\text{PC}_j(G_{\text{POA}}, T_{\text{Mod}}, t)$  of the GPRs  $f_{\text{PC}_j}$  can be interpreted as a predicted Gaussian probability density  $\mathcal{N}(\text{PC}_j(G_{\text{POA}}, T_{\text{Mod}}, t), \sigma_{\text{PC}_j}(G_{\text{POA}}, T_{\text{Mod}}, t))$ . The prediction  $\text{ESPk}(G_{\text{POA}}, T_{\text{Mod}}, t)$  is a linear combination of the predictions  $\text{PC}_j(G_{\text{POA}}, T_{\text{Mod}}, t)$  (compare 3). In the approximation, that the 10 predictions  $\text{PC}_j(G_{\text{POA}}, T_{\text{Mod}}, t)$  are statistically independent for each input of  $[G_{\text{POA}}, T_{\text{Mod}}, t]$  (see Section 2.4), the prediction  $\text{ESPk}(G_{\text{POA}}, T_{\text{Mod}}, t)$  can be interpreted as a Gaussian probability density  $\mathcal{N}(\text{ESPk}(G_{\text{POA}}, T_{\text{Mod}}, t), \sigma_{\text{ESPk}}(G_{\text{POA}}, T_{\text{Mod}}, t))$ , where  $\sigma_{\text{ESPk}}(G_{\text{POA}}, T_{\text{Mod}}, t)$  is determined analogously to 4 from the predictions  $\text{PC}_j(G_{\text{POA}}, T_{\text{Mod}}, t)$  and standard deviations  $\sigma_{\text{PC}_j}(G_{\text{POA}}, T_{\text{Mod}}, t)$ . Considering deviations from the prediction sets the mean of each normal distribution to zero resulting in the prediction of the probability of deviations to  $\mathcal{N}(0, \sigma_{\text{ESPk}}(G_{\text{POA}}, T_{\text{Mod}}, t))$  for each coordinates  $[G_{\text{POA}}, T_{\text{Mod}}, t]$ .



**FIGURE 6** | Histogram (gray) of the absolute deviations  $\Delta\text{ESP} = \text{ESP}_{\text{GPR}} - \text{ESP}_{\text{test}}$  of the from seven GPR predicted PCs reconstructed ESPs from the real test data of ESPs for the exemplary chosen Italy CIGS4 dataset. The histograms standard deviation  $\sigma_{\text{hist}}$  is given in black. The GPR predicted probability density, representing the sum of Gaussian distributions of deviations with standard deviation  $\sigma_{\text{GPR}}$ , is depicted in red.

Our set of  $N$  test data points and the respective  $n = 1, 2, \dots, N$  input vectors  $[G_{\text{POA}}, T_{\text{Mod}}, t]_n$  can essentially be interpreted as a random choice for a subset of coordinates  $[G_{\text{POA}}, T_{\text{Mod}}, t]$ . Exploiting the law of total probability and substituting  $n$  for  $[G_{\text{POA}}, T_{\text{Mod}}, t]_n$  the probability densities for the deviation of each test point  $n$  ( $p_n(\Delta\text{ESP}_k) = \mathcal{N}(0, \sigma_{\text{ESP}_k}(n))$ ) can be summed up to receive the predicted joint probability of deviations from the model prediction

$$P(\Delta\text{ESP}_k) = \frac{1}{N} \sum_{n=1}^N p_n(\Delta\text{ESP}_k) \quad (16)$$

where the sum of the predicted deviations is divided by the number of tests data points  $N$  for normalization. In general, this sum of multiple Gaussian probability densities reflects a Gaussian

mixture model and might be Gaussian in specific scenarios, for example, in case all individual standard deviations are identical. However, regardless of this, the resulting standard deviation can be computed and is given by

$$\begin{aligned} \sigma_{\text{GPR}} &= \sqrt{\int_{-\infty}^{\infty} (\Delta\text{ESP}_k)^2 P(\Delta\text{ESP}_k) d\Delta\text{ESP}_k} \\ &= \sqrt{\int_{-\infty}^{\infty} (\Delta\text{ESP}_k)^2 \frac{1}{N} \sum_{n=1}^N p_n(\Delta\text{ESP}_k) d\Delta\text{ESP}_k}. \end{aligned} \quad (17)$$

Because  $p_n(\Delta\text{ESP}_k)$  is a probability density, it holds  $p_n(\Delta\text{ESP}_k) \geq 0$  and according to the Fubini–Tonelli theorem [21] it follows

$$\sigma_{\text{GPR}} = \sqrt{\frac{1}{N} \sum_{n=1}^N \int_{-\infty}^{\infty} (\Delta \text{ESP}_k)^2 p_n(\Delta \text{ESP}_k) d\Delta \text{ESP}_k} = \sqrt{\frac{1}{N} \sum_{n=1}^N \sigma_{\text{ESP}_k}^2(n)} \quad (18)$$

where we use, that the standard deviation of  $p_n(\Delta \text{ESP}_k)$  is given by  $\sigma_{\text{ESP}_k}(n)$ . In every subplot in Figure 6  $\sigma_{\text{GPR}}$  is given in red for reference.

