Day-ahead probabilistic forecast of solar irradiance: a Stochastic Differential Equation approach

Jordi Badosa, Emmanuel Gobet, Maxime Grangereau, Daeyoung Kim

Day-ahead probabilistic forecast of solar irradiance: a Stochastic Differential Equation approach

Jordi Badosa, Emmanuel Gobet, Maxime Grangereau and Daeyoung Kim

Abstract In this work, we derive a probabilistic forecast of the solar irradiance during a day at a given location, using a stochastic differential equation (SDE for short) model. We propose a procedure that transforms a deterministic forecast into a probabilistic forecast: the input parameters of the SDE model are the Arome deterministic forecast computed at day D-1 for the day D. The model also accounts for the maximal irradiance from the clear sky model. The SDE model is mean-reverting towards the deterministic forecast and the instantaneous amplitude of the noise depends on the clear sky index, so that the fluctuations vanish as the index is close to 0 (cloudy) or 1 (sunny), as observed in practice. Our tests show a good adequacy of the confidence intervals of the model with the measurement.

KEYWORDS: solar power, probabilistic forecast, stochastic differential equation

1 Introduction

Context. As the conventional energy sources such as coal, oil, and gas are considered to be one the main factors responsible for climate change, solar energy has been recognized as a viable alternative. Solar powered plants which utilize photovoltaic (PV) and Concentrated Solar Power (CSP) have penetrated the electricity

Jordi Badosa
LMD/IPSL, Ecole Polytechnique, Université Paris-Saclay, Route de Saclay, 91128 Palaiseau cedex, France, e-mail: jordi.badosa@lmd.polytechnique.fr

Emmanuel Gobet
CMAP, Ecole Polytechnique, e-mail: emmanuel.gobet@polytechnique.edu

Maxime Grangereau
CMAP, Ecole Polytechnique, e-mail: maxime.grangereau@polytechnique.edu

Daeyoung Kim
Ecole Polytechnique, e-mail: daeyoung.kim@polytechnique.edu

grid, and they are contributing to meet the electricity balance between production and consumption. The need for clean energy and energy-generation independence has driven the solar farms to grow drastically in number. This is backed up by policy supports including feed-in tariffs.

One of the challenges of the electricity generation through solar panels is that it's intermittent and driven by meteorological conditions (mainly related to cloudiness), which results in high uncertainty in the final amount of production. Such uncertainty put grid operators at risk since they might have to adjust their production on very short notice and energy storage capacity is limited. Solar irradiance forecast in this context is crucial not only to predict the solar power generation amount but also to save the start and shutdown costs of conventional generators [MBF$^+$16].

Use of day-ahead predictions. Day-ahead predictions are of particular importance for application in the energy market, where the day-ahead auction of power price plays a major role in many countries [PLP$^+$13]. For example, in France, one may take part in the bidding by the noon on the day before the operation[1]. Besides the market participation, the day-ahead prediction can also be useful for unit commitment and energy storage dispatch [DFM15, HKNF14]. Simulation studies have shown that day-ahead forecasts may provide significant cost savings to grid operators and fast start and lower efficiency power plants [MBF$^+$16]. PV-based microgrids also make use of day-ahead predictions for power planning [KLC$^+$11].

Regarding the types of day-ahead solar irradiance forecast, we can distinguish two categories: deterministic forecasts and probabilistic forecasts. In the state of the art of irradiance forecasting, most of the literature relies on deterministic forecasts, also known as point forecasts, since their outputs are specific *values* at given times in the future. On the other hand, most probabilistic forecasts give the full *distributions* of the values of interest at the times considered. Each one of these distributions can be represented by histogram or cumulative density function for example. The representation of uncertainty information takes into account potentially extreme scenarios. It also allows the operators to gain additional trust in the forecasts [OAB$^+$15]. Furthermore, having at hand a probabilistic forecast for the solar irradiance simultaneously at any hour (continuous-time forecast), accounting for the inter-temporal probabilistic dependency allows to properly solve some problems where the full distribution of the inputs plays an important role (see for instance [GG17]). For applications in energy management, see [ACLP14] and for the connection with electricity derivatives, see [Aid15]. The continuous-time forecast may also be updated whenever new data becomes available.

Actually, methods for deriving predictions depend much on both the considered time horizon and the amount of data available at the time when the prediction is made. Moreover, the flows of data and the frequency for updated predictions are tightly related. For horizons of minutes to hours, satellite images (see *Meteosat* for Europe) are much informative on a global area, a new image is available every 15'; on the other hand, the access is not granted to anyone. For very local geographic

[1] see https://www.epexspot.com/en/market-data/dayaheadauction

data, one can be equipped with sky cameras or simply irradiance sensors, the latter being the most common way of collecting data.

In this study, we focus on the day-ahead horizon and we consider that deterministic Numerical Weather Predictions (NWP) of solar irradiance are available at local scale as it is the case of the Meteo-France's AROME NWP data, which has a grid resolution of 1.3 km. This data is available for any individual prosumer and might for instance be used to manage batteries in order to reduce variability of the demand on the grid, as in [GG17].

Deterministic and probabilistic forecasts. In the deterministic forecast of day-ahead solar irradiance, NWP models are widely used, and the model value goes through Model Output Statistics (MOS) before the actual usage. MOS is a post-processing technique used to interpret empirically the outputs of a numerical model and produce site-specific forecasts [HLG06, DLD12]. Statistical learning methods are often used to correct errors in the NWP model outputs and to incorporate knowledge from several models, by appropriately weighing them. These methods allow to correct biases and systematic errors in the forecasts [Kle13]. MOS is known to improve the performance of the raw forecast of NWP by about 10-15% [TZH$^+$15].

While the deterministic forecast has been developed for more than thirty years, the probabilistic forecast seems to be in its infancy yet. However, it seems to have a high potential of usefulness, especially in applications where risk quantification is a crucial factor, for example energy storage in connection with intermittent energy sources.

One noticeable method of probabilistic day-ahead solar irradiance forecast is the analog ensemble approach, which searches the history for similar forecasts and makes corrections to the forecast according to the error in prior forecasts. In this approach, the prior analogs become an ensemble that quantifies the uncertainty of the forecast [ADSC15]. These approach rely on non-parametric techniques and machine learning tools; on the one hand, this is a data-driven approach and therefore it is quite flexible; on the other hand, since it is not aimed at identifying a specific stochastic model, analytical or numerical methods are not applicable, which limits the resolution of some problems (like stochastic control problems). In [ADSC15] an example of ensemble approach using machine-learning based regression models – such as decision tree, K-nearest neighbors, random forests – is presented, showing that each of these models performs better than the Auto-Regressive Integrated Moving Average (ARIMA for short) model [MA16, SLSN15].

Another method is the stochastic modeling, where the evolution of the system is described by a Stochastic Differential Equation (SDE). Knowledge coming from other deterministic forecasts can be incorporated in these models. As explained in [IMMM14], SDE models have several assets: we can incorporate boundedness properties, which are essential for correct modeling of the solar irradiance, and the SDEs are more general than other classical time-series models, like ARIMA processes. Besides, the output of the SDE models can be both a point forecast, by simulation of a single trajectory, or a probabilistic forecast. Indeed, by simulating multiple independent Markovian evolutions of the process, it is possible to infer

the full distribution of irradiance at any time (see Section 2.4). A first attempt for such a SDE modeling is proposed by [SE10] with a simplistic model. The authors of [IMMM14] have designed extensions and more involved models with different degrees of complexity. They assume a parametric SDE model for the solar irradiance and suppose that the observations are noisy. A convenient feature of their SDE model is to account for the maximum irradiance (also known as clear sky model, [BH81]). For the model estimation, they cope with SDE inference with noisy data: for this, they use Kalman-type filtering techniques, which restrict their SDE model to additive Brownian noise. In the current study, we also consider a SDE model but its form is different from [IMMM14] and we assume perfect observation of the data, i.e. we assume there is no noise in the observations, see details below.

