Abstract: The ever-growing penetration of cost-effective photovoltaic (PV) panels within the distribution grid requires a robust and efficient method for PV system monitoring. Especially, the geographical proximity of PV panels can play an important role in lowering the dimension of measurements required for full system observability. Furthermore, the direct impact of variable cloud formation and uncertain propagation necessitates the development and validation of a spatiotemporal model. Accordingly, this study presents the modelling and validation of the spatiotemporal variability of solar power indices at 1 minute resolution for the correlation to estimate PV power output at locations from where measurements are unobserved. The multi-rate feature of the set of unobserved sensors is of paramount importance for performing grid monitoring and management tasks. Hence, a reduced order estimation technique will allow for the accurate knowledge about the instantaneous aggregated distributed PV power output at any time, using measurements from a small subset of the total number of generators. In this regard, we utilise the recursive estimation structure of the Kalman filter [3] due to its proven and widespread applications among others in Fuzzy logic-based system modelling [4, 5], dynamic neural networks [6] and biomedical signal processing [7].

In the context of observing distributed PV generation, the high resolution time series data can be modelled in the form of equal resolution state dynamics, while the measurement can be obtained at a much lower rate through decimation—a down-sampling method widely used in multi-rate signal processing [8]. The measurement updates to the Kalman filter can be ‘event-driven’, i.e. the recursive estimator receives the measurement only when it deviates from previous measurements beyond a pre-specified threshold [9, 10]. Consequently, PV measurements can be recorded at low sampling rates and the temporal dimension of the data is reduced. Furthermore, A key characteristic of distributed PV is its high level of spatial correlation due to geographical proximity. This feature can be exploited by lowering the spatial dimension of...
measurements required for full system observability. In particular, the concept of Kriging utilises the spatial correlation in estimating PV generation at unobserved locations [11]. Thus, measurements are collected from a few nearby locations and used to estimate generation profile of all the distributed PV systems in geographic proximity.

In this paper, we utilise such a Multi-Rate and Event-DRiven Kalman Kriging (MREDRIKK) filter that can recursively estimate distributed and behind-the-meter solar generation in scalable space-time. From the perspective of PV systems, we consider the time series of solar power index (SPI) [12] as the measure and to develop the necessary state-space form. The Oahu Solar Database (described in Section 3) is used to develop a data set of daily SPI time series. The distinctive weather patterns for the entire data set have been identified with the help of k-means clustering. Authors approximately 50% of this data set to fit the spatio-temporal dependence state-space and spatial model for the MREDRIKK filter while the rest is used to investigate the filter performance.

1.1 Literature review

There are two important aspects in the design of a scalable, recursive spatiotemporal PV generation estimation routine: (a) preprocessing of raw time series and (b) development of the necessary state-space form. In this regard, the following prior works on solar irradiance, generation and other weather parameters are important to highlight.

The raw time series of irradiance can be normalised with respect to the clear sky value [13]. However, authors in [14] take additive model approach and decompose the raw measurement into: (a) large scale variability (trend); the seasonal component is based on diurnal patterns. The AR(2) and Larcher models are then combined to produce 1 h ahead forecasts. The combination is based on the positivity test of the difference of deseasoned data and does not involve dynamic estimation using Kalman filter or spatial estimation through Kriging. In [16], solar irradiance is assumed to be an additive model to predict hourly or sub-hourly PV power. The deterministic component is approximated through a sinusoidal equation. Power spectral analysis is then used to find the AR moving average model parameters, and then the state-space model for Kalman filtering. Data for one day is used for model fitting and tested on another day of similar generation pattern. This work, however, demands more comprehensive effort in capturing seasonal PV generation patterns. Authors in [17] use a fusion approach for the regressor and Kalman filter prediction outputs, i.e. short-term 10–20 min horizon nowcasting predictions. One month of solar irradiance data are used to obtain regressor input feature vectors and ramp-down event forecasting through a support vector machine. The ramp-down event forecasting gives the adjusting factor within the state transition matrix, which is used in the Kalman filter. However, the fusion alternatives are data dependent and seem inconclusive. In [18], a third order polynomial is used to define the state-space model, although the selection of an identity matrix to define state transition lacks sufficient justification. Unlike the recursive approach of Kalman filters, ratio methods are used in [19] to estimate 1 min resolution PV power given the PV power recorded at a nearby PV system.

In forming the observation space, a historical data-driven approach is investigated to model the observation matrix and state vector. In this regard, Gaussian weight kernels [20], spline back-fitted kernels [21], and also a probabilistic approach [22] are reported in capturing the mean dynamics of states over a large volume of measurements. Authors in [20], apply Gaussian weight kernels to fit observation model and assume “random walk” state evolution to develop forecast of quarterly mean rainfall and annual mean temperature. However, as with [14], the work needs to be expanded to assess Kriging performance under varying dimension of observation. In [21], space-time and time-space models are developed to fit 1 min resolution GHI data where the spline back-fitted kernels are used to find the model coefficients. The time-dependent state vectors obtained can then be used to train the system evolution model through vector AR analysis [23]. This work, like [16], focuses on day-specific modelling and demands for more comprehensive effort in capturing seasonal irradiance patterns. The pair copula-based versatile probability model is used in [22] to present 15 min resolution PV generation for probabilistic load flow study. However, the consistency in spatial resolution for PV generation data and the power system test case needs to be addressed.