We find standard deviations of the deviation histograms for the current parameters  $I_{\text{SC}}$ ,  $I_{\text{MPP}}$ ,  $I_{\text{qmp+}}$  and  $I_{\text{qmp-}}$  in the range of 14.8 to 23.8 mA (compare  $\sigma_{\text{hist}}$ , black). If we compare these absolute deviations to the mean of the time series data in Figure 2, we find relative deviations in the range between 2.3% and 2.6%. All absolute and relative deviations of the prediction are summarized in Table 1. For the voltage parameters  $V_{\text{OC}}$ ,  $V_{\text{MPP}}$ ,  $V_{\text{qmp+}}$  and  $V_{\text{qmp-}}$ , we find absolute deviation in the range of 1.17 to 1.36 V ( $\sigma_{\text{hist}}$ , black) and compared with the time series at  $G_{\text{POA}} = 750 \text{ Wm}^{-2}$  and  $T_{\text{Mod}} = 40^\circ\text{C}$ , relative deviations in the range of 1.3% to 2.4%.

Comparing the histogram with the GPR predicted probability density (red) for the current and voltage parameters, we find, that all eight deviation distributions are heavy tailed compared with the GPR predicted probability density, where the effect is more pronounced for the four current parameters. Especially for the four voltage parameters, the GPR predicted deviation qualitatively matches the real distribution of deviations, while the standard deviation of the real distribution of deviations  $\sigma_{\text{hist}}$  matches the predicted GPR standard deviation  $\sigma_{\text{GPR}}$  for the four current parameters. Table 1 further summarizes the GPR predicted standard deviation  $\sigma_{\text{GPR}}$  for all parameters. For the 8 current and voltage parameters the GPR predicted standard deviation is in the range of 0.9% to 2.5% relative to the mean of the time series at  $G_{\text{POA}} = 750 \text{ Wm}^{-2}$  and  $T_{\text{Mod}} = 40^\circ\text{C}$ .

For the parameter  $G_{\text{SC}}$ , we find similar to the eight current and voltage parameters a heavy tailed distribution, which is qualitatively well described by the GPR predicted probability density.

**TABLE 1** | The absolute and relative deviations of the GPR ESP prediction.

<b>Parameter</b>	<b><math>\sigma_{\text{hist}}</math></b>		<b><math>\sigma_{\text{GPR}}</math></b>	
	abs.	rel.	abs.	rel.
$I_{\text{SC}}$	23.8 mA	2.3%	24.7 mA	2.4%
$I_{\text{MPP}}$	22.1 mA	2.5%	22.3 mA	2.5%
$I_{\text{qmp+}}$	23.5 mA	2.4%	24.1 mA	2.5%
$I_{\text{qmp-}}$	14.8 mA	2.6%	14.5 mA	2.5%
$V_{\text{OC}}$	1.17 V	1.3%	0.84 V	0.9%
$V_{\text{MPP}}$	1.36 V	1.8%	1.10 V	1.5%
$V_{\text{qmp+}}$	1.32 V	2.4%	1.12 V	2.0%
$V_{\text{qmp-}}$	1.26 V	1.5%	0.96 V	1.1%
$G_{\text{SC}}$	$24.4 \frac{1}{\text{M}\Omega}$	6.9%	$12.4 \frac{1}{\text{M}\Omega}$	3.5%
$R_{\text{OC}}$	59.0 $\Omega$	648%	10.5 $\Omega$	115%

Furthermore, one can see, that the tail of the distribution is more pronounced for negative  $\Delta G_{\text{SC}}$ . The relative deviation of  $\Delta G_{\text{SC}}$ , however, is much higher. we find an absolute deviation of  $G_{\text{SC}}$  of  $24.4 \frac{1}{\text{M}\Omega}$  and a relative deviation of 6.9% compared with the mean of  $G_{\text{SC}}$  in Figure 2. In comparison with the current or voltage parameters,  $G_{\text{SC}}$  is not directly dependent on irradiation and temperature, thus more influenced by more subtle factors, like noise in the IV measurement and such expected to have a higher relative variance. Furthermore, we attribute the observed comparably high relative mean deviation of  $G_{\text{SC}}$  as well as the observed one-sided heavily tailed distribution with the definition of  $G_{\text{SC}}$  as a slope.

For the parameter  $R_{\text{OC}}$ , we find a high discrepancy between the shown histogram (gray) and the GPR predicted probability density (red). This high discrepancy can be explained by (systematic) outliers to high  $R_{\text{OC}}$  values in the dataset, which are not captured with the GPR prediction, as they cannot be described adequately with the assumption of Gaussian noise. The standard deviation prediction of the GPR is, however, clearly affected by the blatant outliers predicting an unreasonable high uncertainty of the prediction, for example, the GPR predicted probability distribution (red) does not match the histogram (gray). Neglecting these outliers would result in a standard deviation of the histogram  $\sigma_{\text{hist}}$  more than one order of magnitude lower, depending on the cut-off deviation, after which outliers are neglected. Because  $R_{\text{OC}}$  also describes a slope (at OC), a similar argument as for  $G_{\text{SC}}$  applies, where a higher relative variance is already expected.

## 5 | Summary and Outlook

This paper has introduced a statistical model for how to analyze PV outdoor data. First the general concept of the model has been discussed. Starting with the compact description of IV characteristics using the ESPs, a suitable filtering approach is applied. Afterward, the ESPs are processed with a principal component analysis (PCA). The main step is then temporally splitting the PC timeseries and training individual Gaussian process regressions (GPRs) using  $G_{\text{POA}}$  and  $T_{\text{Mod}}$  and the time  $t$  as input. Once the GPRs are trained, the model can reproduce and predict the complete IV characteristic at any given time  $t$ ,  $G_{\text{POA}}$  and  $T_{\text{Mod}}$ , incorporating the standard deviation of the ESP prediction based on the data noise and distance from the observations.

