A Robust Forecasting Framework based on The Kalman Filtering Approach with a Twofold Parameter Tuning Procedure: Application to Solar and Photovoltaic Prediction

Ted Soubdhan[a], Joseph Ndong[b], Hanany Ould-Baba[a], Minh-Thang Do[a]

[a] *University of Antilles, Department of Physics, Laboratory LARGE*
[b] *University Cheikh Anta Diop of Dakar, Department of Mathematics and Computer Science, Laboratory LID*

Abstract

This paper presents a framework which relies on the linear dynamical Kalman filter to perform a reliable prediction for solar and photovoltaic production. The method is convenient for real-time forecasting and we describe its use to perform these predictions for different time horizons, between one minute and one hour ahead. The dataset used is a set of measurements of solar irradiance and PV power production measured in a sub-tropical zone: Guadeloupe. In this zone, fluctuating meteorological conditions can occur, with highly variable atmospheric events having severe impact in the solar irradiance and the PV power. In such conditions, heterogeneous ramp events are observed making difficult to control and manage these sources of energy. The present work hopes to build a suitable statistical method, based on bayesian inference and state-space modeling, able to predict the evolution of solar radiation and PV production. We develop a forecast method based on the Kalman filter combined with a robust parameter estimation procedure built with an Auto Regressive model or with an Expectation-Maximisation algorithm. The model is built to run with univariate or multivariate data according to their availability. The model is used here to forecast the univariate solar and PV data and also PV with exogenous data such as cloud cover and air temperature. The accuracy of this technique is studied with a set of performance criterion including the root mean square error and the mean bias error. We compare the results for the different tests performed, from one minute to one hour ahead, to the simple persistence model. The performance of our technique exceeds by far the traditional persistence model with a skill score improvement around 39% and 31%, respectively for PV production and GHI, for one hour ahead forecast.

Keywords: Kalman filter, EM algorithm, AR model, Solar Energy, PV forecast.

1. Introduction

In sub-tropical zone, such as Guadeloupe, a great challenge related to energy production is to perform accurate solar and photovoltaic (PV) forecast for the purpose to control and manage renewable energy production. This important problem is mostly due to the high variability observed in the solar radiation, since the daily atmospheric conditions (weather, temperature, cloud, etc) encounter many changes for very short time scales. The PV production is thus highly dependent on the meteorological conditions which encounter severe variability in various timescales, from seconds to years. In order to have a good estimation of the PV production, the first problem to solve is to build a suitable solar/PV prevision model able to produce reliable forecasts in order to help energy provider to take control and manage carefully their energy production. One of the great challenge is to provide PV production or global solar radiation forecast at a very short timescale under tropical climate. Indeed, large and frequent variations of solar radiation can be observed in tropical climates with amplitudes reaching 800 W/m^2 and occurring within a short time interval, from a few seconds to a few minutes, depending to the geographical location. Such fluctuations can be due,

for example, to the dynamic of clouds which can be very complex and depend on cloud type, size, speed and spatial distribution and, more generally, due to some specific local meteorological conditions [26].

It is then an important matter to have a good knowledge of these rapid fluctuations occurring on very short time scales and to properly predict the evolution of the solar radiation and the PV production. In addition, the Solar/PV forecasts could be performed with the insertion in the model of the exogenous inputs, i.e. the cloud cover and the ambient temperature to render the whole prediction more robust. Furthermore these predictions can be used as inputs for an infrastructure dedicated to control the PV system. So, in this paper, we implement a robust framework to both predict the solar irradiance and the PV production and try to see the impact of the influential features namely the cloud cover and the temperature. For this scope, we build a methodology based on statistical inference techniques, specifically, the Kalman filtering algorithm. We begin by a learning phase which boils down to defining a convenient state-space model which takes into account the main physical properties of the system of solar and PV power. This state-space model is dedicated to describe the system with all the parameters needed to capture the statistical properties and dynamics of the system. Thereafter, as a second phase, we talk about the methodology to estimate the needed parameters which quantify the dynamics of the system. Finally, we apply the Kalman filter and derive some statistics to analyze the performance of the technique.

2. Related works

The literature is full of methodologies dedicated to solar and PV forecasting. However, a few of them are based on the robust statistical Kalman filtering technique, which have been extensively used in various scientific areas including signal processing, financial modeling, biology, medicine, aerospace, etc. We have already developed various models with this technique for the scope of anomaly detection in communication networks [15],[16],[17],[18].

In the domain of solar/PV forecasting, in [2], a method based on a combination of an ARMA model and a Kalman filter is presented for the scope of solar irradiance and temperature forecasting. The ARMA model is built with a high order parameters (p,q) in order to retrieve the state-space parameters needed to run the Kalman filter. The whole methodology is used to do prediction only for a time horizon of 5 minutes. Another similar technique based on Kalman filtering and ARMA model of high order is presented in [7]. Another study, [4] have presented a work which extends previous studies [5],[20], to solve the problem of bias removal from the Kalman forecasts. However the technique was applied only on "persistence" model for a time horizon of 1 hour. Generally, "persistence" forecasts accuracy decrease rapidly with forecast horizon as cloudiness changes severely from one state to another. The works in [5],[13] have developed a linear Kalman filtering model to solve a non-linear forecasting problem. They do so to avoid the parameter tuning operation (estimation of the state system matrix, the observation matrix, and the two covariance matrices for the state and the measurements) which is a crucial need when using this optimization technique. They should use the extended or unscented Kalman filter to solve the problem, but instead they build a high order polynomial model for the state and set the identity matrix to simplify the state equation. We found that the complexity of this model is high enough and it is not necessary to build at least a 3-rd order polynomial for only the scope of bias removal. In the best of our knowledge, we have found any study referred to the linear Kalman filter which gave a robust parameter tuning procedure before running the technique for the purpose of solar/PV forecasting.

In this present work, we build an elaborated method to perform real-time prediction of solar and PV power, based on the knowledge we have on the cloud cover and temperature. The method needs few data for initial parameter estimation. So, all the quantities necessary to run the filter will be identified properly with robust statistical techniques. So an autoregressive (AR) model and a method based on the expectation-maximization algorithm can be suitably used to achieve the parameter settings. On the other hand, our methodology can be applied, with the assumption of the stationarity of the studied process, to do prediction for a various set of time horizons (from 1 minute to 1 hour ahead). The performance level of our methodology is studied with some standard accuracy measures including the root mean square error (RMSE), the mean bias error (MBE) and the mean absolute error (MAE) in their normalized forms.

3. Methodology

In this work, we aim to build a framework to predict solar radiation and PV power production in real-time conditions. The high variability observed in the meteorological conditions (clear sky, cloudy sky, wind speed, temperature,

etc) is the principal source of the different levels of variability inside the measured radiance and the PV power. So, it is necessary to have a model able to take the actual data in time t and predict instantaneously the next data at time $t+1$. One solution to perform this prediction is to consider all the past data up to time time t and perform a prediction for the next time $t+1$. Techniques as vector autoregressive model (VAR), Wiener filter, etc acts in this way, but the complexity is very high since they use all the past information. These techniques, generally, use much memory to save the history of the data to perform the predictions. Other methods, like first-order Markov model use only the recent arrived data at time t to do the prediction at the next time-step. The Kalman filtering technique performs in this second way. The advantage of using linear discrete-time Kalman filter is mainly related to its ability to do future prediction in real-time conditions for on-line systems in a memory-less way. With this approach, we can build a model from which we derive (estimate) an a priori unknown hidden signal (the underlying real-time hidden system state) and re-utilize it to have our predictions w.r.t. the measured data.

We build our forecasting model by assuming that the measured solar/PV power is governed by a hidden signal as in the following linear dynamical system:

$$\mathcal{Y}_t = \mathbf{A}_t \mathcal{X}_t + \mathcal{V}_t \tag{1}$$

where $\mathcal{Y}_t \in \mathbb{R}^m$ denotes a stochastic process containing the real measurement (*observation*) at time t. The stochastic process $\mathcal{X}_t \in \mathbb{R}^n$ is the hidden state of the system which is transformed to the output by the matrix $\mathbf{A}_t \in \mathbb{R}^{m\times n}$ which controls the sources of the measurements. For example, the output $\mathcal{Y}_t$ can be composed of the variables: *solar radiation, ambient temperature, cloud cover, ...*; $\mathcal{Y}_t$ can also be composed of m variables corresponding to the measured solar irradiance or PV power at different points. Generally, due to the intrinsic imperfection of measurement devices, errors more often occur during data collection, then $\mathcal{V}_t$ denotes a stochastic process representing the measurement error.