In summary, the prior methods involve the usage of standard time series models and machine learning to forecast solar irradiance, solar generation, temperature and rainfall. An additive modelling approach is evident in all of these works with the absence of any normalisation efforts. Furthermore, only one work involves both the Kalman filtering and Kriging, while rest of them either use Kalman filter or Kriging or none of them. Finally, none of the prior works addresses the issue of capturing weather-dependent variability in temporal evolution as well as effort in temporal down-scaling for measurement access. On the other hand, it is important to design a real-time, weather-dependent and scalable spatiotemporal estimation platform that not only supports limited communication bandwidth through lesser data aggregation but also provides better visibility to the spatially distributed grid contribution of behind-the-meter PV systems. In this regard, we summarise the contribution of proposed method

- Develop a generic procedure of utilising available information to train and implement a scalable spatiotemporal estimation of distributed PV generation. The procedure is validated using publicly available and highly variable irradiance database.
- Incorporate weather-dependent solar generation patterns with the dynamic estimation algorithm.
- Utilise the multi-rate and event-driven mechanism to collect accessible and available PV power measurements. Exploit spatial correlation to estimate PV generation from unobserved locations. This approach reduces utilities’ dependence on PV developers’ data, preserve user privacy, and can also provide expanded monitoring of ever-growing distributed PVs without additional installation of PV production meters.
- Illustrate estimation performance sensitivity to various parameters. This will serve as guidelines for tuning-parameter adjustment in achieving desired estimation performance.

The rest of the study is organised as follows: Section 2 describes the steps in designing the MREDRIKK filter and provides the algorithm. The solar irradiance database used in this research is described in Section 3. The steps in fitting the state-space and spatial model are described in Section 4. The estimation performance is discussed in Section 5. Finally, the findings and future work are summarised in Section 6.

2 MREDRIKK filter design

In this section, we describe the state-space model formation from a generic AR time series model and introduce the multi-rate and event-driven approach for low-dimensional capturing of temporal measurements. Subsequently, the spatial Kriging will be discussed and the MREDRIKK algorithm will be described for low dimensional spatiotemporal estimation of distributed PV generation.

2.1 System model

We consider a scalar continuous-time phenomena observed by N sensors colocated with geographically distributed PV systems at > 1 1 1 1 = 1,..., N, where the continuous-time series is discretised with a sampling interval T and assumed to follow an AR model of order p.
Here $\mu_t$ is an intercept term reflecting the mean of the time series [24],

$$
\mu_{t,i} = E[x_t] = \frac{\mu_t}{(1 - \sum_{x=1}^{p_t} a_{x,i})}
$$

Defining the $p_t$-dimensional vectors, $w_t = [w_t, 0, \ldots, 0]$ and $x_t = [x_{t,1}, x_{t,2}, \ldots, x_{t,p_t}]$, we have the following model of system dynamics

$$
x_t = F_t x_{t-1} + w_t
$$

Here $\mu_t = E[x_t] = \mu_t A_t$, the state transition matrix

$$
F_t = \begin{bmatrix}
a_{t,1} & a_{t,2} & a_{t,3} & \cdots & a_{t,p_t-1} & a_{t,p_t} \\
1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & 1 
\end{bmatrix}
$$

and $C_{w,t}$ denotes the covariance matrix of $w_t$, the first diagonal being $\sigma_{w,t}^2$, while the rest of entries are zero. We obtain the state covariance matrix $C_{x,t}$, as the solution of the Lyapunov equation

$$
C_{x,t} = F_t C_{x,t} F_t^\top + C_{w,t}
$$

Now, let $\Theta$ denote the set of $m$ active sensors from which measurements are being collected, i.e. $\Theta \subset \mathcal{S}$ and cardinality of $\Theta$ is $m$. Therefore, the set of unobserved sensors, $\mathcal{S}' = \mathcal{S} - \Theta$. Now, the collective observation model for the active sensors

$$
y_t(\Theta) = H_t(\Theta) x_t + v_t; y_t \sim \mathcal{N}(0, C_{y,t})
$$

here the $m \times p_t$ observation matrix $H_t(\Theta)$ consists of the set of observation coefficients $(h_t(s))_{r \in \Theta}$ in the first column while rest of the entries are zero. In the same way, the matrix $H_t(\mathcal{S}')$ is defined for the set of unobserved sensors. The measurement noise covariance matrix $C_{y,t}$ is diagonal in nature indicating the uncorrelated noise variance for the spatial measurements. The parameters of the state-space model just described are time dependent and hence the subscript for discrete time ‘$t$’ is included. Such dependence comes from the time varying weather pattern that constitutes itself by switching from one pattern to another after certain time period. In this research, the weather patterns are captured by a set of disjoint clusters. The frame length of the time series forming the clusters is determined based on the training data sets’ performance in fitting to the AR model. The detailed procedure of finding the number of clusters, frame length and cluster specific model parameter estimation is described in Section 4. Measurements following the observation model (3) are used to estimate the state in the minimum mean squared error (MMSE) sense for a distributed PV system described by the above model. Accordingly, the state estimate at discrete time $t$ is

$$
\hat{x}_{t|t} = E[x_t | y_t(\Theta), y_t(\Theta), \ldots, y_t(\Theta)]
$$

The corresponding error covariance matrix is

$$
M_{t|t} = E[(x_t - \hat{x}_{t|t})(x_t - \hat{x}_{t|t})^\top]
$$

We use this estimate and the spatial covariance information to estimate unobserved PV systems through the Kriging step (Section 2.3).