Our contribution. Our purpose is to turn deterministic forecasts into probabilistic forecasts. Our framework is to consider that a single deterministic forecast is available on the day before (i.e. on D-1): in our case, this comes from Météo-France (AROME NWP data). The precise description of these data is given in Section 2.2. To model the irradiance on the day D, we use a time-dependent SDE, which models the evolutions of the Clear Sky Index (CSI for short), which is the ratio between the observed irradiance and the maximal theoretical irradiance that would be observed in perfect weather conditions. Therefore, the CSI lies in the range $[0,1]$. It will be denoted X in the equations. Some parameters of the SDE for a given day D are fixed (see forthcoming paragraph on the estimation of the parameters), while others are estimated using the Arome data at day D-1. Simulating the SDE gives realistic scenarios of solar irradiance, as demonstrated in Section 3. The SDE is driven by a Brownian motion W and it has the following form:

$$\mathrm{d}X_t = -a(X_t - x_t^{\texttt{forecast}})\mathrm{d}t + \sigma X_t^{\alpha}(1-X_t)^{\beta}\mathrm{d}W_t, \qquad t \in [t_0,t_1] \subset [0,24], \tag{1}$$

for some parameters $a, \sigma, \alpha, \beta, (x_t^{\texttt{forecast}})_{0\leq t\leq 24}$ and where $[t_0,t_1]$ is the period (expressed in hours) of the day D where the sun shines at the location where irradiance is measured. The Brownian motion W allows to model the uncertainty across time. All the above parameters are identified using the Arome data at day $D-1$ and some historical data. In Section 3, we show that simulations from this model generate confidence intervals that are fully consistent with the realizations collected over a period of one year.

We now stress the similarities and differences with [IMMM14]. The form of the SDE is similar in both cases: the drift is a mean-reversion term and the diffusion term vanishes at 0 and 1. However, the form of our diffusion term is more general and allows to take into account more general form of dependence of the fluctuations, while the drift term in [IMMM14] incorporates more parameters. It would be an interesting lead to try to incorporate more parameters in the drift term of the SDE we propose as well, but one needs to make sure to have enough data to estimate them properly. Besides, in our study, the observation is assumed to be direct: by ignoring the observation noise, we do not need anymore to use filtering techniques and we can directly make use of appropriate statistical methods for SDEs [KLS12]. This is

what allows us to consider a more general diffusion coefficient, namely of the form $x^\alpha(1-x)^\beta$ with general exponents (in Section 2.3.2, we find $(\alpha,\beta)\approx(0.8,0.7)$), while the authors of [IMMM14] take $\alpha=\beta=1$ so that, via a Lamperti transformation, they can get back to additive noise model. Last, our SDE directly models the evolution of the clear sky index, while in [IMMM14], it models the evolution of the irradiance (altough the maximal irradiance clear sky model is incorporated in the SDEs from their second model).

The estimation of the parameters of our model turns out to be fairly satisfactory, using historical data for measurements and one-day in advance predictions (Arome data). We are able to reproduce accurately enough the forecast uncertainty (see tests in Section 3); it is quite remarkable, especially because only a small amount of data is used to build the predictions (mainly AROME data).

2 Uncertainty modeling from data

2.1 Definitions and notations

To be accurate, the solar irradiance (or simply irradiance in the following) that we consider refers to Global Horizontal Irradiance (GHI), that is the amount of solar radiation that reaches the surface of the Earth on a horizontal plane (its unit is is W/m^2). The GHI depends on the location; in our study, it is at the latitude and longitude of 48.713°N and 2.208°E, at SIRTA (atmospheric observatory at Ecole Polytechnique, on the Paris-Saclay campus, http://sirta.ipsl.fr).

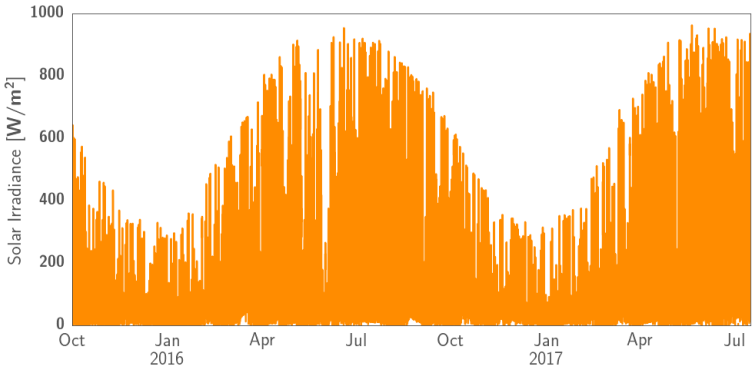


Fig. 1: Measurements of Global Horizontal Irradiance from SIRTA (48.7°N, 2.2°E.) for the considered period.

For different days, we consider the irradiance $I(t)$ (at the above location) as a function of time $t\in[0,24]$ (in hours, for a given day). Theoretically, the irradiance

evolves between two bounds, 0 and $I^{\texttt{clear sky model}}(t)$, where the second is the maximal irradiance obtained under clear sky conditions , that is, a cloudless sky. In our case, the latter has been calculated with an empirically-derived equation that takes into account the Sun-Earth geometry and the day of the year:

$$\forall t \in [t_0, t_1],\ I^{\texttt{clear sky model}}(t) = [83.69 \sin(\frac{2\pi}{365.24}(D+82.07)) + 1130.44] \cos(\theta_z(t))^{1.2}$$

where D is the day of the year $[0, 365]$, and $\theta_z(t)$ is the solar zenith angle. Before the sunrise ($t \leq t_0$) and after the sunset ($t \geq t_1$), $I^{\texttt{clear sky model}}(t)$ is set to 0 and there is no question about uncertainty in the irradiance. From now on, we mainly stick to the case $t \in [t_0, t_1]$ (of course, this interval evolves along the days and seasons). Our aim is to propose a stochastic model for the Clear-Sky-Index (CSI) given by

$$X_t = \frac{I(t)}{I^{\texttt{clear sky model}}(t)}. \tag{2}$$

X close to 1 (resp. 0) corresponds to clear-sky day (resp. dark clouds) while intermediate values relate to a variable sky (or light clouds).

To calibrate the model parameters, we also use the Accumulated CSI of the day D, defined by

$$\text{ACSI}_D = \frac{\sum_{t \in \mathscr{T}} I(t)}{\sum_{t \in \mathscr{T}} I^{\texttt{clear sky model}}(t)}, \tag{3}$$

where $\mathscr{T} := \{\text{measurement times in the day D}\}$. This index is an indicator of the overall clearness on day D.

2.2 Data: AROME forecast and SIRTA measurement

In our study, we consider the forecast GHI values from AROME NWP operated by Météo-France, for the closest grid point to SIRTA site. In particular, we use the run from 12:00 UTC on the previous day (D-1) of the target forecast day (D), which has a time step of 1 hour and covers the whole day D.