Now we aim to find a convenient model for the stochastic process $\mathcal{X}_t$ which is able to track the system traffic features we want to monitor, ie. the estimated solar irradiance and PV production. We propose a simple stochastic model where $\mathcal{X}_t$ at discrete time t is represented by the linear combination of two components:

$$\mathcal{X}_t = \mathbf{C}_t \hat{\mathcal{X}}_t + \xi_t \tag{2}$$

where $\hat{\mathcal{X}}_t$ is a predictable component, ξ_t is a random noise component and $\mathbf{C}_t \in \mathbb{R}^{n\times n}$ is a parameter matrix describing the underlying state dynamics.

Generally the prediction model can have any structure and the noise process can yield from any distribution. However linear stochastic predictive models combined with gaussian noise have a long record of successful applications in a very large spectrum of engineering techniques. Our idea is to relate the predictable components as network *states*. These variables are not directly observable from the measurement devices, so we refer to them as *hidden system states*. Since we want to estimate the variables $\mathcal{X}_t$ as a state, we can use linear stochastic dynamic system based on state-space models. Now we build a temporal model that relates $\mathcal{X}_{t+1}$ to $\mathcal{X}_t$ with the following difference equation:

$$\mathcal{X}_{t+1} = \mathbf{C}_t \mathcal{X}_t + \mathbf{B}_t \mathcal{U}_t + \mathcal{W}_t \tag{3}$$

where the $n \times n$ matrix $\mathbf{C}_t$ relates the state at the previous time step t to the state at the current step $t+1$ and $\mathcal{W}_t$ is the intrinsic noise process. Here, $\mathbf{C}_t$ describes the atmospheric conditions since the value of the solar irradiance, or equivalently the PV power, at time t depend on ambient temperature, cloud cover and wind speed. The value of the solar irradiance between t and $t+1$ might change according to that conditions. So we can estimate $\mathbf{C}_t$ given a set of measured data that contains all types of meteorological conditions to train our forecasting model. If one desires to have forecast of the PV power by incorporating in the model other metrics as the temperature, the cloud cover, the wind, etc in order to learn more about their influence for the measures, he/she may take into account the term $\mathbf{B}_t\mathcal{U}_t$, ($\mathbf{B}_t \in \mathbb{R}^{n\times p}$ and $\mathcal{U}_t \in \mathbb{R}^p$ with p exogenous variables). The equations Eq. 1 and Eq. 3 are now combined to form the complete state-space model for the specification of our forecasting framework:

$$\begin{cases} \mathcal{X}_{t+1} = \mathbf{C}_t \mathcal{X}_t + \mathbf{B}_t \mathcal{U}_t + \mathcal{W}_t \\ \mathcal{Y}_t = \mathbf{A}_t \mathcal{X}_t + \mathcal{V}_t \end{cases} \tag{4}$$

This block of equations is the classic form for a linear dynamical system with inputs. In this model we assume that the state-noise $\mathcal{W}_t$ and the measurement-noise $\mathcal{V}_t$ are uncorrelated zero-mean gaussian white-noise processes with covariance matrices $\mathbf{Q}_t \in \mathbb{R}^{n\times n}$ and $\mathbf{R}_t \in \mathbb{R}^{m\times m}$, respectively.

4. System parameter tuning

Building a framework for multi purpose forecasting might rely on the identification of the needed parameters $\theta = \left\{\mathbf{C}_t, \mathbf{A}_t, \mathbf{Q}_t, \mathbf{R}_t\right\}$ for calibrating the evolving system. To achieve this aim, we use two well-known techniques to perform the identification of these parameters before running the filter for forecasting. The first method is based on an autoregressive (AR) model which uses a step-wise least square algorithm performing a QR factorization (decomposition of a matrix $\mathbf{T}$ into a product $\mathbf{T} = \mathbf{QR}$ of an orthogonal matrix $\mathbf{Q}$ and an upper triangular matrix $\mathbf{R}$) of a data matrix to evaluate, for a sequence of successive orders, some criterion such as Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) for the selection of the model order p. Thereafter we compute the parameters of the AR model of the optimum order and then retrieve the needed parameter θ. So the matrix $\mathbf{C}_t$ can be formed with the p parameters of the AR model. The noise variance of the model is used to fix our system covariance matrix $\mathbf{Q}_t$. This first model does not give us the matrix $\mathbf{R}_t$, so we form it as dependent on the matrix $\mathbf{Q}_t$ since we know that the intrinsic system noise have an influence to the measurement noise. In practice, $\mathbf{R}_t$ is heuristically obtained by dividing the system noise by some constant. To overcome this problem of setting $\mathbf{R}_t$ manually, we use a second parameter identification algorithm based on the expectation-maximization (EM) algorithm, as we describe in the following.

It is important to see that the above parameters θ are time-dependent, so they might be calibrated every time data are collected at time t. In this work, we have assumed a stationary case where the parameters are independent of time, and then we perform the calibration only once, with a small part of the training dataset. A procedure of data normalization is presented to accept this assumption.

Since we have found a model and system equations for our system, next we need to deal properly with the different steps of our optimization algorithm for solar radiance and PV prediction.

5. Prediction equations of the discrete-time Kalman filter

The first problem to solve after establishing a real-time state-space model to do our future predictions, is to find an optimal estimate $\hat{\mathcal{X}}_t$ of our unobservable states $\mathcal{X}_t$, given a set of measurements $\{\mathcal{Y}_1, \ldots, \mathcal{Y}_t\}$. In our dynamical linear system, we refer to $\mathcal{Y}_t$ as the observation at time t. And the state of the system at time t is given by $\mathcal{X}_t$, let also $\hat{\mathcal{X}}_{t|t}$ denotes the estimate of $\mathcal{X}_t$ using all the information available up to time t. $\hat{\mathcal{X}}_{t+1|t}$ denotes the prediction of $\mathcal{X}_{t+1}$ using all the information up to time t, (this constitutes the *phase predictor*). The quantity $\hat{\mathcal{X}}_{t+1|t+1}$ denotes the estimate of $\mathcal{X}_{t+1}$ using all past information and the recently arrived data point at time $t+1$. On the other hand, $\mathbf{P}_{t|t}$ denotes the error covariance of the *state estimate* at time t and $\mathbf{P}_{t+1|t}$ indicates the covariance of *the state prediction* at time $t+1$. If we have the prediction $\hat{\mathcal{X}}_{t+1|t}$, we plug it into Eq. 1 to derive a prediction $\hat{\mathcal{Y}}_{t+1|t}$ at time $t+1$.

In the above formulation and in the rest of the paper, one should always follow the nuance between *estimate* and *prediction*. So, at time t, we refer to the quantity $\hat{\mathcal{X}}_{t|t}$ as an estimate and $\hat{\mathcal{X}}_{t+1|t}$ as a prediction. Estimation is instantaneous while prediction is done for the future, given the present.

As it is shown in its earlier elaboration, the linear Kalman filter addresses the problem of estimating a discrete state vector when the observations are only a linear combination of the underlying state vector. The filter runs as a *predictor-corrector* algorithm. As an iterative algorithm, it estimates the system state using two steps: *prediction* comes in the *time update* phase, and *correction* in the *measurement update* phase.

- **Prediction step** (*time update equations*):
  In this step, the estimated state of the system at time t, $\hat{\mathcal{X}}_{t|t}$, is used to predict the state at next time $t+1$, $\hat{\mathcal{X}}_{t+1|t}$. And, as we know that the noise $\mathcal{W}_t$ influences the evolution of the system at each time t, we compute only the covariance of the prediction, $\mathbf{P}_{t+1|t}$ based on the updated covariance at the previous time t, $\mathbf{P}_{t|t}$, and the noise covariance at the same time, $\mathbf{Q}_t$. The error covariance $\mathbf{P}_{t+1|t}$ provides an indication of the uncertainty associated with the state estimate.

$$\begin{cases} \hat{\mathcal{X}}_{t+1|t} = \mathbf{C}_t \hat{\mathcal{X}}_{t|t} + \mathbf{B}_t \mathcal{U}_{t|t} \\ \mathbf{P}_{t+1|t} = \mathbf{C}_t \mathbf{P}_{t|t} C_t^T + \mathbf{Q}_t \end{cases} \tag{5}$$

- **Correction step** (*measurement update equations*):
  This step updates (corrects) the state and the variance of the estimate in the previous step, using a combination

of their predicted values and the new observations $\mathcal{Y}_{t+1}$. The correctness of this update depends on the Kalman innovation $\mathcal{Y}_{t+1} - \mathbf{A}_{t+1}\hat{\mathcal{X}}_{t+1|t}$.

$$\begin{cases} \hat{\mathcal{X}}_{t+1|t+1} = \hat{\mathcal{X}}_{t+1|t} + \mathbf{K}_{t+1}\left(\mathcal{Y}_{t+1} - \mathbf{A}_{t+1}\hat{\mathcal{X}}_{t+1|t}\right) \\ \mathbf{P}_{t+1} = \left(\mathbf{I} - \mathbf{K}_{t+1}\mathbf{A}_{t+1}\right)\mathbf{P}_{t+1|t}\left(\mathbf{I} - \mathbf{K}_{t+1}\mathbf{A}_{t+1}\right)^T + \mathbf{K}_{t+1}\mathbf{R}_{t+1}\mathbf{K}_{t+1}^T \end{cases} \tag{6}$$

In the measurement equations, $\mathbf{K}_{t+1}$ denotes the Kalman gain and $\mathbf{I}$ the identity matrix. For more details in linear dynamical system, estimation and Kalman filter techniques, we refer the reader to [9, 8, 14]. The above equations with initial conditions of the state of the system $\hat{\mathcal{X}}_{0|0} = E[\mathcal{X}_0]$ and the associated error covariance matrix $P_{0|0} = E\left[(\hat{\mathcal{X}}_0 - \mathcal{X}_0)(\hat{\mathcal{X}}_0 - \mathcal{X}_0)^T\right]$ define the discrete-time sequential recursive algorithm for calculating the linear *minimum variance* estimate known as the *Kalman filter*.

