Short-term irradiance forecastability for various solar micro-climates

Hugo T.C. Pedro, Carlos F.M. Coimbra

Department of Mechanical and Aerospace Engineering, Jacobs School of Engineering, Center of Excellence in Renewable Resource Integration and Center for Energy Research, University of California, San Diego, La Jolla, CA 92093, USA

Received 25 January 2015; received in revised form 18 September 2015; accepted 20 September 2015
Available online 10 November 2015
Communicated by: Associate Editor Christian Gueymard

Abstract

The purpose of this work is to present a simple global solar irradiance forecasting framework based on the optimization of the $k$-nearest-neighbors (kNN) and artificial neural networks algorithms (ANN) for time horizons ranging from 15 min to 2 h. We apply the proposed forecasting models to irradiance from five locations and assessed the impact of different micro-climates on forecasting performance. We also propose two metrics, the density of large irradiance ramps and the time series determinism, to characterize the irradiance forecastability. Both measures are computed from the irradiance time series and provide a good indication for the forecasting performance before any predictions are produced. Results show that the proposed kNN and ANN models achieve substantial improvements relative to simpler forecasting models. The results also show that the optimal parameters for the kNN and ANN models are highly dependent on the different micro-climates. Finally, we show that the density of large irradiance ramps and time series determinism can successfully explain the forecasting performance for the different locations and time horizons.

Keywords: Global horizontal irradiance (GHI); $k$-nearest-neighbors (kNN); Time series features; Forecastability

1. Introduction

Developing high fidelity solar forecasting methods is a nontrivial effort because of the highly nonlinear and generally poorly predictable effects of cloud cover on ground irradiance (Fouquart et al., 1990). The occurrence, shape and optical depth of amorphous clouds and the resulting irradiation attenuation seem, at first sight, random and unpredictable. However, a number of well understood atmospheric phenomena relate cloud patterns to ground topography, preferential wind patterns and recurrent weather systems. Such patterns will be imprinted in the measured irradiance time series and will influence the performance of forecasting models.

Cloud patterns are not always apparent from ground level due to the short time horizon of the observable cloud sky, which with typical cloud speeds is often smaller than 1 h. However, even from the ground it is common to observe periodic cloud formations, especially when they result from pressure fluctuations that arise from geological features on the ground. Fig. 1 shows a few examples of cloud patterning that are frequently observed over both land and ocean surfaces. The left picture in Fig. 1 shows a periodic cloud formation on the background of the Moul n’ga Cirque in the Tadrart region of Algeria. These wave clouds are formed when air flows over ridged land features such as a mountain range or a combination of mesas and valleys, or over a single mount or valley. The clouds, which are nucleated in periodic patterns, form either directly above the land feature or past the ridge in the downstream direction due to horizontal transport. Such cloud
formations produce periodic irradiance fluctuations at the ground level that are easily registered by radiometers, and that affect the production of solar power by all current technologies.

The middle picture in Fig. 1 shows very distinctive trains over the Appalachian Mountains as photographed from a cruising airplane with a smart phone by Mr. Seth Adams. Here the cloud formations follow the periodic ground ridges in a very clear demonstration of how geological phenomena that occurred at very large time scales (the formation of the Appalachians) can influence atmospheric phenomena at much shorter time scales. These cumulus cloud formations are the result of upward thermal flows caused by solar heating of the land surface, which in turn affect the solar irradiance at ground level by imprinting patterns in the radiation time series. In effect patterns derived directly from geological phenomena that occurred millions of years before.

Finally, the right picture in Fig. 1 shows complex pattern formation as photographed by cosmonaut Dr. Fyodor Yurchikhin from the International Space Station. These formations are made of many smaller cloudlets and typically occur below the tropopause, but often above 2 km. A shear layer with periodic vortical structures is clearly visible at the meeting point of two ocean air masses that travel at different speeds.

In this work we aim at producing solar irradiance forecasts that capitalize on local cloud patterns embedded in (and recoverable from) the irradiance data. Furthermore, in order to study the effect of cloud patterns on the ability to produce endogenous solar forecasts we consider global horizontal irradiance (GHI) data from five distinctive locations in the USA: San Diego, Davis, and Merced in California, Bellingham in Washington, and Ewa Beach in Hawaii. The GHI datasets were collected for a period of one year with precision spectral pyranometers (PSPs),
multi-filter radiometers (MFR-7s), and/or rotating shadowband radiometers (RSR2s) that recorded the broadband solar incidence on a horizontal planar surface every 30 s.

One of main objectives of this work is to maximize the forecast performance that can be obtained directly out of the irradiance measurements through times series feature extraction, and to understand how it varies based upon local climate characteristics. Thus, we propose a simple forecasting framework based on the optimization of the kNN and ANN algorithms to enable a detailed search for the best endogenous forecasting model for time horizons ranging from 15 min to 2 h. We analyzed the role of temporal memory, variability and entropy in the irradiance time series, and produce maps of features that best describe each location and forecast horizon. The results obtained are robust and representative of a wide range of micro-climates, which makes the proposed methodology readily applicable to other locations.

The other objective is to propose a method based on the irradiance times series to measure the GHI forecastability, that is to assess the level of forecast error one can expect before any forecast model is implemented and to compare the degree of difficulty in achieving good forecast results for different micro-climates. Such metric could be very helpful in comparing the myriad of forecasting studies in the solar energy literature.

2. GHI data

The five selected locations cover a wide range of micro-climates and ground topologies, thus allowing us to study the effect of local weather patterns on the forecasting performance with a greater degree of generality. These locations cover the solar irradiance variability range more or less evenly. On the low end of this range we have Merced and Davis. These two sites are located in the heart of California’s central valley and enjoy dry, cloudless summers and little solar variability for most of the year. Next, in the solar irradiance variability scale we have San Diego and Bellingham with much more variable irradiance mainly due to their proximity to the Pacific ocean and associated weather phenomena such as the marine layer. On the high end of the range we have Ewa Beach, on the leeward side of Oahu in Hawaii, which enjoys very few cloudless days throughout the year mostly due to local weather formations across the island ridges.

2.1. San Diego

The San Diego high-fidelity solar observatory at the University of California, San Diego campus in La Jolla, CA (latitude: 32.88°; longitude: −117.23°) was established in March 2012 and includes a Yankee Environmental Systems MFR-7 that was used in this study, with several other pyranometers and photodiode sensors that provide redundancy and data quality assurance. The climate in La Jolla is typically semi-arid Mediterranean. Large differences on short distances occur due to the topography of the coastal region, which leads to distinct micro-climates in close proximity. The solar sensors are located within 2 km of the coastline, at 100 m above mean sea level. The observatory lies well within the region influenced by coastal marine layer clouds. Under certain conditions, these marine layer clouds form over the Pacific Ocean and advect inland, where they can remain for days, a distinctive feature of the La Jolla region from May through August. Based on the Köppen climate classification scheme this location near the coastline is either referred to as Csa (C = temperate climate s = dry summer a = hot summer) or Csb (C = temperate climate s = dry summer b = warm summer).