2.2 Multi-rate and event driven approach

The temporal down-scaling is achieved by incorporating the multi-rate and event driven approach to collecting measurements from active sensors (i.e. $y_t(\Theta)$). In this regard, we utilise a D-fold decimation [8] to define the following indicator function

$$
I_D(t) = \begin{cases} 
1; & t = iD; \quad i \in Z, \\
0; & \text{otherwise}.
\end{cases}
$$

The event-driven approach to collecting measurements is based upon a pre-determined disparity threshold $\delta$ [10] and is prioritised over the D-fold decimation [8]. As a consequence, we introduce a binary factor $\phi(t)$ in order to reflect the combined (multi-rate and event-driven) process of measurement collection

$$
\phi(t) = \begin{cases} 
1; & \min_{s \in \Theta} |y_t(s) - y_{t-1}(s)| \geq \theta \\
\max_{s \in \Theta} \sigma(s) \geq \theta; & \text{otherwise.}
\end{cases}
$$

therefore, $\phi(t)$ indicates the measurement collection status for an active sensor. Here, $0 < \theta \leq 1$ is a multiplying factor weighing the sensitivity of the event detection mechanism. The detection becomes more sensitive as the value of $\theta$ decreases. Such design is motivated by the fact that the stability of a dynamic estimator is not affected by the choice of event-triggering threshold $\delta$ [25]. Now, during the sub-sampling interval $[iD, (i + 1)D); i \in Z_+$, measurement from an active sensor is collected only if there is any significant event, otherwise it is estimated using the Kalman filter and the most recent measurement (i.e. collected at $t = iD$). This approach is carried out in two steps. First, at any discrete time $t (D < t < (i + 1)D; i \in Z_+)$, the most recent measurement from active sensors is recorded as

$$
y_t(\Theta) = \phi(t) y_t(\Theta) + (1 - \phi(t)) y_{t-1}(\Theta)
$$

which is then used with the MMSE estimate $\hat{x}_{t|t}$ in order to correct measurements for the active sensor

$$
y_t(\Theta) = \phi(t) y_t(\Theta) + (1 - \phi(t)) H_t(\Theta) \hat{x}_{t|t}
$$

It is worthwhile to mention that the concept of multi-rate and event-driven approach was first introduced in [9] to estimate the range map for a low-altitude helicopter with significant computational reduction. However, the persistence based measurement update and correction proposed in (8) and (9) is not considered. Consequently, the estimation error covariance matrix and Kalman gain based estimation correction take place only when an actual measurement is collected. On the contrary, the proposed approach in (8) and (9) allow the occurrence of such recursive steps independent of measurement collection status inside the MREDRIKK algorithm (discussed in Section 2.4).

2.3 Kriging step

The concept of Kriging is about the utilisation of spatial covariance to make linear MMSE estimates at locations from where no measurements are recorded [11, 26, 27]. As a consequence, Kriging facilitates the spatial downsampling as well as provides estimates for any malfunctioning sensor’s measurement. In this research, we establish the simple Kriging of unobserved sensors based on the active sensors’ measurement update. Let $C_{s,t}$ denote the spatial covariance between sensor locations $s_i$ and $s_j$. Now, given the unobserved sensor location $s_i$, the Kriging weights $(a_{i,k})$ are obtained by solving the following equation

$$
a_{i,k} = \begin{bmatrix} a_{i,1} \\
\vdots \\
a_{i,m} \\
c_{m,k} \end{bmatrix} = \begin{bmatrix} C_{i,1} & \cdots & C_{i,m} \\
\vdots & \ddots & \vdots \\
c_{m,1} & \cdots & C_{m,m} \end{bmatrix}^{-1} \begin{bmatrix} C_{i,k} \\
\vdots \\
c_{m,k} \end{bmatrix}
$$
The spatial covariances, especially those from unobserved sensor locations are obtained from a standard semivariogram model fitted to historical data. In this research, we fit the exponential model to the semivariogram between locations \( s_i \) and \( s_j \) [28],

\[
\gamma(d) = \begin{cases} 
\gamma_0 & d = 0 \\
\gamma_0 + c_0(1 - e^{-cd}) & d \neq 0.
\end{cases}
\]  

(11)

Here \( \gamma(d) = 0.5 \times \text{variance}(\gamma(s_i) - \gamma(s_j)) \) and \( d \) is the Haversine distance between locations \( s_i \) and \( s_j \) in the geographic coordinate system. The choice of exponential fitting is motivated by the exponentially decaying spatial variability of solar generation [29]. Once, the semivariogram model is fitted, the spatial covariance is then calculated as, \( C_{ij} = c_0 + c_1 - \gamma(d) \) and used in (10) to calculate the Kriging weights. Therefore, the estimate for the unobserved sensor \( k \)

\[
\hat{y}_k = h(s_k)\hat{x}_k + \alpha_{ij}(\hat{y}_j - H_i(\Theta)\hat{x}_p)
\]

\[
\hat{y}_k(U) = H_k(\Theta)\hat{x}_k + A(\hat{y}_j(U) - H_i(\Theta)\hat{x}_p)
\]

(12)

Here \( A = [\alpha_1, \ldots, \alpha_{N-m}]^T \). In the next subsection, we constitute the spatiotemporal dynamic estimation algorithm that entails the concepts described so far.