The considered period is 2015 October 1st to 2017 July 16th (see Figure 1). Only days with both measured and forecast GHI available values were considered, which made a final dataset of 473 days. The measured data was taken at 10-minutes resolution. Forecast values with 10-minutes resolution were generated from the forecast with 1-hour resolution using linear interpolation.

2.3 Fitting the SDE model

2.3.1 Heuristic derivation

The stochastic model we propose for the irradiance CSI is aimed at giving a precise meaning to the approximation

$$X_t \approx x_t^{\texttt{forecast}} + \text{error}(t) \tag{4}$$

and of its continuous-time evolution, where $x_t^{\texttt{forecast}}$ is calculated from the AROME forecast

$$x_t^{\texttt{forecast}} := \frac{I^{\texttt{forecast}}(t)}{I^{\texttt{clear sky model}}(t)} \in [0,1]. \tag{5}$$

We assume that $x_t^{\texttt{forecast}}$ is given in continuous time (as mentioned before, we use linear interpolation to compute values at times where no forecast is available in the data). The term $\text{error}(t)$ stands for the unpredictable part in the prediction and reflects the uncertainty of the forecast: observe that, since the CSI lies in the range $[0,1]$, it is certainly not appropriate to assume that the error has a Gaussian distribution.

To begin with, the deterministic forecast $x_t^{\texttt{forecast}}$ is considered as a (time-dependent) mean-reversion level, i.e. when the realized CSI is far from the forecast CSI, it is expected that (in mean) X_t gets closer to $x_t^{\texttt{forecast}}$. In other words, a temporary prediction error is possible but it tends to vanish. We model this feature via a dissipative linear Ordinary Differential Equation that describes the time-evolution[2] $t \mapsto \mathbb{E}[X_t]$,

$$\mathrm{d}\mathbb{E}[X_t] = -a(\mathbb{E}[X_t] - x_t^{\texttt{forecast}})\mathrm{d}t \tag{6}$$

for some mean-reversion speed parameter $a > 0$ that will be estimated later.

We now model the stochastic fluctuations around the above relation, i.e. we write

$$\mathrm{d}X_t = -a(X_t - x_t^{\texttt{forecast}})\mathrm{d}t + \text{noise}(\mathrm{d}t).$$

First, the amplitude of the noise over an infinitesimal interval $[t, t+\mathrm{d}t]$ has to decrease as the CSI gets closer to 0 (cloudy sky) and 1 (clear sky): this is requested to maintain the model values in the unit interval. For this reason, we set

$$\text{noise}(\mathrm{d}t) = X_t^{\alpha}(1-X_t)^{\beta}\widetilde{\text{noise}}(\mathrm{d}t) \tag{7}$$

for two positive parameters α, β. Once the latter will be chosen appropriately, we expect that the amplitude of the noise $\text{noise}(\mathrm{d}t)$ is well tuned so that the newly-renormalized noise $\widetilde{\text{noise}}(\mathrm{d}t)$ does not depend anymore on the CSI (see later Figure 3).

Moreover, as explained in the next paragraph, a statistical analysis of $\widetilde{\text{noise}}(\mathrm{d}t)$ shows that it can be modeled by a Gaussian distribution $\mathcal{N}(0, \sigma^2 \mathrm{d}t)$ and from now

[2] $\mathbb{E}[.]$ is the expectation attached to the forthcoming probabilistic model.

on, we take it as a Brownian increment $\sigma \mathrm{d}W_t$. This is not the only Gaussian model, but it has the advantage of a very low parametric dimension (which is a nice asset when one deals with few data).

Gathering all the previous arguments, we finally obtain a SDE of the form

$$\mathrm{d}X_t = -a(X_t - x_t^{\texttt{forecast}})\mathrm{d}t + \sigma X_t^{\alpha}(1-X_t)^{\beta}\mathrm{d}W_t, \quad t \in [t_0, t_1], \tag{8}$$

with a given initial value $X_{t_0} \in [0,1]$. In the case of $x_t^{\texttt{forecast}}$ constant, and $\alpha = \beta = 1$, we retrieve the well-known Fisher-Wright diffusion used in mathematical genetics. In the following, we will take $(\alpha, \beta) \in [\frac{1}{2}, 1] \times [\frac{1}{2}, 1]$.

From a mathematical point of view, we can justify that this SDE model is well-posed, despite that the coefficients are not globally Lipschitz and that it takes values in the unit interval $[0,1]$ as requested. These properties are rigorously proved in Appendix.

2.3.2 Statistical inference of the parameters

As it is well-know in the statistics for SDEs [Gob02, KLS12], the estimation of the parameters entering in the drift and diffusion coefficients can be made independently, asymptotically as the frequency gets larger and larger.

Estimation of a. The Ito formula gives that, for any $t \geq s$,

$$e^{at}X_t = e^{as}X_s + \int_s^t e^{au} x_u^{\texttt{forecast}} \mathrm{d}u + \int_s^t e^{au}\sigma X_u^{\alpha}(1-X_u)^{\beta}\mathrm{d}W_u.$$

Since the first time integral is deterministic, it readily follows that

$$\mathbf{C}\mathrm{ov}\left(\int_s^t e^{au} x_u^{\texttt{forecast}} \mathrm{d}u, X_s\right) = 0.$$

Moreover the stochastic integral is computed on $[s,t]$ and the SDE solution is adapted to the Brownian filtration, therefore

$$\mathbf{C}\mathrm{ov}\left(\int_s^t e^{au}\sigma X_u^{\alpha}(1-X_u)^{\beta}\mathrm{d}W_u, X_s\right) = 0.$$

To summarize, we get $\mathbf{C}\mathrm{ov}\,(e^{at}X_t, X_s) = e^{as}\mathbf{V}\mathrm{ar}(X_s)$, that is

$$\mathbf{C}\mathrm{ov}(X_t, X_s) = e^{-a(t-s)}\mathbf{V}\mathrm{ar}(X_s), \qquad t \geq s. \tag{9}$$

In (9) we retrieve a nice relation available for any SDE which drift is linear, see [BSS05]. Therefore, it is enough to compute the correlogram of the process X from the data and to extract the parameter a using an exponential fit. In our case, see Figure 2, we obtain $a \approx 0.75\,h^{-1}$.

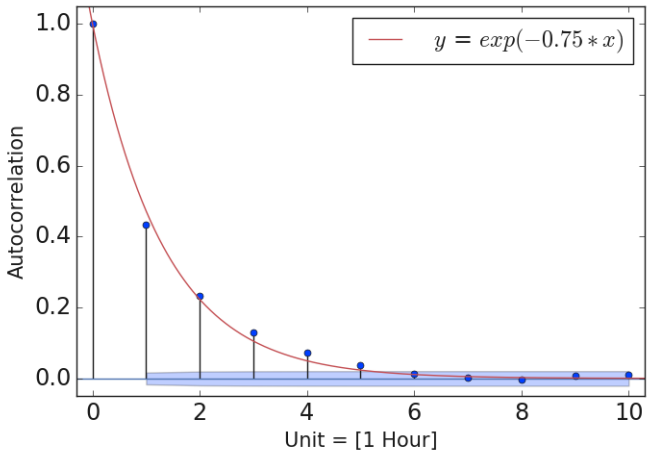


Fig. 2: Correlogram

Estimation of σ, α, β. For high-frequency data (that is when the monitoring frequency is considered high), the increments of $X_t - X_s, t \geq s$ can be approximated by

$$X_t - X_s \underset{t-s \text{ small}}{\approx} \sigma X_s^{\alpha}(1-X_s)^{\beta}(W_t - W_s). \tag{10}$$

Indeed, as $t-s$ is small, the drift term in (8) is of magnitude $t-s$, which can be neglected compared to the order $\sqrt{t-s}$ arising from $W_t - W_s$. This is the usual Euler approximation of SDE in small time [KP10]. Of course, one needs to make sure that the time step used in the simulations is small enough to guarantee that X remains in its domain $[0,1]$.