6. Calibration of the Kalman filter

In Section 3, we have assumed two types of model calibration which can be used to achieve a good prediction. The first method considers that the observation matrix $\mathbf{A}$ is precisely known and then we are only interested to find the remaining quantities $\mathbf{C}, \mathbf{Q}$ and $\mathbf{R}$. To estimate dynamically these parameters, we use an autoregressive (AR) model. A second, more general model can also be developed, where all the needed parameters, $\mathbf{C}$, $\mathbf{A}$, $\mathbf{Q}$ and $\mathbf{R}$, are assumed unknown. This method is based on the Expectation-Maximization (EM) algorithm, [6]. This algorithm is an extension of a previously proposed technique, [24, 25] where the matrix $\mathbf{C}$ is assumed precisely fixed. The work in [6] reinforces this technique by assuming $\mathbf{C}$ undetermined.

6.1. Parameter Estimation via an AR model

The idea behind this parameter estimation technique is to cast an AR model into a state-state form to retrieve the parameters θ. The method is based on a stepwise least squares algorithm which uses a QR factorization of a data matrix to evaluate, for a sequence of successive orders, a criterion (here AIC [1] and BIC [23]) for the selection of the model order p, and to compute the parameters of the AR model of the optimum order. It is sufficient to learn the model's parameters using a sample of the first 20-30 days of measurements from the measured data studied. We run this method based on the notes described in [19], and developed in [22]. We form the matrix $\mathbf{C}$ using the p parameters of the AR(p). The noise variance of the model is used to fix the system noise matrix $\mathbf{Q}$. The measurement matrix $\mathbf{R}$ cannot be estimated by the AR(p), by experience we take it as a function of $\mathbf{Q}$. In practice, $\mathbf{R}$ is obtained by dividing the measurement noise by some constant. For our dataset, which is highly variable, we set manually this constant in the interval $[0.02; 1.2]$ for our different experiences. To discover carefully this interval in order to fix the value of $\mathbf{R}$, we run different scenarios of the Kalman filter and choose the value of $\mathbf{R}$ which minimizes some optimal criteria, namely the root mean-square error (RMSE) prediction and the mean bias error (MBE) defined respectively as:

$$\text{RMSE} = \left(\frac{1}{N}\sum_{k=1}^{N}(\hat{x}_k - x_k)^2\right)^{1/2} \tag{7}$$

and

$$\text{MBE} = \frac{1}{N}\sum_{k=1}^{N}(\hat{x}_k - x_k)\,. \tag{8}$$

The methodology implemented in [22] is based on the following AR(p) model (which is equivalent to the state equation in Eq. 3 with the intercept $\mathcal{W}$ set to zero):

$$\mathcal{V}_r = \mathcal{W} + \sum_{l=1}^{p}\mathbf{A}_l\mathcal{V}_{r-l} + \epsilon_r, \qquad \epsilon_r : \text{is white noise with covariance matrix } \mathbf{C} \tag{9}$$

which can be expressed in the form of a regression model of the form:

$$\mathcal{V}_r = \mathbf{B}\mathcal{U}_r + \epsilon_r \tag{10}$$

where the matrix $\mathbf{B}$ contains the parameters to be estimated:

$$\mathbf{B} = \left(\mathcal{W}, \mathbf{A}_1, \ldots, \mathbf{A}_p\right), \tag{11}$$

and the vectors $\mathcal{U}_r$, of dimension $n_p = mp + 1$ (m is the dimension of $\mathcal{V}_r$), are found as follow:

$$\mathcal{U}_r = \left(\ 1, \mathcal{V}_{r-1}, \ldots, \mathcal{V}_{r-p} \ \right) \tag{12}$$

Transforming an AR model into a regression model is, in fact, a simple approximation. Indeed, for a regression model, the predicators $\mathcal{U}_r$ are supposed to be constant, while in an AR model, these vectors are the realization of a stochastic process. Then, the approximation needs to have an initial predicator $\mathcal{U}_1$ in order to initialize the AR model:

$$\mathcal{U}_1 = \left(\ 1, \mathcal{V}_0, \ldots, \mathcal{V}_{1-p} \ \right) \tag{13}$$

So, the final step of the QR factorization gives us the matrix $\mathbf{C}$ as:

$$\mathbf{C} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{A}_3 & \ldots & \mathbf{A}_p \\ \mathbf{I}_d & 0 & 0 & \ldots & 0 \\ 0 & \mathbf{I}_d & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & \mathbf{I}_d & 0 \end{pmatrix}, \tag{14}$$

where $\mathbf{I}_d$ is the identity matrix and the $\mathbf{A}_i$ are the coefficients of the AR(p). By retrieving a set of quantities, our co-variance matrix $\mathbf{Q}$ can be obtained by the formula:

$$\hat{\mathbf{Q}} = \frac{1}{N - n_p}(-\mathbf{W}\mathbf{U}^{-1}\mathbf{W}') \tag{15}$$

with the following matrices of moments:

$$\mathbf{U} = \sum_{r=1}^{N} u_r u_r', \qquad \mathbf{V} = \sum_{r=1}^{N} v_r v_r', \qquad \mathbf{W} = \sum_{r=1}^{N} v_r u_r'. \tag{16}$$

This AR transformation does not give us directly the observation matrix A needed in equation Eq. 4. After retrieving the states matrices with the AR(p), it is obvious to re-transform this AR model in the form of a state-state model. And then the state equation can be plug into the observation equation in order to retrieve the observation matrix.

6.2. Parameter Estimation via the EM algorithm

Parameter estimation via the EM algorithm boils down, first to combining the state variable $\mathcal{X}_t$ and the state noise $\mathcal{W}_t$ as a single random gaussian process, and doing the same transformation for the output variable $\mathcal{Y}_t$ and the measurement noise $\mathcal{V}_t$. So with equation Eq. (4), we can form the conditional probability densities for the state and the output of the system as follow (in the following we use respectively x_t and y_t for the state and the measurement vectors for notation convenience):

$$\mathbb{P}(y_t|x_t) = \frac{1}{\sqrt{(2\pi)^p|\mathbf{R}|}} \exp\left\{-\frac{1}{2}\left[y_t - \mathbf{C}x_t\right]' \mathbf{R}^{-1}\left[y_t - \mathbf{C}x_t\right]\right\} \tag{17}$$

$$\mathbb{P}(x_t|x_{t-1}) = \frac{1}{\sqrt{(2\pi)^k|\mathbf{Q}|}} \exp\left\{-\frac{1}{2}\left[x_t - \mathbf{A}x_{t-1}\right]' \mathbf{Q}^{-1}\left[x_t - \mathbf{A}x_{t-1}\right]\right\} \tag{18}$$

where the notation $|\mathbf{R}|$ and $|\mathbf{Q}|$ are respectively the determinants of the matrices $\mathbf{R}$ and $\mathbf{Q}$.
Suppose the output of the system is of the form of a multidimensional array with T vectors $(y_1, y_2, \ldots, y_T)$. Let's

denote $\{y\}$ that observation sequence and $\{y\}_{t_0}^{t_1}$ being the sub-sequence $(y_{t_0}, y_{t_0+1}, \ldots, y_{t_1})$. The state equation of the model in Eq. 4 is a first-order Markov process, so:

$$\mathbb{P}\Big(\{x\},\{y\}\Big) = \mathbb{P}(x_1)\prod_{t=2}^{T}\mathbb{P}\Big(x_t|x_{t-1}\Big)\prod_{t=1}^{T}\mathbb{P}(y_t|x_t). \tag{19}$$