To underline its usability, the presented concept is applied to represent the IV characteristic in terms of the 10 ESPs clearly showing superimposed acclimatization, degradation and seasonality effects in the respective parameters. Furthermore, the model is utilized to determine a PV module's performance ratio over time and compared with a classical monthly temperature-corrected performance ratio approach. Especially regarding the uncertainty estimate resulting from the GPR predicted standard deviation we find an improvement of the  $\text{PR}_{\text{GPR}}$  estimate over the classical  $\text{PR}_T$  approach.

Finally, the presented concept is validated using a common 80:20 training to test data ratio, while keeping training and test data temporally separated. The overview of the model's performance measure  $R^2$  over all 10 ESPs and all 45 datasets shows

consistently good accuracy of the model for a wide range of technologies operated in different climates. For one exemplary subset, the accuracy of the model is discussed in more detail, showing low relative deviations in the prediction of the majority of ESPs, while being consistent with the model's prediction for its standard deviations.

The presented model widens the possibilities in the field of PV reliability monitoring and PV degradation analysis. The manageability of big datasets is ensured due to the compact description in terms of the ESPs and a temporal split of the data. The accuracy of the model is validated for different locations (climates) on different module types. Furthermore, only the feasibility to measure in-field *IV* characteristics, module temperature and irradiation is presupposed. As a result, that provides for the concepts presented many applications. Especially the possibility to set the results into context of their standard deviation enables to quantify the confidence of the results. Further, if only parts of the *IV* are available, for example, the SSPs or only AC or DC output voltage, current or power, the concept of training monthly GPRs is still applicable.

With rising accessibility to high computational power and the possibility to utilize Graphics Processing Units (GPUs) for matrix multiplications, GPRs become more and more applicable to big datasets [22]. This way, the presented approach could be updated in the future to bigger subsets or by a single GPR on each timeseries. Furthermore, the model can be utilized to compare different filtering approaches, as the applicability of GPRs to data remains on Gaussian distributed errors in the data and can help to detect the presence of systematic errors.

## Author Contributions

**Timon S. Vaas:** conceptualization, methodology, software, validation, formal analysis, writing – original draft, writing – review and editing, visualization. **Bart E. Pieters:** conceptualization, methodology, validation, formal analysis, writing – review and editing, supervision. **Evgenii Sovetkin:** validation, formal analysis, writing – review and editing. **Andreas Gerber:** resources, writing – review and editing, supervision, project administration, funding acquisition. **Uwe Rau:** resources, writing – review and editing, supervision, project administration, funding acquisition.

## Acknowledgments

This work has been partially funded by the Federal Ministry of Education under the Helmholtz LLEC Project, the Federal Ministry of Economic Affairs and Climate Action, BMWK under the grant FKZ 0325517B (PV-Klima project) and the Ministry of Culture and Science of the State of North Rhine-Westphalia under the grant B1610.01.17. Open Access funding enabled and organized by Projekt DEAL.

## Conflicts of Interest

The authors declare no conflicts of interest.

## Endnotes

<sup>1</sup> In particular, the computational complexity, as GPR optimization scales with  $N^3$ , where  $N$  is the number of data points.

<sup>2</sup> For example, if a lengthscale parameter becomes very large relative to the dataset's variation in the respective direction, further increasing the lengthscale has no visible effect on the GPR outcome.

<sup>3</sup> Note that we use here the formulas for the with the standard deviation weighted average of multiple individual measurements, that is, we treat each prediction from the monthly GPRs as individual predictions with given (predicted) standard deviation.

<sup>4</sup> Note that negative  $R_{OC}$  values do not occur in *IV* characteristics measured under steady conditions. The uncertainty estimate for  $R_{OC}$  thus primarily arises from deviations toward higher  $R_{OC}$  values, resulting in a one-sided tailed distribution. One could argue that the graph and the lower confidence interval in Figure 2i should be constrained to zero as a physically meaningful boundary for  $R_{OC}$ . However, because Gaussian process regression (GPR) predictions follow a Gaussian distribution, the uncertainty output is also Gaussian. As a result, the representation in Figure 2i is mathematically more accurate.

<sup>5</sup> Note that the choice of the reference output power is not arbitrary and might affect the absolute PR.

<sup>6</sup> Arguably, the uncertainty on the reference power output  $P_{MPP,GPR}|_{t \leq 7d}$  might need to be considered as well. Here, we neglect the influence of the uncertainty on the reference power output for two reasons. First, this reference value would be ideally substituted by available manufacturer data on the modules performance, making the task of the estimation of the standard deviation of the reference power output obsolete. Second, a possible deviation of the actual initial power output and the chosen reference affects all computed PR estimates (and the later computed standard deviations) by the same factor. This way, the uncertainty originating from the reference cannot be directly compared with the GPR standard deviation prediction resulting from noise and missing data and would need to be treated separately.

<sup>7</sup> Note that for the determination of  $PR_T$  the dataset is also filtered with the GPR filter as described in Section A, before applying the additional narrow band irradiation filter. Such the presented  $PR_T$  is already corrected from possible influences due to outliers resulting from shading of the module respectively the irradiation sensor.

<sup>8</sup> The contributions of each measurement are weighted differently.

<sup>9</sup> Note that the operation of the CIGS4 module operated in Italy started on the 1st of November 2013.

<sup>10</sup> Note that the incorporation of the covariance would lead to slightly higher standard deviations. The used approximation consequently leads to an underestimation of the standard deviation.