2.2. Merced

The Merced solar observatory is located on the University of California (UC) Merced campus (latitude: 37.36°; longitude: −120.42°). The Köppen climate classification for Merced is Csa. Such climates are characterized by hot, dry summers and wet, cold winters. Merced is located close to the geodesic center of California. A common weather phenomenon in this area is the tule fog, a dense formation of ground fog that is observed normally between November and April and that causes significant ground cooling and drizzling conditions.

2.3. Davis

Data at the location in Davis has been acquired with a MFR-7 (latitude: 38.53°; longitude: −121.78°). Located in the northern portion of the Central Valley of California, the climate in Davis is similar to Merced (also characterized as Csa on the Köppen scheme). Davis is a remote location with no on-site staff, therefore we expect soiling effects to play a role over time, especially during the summer months. Manual inspection of the whole dataset for Davis was performed to remove periods where either misalignment or soiling compromised data quality.

2.4. Bellingham

Bellingham (latitude: 48.76°; longitude: −122.48°) has a marine oceanic climate strongly influenced by the Puget Sound, the Cascade range and the Olympic mountains and its Köppen classification is Dfa (D = continental climate f = moist a = hot summer). Bellingham has one of the most cloud-covered skies in the continental US, and its climate is strongly influenced by Southern winds in the fall, and Northeast winds during the winter. The Bellingham station is equipped with a RSR2 shadow-band radiometer located near the waterfront. The instruments were inspected and maintained throughout the period of performance.
2.5. Ewa Beach

The data for Ewa Beach (latitude: 21.24°; longitude: -158.04°) was collected with a RSR2 shadow-band radiometer. A wide variety of micro-climates can be observed in Oahu, a result from the high relief of the island’s orography and the prevalence of the northeasterly trade winds during most of the year. Ewa Beach Köppen classification is As (A = tropical climate s = wet and dry).

3. Data analysis

Data collected in these locations was filtered by considering only timestamps with solar elevation higher than 5°, thus removing night values. The filtered data was then characterized in several ways. Fig. 2 shows the clear-sky index for GHI, \( k^*_t \) versus normalized time \( \omega \) (where \( \omega \) varies linearly from -1 at sunrise to 1 at sunset, with \( \omega = 0 \) at solar noon) with \( k^*_t \) defined as:

\[
k^*_t = \frac{G_t}{G_{clr}}
\]

where \( G_t \) is the GHI measured at time \( t \) and \( G_{clr} \) is the clear-sky irradiance computed with the model explained in Appendix A. This variable is easier to interpret than GHI since it has a very well defined range \( (k^*_t = 1 \text{ signifies irradiane equal to the clear-sky value and } k^*_t = 0 \text{ signifies no irradiane due to overcast conditions}) \) whereas the GHI’s range varies depending the latitude of the location. Under certain conditions \( k^*_t \) can be higher than 1, for example, on days with bright blue skies and scattered clouds, the solar beam refraction from the cloud edges can concentrate the sunlight resulting in a boost to the irradiance level - a phenomena known as cloud edge effect.

The scatter plots of \( k^*_t \) versus \( \omega \) and the whiskers that identify the 0, 25, 75, 100 percentiles of \( k^*_t \) for a given \( \omega \) in Fig. 2 allow immediately to identify big differences in the GHI for the five locations. Both Merced and Davis (Fig. 2(a) and (b)) show a concentration of \( k^*_t \) values around 1 indicating that a large percentage of the data corresponds to clear-sky periods. The other three locations show much more \( k^*_t \) values below 1 indicating higher frequency of cloudy periods, as indicated by the large gap between the 25 and 75 percentiles. These plots even allow to observe some diurnal trends in the data. That is most clear in the case of San Diego (Fig. 2(c)) in which we can observe that in the morning \( (\omega < 0) \) cloudy periods are much more frequent, a result of the marine layer that affects this location.

Fig. 2. GHI characterization for each one of the five locations in this study. From (a) to (e): Merced, Davis, San Diego, Bellingham and Ewa Beach. (f) identifies the locations in the USA map. For each location (a) to (e) it is shown the (left) GHI clear-sky index \( k^*_t \) versus normalized time \( \omega \). The scatter green dots show all the data used in this work. To create this figure the range \( \omega \in [-1, 1] \) was divided into 50, equally spaced bins. The black dots represent the medians for each bin and the endpoints of the whiskers represent (from bottom to top) the 0, 25, 75 and 100 percentiles for the data in each bin. Also for each location it is shown (right) the scatter plot of daily \( k^*_t \) variability versus daily \( k^*_t \) average and their respective histograms.
The same figure also shows scatter plots of the daily \( k_t \) average and daily \( k_t \) variability (taken as the standard deviation for the \( k_t \) values for each day) and the respective histograms. Again, large differences are visible between the different locations. Merced and Davis (Fig. 2(a) and (b)) show a clear clustering near the \((1,0)\) point, indicating a large percentage of clear-sky days. San Diego and Davis show a much more even data distribution and Ewa Beach shows a concentration of data in the vicinity of \((0.8,0.3)\) which indicates a prevalence of cloudy conditions and high variability in the GHI.

### 3.1. Forecastability

Although the previous analysis is useful to characterize the solar irradiance from the different locations it is incomplete from the point of view of forecasting. This results from the fact that the information in Fig. 2 is a function of the instantaneous data: it provides no information about step changes in the irradiance which affect the performance of forecasting models. Thus, in order to quantify the changes in \( k_t \) we define the variable:

\[
\epsilon_t = \left| \langle k_t \rangle_{[t,t+\tau]} - \langle k_t \rangle_{[t-\tau,t]} \right|,
\]

where \( \langle \rangle_{[t,t+\tau]} \) denotes the average in window \([t,t+\tau]\), where \( \tau \) corresponds to the forecast horizon \((\tau = \{15,30,45,60,90,120\})\) minutes).

If \( \epsilon = 0 \) the average irradiance remains unchanged in the period \([t-\tau,t+\tau]\) and any forecasting model that uses the information from the period \([t-\tau,t]\) should return the perfect forecast. As \( \epsilon \) increases the variability of \( k_t \) increases, the performance of persistence forecast decreases and forecasting in general becomes more challenging.