In our model, we seek a SDE parametrization that, each day, adapts automatically to the Arome D-1 forecast. Our strategy is

- first to identify a pair of exponents (α, β) that will be available in average for all the days (these parameters are somehow day-invariant),
- then to estimate σ that may be day-dependent and that will be estimated from Arome D-1 forecast. In the sequel, the parameter σ for the day D will be written $\sigma_D > 0$.

Because of the relation (10), for measurements collected at a 10'-period ($t_{i+1} - t_i = 10' = \delta$) in the same day D, we expect to have

$$\frac{X_{t_{i+1}} - X_{t_i}}{X_{t_i}^{\alpha}(1-X_{t_i})^{\beta}} \overset{d}{\approx} \mathcal{N}(0, \sigma_D^2 \delta), \tag{11}$$

where $\overset{d}{\approx}$ means *approximation in distribution*, with an independence property between different times t_i (thanks to the properties of the Brownian increments). Therefore, it is enough to estimate empirically the variance (day by day) of the

quantities (11) to get an estimator of $\sigma_D^2\delta$ for that day D, and therefore of σ_D. Actually, instead of considering increments of the irradiance CSI, we take the increments of the forecast errors, i.e.

$$\mathrm{REI} = \frac{\left[X_{t_{i+1}} - I_{t_{i+1}}^{\texttt{clear sky model}}\right] - \left[X_{t_i} - I_{t_i}^{\texttt{clear sky model}}\right]}{X_{t_i}^{\alpha}(1-X_{t_i})^{\beta}} \tag{12}$$

which asymptotic behavior with respect to $\delta \to 0$ is similar to (11) (indeed, the curve of forecast $I^{\texttt{clear sky model}}$ is of finite variation). REI stands for Renormalized Error Increment.

We could identify the parameters $(\alpha, \beta, \sigma_{D_1}, \ldots, \sigma_{D_{473}})$ using a Maximum Likelihood Estimation technique (via the Gaussian approximation (11)), requiring an optimization procedure in dimension 475! We proceed in a simpler way, by choosing α and β so that the σ_D's computed as mentioned (day by day) have a stationary behavior throughout different days. This is a way to have a single pair (α, β) independently of the day of the year. For $\alpha = 0.8$ and $\beta = 0.7$, Figure 3 shows that σ_D are quite decorrelated from ACSI_D (Accumulated Clear Sky Index) defined in (3). Because the ACSI is computed on the measurement (see (3)), it makes its use not

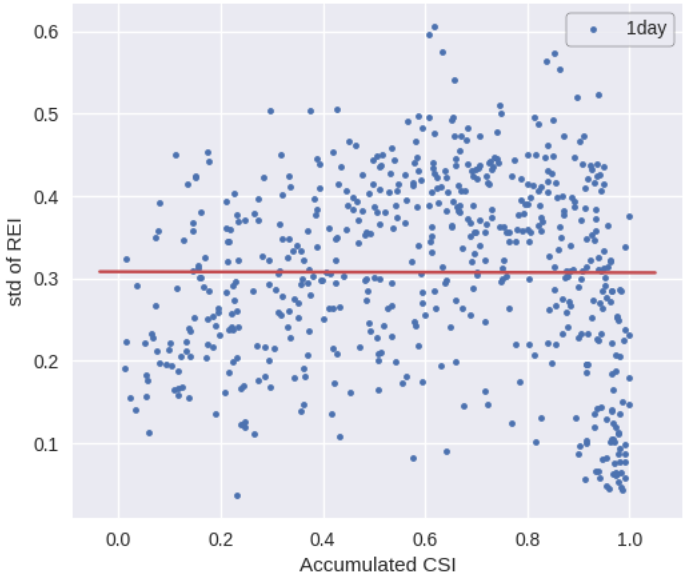


Fig. 3: For each day D in the data set, the standard deviation $\sigma_D\sqrt{\delta}$ of the Renormalized Error Increment as a function of Accumulated Clear Sky Index ACSI_D, when $\alpha = 0.8$ and $\beta = 0.7$.

applicable as it is in the perspective of probabilistic forecast. Therefore, we seek to model σ_D in terms of the Arome forecast on day $D-1$. Since σ_D reflects the amplitude of local variability of the irradiance, we use the Average Time Increment CSI

computed on the Arome forecast as a surrogate:

$$\text{ATICSI}_D = \sum_{t_i \in \mathscr{T}^{\text{forecast}}} \left| \frac{I^{\texttt{forecast}}(t_{i+1})}{I^{\texttt{clear sky model}}(t_{i+1})} - \frac{I^{\texttt{forecast}}(t_i)}{I^{\texttt{clear sky model}}(t_i)} \right|, \tag{13}$$

where $\mathscr{T}^{\text{forecast}} := \{\text{Arome forecast times}\}$. Figure 4 depicts the relation between σ_D and ATICSI_D.

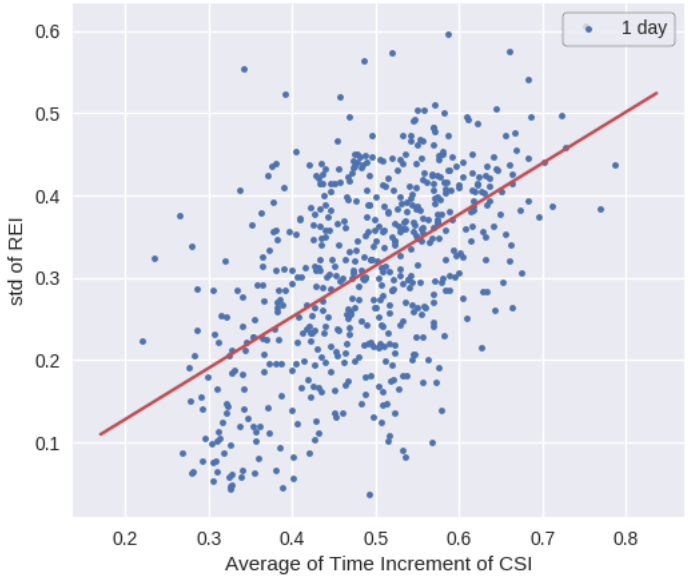


Fig. 4: For each day D in the data set, the standard deviation $\sigma_D\sqrt{\delta}$ as a function of the Average Time Increment CSI defined in (13).

We exhibit a linear relation of the form

$$\sigma_D\sqrt{\delta} = 0.622 \times \text{ATICSI}_D + 0.0004 + \text{error}. \tag{14}$$

In the next experiments, we set the residual error in (14) to 0. Actually, the distribution of this error (see Figure 5) is closed to a Gaussian one, which could be included in our model of σ_D to possibly improve the probabilistic forecast. We haven't gone further in that direction since with the simplified version (error $= 0$), our tests reveal a good accuracy of the probabilistic forecast.