<sup>11</sup> Note that for the predictions there are limitations arising due to the parameter space of  $[G_{POA}, T_{Mod}, t]$  covered by the observations and associated high standard deviations for predictions that need to be extrapolated.

<sup>12</sup> Note that the seasonality appears to have a slightly lower period than 1 year. We attribute this discrepancy to the overlay of effects in the product of  $Nn_{id}$  (showing strong acclimatization) and  $E_a$  (showing an overlay of acclimatization, seasonality and degradation).

## References

1. A. Ndiaye, C. M. Kébé, A. Charki, P. A. Ndiaye, V. Sambou, and A. Kobi, "Degradation Evaluation of Crystalline-Silicon Photovoltaic Modules After a Few Operation Years in a Tropical Environment," *Solar Energy* 103 (2014): 70–77.
2. O. S. Sastry, S. Saurabh, S. Shil, et al., "Performance Analysis of Field Exposed Single Crystalline Silicon Modules," *Solar Energy Materials & Solar Cells* 94, no. 9 (2010): 1463–1468.
3. A. Phinikarides, G. Makrides, B. Zinsser, M. Schubert, and G. E. Georgiou, "Analysis of Photovoltaic System Performance Time Series: Seasonality and Performance Loss," *Renewable Energy* 77 (2015): 51–63.
4. M. Schweiger, J. Bonilla, W. Herrmann, A. Gerber, and U. Rau, "Performance Stability of Photovoltaic Modules in Different Climates,"

- Progress in Photovoltaics: Research and Applications 25, no. 12 (2017): 968–981.
5. R. Saive, “S-Shaped Current Voltage Characteristics in Solar Cells: A Review,” *IEEE Journal of Photovoltaics* 9, no. 6 (2019): 1477–1484, <https://doi.org/10.1109/JPHOTOV.2019.2930409>.
6. B. E. Pieters, “Extended Solar Cell Parameters - General Purpose Descriptive I/V Parameters for Solar Cells,” *Authorea* (2023).
7. C. Hansen Parameter Estimation for Single Diode Models of Photovoltaic Modules. tech. rep., Sandia National Lab.(SNL-NM); Albuquerque, NM (United States): 2015.
8. D. Cotfas, P. Cotfas, and S. Kaplanis, “Methods to Determine the DC Parameters of Solar Cells: A Critical Review,” *Renewable and Sustainable Energy Reviews* 28 (2013): 588–596.
9. A. Ortiz-Conde, F. J. García-Sánchez, J. Muci, and A. Sucre-González, “A Review of Diode and Solar Cell Equivalent Circuit Model Lumped Parameter Extraction Procedures,” *Facta Universitatis, Series: Electronics and Energetics.* 27, no. 1 (2014): 57–102.
10. A. R. Jordehi, “Parameter Estimation of Solar Photovoltaic (PV) Cells: A Review,” *Renewable and Sustainable Energy Reviews* 61 (2016): 354–371.
11. S. Lindig, A. Louwen, D. Moser, and M. Topic, “Outdoor PV System Monitoring Input Data Quality, Data Imputation and Filtering Approaches,” *Energies* 13, no. 19 (2020): 5099.
12. D. C. Jordan and S. R. Kurtz, “The Dark Horse of Evaluating Long-Term Field Performance Data Filtering,” *IEEE Journal of Photovoltaics* 4, no. 1 (2013): 317–323.
13. B. Pieters PV-CRAZE. <https://github.com/PVCRAZE/>; 2024.
14. A. Savitzky and M. J. E. Golay, “Smoothing and Differentiation of Data by Simplified Least Squares Procedures,” *Analytical Chemistry* 36 (1964): 1627–1639.
15. S. Karmalkar and S. Haneefa, “A Physically Based Explicit J-V Model of a Solar Cell for Simple Design Calculations,” *IEEE Electron Device Letters* 29, no. 5 (2008): 449–451.
16. H. Liu, Y. S. Ong, X. Shen, and J. Cai, “When Gaussian Process Meets Big Data: A Review of Scalable GPs,” *IEEE Transactions on Neural Networks and Learning Systems* 31, no. 11 (2020): 4405–4423.
17. P. Virtanen, R. Gommers, T. E. Oliphant, et al., “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” *Nature Methods* 17 (2020): 261–272, <https://doi.org/10.1038/s41592-019-0686-2>.
18. A. Dolara, S. Leva, and G. Manzolini, “Comparison of Different Physical Models for PV Power Output Prediction,” *Solar Energy* 119 (2015): 83–99.
19. G. Cui, X. Yu, S. Iommelli, and L. Kong, “Exact Distribution for the Product of Two Correlated Gaussian Random Variables,” *IEEE Signal Processing Letters* 23, no. 11 (2016): 1662–1666.
20. Photovoltaic Devices Procedures for Temperature and Irradiance Corrections to Measured I-V Characteristics. Standard IEC 60891, International Electrotechnical Commission; Geneva, CH: 2021.
21. G. Fubini and L. Tonelli, “Sulla derivata seconda mista di un integrale doppio. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*,” 40, no. 1 (1915): 295–298.
22. K. Wang, G. Pleiss, J. Gardner, S. Tyree, K. Q. Weinberger, and A. G. Wilson, “Exact Gaussian Processes on a Million Data Points,” *Advances in Neural Information Processing Systems* 32 (2019): 14581–14592.