2.4 MREDRIKK algorithm

We incorporate the spatial (Kriging) and temporal (multi-rate and event-driven) downscaling approach within the recursive structure of Kalman filtering and develop Algorithm 1 (Fig. 1). Step 1 initializes the filtering, and steps 2–7 represent the Kalman filter. The multi-rate and event-driven measurement collection is represented in step 6. Step 8 represents active sensor measurement correction if necessary based on step 6. Finally, Kriging is performed in step 9 for the unobserved locations. \( \hat{x} \) is a \( p \times 1 \) dimensional state vector. Correspondingly, the error covariance matrix \( M \), state covariance matrix \( C_y \), state transition matrix \( F \), and the process noise covariance matrix \( C_y \) are all of dimension \( p \times p \). These matrices are defined in Section 2.1 and associated weather-dependent coefficients are found according to the method described in Section 4.1. The dimensions of observation matrix \( H(\Theta) \) and measurement noise matrix \( C_y \) are \( m \times p \), and \( m \times m \), respectively. In Section 4.2, we describe the least square fitting method to obtain these matrices. The matrices used in Kriging, are \( (N - m) \times p \) dimensional \( H(\Theta) \), and \( (N - m) \times m \) dimensional \( A \), which are obtained using the least square method (Section 4.2) and semivariogram fitting (Section 4.4). The relation between known data, and the models and algorithm described in this section are illustrated in Fig. 2. To summarise, the proposed algorithm adds novelty to the recursive structure of Kalman filter by integrating: (i) time varying weather pattern through distinctive set of parameters for state-space and observation models, (ii) temporal downscaling through an upgraded multi-rate and event-driven approach and (iii) spatial downscaling through Kriging. In the following section, we describe the database used to fit various parameters of the MREDRIKK algorithm and to analyse filter performance.

3 Oahu solar database

The Oahu Database consists of daily GHI data collected from 18 irradiance sensors [30]. The database covers one and a half year of observation (March 2010 through October 2011). Fig. 3 shows the geographic distribution of these sensors in the vicinity of Kalaeloa airport, Oahu, Hawaii – one of the very high solar variability zones [31]. The sensor nearest to the geo-centroid of this layout is the ‘AP01’, marked in Fig. 3. All sensors except ‘AP02’ measure GHI at 1 s resolution. However, the ‘AP02’ sensor collects GHI, temperature, pressure and wind speed at 1 min resolution. Thus, for consistency, the 1 s resolution data set is averaged to 1 min making the sampling interval, \( T = 1 \) min. The maximum separation is 0.7 miles between ‘DH01’ and ‘AP06’ indicating a circular area of...
~0.4 square miles. The number of residences covered by this would be around 1000 and 500 in a dense urban and suburban neighbourhood, respectively [32]. The ‘REST 2’ model [33] based half hourly clear sky GHI is then retrieved at a nearby location from the NSRDB database [34]. With the help of cubic spline interpolation, a 1 min resolution clear sky GHI is then obtained for each day. Since, the goal is to estimate PV generation, we use these spatial–time series of irradiance and weather data in synthesising the PV inverter AC power output profile at the specified locations of $N=18$ sensors. The longitude–latitude pair of these locations form the set $\delta$. At each location, the AC power output from a 2.5 kW inverter (with inverter loading ratio of 1.48 [35]) is calculated by feeding the Oahu Database into the PVLIB toolbox [36]. For 500 residences this would represent a 1.25 MW distributed PV in the dense suburban neighbourhood. Such power profiles obtained are then used to calculate the spatial time series of SPI

$$\text{SPI}(s_r, t) = \frac{P(s_r, t)}{P_{ac}(s_r, t)}, \quad s_r \in \delta$$  \hspace{1cm} (13)

Here, SPI stands for the SPI – a ratio of actual ($P$) and clear sky AC power ($P_{ac}$) [12], and is illustrated in Fig. 4 for the sensor ‘AP01’.

The formation of daily SPI time series is repeated for seven months from April 2010 to October 2010. For each day, the time horizon is kept between 7:07 AM and 5:06 PM totalling 600 time stamps at 1 min resolution. The particular time horizon is chosen so that it cuts off any spurious SPI value caused by the diminishing clear sky AC power in the denominator during sunrise and sunset periods. The aggregate SPI series thus obtained is named as the training data set and used in the next section to fit the system model. It consists of 214 realisations (i.e. 214 days of seven months) of SPI time series at 18 locations. Similarly the test data set is formed for April 2011 through October 2011, which is used in Section 5 to investigate the performance of the fitted model and MREDRIKK algorithm. The procedure of creating the training and test database from known information is summarised in Fig. 2, which also illustrates how these database are being used in the upcoming sections for model fitting, algorithm implementation and performance assessment. It should be noted that, the proposed approach leverages two tuning parameters, which can be adjusted for desired performance while only three types of information are required as input to implement the algorithm. No additional inputs are required for internal data processing to fit the associated state-space model.