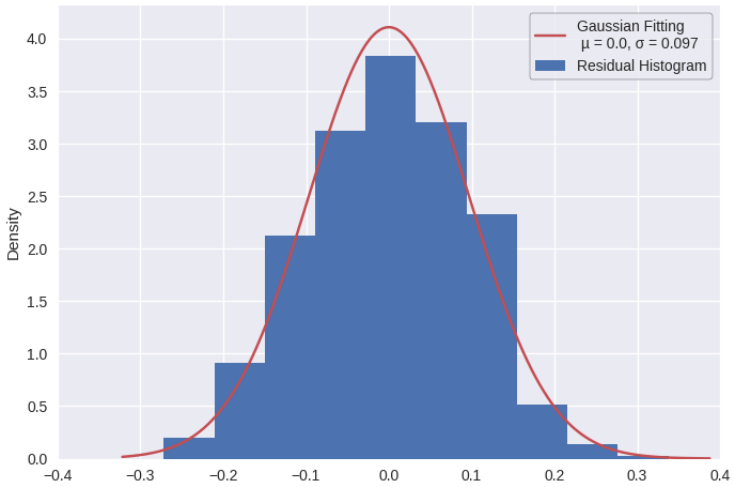


Fig. 5: Histogram of the residual error in (14), with the Gaussian density centered at 0 and with standard deviation 0.097.

2.4 Probabilistic forecast computation

We present how we compute numerically the distribution of irradiance in our model on the day D, given the Arome forecast of day D-1.

Using the values of the parameters derived in the calibration step described earlier, from (8) we are now able to simulate independent trajectories of the clear sky index (and therefore of the solar irradiance accounting for the clear sky model in (2)). In our tests, usually we sample $M = 1000$ or $M = 10000$ paths, in order to produce accurate enough statistics and confidence intervals.

Regarding the simulation scheme itself, once the initial value X_{t_0} is chosen, we just need to simulate independent Markovian evolutions. More precisely, we use the Euler scheme (see [KP10] for an account on the subject of simulating SDEs). In our tests, the time step of the Euler scheme is $\Delta_t = 1'$ and when the value of the Euler scheme is outside $[0,1]$, the value is pushed back to $[0,1]$ (as the exact solution). Set $t_k = t_0 + k\Delta_t$; the i-th sampled path writes finally as $(1 \le i \le M)$

$$\begin{cases} X^{(i)}_{t_{k+1}} = X^{(i)}_{t_k} - a(X^{(i)}_{t_k} - x^{\text{forecast}}_{t_k})\Delta_t + \sigma\,(X^{(i)}_{t_k})^{\alpha}(1-X^{(i)}_{t_k})^{\beta}(W^{(i)}_{t_{k+1}} - W^{(i)}_{t_k}), & k \ge 0, \\ X^{(i)}_{t=t_0} = X^{(i)}_{t_0}, \end{cases} \tag{15}$$

where $(W^{(i)})$ are M i.i.d. trajectories of the Brownian motion, and $(X^{(i)})$ are M i.i.d. trajectories of the CSI.

Initialization. A particular treatment of the independent initial value $X^{(i)}_{t_0}$ for each trajectory of the CSI needs to be implemented. Indeed, one could initialize

all the processes to the same initial value of the deterministic forecast. However, this would not allow us to encompass uncertainty in the initial value of the CSI. We propose another approach, close to Markov Chain Monte-Carlo methods. The initial value $X_{t_0}^{(i)}$ is sampled according to the stationary distribution of the Euler scheme associated to the SDE

$$d\widetilde{X}_t = -a(\widetilde{X}_t - x_{t_0}^{\texttt{forecast}})\mathrm{d}t + \sigma \widetilde{X}_t^{\alpha}(1-\widetilde{X}_t)^{\beta}\mathrm{d}W_t. \tag{16}$$

Note the term $x_{t_0}^{\texttt{forecast}}$ in the above equation. Doing so, we pick a initial point $X_{t_0}^{(i)}$ which reflects in a quite intrinsic way the uncertainty in the forecast at $t = t_0$.

In practice, we simulate the Euler scheme of (16) over a time interval of length T, where T is chosen long enough so that the distribution of the Euler scheme after time T is closed to the target stationary distribution (and quite independent from the initial point of the SDE (16)); see [Tal90] about approximation scheme for ergodic SDEs.

In practice, due to the form of the drift as a mean-reversion term and in view of the fast decorrelation (like in (9)), we set T as three times the characteristic time of the system, that is $T = 3/a \approx 4$ hours. This procedure is repeated independently to sample each $X_{t_0}^{(i)}$.

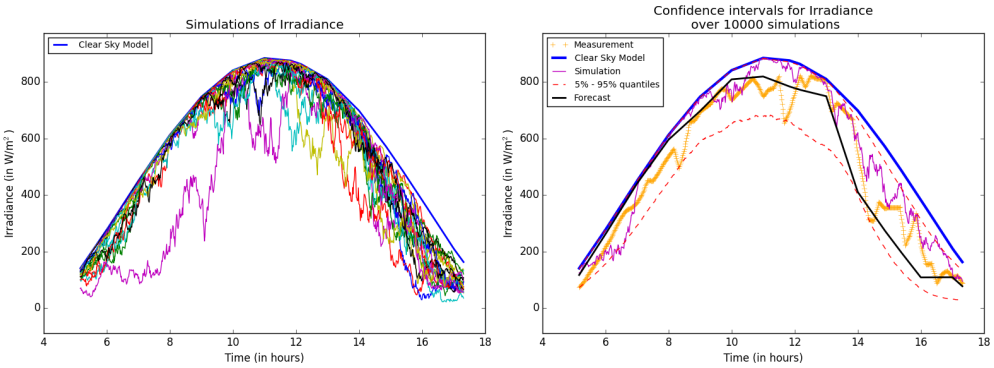
3 Numerical experiments

3.1 Description of the tests

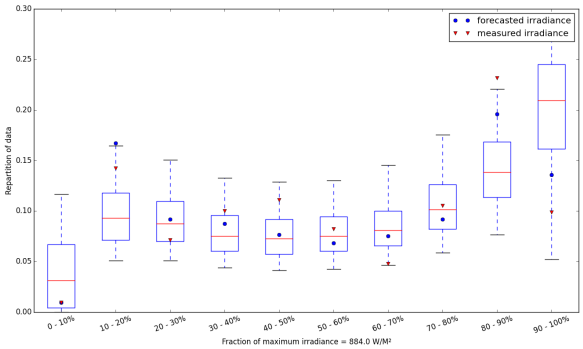
In the following graphs, we represent some features of the results we obtain for several types of days: a day with good weather, when the Clear Sky Index remains close to 1, a day with bad weather, when it remains close to 0 and a day with mitigated weather, when it takes intermediate values. For all those examples, we take days where the deterministic forecast used to establish the probabilistic forecast is fairly good, i.e. close to the real irradiance profile over the day. This is usually the case with Arome forecast.