## Appendix A

### Filtering

#### A.1 | Concept

From the TÜV datasets, we used the time series of  $I_{\text{SC}}$ ,  $T_{\text{Mod}}$ , and  $G_{\text{POA}}$ . The individual datasets consist of up to  $N \approx 65,000$  data points. Assuming rather constant conditions of the PV module and neglecting second-order spectral effects,  $I_{\text{SC}}$  is dependent only on  $T_{\text{Mod}}$  and  $G_{\text{POA}}$ . To describe these dependencies of the short circuit current, we split the dataset into 20 random subsets and fit a GPR  $f_{I_{\text{SC},i}} : \mathbb{R}^2 \rightarrow \mathbb{R}$  model to each of the  $i = 1, 2, \dots, 20$  subset. The splitting of the dataset is necessary as the computation time of GPRs scales with  $N^3$ , that is, we chose an approach, that can be compared with a bootstrap aggregating approach to reduce the computation time. For each subset, we fit the three hyperparameters of a two dimensional (2D) RBF kernel with an added white noise kernel. The kernel of each  $f_{I_{\text{SC},i}}$  is defined as

$$K(X, \tilde{X}) = \exp\left(\frac{\|x_1 - \tilde{x}_1\|^2}{2l_1^2} + \frac{\|x_2 - \tilde{x}_2\|^2}{2l_2^2}\right) + \sigma^2 \mathbb{I} \quad (\text{A1})$$

Where  $X = [x_1, x_2] = [G_{\text{POA}}, T_{\text{Mod}}]$  represents the 2D input space of irradiation and temperature,  $\mathbb{I}$  is the identity matrix and  $l_1$ ,  $l_2$ , and  $\sigma$  are the three hyperparameters to be optimized during training of each  $f_{I_{\text{SC},i}}$ . The hyperparameters are optimized using the limited-memory BFGS algorithm (L-BFGS) implemented in SciPy [17].

Subsequently, we compute the predicted  $I_{\text{SC},i,n}$  for each of the  $n = 1, 2, \dots, N$  points of the dataset from each individual GPR and the, with the standard deviation  $\sigma_{I_{\text{SC},i,n}}$  weighted, overlay of the  $i = 1, 2, \dots, 20$  predictions  $I_{\text{SC},n}$ . We assume here statistical independence of the respective  $I_{\text{SC},i,n}$  predictions for each  $n$ , as they are based on  $i = 1, 2, \dots, 20$  random subsets, to compute the standard deviation  $\sigma_{I_{\text{SC},n}}$  neglecting the covariance between individual predictions. We justify this assumption with the individual optimized hyperparameters and the individual subsets used for training.<sup>10</sup> The overlay of predictions has two advantages. First, individual outliers only affect the prediction of individual  $f_{I_{\text{SC},i}}$  and thus only the individual predictions  $I_{\text{SC},i,n}$  for each point  $n$ . Assuming no aggregation of outliers toward a specific combination of  $G_{\text{POA}}$  and  $T_{\text{Mod}}$  combination, individual outliers only affect 5% of the predictions, whose results are overlaid. Second, the predictions  $I_{\text{SC},i,n}$  of the  $i$ th GPR that are trained with many outliers, as well as predictions  $I_{\text{SC},i,n}$  of the GPRs at points  $n$  with  $T_{\text{Mod}}$  and  $G_{\text{POA}}$  conditions far from the  $i$ th training input, will result in  $I_{\text{SC},i,n}$  predictions with high associated standard deviations  $\sigma_{I_{\text{SC},i,n}}$ . Overlaying multiple individual GPRs will suppress the influence of outliers and lack of information in individual GPRs. Note that each GPR only uses 5% of the available data as training input and might not cover the complete 2D input space of  $T_{\text{Mod}}$  and  $G_{\text{POA}}$ .

To filter the data, we use a filter with a very high threshold deviation of the measured  $I_{\text{SC}}$ , set at 20 standard deviations  $\sigma_{I_{\text{SC},n}}$  from the predicted  $I_{\text{SC},n}$ . Note that because the incorporation of the covariance would lead to slightly higher standard deviations, the used threshold of 20  $\sigma_{I_{\text{SC},n}}$  reflects an effective threshold slightly lower than 20 standard deviations. This filter only excludes data points, which cannot be explained by statistical uncertainty in the measurement. In other words, the GPR predicts a probability density of measuring  $I_{\text{SC}}$  at the respective input conditions  $T_{\text{Mod}}$  and  $G_{\text{POA}}$  and the filter is designed to exclude data points, that show a very high deviation from this predicted probability distribution. Note that this approach inherits the assumption, that all measurements exhibit ideally only statistical noise. This assumption is not valid in the presence of systematic errors. However, in case the data would only exhibit only statistical noise the filter presented would filter a share in the order of  $10^{-87}\%$  (probability of a Normal distribution to find a measurement outside of 20 standard deviations) and such virtually no data. The high threshold is chosen to account for any possible changes over time of  $I_{\text{SC}}$ , which are not described within this filter. Effects of degradation and

seasonality in the short-circuit current typically show low rates and exhibit low amplitudes over the approximately 3-year span of operation in the datasets under consideration. The high threshold ensures that these effects are not filtered out.

As this high threshold only excludes extreme outliers, the filtering process is applied iteratively. In each iteration, the most extreme outliers are detected and removed, which reduces the subsequent standard deviation (as the standard deviation, due to the quadratic weighting of deviations, is strongly affected by outliers). The iterative process can be continued until no extreme outliers are detected, here a threshold of 4 iterations is chosen. Note that the iterative nature of the approach leads to little respectively virtually no influence of adapting the threshold a bit lower or higher. This iterative method effectively rejects the tails of distributions, ensuring that the resulting distributions have no data points beyond 20 standard deviations from the mean. Thus, the filtered distributions become more Gaussian, because systematic errors are filtered out. In this work we limit the iterative process to a fixed number of iterations.