We computed \( \epsilon \) for the different forecast horizons for all locations and determined the probability density function (PDF) and the cumulative density function (CDF) for each case. The results are shown in Fig. 3 for the two extreme cases of Merced and Hawaii. The faster the PDF decays to 0 or, equivalently, the faster the CDF approaches 1 the fewer are the instances of large variations, or ramps as commonly denoted in solar energy literature, present in the irradiance time series. In order to quantify the differences between the locations and forecast horizons we computed the area above the CDF denoted as \( \epsilon^* \). This value measures the density of large ramps in the time series and is defined as:

\[
\epsilon^* = \int_0^\infty 1 - \text{CDF}(\epsilon) \, d\epsilon,
\]

The simple interpretation for this value is that small values for \( \epsilon^* \) indicate a small density of large ramps in GHI which should result in a larger forecastability. Fig. 4(a) plots the \( \epsilon^* \) for all locations and forecast horizons. The figure shows \( \epsilon^* \) increases with forecast horizon, denoting that forecast
becomes more challenging. The most distinguishable feature in this figure relates to the values for Ewa Beach. For windows below 90 min the values for Ewa Beach are clear outliers when compared to other locations. This anticipates that the highly variable irradiance observed there will be challenging to forecast.

Another way to study the characteristics of the time series stems from the concept of recurrence introduced by Poincaré which can be quantified through recurrence plots introduced by Eckmann et al. (1995) and comprehensively reviewed by Marwan et al. (2007). The recurrence plot for the time series is defined as:

\[ R_{ij}(\varepsilon) = H\left(\varepsilon - \|\vec{x}_i - \vec{x}_j\|\right), \quad i, j = 1, \ldots, N \]

where \( H(\cdot) \) is the Heaviside function, \( \vec{x}_i \) the state vector formed by elements from the time series, \( N \) the number of instances, \( \| \cdot \| \) is a norm (the euclidean norm here), and \( \varepsilon \) is a threshold distance (set to 0.01 in this case). If \( R_{ij} = 1 \) it means that \( \vec{x}_i \approx \vec{x}_j \) and we say that the \( i \) state is recurrent. In this case, keeping the focus of this analysis in ramp events, the state vector is defined as:

\[ \vec{x}_i = \left[\{k_i^s\}_{t_i - s_i}, \{k_i^g\}_{t_i - s_i + 1}\right] \]

Once the recurrence plot is determined there are several metrics that convey the characteristics of the time series (Marwan et al., 2007). In this work we use the determinism metric denoted as \( \text{DET}_s \) since it measures the level of predictability of the system:

\[ \text{DET}_s = \frac{\sum_{l=l_{\text{min}}}^{N} P(l)}{\sum_{l=1}^{N} P(l)} \]

where \( P(l) \) is the histogram of diagonal lines of length \( l \) for the recurrence plot \( R_{i,j} \), and \( l_{\text{min}} = 2 \) in this case. \( \text{DET}_s \) varies between 0 and 1, with high \( \text{DET}_s \) corresponding to high determinism. This and other metrics for the recurrence plot can be readily obtained using the “Cross Recurrence Plot Toolbox” available at Potsdam Institute for Climate Impact Research (2015).

Using this toolbox we computed \( \text{DET}_s \) for all locations and forecast horizons and plotted them in Fig. 4(b). The values in this figure show that, although the variable \( \text{DET}_s \) is not very sensitive to the forecast horizon/averaging window, it appears to be a very good metric to characterize the GHI from the different locations: the values for each location fall into very narrow ranges regardless of \( \tau \). Conversely, \( \epsilon^s \) depends strongly upon the averaging window, and it also allows to differentiate from location to location but in a much less clear manner than \( \text{DET}_s \). For instance, the relative position between \( \epsilon^s \) for San Diego and Bellingham switches with increasing \( \tau \). Thus, with this two variables we have \( \text{DET}_s \), that characterizes the GHI depending on the location and \( \epsilon^s \) that characterizes the type of ramps that the forecast models must predict. As we show below in the results section these variables can be used as a good indicator for the forecast performance.

### 4. Forecasting models

Having concluded the above data analysis we then explored techniques to improve the forecasting of GHI for short forecast horizons ranging from 15 to 120 min. The data for the five locations were split into three different subsets: training, validation and testing. The training subset was used to create a database of features for the kNN and ANN, the validation subset was used to determine the optimal set of free parameters for the different forecasting models (discussed below) and the testing subset was used as an independent dataset for evaluating the forecast accuracy. The results and discussion below are based on the predictions for the testing dataset. In order to guarantee that the three datasets include data from similar irradiance periods we selected the first week of each month as testing data, the second week as validation data and the remaining as training (or historical) data, resulting approximately in a typical 25%, 25% and 50% partition of the data. Table 1 lists the metadata for the different datasets and locations. In the remainder of the article the three datasets are referenced by the superscripts \( T \), \( V \) and \( H \) respectively.

Predictions for the different locations and forecast horizons \( \tau = \{15, 30, 45, 60, 90, 120\} \) were obtained with four models: persistence, optimized persistence, optimized kNN and optimized ANN. These models use no exogenous variables such as other irradiance components, temperature

---

**Fig. 4.** (a) \( \epsilon^s \) computed with Eq. (3) and (b) \( \text{DET}_s \), for all locations and forecast horizons.
or sky images. All these models can be expressed in the general form:
\[
\hat{G}_t^f = r^f \langle G_{clr} \rangle_{[t,t]}^f
\]  
(6)

where \( \langle G_{clr} \rangle_{[t,t]} \) is the average GHI under clear-sky conditions for the interval \([t, t + \tau]\) where \( \tau \) represents the forecast horizon, and \( \hat{G}_t^f \) is the forecasted GHI obtained with model \( f \). The factor \( r^f \), described below, depends on the forecasting model, and represents the attenuation/intensification of the clear-sky GHI.

The forecasting performance is measured based on the forecast error, which for a model \( f \) and forecast horizon \( \tau \) is given by
\[
\varepsilon_t^f = \hat{G}_t^f - \langle G \rangle_{[t,t]}.
\]  
(7)

Several error metrics were used to assess the performance of the different models: the root mean square error (RMSE):
\[
\text{RMSE}_t^f = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\varepsilon_{t,i}^f)^2};
\]  
(8)

the mean absolute error (MAE):
\[
\text{MAE}_t^f = \frac{1}{N} \sum_{i=1}^{N} | \varepsilon_{t,i}^f |;
\]  
(9)

and the mean bias error (MBE):
\[
\text{MBE}_t^f = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_{t,i}^f.
\]  
(10)

A fourth metric, the forecast skill, measures the percentage improvement in RMSE relative to a reference model, in this case the persistence model:
\[
s_t^f = \left(1 - \frac{\text{RMSE}_t^f}{\text{RMSE}_t^f}\right) \times 100.
\]  
(11)