4 Fitting of state-space model

From the training data set, we pick the 214 realisations of SPI$(s_{\text{min}}, t)$ to estimate various parameters of the state-space model. This choice is motivated by the closest vicinity of ‘AP01’ to the geo-centroid (Fig. 3). Each of these highly variable and non-stationary SPI time series is then broken into (600/$l_{\text{frame}}$) time frames for a given frame length, $l_{\text{frame min}}$. Thus, a total of (214 × 600/$l_{\text{frame}}$) time frames are used to perform cluster analysis.

The clustering is performed to identify distinctive weather patterns as represented by different clusters. In this regard, the squared Euclidean distance based $k$-means++ algorithm is chosen to perform unsupervised clustering of the entire set of time frames just formed. Such choice is motivated by its proven application in weather-dependent solar irradiance classification [37]. Now, due to the random initialisation, the algorithm is run multiple times and the clusters with the lowest within-cluster sums of point-to-centroid distances are identified [38]. For a given number of clusters being sought, the performance of $k$-means clustering is analysed by calculating the average silhouette coefficient (ASC) – a measure of a time frame's similarity to its own cluster as compared to other clusters. The value of ASC lies in between $-1$ and $+1$ and closer to the later indicates higher quality of clustering performed [39]. The entire process of $k$-means clustering, and ASC calculation is repeated for different number of clusters. Furthermore, the frame length $l_{\text{frame min}}$ is varied to create new set of time frames and fed to the clustering and ASC calculation process. It has been found that the ASC value becomes the highest for partitioning into two clusters, irrespective of the frame length, $l_{\text{frame min}}$. This indicates the existence of two possible weather patterns for the data set under consideration – a direct consequence of continuous intermittent cloud movement over such a high solar variability zone [31]. Consequently, we continue with two clusters to fit various models in the following subsections.

4.1 Fitting of system dynamics

We denote the set of parameters to be estimated for the dynamic model (1) as $\beta = \{\beta_0, \beta_1, \beta_2, \ldots, \beta_p, \ldots, \beta_p\}$. The partitioning into two clusters is repeated by varying the length of time frames and the corresponding ASC is shown in Fig. 5a. For each cluster and frame length, the ARFIT toolbox [40] is used to fit an AR model to the training data set while capturing the approximate 95% confidence intervals for the estimated parameters [24]. The set of intervals is denoted by $\beta_0$ and is normalised with respect to estimated parameters to calculate the metric, $\gamma \equiv 100 \times \beta_0 / \beta$. Fig. 5b shows the maximum and minimum values of $\gamma$ for each cluster. The narrowest range of confidence interval is observed for both clusters when the frame length is set to 12 min. Based on these observations, $l_{\text{frame}}$ is set to 12 min and both the training and test data set are broken into 12 min time frames and two clusters are obtained for each data set. Intuitively, the weather switches between two possible patterns after every 12 min and the subscript $t$ in the system model and Algorithm 2 captures this time varying behaviour. The clustered 12 min long training data set is used to fit the models (1), (2) and the cluster labels of the test data set are used as weather forecast while investigating the performance of proposed algorithm in Section 5.
4.2 Fitting of observation model

The observation model (3) is fitted to the clustered training data set through least square fitting. For each cluster \( C \), a concatenated time series is formed as a row vector using all the 12 min long cluster members. We denote its transpose as \( \mathbf{R}(\mathbf{s}_{AP1}, C) \). At the second step, the spatial observation from the remaining 17 sensors are broken into 12 min frames and labelled with the cluster of concurrent 12 min frame of \( \mathbf{SPI}(\mathbf{s}_{AP1}, t) \). Thus, the set of 12 min SPI series labelled as cluster \( C \) is denoted by \( \mathbf{R}(s, C) \) for the sensor located at \( s \). The concatenation of all series in \( \mathbf{R}(s, C) \) results in a row vector, whose transpose is denoted as \( \mathbf{b}(s, C) \). Therefore, the least square estimate of the observation coefficients for cluster \( C \) is given by [41]

\[
\mathbf{h}(s, C) = \frac{\mathbf{b}(s, C)^{\top} \mathbf{R}(\mathbf{s}_{AP1}, C)}{\mathbf{R}(\mathbf{s}_{AP1}, C)^{\top} \mathbf{R}(\mathbf{s}_{AP1}, C)}; C = \{1, 2\}
\]

(14)

The measurement noise variance is then modelled as the error variance of this fitting

\[
\sigma^2(s, C) = \text{VAR}(\mathbf{b}(s, C) - \mathbf{h}(s, C) \mathbf{R}(\mathbf{s}_{AP1}, C))
\]

(15)

With the prior knowledge about the set of active sensors, the observation matrices \( \mathbf{H}(C), \mathbf{H}(\mathit{C}) \) and noise covariance matrix \( \mathbf{C}_{\delta t} \) are then formed using the fitted coefficients

4.3 Calculating threshold \( \delta \)

We utilise the training data set to calculate the disparity threshold for each location \( s \) and cluster \( C \) as follows:

\[
\delta(s, C) = \max_{z \in \mathbf{R}(s, C)} \max_{1 \leq t \leq 12} \left| z(t) - z(t-1) \right|
\]

(16)

The threshold values are then chosen according to the cluster type for any given 12 min frame. It should be noted that the threshold calculation in (16) is rather comprehensive as compared to previous work of determining the event detection threshold in [10].

