A Step-by-step Methodology to Solar Power Forecasting using ARMA Models

Bismark Singh, and David Pozo

***Abstract*—We describe a simple and succinct methodology to develop hourly auto-regressive moving average (ARMA) models to forecast power output from a photovoltaic generator. We illustrate how to use statistical tests to validate the model and construct hourly samples. These samples can be used as scenarios for energy planning or in stochastic optimization models.**

***Index Terms*—ARMA, solar power, photovoltaic, forecasting, scenario generation**

I. Introduction

INCREASING penetration of renewable energy sources, such as wind and solar, in the electricity grid requires good day-ahead power forecasts. Solar power differs from wind power due to its diurnal nature, and can have much greater ramps than wind [**?**]. In this letter, we focus on forecasting hourly solar power generation, in particular from photovoltaic (PV) technology.

Forecasting methods for solar power are broadly divided into two categories: (i) physics-based models—these models predict solar power from numerical weather predictions and solar irradiation data, and (ii) statistical models—these models forecast solar power directly from historical data. Comparisons of these two methods are also available; see, e.g., [**?**], [**?**]. There are other approaches available as well which combine these two methods [**?**]. In this letter we center on statistical methods alone, and specifically the use of auto-regressive moving average (ARMA) models to develop our forecasts. ARMA models are widely used for forecasting many economic and planning processes; see, e.g., [**?**]. They have also been used to forecast wind power [**?**], [**?**], as well as solar power [**?**], [**?**]. Yet, accurate and fast methods to generate solar power scenarios are often unavailable or significantly complex, and normal approximations are frequently used; see, e.g., [**?**]. Here, we describe a summary of the methodology to forecast solar power using ARMA models. The generated scenarios are available on request. The presented models can be applied either to a local PV generating plant or at the regional level.

the

The main contribution of this letter is to provide a step-by-step approach and easy-to-implement ARMA model to forecast PV solar power generation. The proposed model is able to capture the important statistical features of the parameters, while maintaining simplicity. The model allows modelers to embed it into more complex decision-making structures, statisticians to have an all-in-one place ARMA model design for PV power generation, and policy makers and electrical engineers to have a scenario generation tool.

B. Singh is with the Discrete Mathematics & Optimization, Sandia National Laboratories, Albuquerque, NM 87185, USA e-mail: bsingh@sandia.gov.

D. Pozo is with the Center for Energy Systems, Skolkovo Institute of Science and Technology, Moscow, Russia.

II. Methodology

We take hourly year-long historical solar power output from a site described in [**?**]. This zone is at an altitude of 595m, has a nominal power of 1560 MW, a panel tilt of 36°, and a 38° clockwise panel orientation from the north. Further installation specifics are available in zone 1 from Table 1 of [**?**], while technical specifications are available in [**?**]. We use approximately nine months of data for training. The data does not have any solar power for the ten hours [20:00-5:00], and hence we restrict the forecasts in these hours to be zero as well. Equivalently, a criteria based on the solar zenith angle can be used; i.e., 0° at sunrise and 90° when the sun is directly overhead. For each of the remaining 14 hours of the day, we build an ARMA(p, q) model. Further, for each hour, we verify the stationarity of the time series and test a number of ARMA(p, q) models to find the best one. We use statistical tests on the residuals to validate the models. Finally, we use Monte Carlo sampling from the best ARMA model, for each hour, to create hourly scenarios. Below we provide more details.

A. Stationarity

An ARMA model may be suitable if a time-series is stationary. We test the hourly data for stationarity using the Augmented Dickey-Fuller (ADF) test [**?**]. The ADF test has a null hypothesis that the series includes a unit root (or, is non-stationary). We reject the null hypothesis at a level 0.05 if the test-statistic exceeds its 0.95 level quantile. For all the 14 hours of the day, the null hypothesis is rejected suggesting the series may be stationary, and hence an ARMA model may be suitable. If the series were not stationary, an ARIMA model may be suitable; see, e.g., [**?**].

B. Selecting parameters of the ARMA model

Next, we estimate the parameters of the ARMA model, p, the order of the autoregressive part and, q, the order of the moving average part. For each hour, we construct 16 models with both p and q between one and four, and compute the log-likelihood objective function value. Next, for each hour, we calculate the Bayesian information criteria (BIC) for the 16 models using $p + q + 1$ parameters. The BIC penalizes for models with more parameters, and the smallest value of the

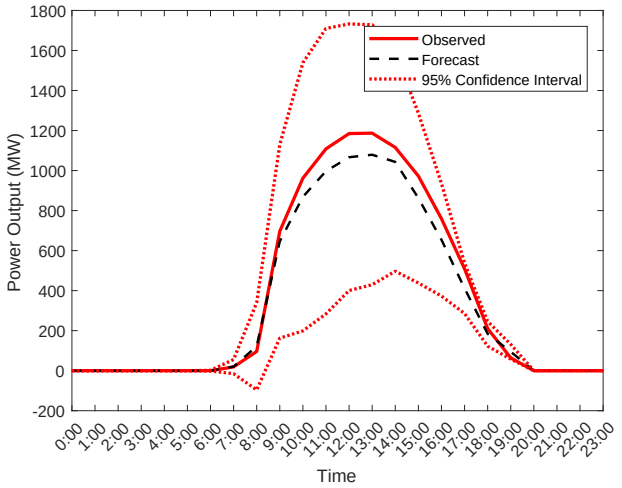

Fig. 1. Day-ahead actual and predicted values using ARMA models from Table **??**