4.1. Persistence model

The persistence model is the simplest method for predicting the future behavior of a time series. As the name indicates, persistence implies that future values of the time series are calculated assuming that current conditions, the conditions at the time the forecast is issued, remain constant in the forecast window \([t, t + \tau]\). Here, these conditions are defined as the average \( k^* \) for GHI over the window \([t - p, t]\), where \( p \) denotes the history or memory length. For a stationary time series – a time series whose mean and variance do not change over time - a straightforward implementation of the persistence model is the naive persistence model:
\[
\hat{G}_t^f = \langle G \rangle_{[t-p,t]}
\]  
(12)

with the history length equal to the forecast horizon \( (p = \tau) \). However, solar irradiance at the ground level is clearly non-stationary because of diurnal and seasonal cycles. A simple and effective approach to circumvent this issue is to describe the variable with respect to clear-sky conditions using the clear-sky index \( k^* \). The average value of the clear-sky index for the window \([t - \tau, t]\) is calculated as
\[
\langle k^*_t \rangle_{[t-\tau,t]} = \frac{\langle G \rangle_{[t-\tau,t]}^*}{\langle G_{clr} \rangle_{[t-\tau,t]}}.
\]  
(13)

Using this variable we can define a new condition for persistence that assumes: \( \langle k^*_t \rangle_{[t-\tau,t]} = \langle k^*_t \rangle_{[t-\tau,t]}^* \). With this assumption and using (13) the factor in (6) becomes:
\[
r^p = \langle k^*_t \rangle_{[t-\tau,t]}
\]  
(14)

A question that arises from this model is whether \([t - \tau, t]\) is the best window to characterize the persistence conditions. Given that there is no reason why the memory length \([t - \tau, t]\) should be coupled to the forecast window \([t + \tau, t]\) we explored an optimized persistence model
\[
r^{p*} = \langle k^*_t \rangle_{[t-p,t]}
\]  
(15)

where the variable \( p^* \) is optimized via an exhaustive search in order to minimize the RMSE for the validation dataset:
\[
\arg\min_{p^*} \sqrt{\frac{1}{N^v} \sum_{i=1}^{N^v} (\hat{G}_i^f(p^*) - \langle G \rangle_{[t_i, t_i + \tau]}^*)^2},
\]  
(16)

with \( p^* \in \{5, 10, \ldots, 120\} \). The RMSE is chosen as the objective function for this problem since it emphasizes the largest errors.

4.2. k-Nearest neighbors

The kNN method is one of the simplest and most robust methods of pattern recognition in machine learning. In contrast to auto-regressive statistical methods such as AR and ARMA that attempt to find models from training
data, kNN uses the training data as the model. The kNN algorithm was originally developed for pattern classification but its simplicity allows it to be easily applied to regression problems for time series (Yakowitz, 1987), such as this case. Although there are very few articles in literature that apply kNN to the forecasting of solar irradiance (see for instance Paoli et al., 2010; Pedro and Coimbra, 2012), kNN has been extensively applied to forecast electricity load and electricity market price (Lora et al., 2002, 2003), water temperature in rivers (St-Hilaire et al., 2012), water inflow (Akbari et al., 2011), and weather (Bannayan and Hoogenboom, 2008).

For the purpose of forecasting time series the kNN model consists of identifying the instances in the past that resemble the current conditions most closely – the nearest neighbors. Once they are determined the forecasting is computed from the time series values subsequent to the kNN search.

The first step in developing a kNN model is to create a database of features that will be used in the comparison with current conditions. As mentioned before, in this work, the only available information is historical GHI data, thus all features were derived from that variable. In this work we used three features computed from the \( k_t \) time series to determine the nearest-neighbors: the cumulative backward average

\[
B_{ij} = \frac{1}{N} \sum_{i \in [k-p_t,t]} B^*_i, \quad (17)
\]

variability

\[
V_{ij} = \sqrt{\frac{1}{N} \sum_{i \in [k-p_t,t]} (\Delta k^*_i)^2}, \quad (18)
\]

and entropy

\[
E_{ij} = -\frac{N_{max}}{k-1} \sum_{k=1}^{N_{max}} f_k \log(f_k) \quad (19)
\]

where \( N \) is the number of data points in the interval \([t_i - p_j, t_i]\), and \( f_k \) is the relative frequency of \( k^*_i \) in the same interval \([t_i - p_j, t_i]\) for the \( k \)th bin out of 100 bins in the range [0,1.5]. The variable \( p_j \) specifies the window for which these features were calculated, ranging from 5 to 120 min in 5 min steps. For a fixed \( t_i \) these equations result in 24 values that are assembled row-wise into matrices with as many rows as the timestamps in the dataset. Fig. 5 shows these features for three windows in a day in San Diego.

These three features were computed for the three datasets (training, validation and testing) and then concatenated into a single matrix of features:

\[
F = \begin{pmatrix}
  w^B_{ij} B_{ij} & w^V_{ij} V_{ij} & w^E_{ij} E_{ij}
\end{pmatrix}, \quad \forall i \in \{1, \ldots, N\} \land j \in \{1, \ldots, 24\}
\]

\[
(20)
\]

where the symbol \( \vec{\cdot} \) indicates the column-wise normalized feature matrices, and \( N \) is the number of rows in the feature matrices. For instance, feature \( B_{ij} \) is normalized as

\[
\tilde{B}_{ij} = \frac{B_{ij} - \min(B_{ij})}{\max(B_{ij}) - \min(B_{ij})}.
\]

where \( B_{ij} \) denotes the entire column \( j \) of matrix \( B \). The normalization is necessary given that, as seen in Fig. 5, the features have very distinct ranges. Without normalizing the matrix of features the kNN would be biased towards features with larger magnitudes. The weights \( w^B, w^V, \) and \( w^E \) are vectors with as many elements as the number of columns in each one of the matrices of features, which is 24 in this case. The elements of the weight vectors vary from 0 to 1 controlling the relative importance of different features in the kNN search.

Once the list of \( k \) nearest-neighbors is obtained from computing the distance – the euclidean distance \( d_i \), in this case, of the rows in the validation or testing matrix of features relative to the training matrix of features, the factor from Eq. (6) can be calculated as:

\[
r_{kn}^{\text{kn}} = \frac{\sum_{i=1}^{k} x_i \langle k^*_i \rangle_{[d_i,z]}^m}{\sum_{i=1}^{k} z_i}, \quad i = 1, \ldots, k
\]

where \( \langle k^*_i \rangle_{[d_i,z]} \) are the instances in the training data that correspond to the selected neighbors, sorted in ascending order in terms of the corresponding distance \( d_i \). The values \( x_i \) are used to increase the weights of neighbors associated with smaller distances, and are computed as:

\[
x_i = \left( \frac{\max(d) - d_i}{\max(d) - \min(d)} \right)^m, \quad i = 1, \ldots, k
\]

with \( m \) as a free integer parameter that controls the weight distribution. For instance, for \( m = 0 \) the weights are 1 for all values and for \( m = 1 \) the weights decrease linearly from 1 to 0 with increasing distance. This algorithm was implemented in Python using the package scikit-learn (Pedregosa et al., 2011) and the K-D Three algorithm (Bentley, 1975) with leaf-size = 30.