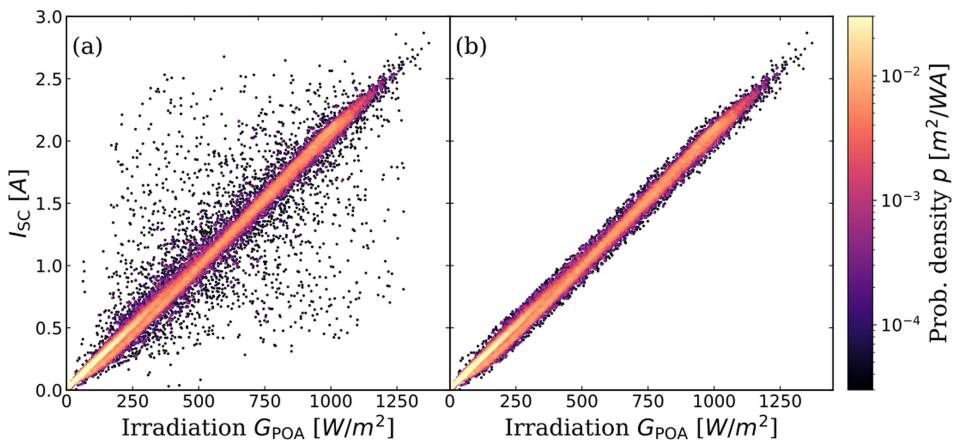
#### A.2 | Gaussian Process Regression Filtering Results

The presented filtering concept is applied to all 45 TÜV datasets. The share of filtered data in the individual datasets varies between 1.6% and 4.4%, where in total 2.6% of the  $\approx 2,200,000$  data points are filtered out. It is noticeable, that the share of filtered data does not vary much among one location. As the modules operated at the same location are positioned close to each other, the probability of different effective irradiation of the PV module and the irradiation sensor leading to outliers is correlated. Furthermore, the same irradiation sensor is used for all datasets at the same location, resulting in (partial) shading of the irradiation sensor affecting all datasets of the location in the same way.

Figure A1 shows a scatter density plot of measured  $G_{\text{POA}}$  versus  $I_{\text{SC}}$  for the exemplary chosen unfiltered (a) and filtered (b) Italy CdTe1 dataset after four iterations of the applied GPR filter. As expected, the short circuit current is linear in the POA irradiation. It is observed, that the density of measured  $[G_{\text{POA}}, I_{\text{SC}}]$  pairs decreases with increasing deviation from the linear trend, that is, slightly different effective irradiation on the PV module and the irradiation sensor is more likely than large discrepancies. Furthermore, we find no clusters with a high scatter density deviating from the observed linear relationship.

The first iteration of applying the GPR filter clearly removes only extreme outliers with a share of 1.0% of the dataset. The second and third iteration of applying the GPR filter remove 1.1% and 0.8% of the exemplary chosen Italy CdTe1 dataset, respectively. The fourth and final iteration removes 0.4% of the Italy CdTe1 dataset. From Figure A1, one can see that the filter reliably removes outliers with high deviations from the expected relationship between  $I_{\text{SC}}$  and  $G_{\text{POA}}$ . This trend of a decreasing share of removed data with more iterations of applying the GPR filter is observable across all 45 datasets. This occurs as the distributions become more normal and less tailed. As the distributions become more Gaussian, the share of data that can be accurately predicted using GPRs increases, resulting in fewer points being filtered out with each iteration. The applied GPR filter removes reliably data points where the measured  $G_{\text{POA}}$ ,  $T_{\text{Mod}}$  and  $I_{\text{SC}}$  show a high deviation from the expected relationship. In general, the GPR filter presented here is applicable on any dataset, where  $G_{\text{POA}}$ ,  $T_{\text{Mod}}$  and  $I_{\text{SC}}$  are available. IV characteristics, where the measured  $I_{\text{SC}}$  does not match (in a given uncertainty) the expectation based on the measured  $G_{\text{POA}}$  and  $T_{\text{Mod}}$  are hard to interpret, because an interpretation w.r.t. the PV modules performance or efficiency is only possible taking the conditions the IV is measured at into account. On one hand, the filter removes outliers that originate from different irradiation levels on the PV module and the irradiation sensor reliably. On the other hand, the filter only accounts for this particular cause of systematic errors in the dataset.

The major limitation of the filter is the inability to differentiate between a signal in the data and a systematic deviation, that is, the GPRs do not incorporate the temporal development of the  $I_{\text{SC}}$  due to acclimatization, seasonality and degradation effects. These effects (the signal in the time



**FIGURE A1** | Scatter density plot of  $I_{\text{sc}}$  versus  $G_{\text{POA}}$  for the (a) unfiltered and (b) filtered Italy CdTe1 dataset after four iterations of applying the GPR filter.

series of  $I_{\text{sc}}$ ) are overlaid with statistical noise and systematic deviations, which leads to further broadening of the probability distribution measuring  $I_{\text{sc}}$  at conditions  $G_{\text{POA}}$  and  $T_{\text{Mod}}$ . A reduced threshold for the filter, as well as each iteration of applying the filter, increases the probability of removing valid data points. Due to this limitation, the filter is utilized to remove only outliers with high deviations from the expectation, which cannot be explained with any mechanism relevant for the further analysis. Using the scatter density plot, we further demonstrate that the data, which is filtered out, does not show any clustering regarding high scatter densities at specific  $I_{\text{sc}}$  and  $G_{\text{POA}}$  combinations deviating from the linear relationship. As the effects of acclimatization, degradation and seasonality exhibit, in general, low amplitudes, that is, long lengthscales for changes in  $I_{\text{sc}}$  are observed, the signal in  $I_{\text{sc}}$  cannot explain unclustered deviations from the predicted probability density targeted and excluded with the applied filter.