For each day with fairly good deterministic forecasts (Figures 7-9-11), we plot 3 graphs. The first in the top left corner features the clear sky irradiance as well as several simulated trajectories of the irradiance (at the time-scale of 1' as mentioned before). The clear sky irradiance is the irradiance we would observe if the sky were perfectly clear (this is the clear sky model as explained at the beginning of Section 2.1). The simulated trajectories of the irradiance are obtained by multiplying the simulated trajectories of the Clear Sky Index, obtained by Euler scheme of a SDE (see previous paragraph), by the clear sky irradiance (see the relation (2)).

The second graph in the top right corner features the clear sky irradiance, the deterministic Arome forecast irradiance, the irradiance measured on that day, a simulated trajectory of the irradiance obtained using our model, as well as confidence



Results for the probabilistic forecast

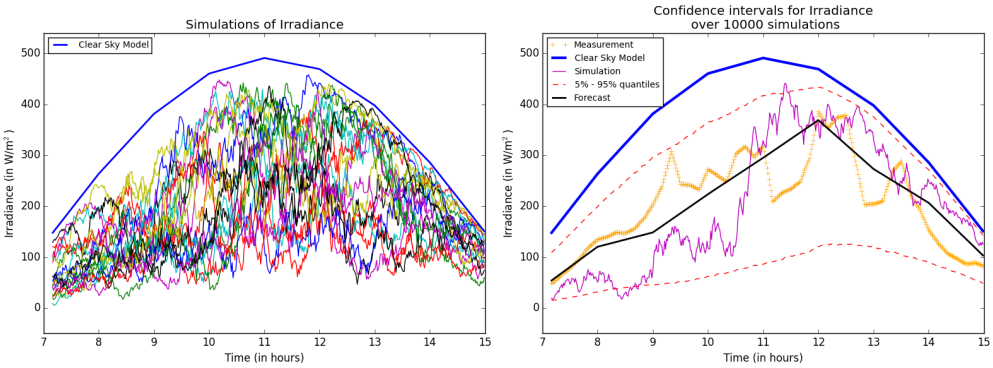


Validation and accuracy of the model for the distribution of the irradiance over the day

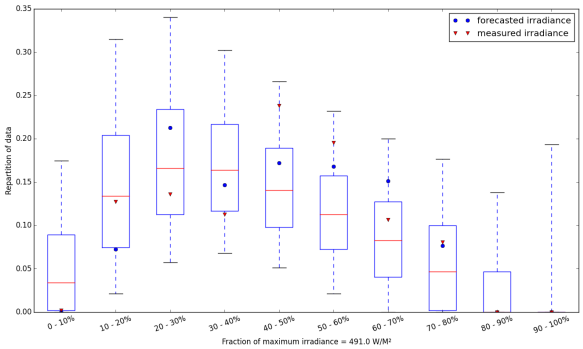
Fig. 7: Results for a day with good weather (May 2^{nd}, 2016)

intervals for the irradiance forecast. This confidence region is obtained by Monte-Carlo methods: by simulating $M = 100000$ i.i.d. trajectories of the irradiance (using our SDE model), we can estimate accurately the 5% and 95% quantiles. If our model is accurate, we expect that the measured irradiance remains inside the confidence intervals most of the time. However, having a measurement outside this confidence area does not necessarily mean poor performance of the probabilistic forecast, because of the definition of quantiles. Of course, the deterministic forecast staying inside this confidence area is an intrinsic property of the model we have designed: we take an SDE with a drift which is a mean-reversion term, the mean being this deterministic forecast.

In the bottom graph, we represent several vertical box-plots, as well as other information. Let us explain in more details. For each day studied, we can derive the number of points of a discrete trajectory (observation, forecast or simulation) for which the irradiance value lies in each of the 10 intervals of length 10% of the



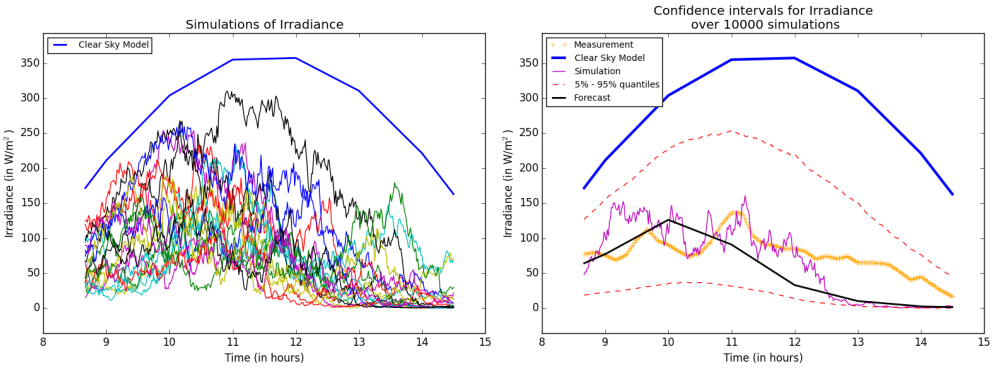
Results for the probabilistic forecast



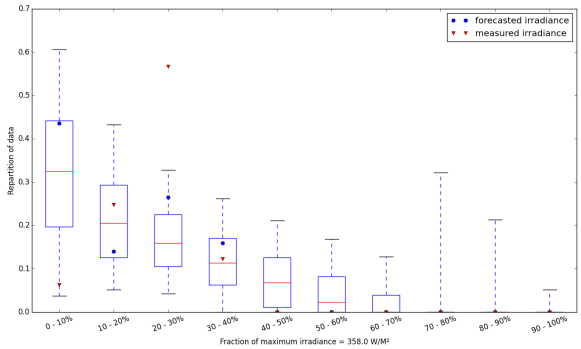
Validation and accuracy of the model for the distribution of the irradiance over the day

Fig. 9: Results for a day with mitigated weather (October 24^{th}, 2015)

maximal value of the clear sky irradiance (given by the clear sky model). Using this procedure for the measured irradiance, we obtain by renormalization an estimation of the proportion of time spent by the irradiance in a given interval over the day. A possible application is to derive an estimation of the PV energy than can be produced on that specific day. For example, a red triangular point with $0-10\%$ on the x-axis and 0.4 on the y-axis means that the measured irradiance was smaller than 10% of the maximal theoretical irradiance about 40% of the time of sun exposure over this day. We can do the same thing for the forecast irradiance and for each of the simulations. The simulations allow to estimate the distribution of these proportions of time spent in each subinterval. Indeed, using the values obtained for each simulation, we can draw box-plots which show how the irradiance was distributed over the whole day. These statistical outputs can for instance give a clear indication about the probability distribution of the energy that can be produced using PV panels. The red line represents the median, while the box extends from the first quartile



Results for the probabilistic forecast



Validation and accuracy of the model for the distribution of the irradiance over the day

Fig. 11: Results for a day with bad weather (January 22nd, 2016)

to the third quartile, and the whiskers extend from the 5% to the 95% quantiles. For example, for the interval $0-10\%$, which regroups all data with a value corresponding to less than 10% of the theoretical maximal irradiance observable on this day, a red line with y-coordinate equal to 0.33 means that for half of the simulations, less than one third of the points of the corresponding trajectory have a value in this interval, while for the other half of the simulations, more than one third of the points of the trajectory have a value in this interval.

A good probabilistic forecast would therefore have several characteristics:

- The trajectory of the measured irradiance would lie in the confidence area obtained by Monte-Carlo simulations.
- In the box-plots, the points for irradiance measured would lie in high density regions (between the whiskers and often inside the boxes, see figure 12).