If we suppose that the initial state is a gaussian random process of expectation π_1 and co-variance matrix $\mathbf{V}_1$ given by:

$$\mathbb{P}(x_1) = \frac{1}{\sqrt{(2\pi)^k|\mathbf{V}_1|}}\exp\left\{-\frac{1}{2}\left[x_1-\pi_1\right]'\mathbf{V}_1^{-1}\left[x_1-\pi_1\right]\right\}, \tag{20}$$

we can write the logarithm of the jointly probability density between $\{x\}$ and $\{y\}$, as a quadratic sum of terms in the form:

$$\begin{aligned}\log\mathbb{P}\Big(\{x\},\{y\}\Big) = &-\sum_{t=1}^{T}\left(\frac{1}{2}[y_t-\mathbf{C}x_t]'\mathbf{R}^{-1}[y_t-\mathbf{C}x_t]\right)-\frac{T}{2}\log|\mathbf{R}|\\ &-\sum_{t=2}^{T}\left(\frac{1}{2}[x_t-\mathbf{A}x_{t-1}]'\mathbf{Q}^{-1}[x_t-\mathbf{A}x_{t-1}]\right)-\frac{T-1}{2}\log|\mathbf{Q}|\\ &-\frac{1}{2}\left[x_1-\pi_1\right]'\mathbf{V}_1^{-1}\left[x_1-\pi_1\right]-\frac{1}{2}\log|\mathbf{V}_1|-\frac{T(p+k)}{2}\log 2\pi\end{aligned} \tag{21}$$

Our linear dynamical system can be seen as a continuous state of a hidden Markov model (HMM) [21]. Then, the "forward" part of the "forward-backward" algorithm for a HMM can be established by a Kalman filter, whereas the "backward" part is solved by a recursive "Kalman smoother" [14]. Finally, if we take into account the measurements, we run the forward-backward algorithm and we solve the problem of inferring the above probability densities for the system state; this operation is done via the Kalman smoother and constitutes the "E" step of the EM algorithm. To derive this step, we need the three following quantities:

$$\hat{x}_t \equiv \mathrm{E}\left[x_t|\{y\}\right], \tag{22}$$

$$\mathbf{P}_t \equiv \mathrm{E}\left[x_t x_t'|\{y\}\right] \tag{23}$$

$$\mathbf{P}_{t,t-1} \equiv \mathrm{E}\Big[x_t x_{t-1}'|\{y\}\Big] \tag{24}$$

The term $\hat{x}_t$ gives an estimate of the system state given the past and future observations, then it differ to the state estimate given by a standard Kalman filter, where this estimate is given by $(\mathrm{E}[x_t|\{y\}_1^t)$. Thereafter, we can calculate the log-likelihood defined by:

$$\mathbb{L}(\theta) = \mathrm{E}\Big[\log\mathbb{P}\left(\{x\},\{y\}\right)|\{y\}\Big]. \tag{25}$$

By setting the quantities $x_t^\tau = \mathrm{E}\Big[x_t|\{y\}_1^\tau\Big]$ and $\mathbf{V}_t^\tau = \mathrm{Var}\Big[x_t|\{y\}_1^\tau\Big]$, we can establish the Kalman filter equations for the estimation of our parameters, as follow:

$$\begin{aligned}x_t^{\tau-1} &= \mathbf{A}x_{t-1}^{\tau-1}, &(26)\\ \mathbf{V}_t^{\tau-1} &= \mathbf{A}\mathbf{V}_{t-1}^{\tau-1}\mathbf{A}' + Q, &(27)\\ \mathbf{K}_t &= \mathbf{V}_t^{\tau-1}\mathbf{C}'(\mathbf{C}\mathbf{V}_t^{\tau-1}\mathbf{C}' + \mathbf{R})^{-1}, &(28)\\ x_t^\tau &= x_t^{\tau-1} + \mathbf{K}_t(y_t - \mathbf{C}x_t^{\tau-1}), &(29)\\ \mathbf{V}_t^\tau &= \mathbf{V}_t^{\tau-1} - \mathbf{K}_t\mathbf{C}\mathbf{V}_t^{\tau-1}. &(30)\end{aligned}$$

To find $\hat{x}_t \equiv x_t^T$ and $\mathbf{P}_t \equiv \mathbf{V}_t^T + x_t^T(x_t^T)'$, we need the variables:

$$\begin{aligned} J_{t-1} &= \mathbf{V}_{t-1}^{t-1}\mathbf{A}'(\mathbf{V}_t^{t-1})^{-1}, && (31)\\ x_{t-1}^T &= x_{t-1}^{t-1} + J_{t-1}(x_t^T - \mathbf{A}x_{t-1}^{t-1}), && (32)\\ \mathbf{V}_{t-1}^T &= \mathbf{V}_{t-1}^{t-1} + J_{t-1}(\mathbf{V}_t^T - \mathbf{V}_t^{t-1})J_{t-1}'. && (33)\\ & && (34) \end{aligned}$$

The quantity $\mathbf{P}_{t,t-1} \equiv \mathbf{V}_{t,t-1}^T + x_t^T(x_{t-1}^T)'$ is needed and can be obtained by deriving:

$$\mathbf{V}_{t-1,t-2}^T = \mathbf{V}_{t-1}^{t-1}J_{t-2}' + J_{t-1}(\mathbf{V}_{t,t-1}^T - \mathbf{A}\mathbf{V}_{t-1}^{t-1})J_{t-2}'. \quad (35)$$

The "M" Step maximizes the likelihood of observing the parameter $\theta = (\mathbf{C}, \mathbf{A}, \mathbf{R}, \mathbf{Q}, \pi_1, \mathbf{V}_1)$. Each term of this parameter is derived by considering the partial derivative corresponding to the logarithm of the likelihood. It suffice to put this derivative to zero and resolve the resulting equation, and finally, we find the value of the given parameter:

- the observation matrix is given by:

$$\frac{\partial \mathbb{L}(\theta)}{\partial \mathbf{A}} = -\sum_{t=1}^{T}\mathbf{R}^{-1}y_t\hat{x}_t' + \sum_{t=1}^{T}\mathbf{R}^{-1}\mathbf{A}\mathbf{P}_t = 0 \quad (36)$$

$$\mathbf{A}^{\text{new}} = \left(\sum_{t=1}^{T} y_t\hat{x}_t'\right)\left(\sum_{t=1}^{T}\mathbf{P}_t\right)^{-1} \quad (37)$$

- the co-variance of the observation noise is derived as:

$$\frac{\partial \mathbb{L}(\theta)}{\partial \mathbf{R}^{-1}} = \frac{T}{2}\mathbf{R} - \sum_{t=1}^{T}\left(\frac{1}{2}y_t y_t' - \mathbf{C}\hat{x}_t y_t' + \frac{1}{2}\mathbf{C}\mathbf{P}_t\mathbf{C}'\right) = 0 \quad (38)$$

$$\mathbf{R}^{\text{new}} = \frac{1}{T}\sum_{t=1}^{T}(y_t y_t' - \mathbf{C}^{\text{new}}\hat{x}_t y_t') \quad (39)$$

- the system state matrix is equal to:

$$\frac{\partial \mathbb{L}(\theta)}{\partial \mathbf{C}} = -\sum_{t=2}^{T}\mathbf{Q}^{-1}\mathbf{P}_{t,t-1} + \sum_{t=2}^{T}\mathbf{Q}^{-1}\mathbf{C}\mathbf{P}_{t-1} = 0 \quad (40)$$

$$\mathbf{C}^{\text{new}} = \left(\sum_{t=2}^{T}\mathbf{P}_{t,t-1}\right)\left(\sum_{t=2}^{T}\mathbf{P}_{t-1}\right)^{-1} \quad (41)$$

- and the state noise co-variance is given by:

$$\begin{aligned} \frac{\partial \mathbb{L}(\theta)}{\partial \mathbf{Q}^{-1}} &= \frac{T-1}{2}\mathbf{Q} - \frac{1}{2}\sum_{t=2}^{T}\left(\mathbf{P}_t - \mathbf{C}\mathbf{P}_{t-1,t} - \mathbf{P}_{t,t-1}\mathbf{C}' + \mathbf{C}\mathbf{P}_{t-1}C'\right)\\ &= \frac{T-1}{2}\mathbf{Q} - \frac{1}{2}\left(\sum_{t=2}^{T}\mathbf{P}_t - \mathbf{C}^{\text{new}}\sum_{t=2}^{T}\mathbf{P}_{t-1,t}\right) \quad (42)\\ &= 0 \end{aligned}$$

$$\mathbf{Q}^{\text{new}} = \frac{1}{T-1}\left(\sum_{t=2}^{T}\mathbf{P}_t - \mathbf{C}^{\text{new}}\sum_{t=2}^{T}\mathbf{P}_{t-1,t}\right) \quad (43)$$