4.3. Artificial Neural Networks

Artificial Neural Networks have been widely employed in forecasting problems (Marquez and Coimbra, 2011; Mellit and Pavan, 2010; Chu et al., 2013). Extensive reviews on the forecasting with ANNs can be found in Inman et al. (2013) and Mellit (2008).

The fundamental element of ANNs is the neuron that maps the input variables onto the output. Neurons are arranged in layers: the input layer that receives the input variables, the output layer that produces the output and the layers in between, denoted as hidden layers. A neuron receives a weighted and biased sum of the inputs and produces the output by applying an activation function to the weighted sum. Inputs to a neuron originate externally or internally from the output of the other neurons. Once the
ANN structure (number of layers, number of neurons, activation functions and so forth) is set the ANN undergoes a training process in which the weights and biases are adjusted so that the minimization of a performance function is achieved, usually the mean square error (MSE) between targets and outputs.

In this case the ANN is used to determine the factor in Eq. (6)

\[ r_{\text{ann}} = f \left( \mathbf{F} (w^\theta, w^\gamma, w^\xi) \right) \]

using as inputs the matrix of features \( \mathbf{F} \) as defined by the weights \( w^\theta, w^\gamma, \) and \( w^\xi \) (Eq. (20)). The function \( f \) is determined by the ANN specifications. In this work we used a feedforward network, with a variable number of inputs (columns of \( \mathbf{F} \) with weight 0 are removed), a single value output layer (the factor \( r_{\text{ann}} \)) that uses the linear transfer function, and a single hidden layer with 10 neurons that uses the hyperbolic tangent sigmoid transfer function. The ANN is trained with the Levenberg–Marquardt algorithm with early stopping. The target data for training the ANNs consists of the measured values \( \{ k^*_j \}_{k=1}^{K} \) for the training dataset. The ANN model was implemented in Matlab using the Neural Network toolbox 8.2.

4.4. kNN and ANN optimization

In the kNN and ANN algorithms explained above there are several free parameters. In the case of kNN these consist of the number of nearest-neighbors \( k \), the exponent \( m \) in the nearest-neighbors averaging, and the features to include in the kNN search, which are specified by the weight vectors \( w^f, w^g, \) and \( w^e \). In the case of ANN the free parameters consist of the weights \( w^\theta, w^\gamma, \) and \( w^\xi \). These parameters are determined via an optimization algorithm that seeks to minimize the forecast RMSE over the validation dataset:

\[
\hat{x} = \arg\min_{\hat{x}} \sqrt{\frac{1}{N^v} \sum_{i=1}^{N^v} \left( \hat{G}_{n}^f(\hat{x}) - \langle G_{n} \rangle \right)^2},
\]

where the vector \( \hat{x} \) consists of the search space for the optimization algorithm and contains the aforementioned free parameters. The optimization method used to determine \( \hat{x} \) is based on the genetic algorithm (GA) (Mitchell, 1998) implementation available in the package Distributed Evolutionary Algorithms in Python (Fortin et al., 2012). We used a heuristic optimization tool for this task since gradient-based algorithms are not well suited for this type of search space. The drawback is that the GA does not guarantee obtaining a global optimum.

The elements of the vector \( \hat{x} \) and the different GA settings are listed in Table 2. The GA computations were run in parallel using the package Scalable Concurrent Operations in Python (Hold-Geoffroy et al., 2014).

5. Results and discussion

5.1. Model optimization

The memory length \( p^* \) for the optimized persistence model is shown in Table 3 as a function of the location and forecast horizon. The table also shows the improvement relative to the original persistence model for the validation and testing sets. In most cases there is a small improvement in the forecasting for the testing set. In a
few cases the optimized persistence is actually worse than the original model. Appreciable forecast improvement was only obtained in the San Diego case. With the exception of Ewa Beach the optimal persistence is always smaller than the forecast horizon, indicating that most of the times shorter memory lengths describes the current conditions more accurately than longer ones. In this respect the case of San Diego is particular noteworthy: for all forecast horizons the optimal length is never higher than 10 min. In the case of Ewa Beach the opposite is true, possibly a consequence of the much higher GHI variability observed in that location.

Regarding the GA optimizations for the kNN and ANN models, we observed that they converge very quickly. Fig. 6 illustrate this fact for the optimization of the ANN-based model for Ewa Beach for the forecast horizons 15, 45, and 120 min. The figure shows the fitness range (colored areas) and the average fitness (solid lines) for the three cases. The convergence plots show that after 10 generations bad performing individuals have been eliminated from the population and that after 20–30 generations the average fitness of the population improves very little with increasing number of generations.

The optimal set of parameters for the kNN and ANN is taken from the best individual in the last generation available. These parameters are depicted in Fig. 7. For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.

Table 2
<table>
<thead>
<tr>
<th>GA search space</th>
<th># elem.</th>
<th>Type</th>
<th>Range</th>
<th>Description</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Integer</td>
<td>5 to 200</td>
<td>Number of nearest neighbors</td>
<td>kNN</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Integer</td>
<td>0 to 5</td>
<td>Weight in the kNN averaging</td>
<td>kNN</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Real</td>
<td>0 to 1</td>
<td>Weights for feature $B$</td>
<td>kNN, ANN</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Real</td>
<td>0 to 1</td>
<td>Weights for feature $V$</td>
<td>kNN, ANN</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Real</td>
<td>0 to 1</td>
<td>Weights for feature $E$</td>
<td>kNN, ANN</td>
<td></td>
</tr>
</tbody>
</table>

Ga settings Value/description

Population Number of individuals in a given generation. Set to 100
Generation Number of maximum generations. Set to 100
Fitness Function Minimization of forecast RMSE for the Validation dataset
Cross-over 2-point Crossover. Swaps elements chosen randomly from 2 parent
Mutation Gaussian mutation. Changes random elements based on a Gaussian distribution
Selection Tournament with 5 individuals chosen randomly
Stopping Criteria Max number of generation or when no improvement is observed for 20 generations