## Appendix B

### Physical Analysis Using the One-Diode Model

In Figure 2, we show one possible output of the GPR *IV* model, where the *IV* characteristic respectively the ESPs are shown over time for one constant set of irradiation and temperature conditions. However, from the trained GPRs, one can reproduce the ESPs at arbitrary conditions, enabling multiple applications.<sup>11</sup> One possible application is to fit a physical model to a given produced output. To cover the range of irradiation intensities and module temperatures occurring in operation, we compute the timeseries prediction of the ESPs for the exemplary chosen Italy CIGS4 dataset for multiple combinations of the two conditions, where we use  $T_{\text{Mod}} = 10, 20, 30, 40, 50, 60^\circ\text{C}$  and  $G_{\text{POA}} = 100, 250, 500, 750, 1000 \text{ W m}^{-2}$ . With the output of a total of 30 timeseries of ESPs at constant conditions, the temperature and irradiation dependent one diode model is fitted to the predicted *IV* data.

Utilizing the well-known one-diode model equation for the current-voltage relation in a solar cell, accounting for  $N$  number of cells connected in series and assuming a temperature-activated recombination current results in

$$I(V) = I_{00} \left( \frac{T}{T_{\text{STC}}} \right)^3 \exp \left( \frac{(T - T_{\text{STC}})E_a}{kT_{\text{STC}}T} \right) \\ \left[ \exp \left( \frac{e(T_{\text{STC}} - T)(V - I(V)R_s)}{Nn_{\text{id}}kT_{\text{STC}}T} \right) - 1 \right] + \frac{V - I(V)R_s}{R_{\text{sh}}} - I_{\text{ph}}(G, T), \quad (\text{B2})$$

where  $E_a$  is an activation energy and

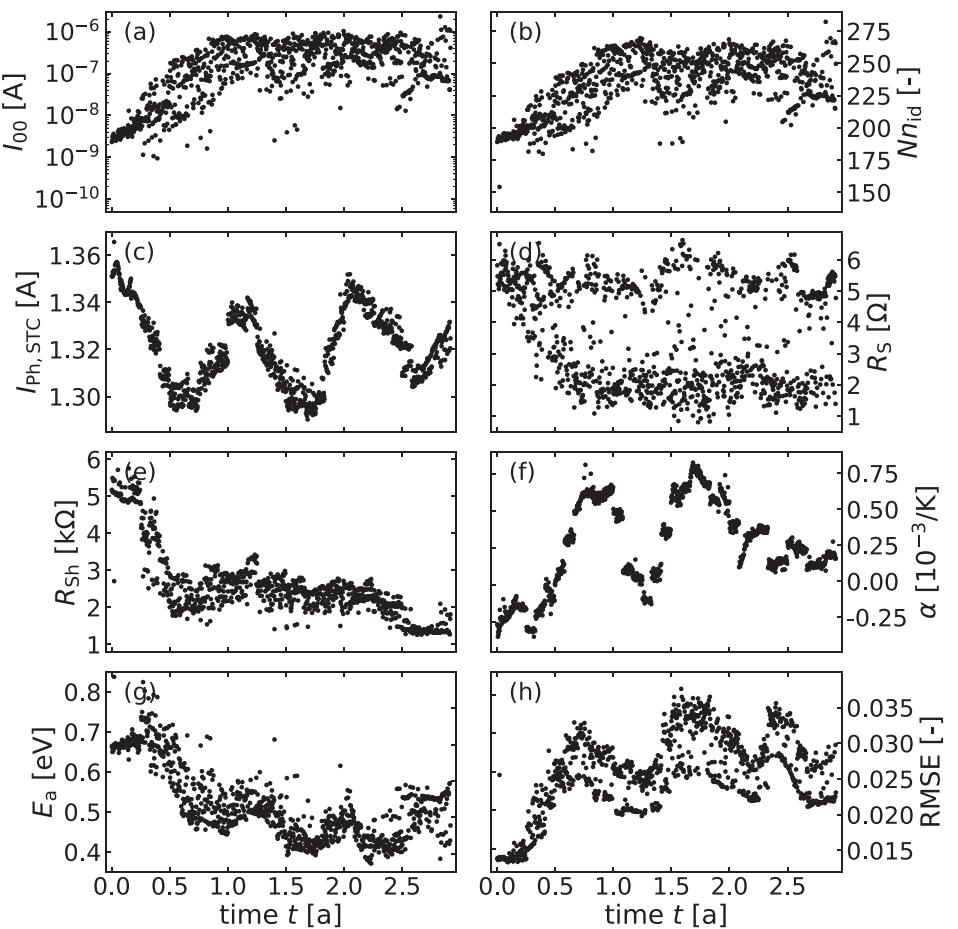
$$I_{\text{ph}}(G, T) = \frac{G}{G_{\text{STC}}} (1 + \alpha(T - T_{\text{STC}})) I_{\text{ph,STC}} \quad (\text{B3})$$