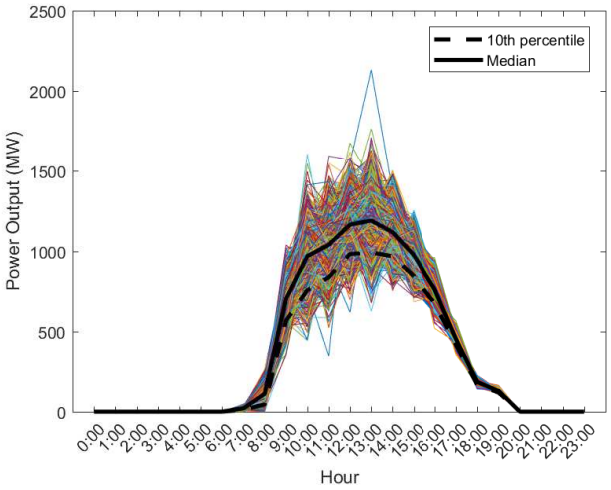

Fig. 2. 2000 hourly scenarios for solar power generated using the ARMA models from Table **??**. The dashed black line is the median hourly value, and the solid black line is the 10 percentile solar power value.

BIC gives the best model, for each hour. Table **??** provides our estimated p and q values for the 14 hours of the day. We note that none of the hours have an order value exceeding two.

TABLE I
ESTIMATED p AND q VALUES FOR ARMA(p, q) MODELS FOR 14 HOURS OF THE DAY

Hour	6:00	7:00	8:00	9:00	10:00	11:00	12:00
p	1	1	1	1	2	1	1
q	1	1	1	1	1	1	2
	13:00	14:00	15:00	16:00	17:00	18:00	19:00
p	1	1	1	1	1	1	2
q	1	1	1	1	1	2	1

C. Prediction

Figure **??** plots a day-ahead prediction using the above constructed ARMA models; i.e., one hour ahead predictions from the 14 ARMA models. A number of metrics are available to evaluate the prediction; see, e.g., [**?**]. We use a few of them here. The mean absolute error between the actual and the predicted series is 39.6 MW, or 3.3% of the maximum actual value. The root mean square error between the actual and the predicted series is 61.0 MW, or 5.1% of the maximum actual value.

We further verify autocorrelation in the series, for each hour, using the Ljung-Box test [**?**] on the residuals for lags of 5, 10, and 15. The Ljung-Box test has a null hypothesis that the residuals are uncorrelated up to a given lag. We reject the null hypothesis at a level 0.05 if its test-statistic exceeds its 0.95 level quantile. For all the 14 hours of the day, the null hypothesis is not rejected suggesting a zero autocorrelation in the series, or the model choice may be appropriate.

With increasing penetration of solar power in the electricity grid, a number of stochastic optimization models for bidding, storage, and generation have been developed; see, e.g., [**?**], [**?**]. Stochastic optimization models rely on the availability of a large number of scenarios. We can use Monte Carlo sampling to generate hourly solar power scenarios. The output from an ARMA model is real valued, and hence can be negative. In our analysis, we truncate the negative powered outputs to 0. For the 14 hours of the day, this sampling resulted in 1.6% of the outputs with estimated power output below -5MW. Figure **??** plots 2000 day-ahead scenarios as well as the median and 10 percentile values.

III. CONCLUSIONS

We present a simple step-by-step scheme for fitting an ARMA model to historical solar power data, and use it to forecast future hourly scenarios. We present statistical tests to check the applicability of various models, identify model parameters, and to finally forecast scenarios. If significant statistical evidence is not present in support of a test, the chosen model is not expected to perform well. This can lead to erroneous conclusions; see, e.g., [**?**], [**?**]. The methodology in this letter can be directly applied to historical data, both for a single PV source and a PV site, to create future scenarios for use in a stochastic model for power system operation and planning.

ACKNOWLEDGMENT

B. Singh thanks Jean-Paul Watson and Andrea Staid at Sandia National Laboratories for helpful discussions and for sharing data. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. The work of D. Pozo was supported by Skoltech NGP Program (Skoltech-MIT joint project).

keepaspectratio]mshell]Michael Shell

Bismark Singh received the B.Tech. degree in chemical engineering from the Indian Institute of Technology (IIT), Delhi, India in 2011, the M.S.E. and Ph.D. degrees in operations research and industrial engineering from The University of Texas at Austin, Austin, TX, USA in 2013 and 2016, respectively. He is currently a postdoctoral appointee at Sandia National Laboratories in Albuquerque, New Mexico, USA.

David Pozo