Table 3
<table>
<thead>
<tr>
<th>Forecast Horizon</th>
<th>15 min</th>
<th>30 min</th>
<th>45 min</th>
<th>60 min</th>
<th>90 min</th>
<th>120 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>$s^{O}$</td>
<td>$s^{T}$</td>
<td>$p$</td>
<td>$s^{O}$</td>
<td>$s^{T}$</td>
<td>$p$</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Merced</td>
<td>15</td>
<td>0.0</td>
<td>0.0</td>
<td>15</td>
<td>2.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Davis</td>
<td>5</td>
<td>2.0</td>
<td>−1.4</td>
<td>15</td>
<td>1.4</td>
<td>−0.5</td>
</tr>
<tr>
<td>San Diego</td>
<td>5</td>
<td>2.6</td>
<td>5.3</td>
<td>10</td>
<td>6.8</td>
<td>7.2</td>
</tr>
<tr>
<td>Bellingham</td>
<td>5</td>
<td>4.8</td>
<td>2.0</td>
<td>10</td>
<td>3.0</td>
<td>3.7</td>
</tr>
<tr>
<td>Ewa Beach</td>
<td>35</td>
<td>1.2</td>
<td>3.6</td>
<td>40</td>
<td>0.7</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Fig. 6. Illustration of the GA convergence for the optimization of the ANN-based model for the Ewa Beach location and the 15, 45 and 120 min forecast horizons. The solid lines and the color bands show the average RMSE and the RMSE range for all the individuals in a given generation, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Fig. 7. Graphical representation of the optimized weights in Eq. (20) determined by the GA for the kNN (left column) and ANN (right column) models for the five locations and the six forecast horizons. Each row corresponds to a different location: (a,b) Merced, (c,d) Davis, (e,f) San Diego, (g,h) Bellingham, and (i,j) Ewa Beach. The colors identify the type of features: Blue = backward cumulative average, Red = variability, and Green = entropy). The weights are multiplied by 10 to make the figure more readable. The dash “-” identifies 0 weight for the respective column of matrix $F$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
The plots in this figure show that the cumulative average is the most predominate feature in every case, both for the kNN and the ANN models. The plots also show that there is much diversity between the best set of features for the different locations and models. For instance, in the ANN models variability features are selected more often than entropy features, the opposite happens in the case of the kNN models. Additionally we observe that the ANNs use less input data (fewer columns from the matrix of features) than the kNN. Despite using much less information the performance of the ANN models (shown in the next section) is very similar to the one by the kNN models. We attribute this fact to the nonlinearity of the ANNs and to the fact that increasing the number of input variables for the ANNs has detrimental effects to the mapping between the input data and the target data. Specially if the number of neurons in the hidden layer remains constant as in this case. The kNN does not suffer this limitation as all the input variables are lumped into a single value when the euclidean distance is calculated.

From these plots we can also observe that, in general, the weights for the features that correspond to longer memory increase with increasing forecast horizon. That dependence is clearest in the cases of Davis and Ewa Beach (Fig. 7(c) and (i)). We can also see that the level of complexity in the best set of patterns is quite different from location to location. San Diego, for instance, only requires a few columns from the Backward Average and Entropy to define the best kNN, whereas the kNN model for Ewa Beach uses almost all columns from the three features. That is consistent with the results for the ANN model, in which much more features are selected for Ewa Beach than for any other location.

As shown before the weather conditions are very different from location to location, resulting in very distinct GHI time series. Thus it is not surprising that the selected features for best kNN and ANN vary so much from location to location.

5.2. Error metrics

The parameters presented above define the different forecast models completely. The first step to analyze their performance is to study the respective error metrics. These are listed in Table 4 as a function of forecast horizon and location. The bold values in the table indicate the best model for a given metric.

The values in this table show that in most cases the forecasts show small bias, measured by the MBE. Only Davis and San Diego show appreciable biases, with San Diego a clear outlier. In both cases the optimized models substantially reduce the bias relative to the persistence model. From these results we can explain the improvement observed for the optimized persistence model reported in Table 3: it results from correcting the bias in the original persistence model.

In terms of MAE the optimized kNN and ANN models show the best performance, with the exception of the 15, 45, 60 and 90 min forecasts for San Diego, where the MAE achieved by the optimized persistence is marginally smaller. In the case of RMSE, the variable for which the models were optimized, the kNN and ANN models shows the smallest values regardless of location and forecast horizon. Consequently they always achieve the highest forecast skill.

Comparing only the results for the kNN and ANN models, we observe that the performances for these models are comparable. In general there seems to be a trend suggesting that the ANN model performs better than the kNN model for the shorter forecast horizons ($\tau \leq 60$ min) and worse for the longer horizons. Using values from Table 4 we created Fig. 8 that shows the forecast skill versus the RMSE for each locations. This figure allows for a visual assessment of the evolution of forecast skill and RMSE with forecast horizon and location for the four models used. In this representation the best models are the ones approaching the top left corner of the plot, that is, the ideal case of RMSE = 0 and forecast skill = 100%.

The first conclusion from this figure is that, with the exception of Ewa beach, the RMSE increases with the forecast horizon. This is evidenced by the lines ordered from black, 15 min forecast, to light blue, 120 min forecast in the ascending order of RMSE. In the case of Ewa Beach this order is altered and the smallest RMSEs are achieved for the 45 and 60 min forecasts and the highest for the 15 min forecast. This shows that in some cases the forecast error does not increase monotonically with the forecast horizon. A second observation is that the range for the RMSE for Merced, Davis is $\approx 20$ Wm$^{-2}$, in the cases of Bellingham and Ewa Beach it increases to $\approx 30$ Wm$^{-2}$ and in the case of San Diego it increases to $\approx 60$ Wm$^{-2}$. These observations demonstrate that the forecast performance as a function of the forecast horizon depends on the location.

5.3. Forecastability

The previous analysis shows that, although there are similarities in the forecasting performance between the different locations, there are also plenty of dissimilarities. In this section we attempt to explain the dissimilarities using the knowledge acquired from the study of the time series and the resulting values of $c_i$ and $DE$. To do so, we plotted RMSE for the four forecasting models as a function of $c_i$ (Fig. 9(a)). Each color identifies a different location, the symbols identify the different models, and the size of the symbols indicates the forecast horizon. The shaded areas enclose all the data points belonging to a given location. The figure shows that, as expected, higher values of $c_i$ correspond to higher errors. However, we can see that for a given value of $c_i$ there is a wide range for the RMSEs...
Table 4
MBE, MAE, RMSE and forecast skill $s$ for the different locations and forecast horizons. MBE, MAE and RMSE values are in Wm$^{-2}$ and the skill $s$ is in percentage. The labels $P$, $OP$, $OK$, and $OA$ identify the models used: $P =$ Persistence, $OP =$ Optimized Persistence, $OK =$ Optimized kNN, and $OA =$ Optimized ANN. The bold values identify the best metric for each pair (location, forecast horizon).