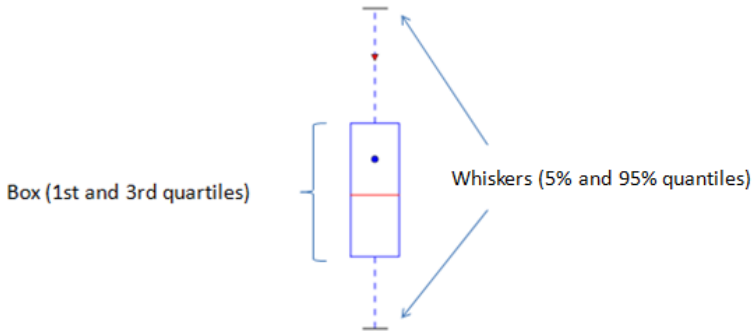


Fig. 12: Elements of a box-plot: explanation

These last points are not a guarantee that the probabilistic forecast was good. Indeed, one could imagine the following situation: if on one day, one forecasts a clear sky in the morning and a cloudy sky in the afternoon, but in practice, these events occur in the other order, the box-plots might be consistent (i.e. the measurements should lie inside the confidence area represented by the box and the whiskers), whereas the forecast and measured irradiance paths would be radically different. Even the forecast and measurements having close trajectories does not necessarily mean that all statistical properties of the irradiance are captured by our model.

3.2 Analysis of the results

From these 3 figures (that are quite representative of what we have observed throughout our tests on the full history), we have general observations. First, as expected, the forecast lies in the confidence area: this is mathematically consistent with the model and in accordance with the intuition. Second, generally speaking, the points drawn using the measured irradiance data lie in the confidence areas obtained with our Monte-Carlo procedure; this can be observed in two ways.

- From the second graph, one checks that the measured irradiance (at any given time) remains most of the time in the confidence area. It helps to answer positively the question whether the profile of the irradiance observed over the day is in accordance with the forecast distribution that is induced by our SDE model.
- From the third graph, we analyse whether the repartition of irradiance over the day is correctly predicted by our model, regardless of when phenomenon occurred. The answer is yes. This criterion is of course less severe than the previous one, but it has the merit to accept events in the days which were predicted but occurred at another time thant what the forecast predicted. For example, if the

forecast predicts a sunny morning and a cloudy afternoon, the time frame when the transition happens may not be accurately predicted by Arome models.

These results constitute empirical evidence of the good performance of our model: in a way, we are able to correctly reproduce the distribution of the irradiance over the day (either by time interval – second graph – or by repartition of irradiance – third graph). Further measures of performance will be investigated in future works.

3.3 Limits

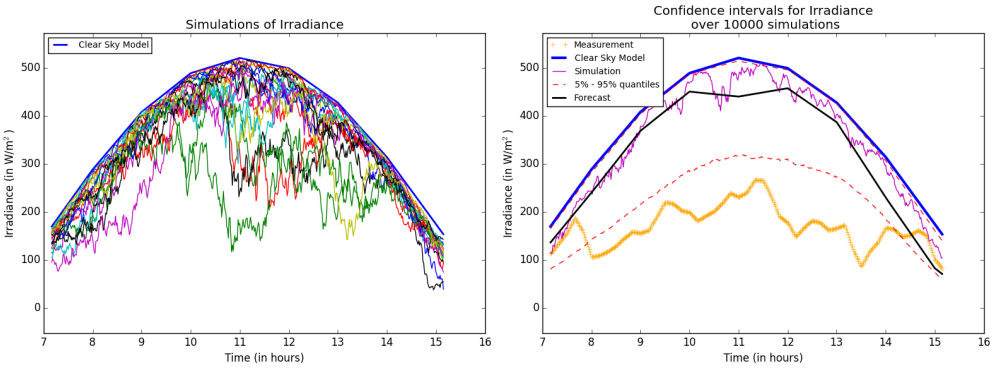
We show in the next figure 14 that if the deterministic forecast is completely erroneous, i.e. if the irradiance profile observed is completely different from the predicted one, then the probabilistic forecast performs poorly as well. It is not surprising at all, since the probabilistic forecast is built upon the deterministic one.

In the graph at the top right corner, we see that due to a poor forecast, the measured irradiance fails to be inside the confidence area over the whole day. In the box-plots at the bottom, it is also clear that the distribution of the proportions of time spent in each subinterval of irradiance is not estimated correctly.

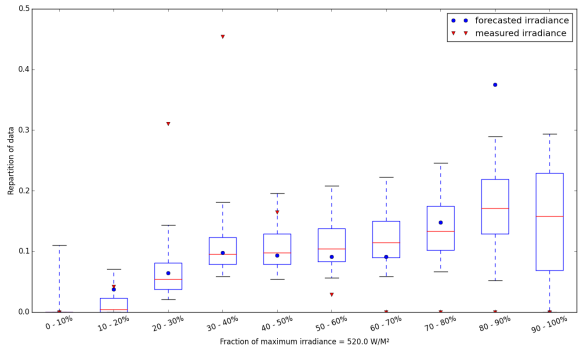
4 Conclusion

We have designed a stochastic differential equation that models the solar irradiance for a given day D. This gives rise to a probabilistic forecast. The parameters of the model change from day to day and can be tuned automatically: they depend only the irradiance from clear sky model on the day D and the deterministic Arome forecast computed on day D-1 for the day D. By simulating the SDE in a Monte-Carlo framework, we obtain the distribution of the forecast and related statistics. Despite its apparent simplicity, the model is able to produce quite accurate confidence intervals for the irradiance at a given time, and for the repartition of irradiance during the day.

Acknowledgements This research is part of the Chair *Financial Risks* of the *Risk Foundation*, the *Finance for Energy Market Research Centre* and the ANR project *CAESARS* (ANR-15-CE05-0024). The work benefits from the support of the Siebel Energy Institute and it was conducted in the frame of the TREND-X research program of Ecole Polytechnique, supported by Fondation de l'Ecole Polytechnique. The authors acknowledge Météo-France and the Cosy project for the numerical weather prediction data used in the study.



Results for the probabilistic forecast



Validation and accuracy of the model for the distribution of the irradiance over the day

Fig. 14: Results for a day with bad forecast (October, 19^{th}, 2015)

Appendix: Proof of well-posedness of the SDE model (8), when $\frac{1}{2} \leq \alpha \leq 1$ and $\frac{1}{2} \leq \beta \leq 1$.

Because of the exponents α and β possibly smaller than 1 in the definition of (8), the signs of X_t and $1 - X_t$ may be an issue. Therefore, we start with a modification of the SDE model (8) avoiding the sign problems:

$$\mathrm{d}X_t = -a(X_t - x_t^{\text{forecast}})\mathrm{d}t + \sigma |X_t|^{\alpha} |1 - X_t|^{\beta} \mathrm{d}W_t, \tag{17}$$

where $X_0 \in [0,1]$ is a given deterministic initial value.

Existence/uniqueness. A direct application of [RY99, Chapter IX, Theorem 3.5-ii) pp.390 and Theorem 1.7 pp.368] shows that the model (17) is well-posed, in the sense that there is an unique strong solution on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ where the filtration is the natural filtration of the Brownian motion completed as

usually with the $\mathbb{P}$-null sets. In [RY99, Chapter IX, Theorem 3.5-ii) pp.390] we have used $\frac{1}{2} \leq \alpha \leq 1$ and $\frac{1}{2} \leq \beta \leq 1$.