similar to the IEC60891 temperature dependency of the currents in [20] substituting  $I_1 = I_{\text{ph,STC}}$ . With the fit of B2 and B3 to the timeseries of ESPs at the distinct conditions using the PV-CRAZE library [13] the result is a timeseries of the parameters  $R_s$ ,  $R_{\text{sh}}$ ,  $I_{\text{ph,0}}$ ,  $Nn_{\text{id}}$ ,  $I_{00}$ ,  $\alpha$  and  $E_a$ . We find a better fit is obtained with PV-CRAZE, if the fitting is repeated several times and the best optimum is selected (some fitting algorithms used by PV-CRAZE are probabilistic). In this work, the fit was repeated 23 times in a trade off between computation time and fit error. Figure B1 shows the development of the extracted parameters  $R_s$ ,  $R_{\text{sh}}$ ,  $I_{\text{ph,0}}$ ,  $Nn_{\text{id}}$ ,  $I_0$ ,  $\alpha$  and  $E_a$  over time, enabling to physically interpret the temporal development of the performance of the operated CIGS module. Furthermore, the root mean squared error (RMSE) of the fit is shown.

First of all, we observe a strong acclimatization effect most prominent in the shunt resistance  $R_{\text{sh}}$  (Figure B1e), the activation energy  $E_a$  (Figure B1g), as well as in  $I_{00}$ ,  $Nn_{\text{id}}$  and  $R_s$  (Figure B1a,b,d). In the first few months of operation, the  $R_{\text{sh}}$  is reduced from over 5 kΩ to less than 2 kΩ and the activation energy drops from 0.7 to 0.5 eV. Comparing the temporal development of  $I_{00}$  and  $Nn_{\text{id}}$  (Figure B1a,b), we find a strong correlation between the two parameters. The observed increase in dark saturation current  $I_{00}$  is associated with an enhanced ideality factor ( $Nn_{\text{id}}$  increases and  $N$  is constant). This suggests, that there is a change in the most prominent recombination mechanism, that is, either a recombination process with low ideality is reduced or a recombination process with high ideality enhanced. Furthermore, the series resistance shows an acclimatization effect, where the diode model fit finds two solutions, one close to the initial  $R_s$  between 5 and 6 Ω and one solution between 1 and 2 Ω. This effect where the solution alternates between two values is also visible in the RMSE, indicating the solution alternates between two local minima with distinct  $R_s$  values. Inspection shows the slightly smaller errors are associated with the more constant  $R_s$  solution between 5 and 6 Ω.

Besides the described acclimatization effects, a strong seasonality is visible for the photo-generation current  $I_{\text{ph,STC}}$  (Figure B1c). With the metastable nature of CIGS solar cells and modules in mind this clear difference in performance in winter and summer is not surprising. Furthermore, a seasonality contribution is visible in  $E_a$  (Figure B1g), the temperature coefficient  $\alpha$  (Figure B1f) as well as in the RMSE (Figure B1h), which implies a slightly better accuracy in winter.

We find contributions of a linear degradation for the parameter  $R_{\text{sh}}$  and  $\alpha$  (Figure B1e,f). A linear degradation over the complete operation period is further observed for the activation energy  $E_a$ , where a slight linear reduction is overlaid with the seasonality and the acclimatization. Assuming SRH recombination, the activation energy is expected to be dependent on the ideality factor, where the activation energy can be replaced by the quotient of the band gap energy and the ideality factor ( $E_a \approx \frac{E_g}{n_{\text{id}}}$ ). To this end, we correct the activation energy for the ideality factor influence. Figure B2 shows the temporal development of the

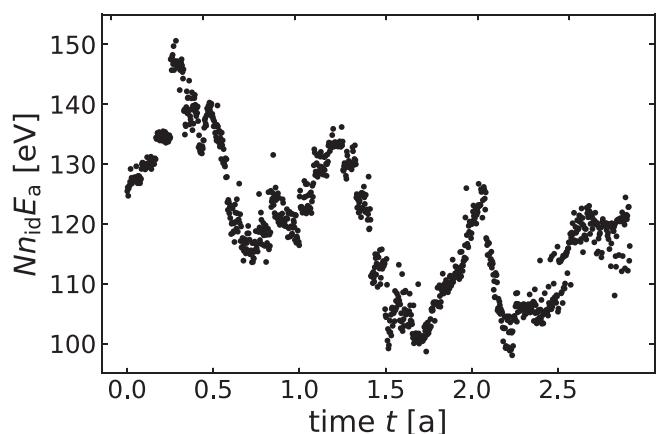


**FIGURE B1** | One diode model parameters timeseries prediction fitted to a GPR predicted timeseries matrix at different irradiation and temperature conditions for the Italy CIGS4 dataset.

product  $Nn_{id}E_a$ . We find a clear linear degradation overlaid with a seasonal variation.<sup>12</sup> This further indicates a more complex change in the prominent recombination mechanism over time.

Comparing the qualitative development of the diode parameters with the PR<sub>GPR</sub> in Figure 3a,b, we attribute the acclimatization effect to a change in the prominent recombination mechanism overlaid with a reduced shunt resistance. The relative amplitude of the seasonality

matches well with the seasonality observed for the photocurrent and the long-term degradation is influenced by multiple factors, such as the further reduced activation energy and shunt resistance.



**FIGURE B2** | Product of the one diode model parameters timeseries prediction  $Nn_{id}E_a$  fitted to a GPR predicted timeseries matrix at different irradiation and temperature conditions for the Italy CIGS4 dataset.