<table>
<thead>
<tr>
<th></th>
<th>MBE</th>
<th>MAE</th>
<th>RMSE</th>
<th>$s$</th>
<th>MBE</th>
<th>MAE</th>
<th>RMSE</th>
<th>$s$</th>
<th>MBE</th>
<th>MAE</th>
<th>RMSE</th>
<th>$s$</th>
<th>MBE</th>
<th>MAE</th>
<th>RMSE</th>
<th>$s$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P$</td>
<td>$OP$</td>
<td>$OK$</td>
<td>$OA$</td>
<td>$P$</td>
<td>$OP$</td>
<td>$OK$</td>
<td>$OA$</td>
<td>$P$</td>
<td>$OP$</td>
<td>$OK$</td>
<td>$OA$</td>
<td>$P$</td>
<td>$OP$</td>
<td>$OK$</td>
<td>$OA$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15 min</td>
<td></td>
<td></td>
<td></td>
<td>30 min</td>
<td></td>
<td></td>
<td></td>
<td>45 min</td>
<td></td>
<td></td>
<td></td>
<td>60 min</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merced</td>
<td>-0.1</td>
<td>-0.1</td>
<td>-0.2</td>
<td>-0.5</td>
<td>19.3</td>
<td>19.3</td>
<td>18.7</td>
<td>17.6</td>
<td>47.9</td>
<td>47.9</td>
<td>42.7</td>
<td>42.1</td>
<td>0.0</td>
<td>10.7</td>
<td>12.0</td>
<td></td>
</tr>
<tr>
<td>Davis</td>
<td>1.9</td>
<td>1.2</td>
<td>-0.7</td>
<td>-1.2</td>
<td>19.5</td>
<td>18.2</td>
<td>17.3</td>
<td>17.0</td>
<td>47.7</td>
<td>48.3</td>
<td>41.9</td>
<td>41.3</td>
<td>-1.4</td>
<td>12.0</td>
<td>13.3</td>
<td></td>
</tr>
<tr>
<td>San Diego</td>
<td>3.4</td>
<td>2.2</td>
<td>1.6</td>
<td>1.6</td>
<td>29.4</td>
<td>26.9</td>
<td>27.2</td>
<td>27.9</td>
<td>60.7</td>
<td>57.5</td>
<td>53.0</td>
<td>52.6</td>
<td>5.3</td>
<td>12.7</td>
<td>13.3</td>
<td></td>
</tr>
<tr>
<td>Bellingham</td>
<td>0.8</td>
<td>0.5</td>
<td>1.8</td>
<td>0.6</td>
<td>29.3</td>
<td>27.2</td>
<td>26.9</td>
<td>26.1</td>
<td>61.8</td>
<td>60.6</td>
<td>54.1</td>
<td>53.9</td>
<td>2.0</td>
<td>12.4</td>
<td>12.8</td>
<td></td>
</tr>
<tr>
<td>Ewa Beach</td>
<td>-0.6</td>
<td>-1.0</td>
<td>-2.8</td>
<td>-1.7</td>
<td>77.0</td>
<td>76.9</td>
<td>69.9</td>
<td>69.7</td>
<td>121.9</td>
<td>117.5</td>
<td>103.9</td>
<td>104.2</td>
<td>3.6</td>
<td>14.7</td>
<td>14.5</td>
<td></td>
</tr>
<tr>
<td>Merced</td>
<td>-0.6</td>
<td>-0.4</td>
<td>-0.3</td>
<td>-0.3</td>
<td>22.8</td>
<td>22.0</td>
<td>20.9</td>
<td>20.4</td>
<td>51.4</td>
<td>50.7</td>
<td>45.6</td>
<td>46.0</td>
<td>1.4</td>
<td>11.4</td>
<td>10.5</td>
<td></td>
</tr>
<tr>
<td>Davis</td>
<td>6.5</td>
<td>4.6</td>
<td>-1.6</td>
<td>0.0</td>
<td>24.9</td>
<td>23.2</td>
<td>21.5</td>
<td>20.0</td>
<td>49.9</td>
<td>49.3</td>
<td>44.2</td>
<td>41.8</td>
<td>1.3</td>
<td>11.5</td>
<td>16.2</td>
<td></td>
</tr>
<tr>
<td>San Diego</td>
<td>11.6</td>
<td>6.7</td>
<td>3.3</td>
<td>3.5</td>
<td>42.4</td>
<td>37.0</td>
<td>37.7</td>
<td>37.8</td>
<td>77.5</td>
<td>71.7</td>
<td>65.9</td>
<td>65.0</td>
<td>7.5</td>
<td>15.0</td>
<td>16.2</td>
<td></td>
</tr>
<tr>
<td>Bellingham</td>
<td>2.7</td>
<td>1.5</td>
<td>3.9</td>
<td>3.9</td>
<td>37.6</td>
<td>35.2</td>
<td>32.7</td>
<td>32.2</td>
<td>73.4</td>
<td>70.3</td>
<td>60.1</td>
<td>60.0</td>
<td>4.1</td>
<td>18.0</td>
<td>18.1</td>
<td></td>
</tr>
<tr>
<td>Ewa Beach</td>
<td>-1.3</td>
<td>-1.3</td>
<td>-2.4</td>
<td>-2.5</td>
<td>71.7</td>
<td>71.7</td>
<td>64.6</td>
<td>64.7</td>
<td>105.6</td>
<td>105.6</td>
<td>92.1</td>
<td>92.7</td>
<td>0.0</td>
<td>12.8</td>
<td>12.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>90 min</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merced</td>
<td>-1.6</td>
<td>-0.9</td>
<td>-0.9</td>
<td>-0.8</td>
<td>27.8</td>
<td>24.9</td>
<td>22.6</td>
<td>22.8</td>
<td>57.9</td>
<td>54.5</td>
<td>48.8</td>
<td>49.8</td>
<td>6.0</td>
<td>15.8</td>
<td>14.1</td>
<td></td>
</tr>
<tr>
<td>Davis</td>
<td>13.5</td>
<td>8.1</td>
<td>-1.0</td>
<td>0.9</td>
<td>30.7</td>
<td>26.8</td>
<td>23.4</td>
<td>23.1</td>
<td>54.3</td>
<td>52.8</td>
<td>45.7</td>
<td>45.9</td>
<td>2.8</td>
<td>16.0</td>
<td>15.6</td>
<td></td>
</tr>
<tr>
<td>San Diego</td>
<td>25.8</td>
<td>13.1</td>
<td>8.3</td>
<td>5.9</td>
<td>56.8</td>
<td>46.0</td>
<td>46.5</td>
<td>46.7</td>
<td>98.0</td>
<td>84.6</td>
<td>75.1</td>
<td>77.0</td>
<td>13.6</td>
<td>23.3</td>
<td>21.4</td>
<td></td>
</tr>
<tr>
<td>Bellingham</td>
<td>5.5</td>
<td>3.8</td>
<td>7.2</td>
<td>6.1</td>
<td>45.7</td>
<td>42.1</td>
<td>37.9</td>
<td>37.5</td>
<td>81.3</td>
<td>78.2</td>
<td>65.0</td>
<td>65.2</td>
<td>3.7</td>
<td>20.0</td>
<td>19.7</td>
<td></td>
</tr>
<tr>
<td>Ewa Beach</td>
<td>-2.2</td>
<td>-2.4</td>
<td>-2.2</td>
<td>-2.6</td>
<td>75.5</td>
<td>75.8</td>
<td>63.5</td>
<td>64.1</td>
<td>109.7</td>
<td>110.2</td>
<td>90.1</td>
<td>91.1</td>
<td>-0.4</td>
<td>17.9</td>
<td>17.0</td>
<td></td>
</tr>
<tr>
<td>Average GHI values for the testing set (all values in Wm$^{-2}$): Merced = 473.2, Davis = 422.9, San Diego = 399.2, Bellingham = 353.4, Ewa Beach = 464.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
obtained by the same forecasting model. For instance, $\epsilon_{C3}^r \approx 0.08$ returns a \( \text{RMSE}^{kNN} \approx 40 \text{ Wm}^{-2} \) for Davis and \( \text{RMSE}^{ANN} \approx 60 \text{ Wm}^{-2} \) for San Diego. Hence, we must conclude that the $\epsilon_{C3}^r$ by itself is not a good indicator for the time series forecastability.