- the initial system state can also be estimated by the quantity:

$$\frac{\partial \mathbb{L}(\theta)}{\partial \pi_1} = (\hat{x}_1 - \pi_1)\mathbf{V}_1^{-1} = 0 \tag{44}$$

$$\pi_1^{\text{new}} = \hat{x}_1. \tag{45}$$

- and the initial co-variance given by:

$$\frac{\partial \mathbb{L}(\theta)}{\partial \mathbf{V}_1^{-1}} = \frac{1}{2}\mathbf{V}_1 - \frac{1}{2}\left(\mathbf{P}_1 - \hat{x}_1\pi_1^{'} - \pi_1\hat{x}_1^{'} + \pi_1\pi_1^{'}\right) = 0 \tag{46}$$

$$\mathbf{V}_1^{\text{new}} = \mathbf{P}_1 - \hat{x}_1\hat{x}_1^{'}. \tag{47}$$

7. Forecasting criterion of accuracy

To evaluate the performances of our model, we use several accuracy measures defined to evaluate solar and PV forecasts. Benchmarking of solar forecasts has been examined by the International Energy Agency Solar Heating and Cooling Program Task 36 on "Solar Resource Knowledge Management" and the project "Management and Exploitation of Solar Resource Knowledge," which have suggested guidelines for performance analysis of the forecasting models [12]. Solar and PV forecast accuracy was assessed in terms of RMSE, mean absolute error (MAE), and mean bias error (bias) MBE. RMSE gives more weight to large errors, whereas MAE, less sensitive to large errors, reveals the average magnitude of the error, and MBE indicates whether there is a significant tendency to systematically over-forecast or under-forecast. When comparing between different models in the training year, RMSE was used as the metric for minimization, that is, forecasts were trained with the goal of reducing the largest errors. These performance criterion are used here in the normalized forms defined as follow:

$$\text{nRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{P}_i - \hat{\text{P}}_i)^2}}{\max(\text{P}_i) - \min(\text{P}_i)} \tag{48}$$

$$\text{nMAE} = \frac{\frac{1}{N}\sum_{i=1}^{N}|\text{P}_i - \hat{\text{P}}_i|}{\max(\text{P}_i) - \min(\text{P}_i)} \tag{49}$$

$$\text{nMBE} = \frac{\frac{1}{N}\sum_{i=1}^{N}(\text{P}_i - \hat{\text{P}}_i)}{\max(\text{P}_i) - \min(\text{P}_i)} \tag{50}$$

We also include a new metric: the forecast skill score (SS) parameter. The latter proposed by Coimbra et al.,[3]:

$$\text{SS} = 1 - \frac{\text{RMSE}_{\text{forecast}}}{\text{RMSE}_{\text{sc-pers}}} \tag{51}$$

The skill score gives the fractional improvement in the mean square error of the proposed forecasting model over the reference model (here the "persistence" model): a skill score of 1 indicates a perfect forecast, a score of 0 indicates no improvement against the reference, and a negative skill means that the forecast model tested performs worse than the reference.

8. Forecast methods

8.1. Basic persistence

The Kalman filter forecasting technique is compared against the classical scaled persistence model. It is based on the assumption that the current conditions will persist so that,

$$\hat{I}_p(t+T) = I_p(t) \tag{52}$$

where $\hat{I}_p(t+T)$ represents a GHI prediction from the persistence model, T represents the forecast horizon, and $I_p(t)$ is the measured GHI value at time t.

8.2. Smart persistence

Our forecasting approach is also compared with the smart persistence. The smart persistence model is based on the same assumption as persistence model but it corrects for the deterministic diurnal variation in solar irradiance. It is defined as,

$$\hat{I}_{sp}(t+T) = k_c(t) I_{cs}(t+T) \tag{53}$$

where $\hat{I}_{sp}(t)$ represents a GHI prediction at time t, $I_{cs}(t)$ represents the estimated clear sky solar irradiance [11].

9. Validation of the approach

9.1. Experimental data

9.1.1. Solar and weather data

The solar data was measured at the campus Fouillole of the University of the French West Indies and Guiana with the interval of 1 second from 2010 to 2014. In this paper we show only the results concerning one year (2011) which constitutes a complete record of measurements. Before applying the whole methodology and in order to perform forecasts with many time horizons, we have build a pre-processing normalization technique to organize the data into different time-scales with time horizons including 1, 5 and 10 minutes bin increment. The normalization method is exposed in section 9.2. We retrieve then the deterministic component of the signal and obtain consequently a stationary process.

9.1.2. PV data

The PV system studied in this paper is installed on the roof of a storehouse in Guadeloupe. The system has 238 flexible PV modules Unisolar of 136Wp each (total of 32,3 kW installed) . The system consists of two inverters of 15 kW and 20 kW. The data logging of PV output power (in W) is integrated in these inverters (collected every 5 min). For the objective of forecasting, a weather station has been also installed on the roof of the storehouse in 2014. However the duration of data is not long enough and therefore will not be used in this study. The time series data is collected in 2 years: 2012 and 2013. However, due to some errors in the data acquisition of the PV output power, only a collection of data covering 696 days is available. As during the night, the forecast is not necessary because the output power of PV is zero, the study takes into account only the data between 0700 and 1700 LST every day.

The two exogenous variables, cloud cover (in octa) and ambiance temperature (in °C) are measured on an hourly basis at the Meteo France meteorological station of le Raizet (16°26N, 61°24W, 11m asl). The daily average for the solar load on a horizontal surface is around 5 kWh/m^2. A constant sunshine combined with the thermal inertia of the ocean makes the air temperature variation quite weak, between 17°C and 33°C, with an average of 25°C to 26°C. Relative humidity ranges from 70% to 80%, and the trade winds are relatively constant all along the year. Two main regimes of cloudiness are superposed: the clouds driven by the synoptic conditions over the Atlantic Ocean and the orographic cloud layer generated by the local reliefs.

9.2. Procedure of the data normalization

As the irradiation and the PV production depend on the sun path during the day, a normalization is needed to eliminate this subordination and accept the assumption we have made in the introduction about the stationarity of the signal.

In this study, the irradiation will be normalized by the the theoretical clear sky, GHI_{csk} curve. The Global Horizontal Irradiance (GHI) is the total amount of shortwave radiation received by a horizontal surface on the ground, which consists of the direct irradiance and the diffuse irradiance. The GHI_{csk} is the GHI calculated in the condition of clear sky, using the Kasten clear sky model. This model accounts for atmospheric turbidity and solar elevation angle. The inputs to this model are air mass, Linke Turbidity, and elevation [10]. The Linke turbidity factor is a very convenient approximation to model the atmospheric absorption and scattering of the solar radiation under clear skies. It describes the optical thickness of the atmosphere due to water vapor and the aerosol particles relative to a dry and clean atmosphere. With larger Linke turbidity, there is more attenuation of the radiation by the clear sky atmosphere. We obtain then the clear sky index, k_c, defined as:

$$k_c(t) = \frac{\mathrm{GHI}(t)}{\mathrm{GHI}_{csk}(t)} \tag{54}$$

The input of the PV production at time (t) is then normalized into $\bar{P}(t)$, the normalized value of the PV production with respect to the maximum value at time (t), $P_{\max}(t)$.

$$\bar{P}(t) = \frac{P(t)}{P_{\max}(t)} \tag{55}$$

This maximum value can be evaluated from the $\mathrm{GHI}_{\max}$ curve with the following equation:

$$P_{\max}(t) = \frac{\mathrm{GHI}_{csk}(t)}{\max(\mathrm{GHI}_{csk})}\mathrm{PV}_{\text{installed}} \tag{56}$$

9.3. Summary of the results for the different forecast schemes

This work aims to show the ability of the Kalman filter to perform reliable predictions of solar irradiance and PV power production. We have proposed two parameter identification techniques, namely the AR model and the EM algorithm, which can be used successfully to calibrate the optimization method for the forecasting.

9.3.1. Initialization phase

First, in order to run the AR and the EM initialization, we need only a few set of training data. The Kalman filter is an algorithm which performs prediction in the sense of a first-order Markov process. This means that it does not consider all the past information up to time t to give an estimate for time $t + 1$. Instead, we need only the arrived data at time t to have a prediction at the next time step, so we say that the algorithm performs a one-step ahead prediction. Another interesting feature about the Kalman filter is in its feedback nature; it is able to do robust prediction with a strong potential to reducing error estimates by minimizing the variance of this error (with the help of the Kalman gain). So, for our training data, for all variables, we consider a short time-interval formed with the first 20-30 days of year of the data. And the initial parameters are set at the beginning once a time (since we assume stationary of the underlying process) with this part of the dataset. For all our implementations, we have built an $AR(1)$ which is convenient to perform our forecasts, the measurement dataset consisting of a single vector. The rest of the data serves as test data.

The obtained initial parameters are then used to perform the calibration of the Kalman filter by running the predictor-corrector phases as described in Section 5.

We plot, in Figure 1, the measured GHI data and the clear sky index, for four days, to give an overview of the fluctuations of the solar irradiance, observed at 1 second time horizon.