4.4 Semivariogram fitting

The sample semivariograms for each cluster \( C \) are calculated using the series set \( \{b(s, C), s \in S\} \). In fitting the exponential model (11), we vary the sum \( (c_0 + c_1) \) and use the calculated semivariograms to find the lowest fitting error. Fig. 6 shows the performance for both clusters and suggests the following choice for \( (c_0 + c_1) \)

\[
c_0 + c_1 = \begin{cases} 0.12; & \text{Cluster1} \\ 0.08; & \text{Cluster2}. \end{cases}
\]

(17)

Using these values, we obtain the exponential fitting of sample semivariograms of each clusters, which is illustrated in Fig. 7. Table 1 shows the fitted coefficients rounded to two decimal places. These coefficients are then used to calculate the spatial covariance matrix for simple Kriging expressed in (12).

5 MREDRIKK performance analysis

The fitted state-space models are applied to the MREDRIKK filter according to the cluster sequence of the test data set (April 2011 through October 2011). As mentioned in Section 4.1, for any time frame \( (j - 1)h_{\text{frame}} + 1 \leq t \leq jh_{\text{frame}}, j \in \mathbb{Z} \), the associated model parameters tagged as Cluster 1(or 2) are selected, if the jth cluster is 1(or 2). As time evolves, this parameter selection for Algorithm 1 is repeated every 12 min – the value of \( h_{\text{frame}} \) set for the database being used. With such formation, Fig. 8 illustrates the estimation and Kriging of SPIs for one of the days within the test data set, at locations of ‘DH07’ and ‘DH11’, respectively. In this example, ‘DH07’ represents one of the nine active sensors and ‘DH11’...
represents one of the locations to be Kriged. Now, for each day of the test data set, we calculate the mean absolute error (MAE) in estimation and Kriging, for the active and unobserved sensors, respectively

$$\text{MAE}(s_r) = \frac{1}{600} \sum_{t=1}^{600} (\hat{y}_t(s_r) - y_t(s_r))$$  \hspace{1cm} (18)

The above metric represents performance of estimation when $r \in O$ and Kriging when $r \in U$. The spatial average of MAE is then obtained for each day while varying the sets $O$ and $U$. The cardinality of $O$ is chosen from $\{3,6,9,12,15\}$, while $|U|$ is set to $9$. The gradual right shifts of CDFs in Fig. 9b indicates the increase in spatially averaged MAEs with the increase in decimation factor $D$ – a direct consequence of not allowing the measurement update for longer duration and utilising only the MMSE estimate.

5.1 Effect of active sensor volume

The estimation and Kriging performance for varying size of active sensor set is shown in Fig. 9a. For this investigation, $D$ and $\theta$ is set to 5 and 1, respectively. It is observed that the estimation and Kriging performance improve as more sensors are active while less sensors are needed to be Kriged. The increase in the number of active sensors leads to a better observability of state in (3) and hence better estimation for the active sensors can be achieved. Consequently, Kriging has to be done for fewer sensors from improved estimates, resulting in better Kriging performance.

5.2 Effect of decimation

Fig. 9b shows the effect of decimation on the MREDRIKK performance for a given number of active and Kriged sensors. The effect of detection sensitivity is neutralised by fixing $\theta = 1$ while $|O|$ is set to 9. The gradual right shifts of CDFs in Fig. 9b indicates the increase in spatially averaged MAEs with the increase in decimation factor $D$ – a direct consequence of not allowing the measurement update for longer duration and utilising only the MMSE estimate.

5.3 Effect of event detection

The effect of event detection sensitivity is shown in Fig. 9c by varying $\theta$ while keeping the decimation factor and active sensor volume at $D = 15$ and $|O| = 9$, respectively. It can be observed that both the estimation and Kriging performance improve as the event detection becomes more sensitive (i.e. $\theta$ is reduced). According to (7), a smaller value of $\theta$ causes the detection of greater number of SPI significant events by overriding decimation in measurement update. As a result, the estimation for the active sensors improves by capturing more frequent measurements leading to further betterment in Kriging. However, a careful observation of Algorithm 1 reveals that unlike Kriging (Step 9), the estimation for active sensors (Step 8) have direct impact from the measurement collection status $\phi(t)$. This characteristic can be observed in Fig. 9c as the improvement in estimation for active sensors is more significant than that for Kriging.

It is expected that, the number of SPI events detected as significant would increase when the event detection mechanism becomes more sensitive. However, the observability of active sensors can mitigate such detection based measurement augmentation. The box-plots in Fig. 10 show that this phenomena where the number of significant events reduces with the increase of active sensors. This is a direct consequence of the state observability, which is enhanced as more sensors are actively
our prior observation in Fig. 9c about more significant variation in estimation performance compared to Kriging.

- While the decimation has similar impact on both estimation and Kriging, the rate of performance degradation drops with the increase of decimation.