If instead of $\epsilon_{C3}^r$ we use the variable DET, as the coordinate variable we obtain Fig. 9 (b). The figure shows that in general a higher determinism corresponds to a lower forecast error. However since the DET, values for each location vary very little with $\tau$ whereas the corresponding RMSEs vary substantially this variable explains the forecastability even less than $\epsilon_{C3}^r$.

Combining these two metrics we obtain Fig. 10 that shows the RMSE for the testing set obtained with the kNN model as a function of $\epsilon_{C3}^r$ and DET, $\tau$. The lighter symbols show the RMSE projection on the DET, $\tau$ and $\epsilon_{C3}^r$ planes. Contrary to the two previous figures, the information conveyed by this figure allows for a better assessment of the forecastability for the different locations and horizons. Now it is possible to explain that the reason why Davis RMSE is lower than the RMSE for San Diego and Bellingham for the same value of $\epsilon_{C3}^r$ it is due to the higher values of DET, $\tau$ for that location. It is also possible to see that, although, the determinism for San Diego is much higher than the one for Ewa Beach, the forecast errors for the 120 min horizon are similar for these two locations due to a similar density or large ramps.

Based on this experiments, we proposed that the forecastability can be defined using both metrics $\epsilon_{C3}^r$ and DET, $\tau$. It is not possible, however, to achieve a general formula for the forecast performance as a function of DET, and $\epsilon_{C3}^r$ due to the lack of coverage in the domain. But it is possible discern some features from this hypothetical relation.
For instance, the figure shows that the forecast error increases steeply in the range $\text{DET}_s \in [0.75, 1]$, whereas the variation with $\epsilon_r^*$ is more gradual. Also, for low values of $\text{DET}_s$ and high values of $\epsilon_r^*$, as in the case of Ewa Beach, it appears that the forecast error reaches a plateau.

To fully characterize the relation between forecastability and $\text{DET}_s$ and $\epsilon_r^*$ other locations must be included such that the domain $\text{DET}_s \times \epsilon_r^*$ is properly covered. Finally, it must be remarked that this relationship depends on the forecasting model, but the general characteristics remain unchanged. Had we used the RMSE from the persistence model in Fig. 10 we would only observe an upward translation of the data points.

6. Conclusions

In this work we propose simple feature extraction methodology to optimize GHI forecasts out to 2 h. We also proposed a simple technique to estimate the forecastability of the GHI time series. Furthermore, we studied the impact of different micro-climates into the forecast error and forecastability by applying the proposed methodology to five locations in the USA. The work presented allows to draw the following conclusions:

- The type of climate and the associate cloud patterns have a profound impact in the forecasting performance. The forecast error more than doubles between Merced, a location with low GHI variability, and Ewa Beach, a location with high GHI variability. This result indicates that low forecast errors are more difficult to obtain for locations where clouds are formed locally due to the orography, as in Ewa Beach. In those cases, cloud patterns will not be as well behaved as the ones illustrated in the introduction making the approach here proposed less effective.
- The best set of endogenous features for the kNN and ANN model are highly dependent on the location. Therefore, a proper feature selection is necessary to maximize the forecasting performance. We demonstrated that genetic algorithms can be used for that task.
- Forecastability can be explained using the large ramp density $\epsilon_r^*$ and the time series determinism $\text{DET}_r$. Such knowledge is important to estimate the forecasting performance before any predictions are produced. This can be an important tool for benchmarking forecasting models, and an effective way to compare the performance of different models proposed in the solar forecast literature.
- It was not possible to obtain a general relation for the forecastability because the data in this study does not cover evenly the $\text{DET}_r \times \epsilon_r^*$ domain. Other locations, with different climates, must be studied to fully assess this relationship.

Acknowledgments

Both authors gratefully acknowledge the partial support by the National Science Foundation (NSF) EECS (EPAS) award N. 1201986, which is managed by Dr. Paul Werbos, and the partial support by the California Energy Commission PIER PON-13-303 program, which is managed by Dr. Zhiqin Zhang.

Appendix A. Clear-sky model

The clear-sky model returns irradiance for a given geographical location and time when clouds are absent. However, even in this condition the broadband irradiance at ground level is influenced by other environmental conditions such as the aerosol content and precipitable water, primarily, and atmosphere ozone, nitrogen dioxide and site pressure secondarily. The presence of aerosols and water in the atmosphere can be quantified through the atmospheric turbidity. Several turbidity parameters have been introduced with the most popular being the Linke turbidity factor. Many models have been developed to estimate the clear-sky irradiance and comprehensive comparisons of some of the most popular models can be found in Ineichen (2006) and Gueymard (2012). In this work we
used the clear-sky model proposed by Ineichen and Perez (2002), as it is one of the best performing models, and one of the simplest. It requires as inputs the solar elevation and the Linke turbidity. The solar elevation for a given latitude and longitude is computed from the sun’s orbital elements. The Linke turbidity is obtained from the worldwide latitude and longitude is computed from the sun’s orbital elements. The Linke turbidity is obtained from the worldwide

## References