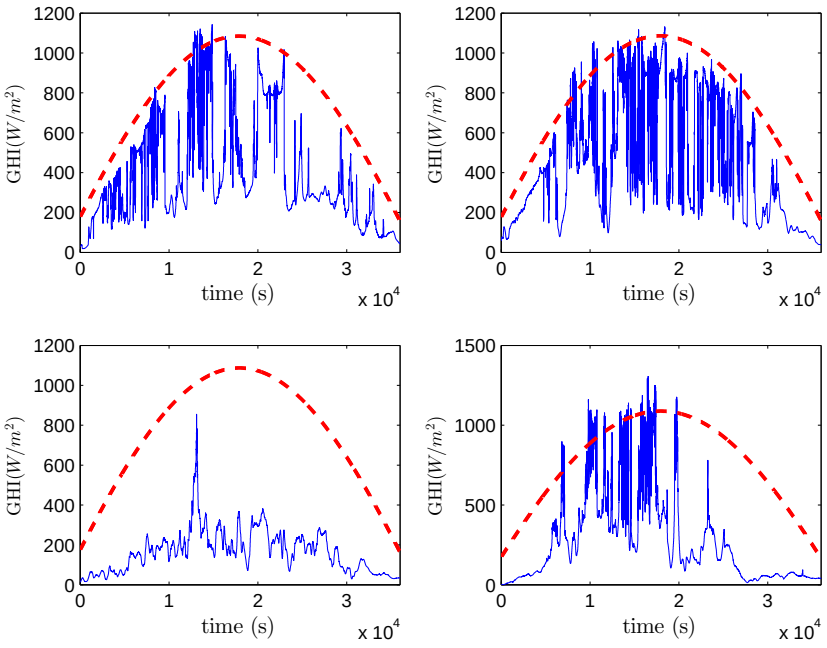


Figure 1. Variability of the GHI data, for 4 typical days. The blue line is the measured GHI and the dashed red line is the estimated clear sky with the Kasten model. Time Horizon is 1 second. These figures shows how fluctuating is the solar radiation, with ramps rates of about 800 W/m^2.

9.3.2. *Data preprocessing*

Before applying the whole procedure, we process the data to ensure it meets all the necessary assumption, i.e. the stationnarity of the underlying process. So, we apply the normalization procedure as described in Section 9.2. The PV data set is collected at 1 second timestamp and after we average it at a 1 hour bin increment. For the solar data, we have collected it also at 1 second and then, we make several averages for 1, 5, 10, 30 and 60 min. So, the Kalman filter is applied on the normalized data and thereafter, we reconsider the original signal without normalization, to build the final result and compute the performance measures. The exogenous variables are available only for the PV data.

9.3.3. *Results obtained for the PV data without the exogenous variables*

First, we deal with the PV forecast without the knowledge we have on the two exogenous variables, the ambient temperature and the cloud cover. The results are shown in the graphs of Figure. 2(a) and the asbolute error is shown on figure 3(a). For the scope of comparison, we plot the forecasts performed with the parameter tuning done by the AR and the EM algorithms. The Kalman filter is, first, applied to the normalized data and we apply the results to the real measurements i.e the not-normalized dataset. We can see that the predicted signal is a good forecast of the real PV at future horizon. The accuracy measures are confined in Table 1, with a nRMSE of 8.29% and 8.87% respectively for the EM and the AR calibration models. Globally the table says that the initialization method based on the EM algorithm is a little bit more efficient than the AR model calibration, but the mean absolute error introduced by the EM technique is more important than for the AR case, with the nMAE reaching 4.72% for EM versus 4.61% for AR. Another important issue is related to the performance gained with these technique above the classical persistence model, which gives us a nRMSE of 13.35%. Comparing the method against the basic persistence model, the two initialization methods give a skill score of 33.56%, for the AR and, 37.90% for the EM. Against the smart persistence, our forecasts gives a skill score slightly less, with 32.39% for the AR model and 36.81% for the EM technique. The difference between the two persistence models is not large since, the smart persistence model performs with slightly less error (nRMSE) than the basic persistence. In Figure. 2(d), we plot the measured versus predicted PV power and the linear polynomial fitting to show globally the modest performance of the EM initialization above the AR model.

Both for all cases, at least 80% of the data are in accordance with the fit, thus we can consider that the two methods are satisfactory.

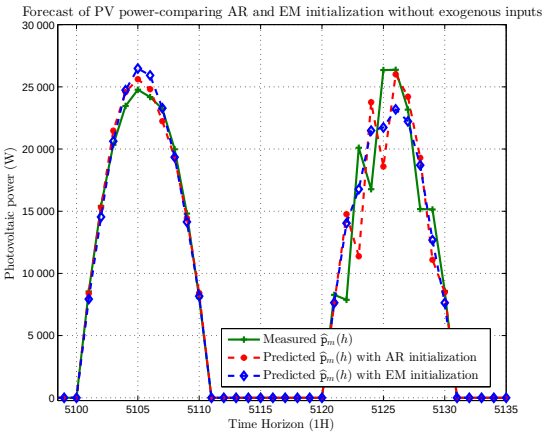
9.3.4. Results obtained for the PV data with the exogenous variables

In this study we were also interested in doing forecast for the PV with the knowledge of the exogenous variables, the temperature and the cloud cover which are taken as inputs. So we consider our system of Eq. 4 with the input B_tU_t containing the two variables. The matrix B_t is obviously set to the identity matrix since each variable is collected in a single point as an unique data vector. If, for example, the temperature were measured at different geographical areas, thus, the matrix B_t would be used properly to identify the source of the data each time step. The results are depicted in graphs of Figure. 2(b) and Figure. 2(c). Here, we plot the results to show how each given initialization procedure performs individually with and without the temperature and the cloud cover. The accuracy measures are depicted in Table 1. Here we obtain a nRMSE of 8.03% and 8.77% respectively for the EM and the AR initialization techniques. Also we show that each of the techniques performs slightly better with the inputs in terms of root mean square error, than without these exogenous variables. Nevertheless, the results are more sensitive to bias but, the mean absolute error is less than for the case we run the algorithm without the inputs features. For example, for the AR model without inputs, the nMBE and nMAE values reach respectively 0.45% and 4.61% vs 1.81% and 4.18% with the exogenous inputs. The EM algorithm gives, for the model with inputs, a nMBE of 0.42% and a nMAE of 4.57% against a nMBE of -0.08% and a nMAE of 4.72% for the model without inputs. Since the PV production depends strongly to the atmospheric conditions, it is very important to be able to do the prediction by incorporating this features in the model. We can accept that the results with inputs are more biased since the temperature and the cloud cover does not evolve in the same fashion and they might influence the production differently making the bias more important. Globally the results found by the two initialization procedures are good with a not negligible benefit for the EM which give less errors than the AR model and than with the same technique used without the exogenous variables. However the EM method is always more sensitive to bias. This model with exogenous inputs is also better than the basic persistence model with an improvement of the skill score which belongs between 34.31% and 39.85%. Also, we obtain better results against the smart persistence with a skill score of 33.16% and 38.80%, respectively for the AR and EM initialization models. The only point where the persistence model outperforms our model is about the bias, which is casually nonexistent.

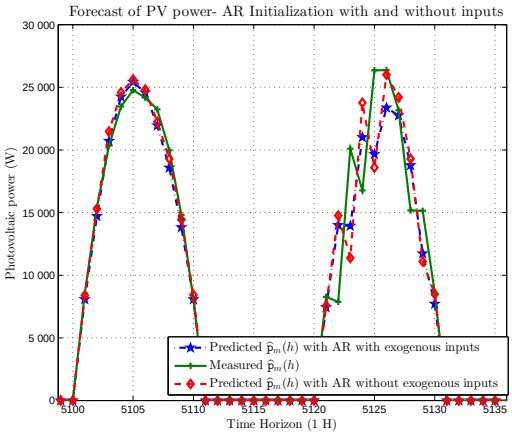
9.3.5. Results obtained for the solar data (GHI) without exogenous variables

For the solar data, we have not at hand the exogenous variables. These variables are only available for the PV data. For this data, we have done the calibration for time horizons 1, 5, 10, 30 and 60min but, for the scope of illustration, we plot,in graphs of Figures 4(a),4(b), only the results for time horizons 30 and 60 minutes, respectively. On figure 3(b) the asbolute error is shown for a time horizon of one hour. As for the previous results about the PV forecasts, the EM and AR techniques give good performance with a nRMSE in the interval [7.61%; 12.58%] against a nRMSE in the interval [7.79%; 16.94%] for the persistence model. We have almost the same results by comparing the obtained results with the smart persistence model. Comparing our two tuning algorithms, as we observe in Table 2, the three accuracy measures gives better for the EM than for the AR in terms of nRMSE. In Figure. 4(c), we find the trade-off between the measured and the predicted solar irradiance and their polynomial fit of degree 1. More generally, the two different initialization techniques give globally good results and they can be used according to the need and objective of the engineer, to do forecasting with several time scales. The results obtained for very short time scales (1, 5, 10min) has a skill score between 5.35% and 20.40%. These results are interesting since for short time scale we have very strong flcutuating conditions.