To summarise, the estimation and Kriging performance vary significantly with respect to active sensor volume, decimation as well as event detection threshold, which are validated through the KSTS values.

### 5.4 MREDRIKK versus persistent estimation

In this section, we compare the performance of MREDRIKK filter with the persistent estimation of distributed PV generation. The concept of ‘persistence’ is widely used to assess performance of solar forecasting methods [12], which motivates the derivation of persistent estimation in this research. This applies both to the active sensor measurement collection as well as estimation for unobserved sensors. For the active sensors, the persistence implies no change in the measurement during the decimation interval unless there is a significant event, and hence no measurement correction is required (in contrast to step 8 of Algorithm 1). Thus, from (8), the persistence in active sensor measurement is given by

\[
y_{\text{persist}}(x) = y(x)
\]

For the unobserved sensors, the persistence implies that the least-square fitted observation coefficients are sufficient to capture distributed PV generation from the intermediary states (in contrast to step 9 of Algorithm 1). Accordingly, (12) is modified as follows to give persistent estimation at unobserved sensors

\[
y_{\text{persist}}(x) = H(x)y_{\text{Krig}}
\]

Now, we obtain persistent estimation algorithm by replacing steps 8 and 9 of Algorithm 1 by (19) and (20), respectively. Next, the performance of MREDRIKK filter is compared with persistent estimation. In this regard, the global average MAE is obtained from the daily spatial MAEs (pre-calculated using (18)) for the entire test database. The global performance in estimation and Kriging thus obtained are shown in Fig. 11 for different size of active sensor set. It can be observed, that the MMSE estimation performs better when the filter receives measurement from at least nine sensors. On the other hand, the Kriging performs better irrespective of the number of sensors when the event detection threshold is set to \( \theta = 0.01 \).

### 6 Conclusion

A MREDRIKK filter is utilised for dynamic estimation of distributed PV generation in this research. The scalable estimation is performed in temporal domain through the multi-rate and event-driven measurement update, whereas the Kriging enables a scalable spatial estimation. A time series of SPI is used to define the behind-the-meter PV states for a set of irradiance sensors located in the proximity corresponding to a high PV penetrated residential neighbourhood. The weather-dependent state-space and semivariogram model have been fitted to publicly available high resolution irradiance database and the performance of the MREDRIKK filter has been thoroughly investigated by varying (i) the size of active sensor set, (ii) decimation factor and (iii) event detection sensitivity. The MREDRIKK filter shows improvement in both estimation and Kriging with the increase of active sensors while the Kriging is found to be less sensitive to event detection mechanism as compared to the decimation. The enhanced observability from increased number of active sensors is further illustrated through the reduction in significant SPI events irrespective of the sensitivity of event detection. The proposed filter would not only reduce utilities’ dependence on PV developers provided costly data service, or avoid incentives to access residential customers’ PV readings, but would also enable the best utilisation of available information in maintaining voltage stability.

---

**Table 2** KSTS for the estimation MAE CDFs

| Number of active sensors, [θ]| Effect of decimation \((θ = 1)\) | Effect of event detection \((D = 15)\) |
|-----------------------------|-----------------------------|-----------------------------|
| \(D = 5\) versus \(D = 10\) | \(θ = 0.5\) | \(θ = 0.1\) |
| \(D = 10\) versus \(D = 15\) | \(θ = 0.1\) | \(θ = 0.01\) |

- The average MAE performance does not change.

| Number | Effect of decimation \((θ = 1)\) | Effect of event detection \((D = 15)\) |
|--------|-----------------------------|-----------------------------|
| 3      | 0.3857 0.2048 0.919 0.5048   |
| 6      | 0.3762 0.2048 1 0          |
| 9      | 0.3619 0.2095 1 0          |
| 12     | 0.3524 0.2095 1 0          |
| 15     | 0.3619 0.219 1 0           |

**Table 3** KSTS for the Kriged MAE CDFs

| Number of sensors kriged, [θ]| Effect of decimation \((θ = 1)\) | Effect of event detection \((D = 15)\) |
|-----------------------------|-----------------------------|-----------------------------|
| \(θ = 0.1\) | \(θ = 0.5\) | \(θ = 0.1\) |
| \(θ = 0.5\) | \(θ = 0.1\) | \(θ = 0.01\) |

- The average MAE performance are different at significance level 0.02.

**Fig. 11** Global average MAEs for MREDRIKK filter and persistence

Collecting measurements, resulting in less override of the decimated measurement update. The impacts of decimation and event detection are further quantified by the Kolmogorov-Smirnov test statistic (KSTS) among the MAE CDFs while varying \([θ]\) and \([θ]\). The associated critical values for 64-bin histogram are 0.189 and 0.204 at significance levels of 0.02 and 0.01, respectively [42]. A pair of CDFs are significantly different if the corresponding KSTS is greater than the critical value. In order to compare the MREDRIKK filter performance under various settings of active sensor volume, decimation and event detection threshold, the KSTS values are obtained for both estimation and Kriging. Tables 2 and 3 represent the respective KSTS and lead to following observations:

- A zero value for KSTS in the last column of Table 2 indicates consistent estimation performance when at least six sensors are active. This is a direct consequence of achieving the best estimation performance (i.e. average MAE reaches zero) when \(θ \leq 0.1\). Fig. 9c illustrates this finding for \([θ]\) = 6. The rest of the KSTS values in Tables 2 and 3 are greater than the critical values indicating significant variation in both the estimation and Kriging performance.