The solution (17) *takes values in* $[0,1]$. We invoke a comparison theorem for SDEs. Denote $b^X(t,x) = -a(x - x_t^{\texttt{forecast}})$ the drift coefficient of X and now, consider the solution to

$$\mathrm{d}Y_t = -aY_t\mathrm{d}t + \sigma|Y_t|^\alpha|1-Y_t|^\beta \mathrm{d}W_t, \qquad Y_0 = 0. \tag{18}$$

Its initial condition fulfills $X_0 \geq Y_0$, its drift $b^Y(t,y) = -ay$ is globally Lipschitz in space (and b^X too) and last, we have $b^Y(t,x) - b^X(t,x) = -ax_t^{\texttt{forecast}} \leq 0$. Therefore, [KS91, Chapter V, Proposition 2.18 pp.293] shows that $X_t \geq Y_t$ for any t with probability 1. But since THE solution to (18) is 0, the above proves that X remains positive.

Similarly, set

$$\mathrm{d}Y_t = -a(Y_t - 1)\mathrm{d}t + \sigma|Y_t|^\alpha|1-Y_t|^\beta \mathrm{d}W_t, \qquad Y_0 = 1. \tag{19}$$

Clearly, $Y_0 \geq X_0$, $b^X(t,x) - b^Y(t,x) = -a(1 - x_t^{\texttt{forecast}}) \leq 0$, $Y_t = 1$ and we conclude that $X_t \leq 1$. This justifies why we can remove the absolute values in (17) to get (8) . □

References

[ACLP14] R. Aïd, L. Campi, N. Langrené, and H. Pham. A probabilistic numerical method for optimal multiple switching problem in high dimension. *SIAM J. on Financial Mathematics*, 5(1):191–231, 2014.

[ADSC15] S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone. An analog ensemble for short-term probabilistic solar power forecast. *Applied energy*, 157:95–110, 2015.

[Aid15] R. Aid. *Electricity derivatives*. Springer Briefs in Quantitative Finance, 2015.

[BH81] R.E. Bird and R.L. Hulstrom. Simplified clear sky model for direct and diffuse insolation on horizontal surfaces. Technical report, Solar Energy Research Inst., Golden, CO (USA), 1981.

[BSS05] B.M. Bibby, I.M. Skovgaard, and M. Sorensen. Diffusion-type models with given marginal distribution and autocorrelation function. *Bernoulli*, 11(2):191–220, 2005.

[DFM15] M. Delfanti, D. Falabretti, and M. Merlo. Energy storage for PV power plant dispatching. *Renewable Energy*, 80:61–72, 2015.

[DLD12] H.M. Diagne, P. Lauret, and M. David. Solar irradiation forecasting: state-of-the-art and proposition for future developments for small-scale insular grids. In *WREF 2012-World Renewable Energy Forum*, 2012.

[GG17] E. Gobet and M. Grangereau. Optimal management under uncertainty of microgrid equipped with PV panels and battery: resolution using McKean-FBSDE. *preprint*, 2017.

[Gob02] E. Gobet. LAN property for ergodic diffusion with discrete observations. *Ann. Inst. H. Poincaré Probab. Statist.*, 38(5):711–737, 2002.

[HKNF14] R. Hanna, J. Kleissl, A. Nottrott, and M. Ferry. Energy dispatch schedule optimization for demand charge reduction using a photovoltaic-battery storage system with solar forecasting. *Solar Energy*, 103:269–287, 2014.

[HLG06] D. Heinemann, E. Lorenz, and M. Girodo. Solar irradiance forecasting for the management of solar energy systems. *Energy and Semiconductor Research Laboratory, Energy Meteorology Group, Oldenburg University*, 2006.

[IMMM14] E.B. Iversen, J.M. Morales, J.K. Møller, and H. Madsen. Probabilistic forecasts of solar irradiance using stochastic differential equations. *Environmetrics*, 25(3):152–164, 2014.

[KLC+11] H. Kanchev, D. Lu, F. Colas, V. Lazarov, and B. François. Energy management and operational planning of a microgrid with a PV-based active generator for smart grid applications. *IEEE Transactions on Sustainable Energy*, 58(10), 2011.

[Kle13] J. Kleissl. *Solar energy forecasting and ressource assessment*. Academic Press, 2013.

[KLS12] M. Kessler, A. Lindner, and M. Sorensen. *Statistical methods for stochastic differential equations*. CRC Press, 2012.

[KP10] P.E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations. 4th corrected printing*. Applications of Mathematics 23. Berlin: Springer, 2010.

[KS91] I. Karatzas and S.E. Shreve. *Brownian motion and stochastic calculus*. Springer Verlag, second edition, 1991.

[MA16] A.A. Mohammed and Z. Aung. Ensemble learning approach for probabilistic forecasting of solar power generation. *Energies*, 9(12):1017, 2016.

[MBF+16] C.B. Martinez-Anido, B. Botor, A.R. Florita, C. Draxl, S. Lu, H.F. Hamann, and B.M. Hodge. The value of day-ahead solar power forecasting improvement. *Solar Energy*, 129:192–203, 2016.

[OAB+15] K.D. Orwig, M.L. Ahlstrom, V. Banunarayanan, J. Sharp, J.M. Wilczak, J. Freedman, S.E. Haupt, J. Cline, O. Bartholomy, H.F. Hamann, B.M. Hodge, C. Finley, D. Nakafuji, J.L. Peterson, D. Maggio, and M. Marquis. Recent trends in variable generation forecasting and its value to the power system. *IEEE Transactions on Sustainable Energy*, 6(3):924–933, 2015.

[PLP+13] R. Perez, E. Lorenz, S. Pelland, M. Beauharnois, G. Van Knowe, K. Hemker Jr, D. Heinemann, J. Remund, S.C. Müller, W. Traunmüller, G. Steinmauer, D. Prozo, J.A. Ruiz-Arias, V. Lara-Fanego, L. Ramirez-Santigosa, M. Gastin-Romero, and L.M. Pomares. Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. *Solar Energy*, 94:305–326, 2013.

[RY99] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*. Comprehensive Studies in Mathematics. Berlin: Springer, third edition, 1999.

[SE10] T. Soubdhan and R. Emilion. Stochastic differential equation for modeling global solar radiation sequences. In *Proceedings of the IASTED International Conference: Modelling, Identification, and Control (AsiaMIC 2010)*, pages 14–17, 2010.

[SLSN15] C.V.A. Silva, L. Lim, D. Stevens, and D Nakafuji. Probabilistic models for one-day ahead solar irradiance forecasting in renewable energy applications. In *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on*, pages 1163–1168. IEEE, 2015.

[Tal90] D. Talay. Second-order discretization schemes of stochastic differential systems for the computation of the invariant law. *Stochastics and Stochastics Reports*, 29:13–36, 1990.

[TZH+15] A. Tuohy, J. Zack, S.E. Haupt, J. Sharp, M. Ahlstrom, S. Dise, E. Grimit, C. Mohrlen, M. Lange, M.G. Casado, J. Black, M. Marquis, and C. Collier. Solar forecasting: methods, challenges, and performance. *IEEE Power and Energy Magazine*, 13(6):50–59, 2015.