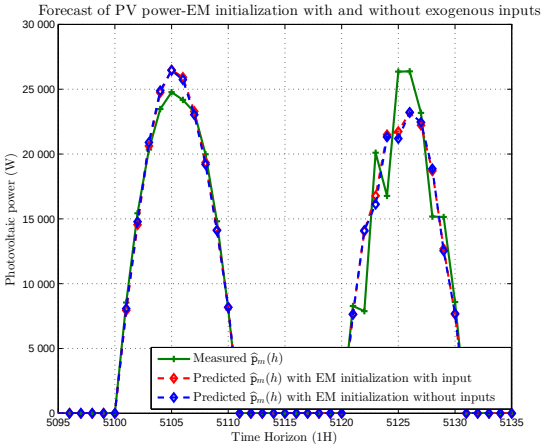
The performances with the time horizons 30 and 1H are much better than the persistence models; we can achieve an improvement with a skill score of 32.16% and 31.64% respectively for the basic and the smart persistence models. This results are promising since we are in presence of high fluctuating meteorological conditions for which it is not obvious to gather all the knowledge in order to reduce subsequently the error and the bias of the evolving process being modeled.



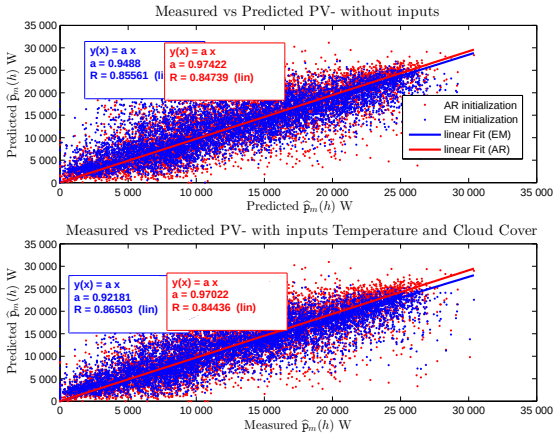
(a) Comparison of the two models AR and EM. The EM initialization model gives more reliable results than the AR.



(b) The AR model performs slightly better with exogenous variables than without these features.

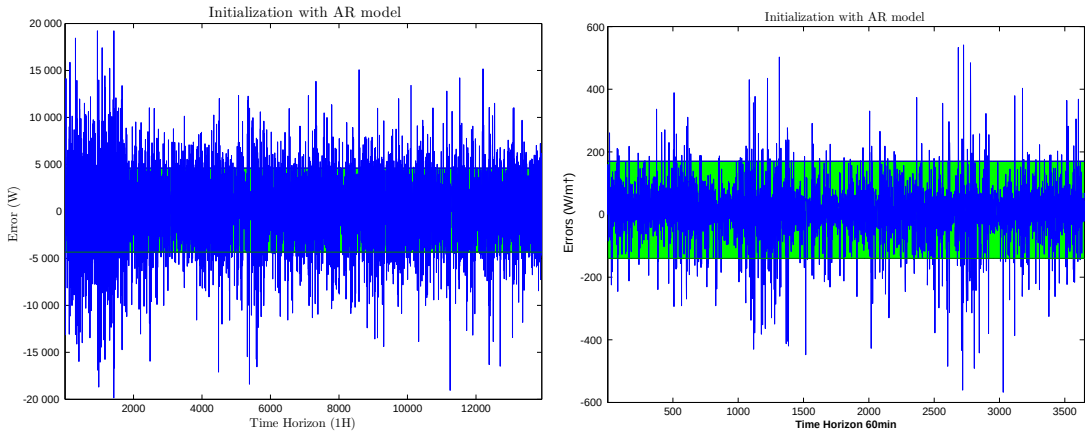


(c) The EM model performs slightly better with exogenous variables than without these features.



(d) Plot of the measured vs predicted PV data to show the performances of the two tuning algorithms, namely, AR and EM.

Figure 2. The graphs show the Kalman filter forecast for the PV data at time horizon 1H. In Figure (2(a)), we compare the AR and EM algorithms to show which one gives better results. In Figure (2(b)) and (2(c)) we plot the results for, respectively, the AR and EM models for parameter initialization, when we take into account the exogenous inputs (cloud cover and ambient temperature). In graph of Figure (2(d)) we measure and compare the performance of each initialization method w.r.t. the fit along with a 1-st order polynomial function.



(a) Forecast errors of the PV at Time Horizon 1H for one year of data. AR model.

(b) Forecast errors of the GHIat Time Horizon 1H for one year of data.

Figure 3. Kalman filter prediction absolute errors obtained with AR initialization methods for time horizon 1H, PV and solar. We have calculated and represented (green interval) the 95% confidence interval around the mean.

Table 1. Comparison of the different performance criterion for the different prediction models for the real PV with and without the exogenous inputs Temperature and the cloud cover. SS_p is the skill score calculated for the basic persistence model and SS_{sp} the skill score for the smart persistence. Time Horizon 1H.

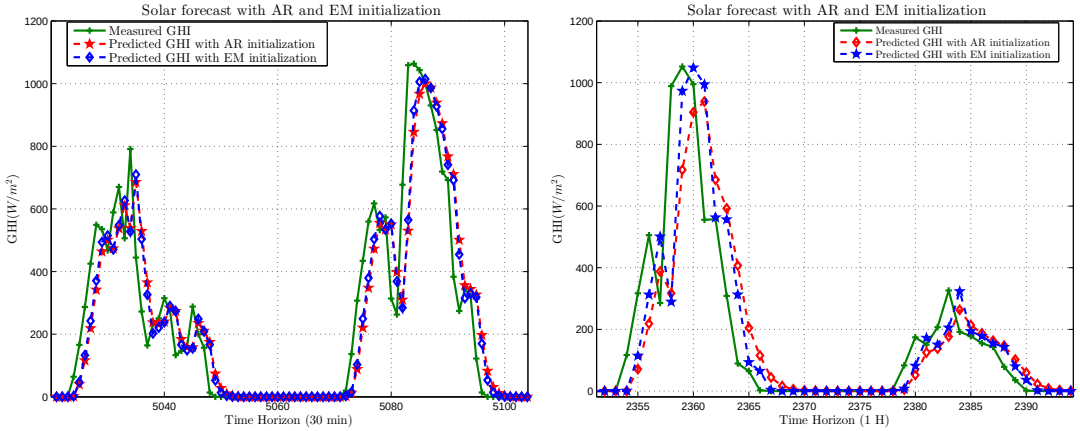
Without exogenous inputs					
	nRMSE (%)	nMBE (%)	nMAE (%)	SS_p(%)	SS_{sp}(%)
EM	8.29	-0.08	4.72	37.90	36.81
AR	8.87	0.45	4.61	33.56	32.39
Persistence	13.35	0.00024	8.13		
Smart Persistence	13.12	0.0084	8.04		
With exogenous inputs					
EM	8.03	0.42	4.57	39.85	38.80
AR	8.77	1.81	4.18	34.31	33.16

10. Conclusion

We have presented in this work a robust forecasting method based on the Kalman filter combined with a probabilistic initialization, Expectation Maximisation (EM) or Auto Regressive (AR) based. The model is built to be performed with both univariate or multivariate data. We test here it's ability to forecast solar radiation from 1 minute to one hour ahead, and photovoltaic power production for one hour ahead. The influence of exogenous inputs on the forecasting error is also evaluated. And finally we compare the forecasting errors to those obtained with the naive persitence model.

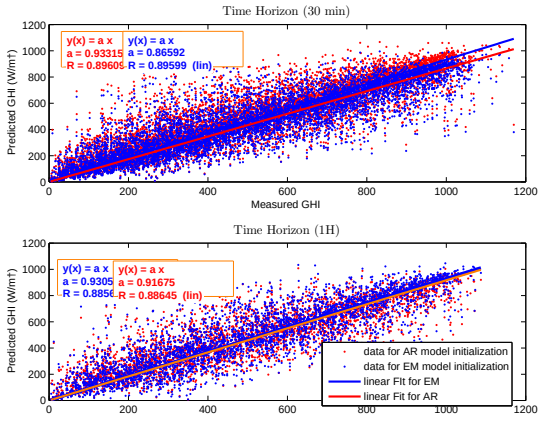
The proposed technique begins with the establishment of a dynamical state-space model able to capture the dynamics of the system we want to monitor. Thereafter, we propose different algorithms for the scope of parameter identification, since a great challenge in order to run properly the Kalman filter is the discovery of the initial values of the system quantities. A thorough description is performed about the usage of the defined system in its state-space form formulation which could help for system identification.