- Under the effect of event detection, the KSTS values for estimation are greater than that for Kriging. This finding verifies...
7 Acknowledgment

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

8 References

[1] Mai, T., Wiser, R.H., Barbose, G.L., et al. ‘A prospective analysis of the costs, benefits, and impacts of US renewable portfolio standards’. 2016
[2] Enayati, J.: ‘Renewable distributed energy resources - technical benefits and challenges’, (IEEE Smart Grid Newsletters, 2017, Available at http://resourcescenter.smartgrid.ieee.org/tp/product/publications/SGNL0427
[3] Kalman, R.E.: ‘A new approach to linear filtering and prediction problems’, J. Basic Eng., 1960, 82, (1), pp. 35–45
[4] Rubio, J.D.J.: SOFMISL: online self-organizing fuzzy modified square networks’, IEE Trans. Fuzzy Syst., 2009, 17, (6), pp. 1296–1309
[5] Rubio, J.D.J., Lughofer, E., Meda Campaa, J.A.: ‘ ‘Estimation of parameters and eigenmodes of multivariate autoregressive models’, ACM Trans. Math. Softw., 2001, 27, (1), pp. 27–57, Available at http://doi.acm.org/10.1145/382043.382094
[6] Zou, L., Wang, Z., Guo, H., et al.: ‘Event-triggered state estimation for complex networks with mixed time delays via sampled data information: the continuous-time case’, IEE Trans. Cybern., 2015, 45, (12), pp. 2804–2815
[7] Sahu, S.K., Mardia, K.V.: ‘A Bayesian kriged kalman model for short-term forecasting of air pollution levels’, J. R. Stat. Soc.: Series C (Appl. Stat.), 2005, 54, (1), pp. 223–244, Available at http://dx.doi.org/10.1111/j.1467-9868.2005.00480.x
[8] Saengtupa, M., Andreas, A.: ‘Oahu solar measurement grid (1-year archive): 1-second solar irradiance upstate, Hawaii (data)’, (National Renewable Energy Laboratory, 2010, DA-5500-5650
[9] PV variability datasets’ [Dataset]. (Sandia National Laboratory, Available at https://pvpmc.sandia.gov/data/variability-datasets
[10] Sridhar, B., Smith, P., Suorsa, R.E.: ‘Global horizontal irradiance clear sky models: implementation and analysis’ (Sandia National Laboratories, NM, USA, 2012)
[11] ‘National solar radiation database (NSRDB)’ [Dataset]. (National Renewable Energy Laboratory, Available at http://nsrdb.nrel.gov/
[12] Bolinger, M., Seel, J.: ‘Utility-scale solar 2015 utility-scale solar 2015: an empirical analysis of project cost, performance, and pricing trends in the United States’ (Lawrence Berkeley National Laboratory, CA, USA, 2016), LBNL-100637
[13] Stein, J.S., Holmgren, W.F., Forbes, J., et al.: ‘PVLIB: open source photovoltaic performance modeling functions for Matlab and python’. 2016 IEEE 43rd Photovoltaic Specialists Conf. (PVSC), 2016, pp. 3425–3430
[14] Bae, K.Y., Jung, H.S., Sung, D.K.: ‘Hourly solar irradiance prediction based on support vector machine and its error analysis’, IEEE Trans. Power Syst., 2017, 32, (2), pp. 935–946, Available at http://dx.doi.org/10.1109/TPWRS.2016.2584377
[15] Arthur, D., Vassilvitskii, S.: ‘K-means++: the advantages of careful seeding’ Proc. 18th Annual ACM-SIAM Symp. on Discrete Algorithms. SODA ’07. (Philadelphia, PA, USA, Society for Industrial and Applied Mathematics, New Orleans, LA, USA, 2007, pp. 1027–1035, Available at http://dl.acm.org/citation.cfm?id=1283383.1283494
[16] Rousseau, P.J.: ‘Silhouettes: A graphical aid to the interpretation and validation of cluster analysis’, J. Comput. Appl. Math., 1987, 20, pp. 53–65, Available at http://www.sciencedirect.com/science/article/pii/0377042787901257
[17] Schneider, T., Neumaier, A.: ‘Algorithm 808: Arfit—a Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models’, ACM Trans. Math. Softw., 2001, 27, (1), pp. 58–65, Available at http://doi.acm.org/10.1145/382043.382316
[18] Demmel, J.: ‘Applied numerical linear algebra’, (Society for Industrial and Applied Mathematics, 1997), Available at http://epubs.siam.org/doi/abs/10.1137/1.9781611971446
[19] O'Connor, P., Kleyner, A.: ‘Appendix 3: Kolmogorov-Smirnov tables’ (John Wiley & Sons, Ltd., United Kingdom, 2012), pp. 455–456

and reliable grid. In future work, we plan to further investigate and compare the performance of MREDRIKK filter with a much larger spatial irradiance database of comparable resolution and variability. Furthermore, the data-driven design of the MREDRIKK filter for distributed PV generation estimation will be extended to a distribution feeder scale with the incorporation of standard decentralised optimisation approaches.