On the other hand, our state-space model is defined in a more elaborated form, convenient to including all the matrices which govern the whole system and which have a great importance to incorporate the relevant knowledge (features) to perform a suitable forecasting operation. Thereafter, the difficult and challengeable task of parameter tuning is overcome with two robust optimization techniques, namely the EM algorithm and the AR model, based on



(a) Forecast of the GHI at Time Horizon 30min. We plot on the same grpah the results obtained by the AR and EM initialization models.

(b) Same comparison as in Figure 4(a) for Time Horizon 60min.



(c) Real vs predicted Solar data for the AR and EM.

Figure 4. Kalman filter forecast for the Solar data (GHI) at time horizon 30min (4(a)) and 1H (4(b)). We compare the AR and EM algorithms. In graph (4(c)) we measure and compare the performance of each initialization method wrt. the fit along with a 1-st order polynomial function.

Table 2. Comparison of the different performance criterion for the different prediction models for the real Solar data (GHI) without exogenous inputs. SS_p is the skill score calculated for the basic persistence model and SS_{sp} the skill score for the smart persistence.

	nRMSE (%)	nMBE (%)	nMAE (%)	SS_p(%)	SS_{sp}(%)	Time Horizon
EM	7.61	0.81	3.91	7.42	6.51	1*min*
AR	7.78	0.40	4.02	5.35	4.42	
Persistence	8.22	0.00008	4.02			
Smart Persistence	8.14	0.00074	3.52			
EM	9.12	0.56	5.85	19.51	17.99	5*min*
AR	9.61	0.69	6.02	15.18	13.58	
Persistence	11.33	0.0006	6.96			
Smart Persistence	11.12	0.0032	5.98			
EM	10.22	2.68	7.37	20.40	19.15	10*min*
AR	10.36	0.84	6.85	19.31	18.04	
Persistence	12.84	0.0022	8.39			
Smart Persistence	12.64	0.047	8.01			
EM	11.23	4.31	9.19	31.06	30.29	30*min*
AR	11.46	1.18	8.18	29.65	28.86	
Persistence	16.29	0.0057	11.50			
Smart Persistence	16.11	0.0071	11.03			
EM	11.54	1.68	9.17	32.16	31.64	60*min*
AR	12.08	1.70	9.47	28.98	28.44	
Persistence	17.01	0.0057	12.46			
Smart Persistence	16.88	0.0086	12.21			

the maximum likelihood framework, despite the methods already presented in the literature. In addition, despite the fact that there's an important level of variability in the dataset, our methodology gives good results when accepting the stationarity of the process, this assumption having as positive consequence the calibration of the filter only once a time with a small amount learning dataset. All these aspects make the whole framework robust enough to perform the forecasts for heterogeneous dataset for several time scales.

The level of performance of the approach is studied with a set of standard accuracy measures largely used in the literature, named the root means square error, the mean bias error and the mean absolute error, in their normalized form. We have shown that goods results of the PV forecast can be obtained by taking into consideration some important exogenous variables, cloud cover and ambient temperature, that have a great influence in the production of energy. In this work, these features have been used as inputs.

Regarding the global solar radiation forecast, the results show that our methodology outperforms by far the classical persistence model with a skill score improvement reaching 39.85%. They might be used as variables to be estimated as we have done for the PV. Also, other variables should be incorporated in the model, as the speed of the wind, to reinforce the prediction. We will address this issue soon in our research.

The paper tackles an important problem related to the forecast of solar and PV production. We have proposed several models for parameter tuning because we believe that in high fluctuating meteorological conditions, it is not obvious to build a single model which is reliable to give accurate performances for many time scales. We think that it is a necessity to learn more about the variability of the solar irradiance in order to describe the whole system by a set of meaningful behavioral classes. So, one could apply in parallel each model to a set of classes according to their underlying characteristics. As a final benefit, the whole methodology could be implemented as a single framework to do at the same time solar irradiance and PV production prediction in order to help energy provider to control and manage carefully their industry.

References

[1] Akaike, H., 1971. Autoregressive model fitting for control. Annals of the Institute of Statistical Mathematics 23 (1), 163–180.

[2] Chaabene, M., Ammar, M. B., 2008. Neuro-fuzzy dynamic model with kalman filter to forecast irradiance and temperature for solar energy systems. Renewable Energy 33 (7), 1435–1443.

[3] Coimbra, Carlos, F. M., Kleissl, J., Marquez, R., 2013. Solar Energy Forecasting and Resource Assessment. Elsevier, Ch. Chapter 8 Overview of Solar-Forecasting Methods and a Metric for Accuracy Evaluation.

[4] Diagne, M., David, M., Boland, J., Schmutz, N., Lauret, P., 2014. Post-processing of solar irradiance forecasts from wrf model at reunion island. Solar Energy 105, 99–108.

[5] Galanis, G., Louka, P., Katsafados, P., Pytharoulis, I., Kallos, G., 2006. Applications of kalman filters based on non-linear functions to numerical weather predictions. In: Annales Geophysicae. Vol. 24. Copernicus GmbH, pp. 2451–2460.

[6] Ghahramani, Z., Hinton, G. E., 1996. Parameter estimation for linear dynamical systems. ACM Transactions on Mathematical Software (TOMS) TOMS Homepage archive.

[7] Hassanzadeh, M., Etezadi-Amoli, M., Fadali, M, S., 2010. Practical approach for sub-hourly and hourly prediction of pv power output. In: North American Power Symposium (NAPS), 2010. IEEE, pp. 1–5.

[8] Kailath, T., Sayed, A. H., Hassibi, B., 2000. Linear estimation. Vol. 1. Prentice Hall Upper Saddle River, NJ.

[9] Kalman, R. E., Bucy, R. S., 1961. New results in linear filtering and prediction theory. Journal of Fluids Engineering 83 (1), 95–108.

[10] Kasten, F., 1980. A simple parameterization of two pyrheliometric formulae for determining the linke turbidity factor. Meteorol. Rundsch. 33, 124–127.

[11] Kaur, A., Nonnenmacher, L., Pedro, H, T. C., Coimbra, C, F. M., ???? Benefits of solar forecasting for energy imbalance markets.

[12] Lorenz, E., Remund, J., Müller, S. C., Traunmüller, W., Steinmaurer, G., Pozo, D., Ruiz-Arias, J., Fanego, V. L., Ramirez, L., Romeo, M. G., et al., 2009. Benchmarking of different approaches to forecast solar irradiance. In: 24th European photovoltaic solar energy conference, Hamburg, Germany. Vol. 21. p. 25.

[13] Louka, P., Galanis, G., Siebert, N., Kariniotakis, G., Katsafados, P., Pytharoulis, I., Kallos, G., 2008. Improvements in wind speed forecasts for wind power prediction purposes using kalman filtering. Journal of Wind Engineering and Industrial Aerodynamics 96 (12), 2348–2362.

[14] Mayne, D, Q., 1966. A solution of the smoothing problem for linear dynamic systems. Automatica 4 (2), 73–92.

[15] Ndong, J., 2014. A new approach to anomaly detection based on possibility distributions. In: INTERNET 2014, The Sixth International Conference on Evolving Internet. pp. 1–8.

[16] Ndong, J., 2014. Using sub-optimal kalman filtering for anomaly detection in networks. Proceeding of the Electrical Engineering Computer Science and Informatics 1 (1), 408–414.

[17] Ndong, J., Salamatian, K., 2011. A robust anomaly detection technique using combined statistical methods. In: CNSR. IEEE Computer Society, pp. 101–108.

[18] Ndong, J., Salamatian, K., 2011. Signal processing-based anomaly detection techniques: a comparative analysis. In: INTERNET 2011, The Third International Conference on Evolving Internet. pp. 32–39.

[19] Neumaier, A., Schneider, T., 2001. Estimation of parameters and eigenmodes of multivariate autoregressive models. ACM Transactions on Mathematical Software (TOMS) TOMS Homepage archive 27 (1), 27–57.

[20] Pelland, S., Galanis, G., Kallos, G., 2013. Solar and photovoltaic forecasting through post-processing of the global environmental multiscale numerical weather prediction model. Progress in Photovoltaics: Research and Applications 21, 284–296.

[21] Rabiner, L., 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77 (2), 257–286.

[22] Schneider, T., A., N., 2001. Algorithm 808: Arfit - a matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. ACM Transactions on Mathematical Software (TOMS) TOMS Homepage archive 27, 58–65.

[23] Schwarz, G., 1978. Estimating the dimension of a model. The annals of statistics 6 (2), 461–464.

[24] Shumway, R. H., Stoffer, D. S., 1982. An approach to time series smoothing and forecasting using the em algorithm. Journal of time series analysis 3 (4), 253–264.

[25] Shumway, R. H., Stoffer, D. S., 1991. Dynamic linear models with switching. Journal of the American Statistical Association 86 (415), 763–769.

[26] Soubdhan, T., Emilion, R., Calif, R., 2009. Classification of daily solar radiation distributions using a mixture of dirichlet distributions. Solar energy 83 (7), 1056–1063.