Energy production from solar photovoltaic (PV) plants is unpredictable, mainly due to the stochastic formation and movement of clouds or aerosol - dust particles which scatter or disperse solar radiation. Accurate forecasts of PV output are essential to Distribution and Transportation System Operators as they assist efficient solar energy trading and management of electricity grids. This work evaluates an autoregressive, computationally-light KNN-regression scheme (TSFKNN) for hourly, day-ahead forecasts of solar irradiance and energy yield of various PV technologies. The model is being tested and validated using data measured in Thuwal, Saudi Arabia. The available measured records span a 60-month period. The developed forecasting models are designed for online systems and provide increased levels of accuracy and low computational cost. Several parametric and nonparametric specifications, coupled with conventional versus outlier-robust estimation procedures are tested, in order to derive an optimal month-specific daily profile (MDP). Current results demonstrate that including intraday variability to the monthly-based irradiance models achieve improved predictive accuracy between 10% and 25% on average.

I. INTRODUCTION

Solar plant power production is uncertain, mainly due to the stochastic formation and movement of clouds. Accurate forecasts of photovoltaic (PV) output are essential to Distribution System Operators and Transportation System Operators [1] as they assist efficient solar energy trading and management of electricity grids [2]. The main aim of this work is to develop specifications for day-ahead forecasting of PV energy output, which are solely based on historical information and combine satisfactory levels of forecasting accuracy with low computational and monetary costs. To this purpose, we evaluate a recently proposed, computationally-light time series KNN-regression scheme (KNNTS) [3], using observed energy yields (EY) from 2 PV module technologies: Aluminum Back Surface Field (Al-BSF) and Back Contact (BC).

KNNTS is straightforward to implement online as it is based solely on historical data and does not require next-day meteorological predictions as input. Advanced predictive models (e.g. penalized regressions, random forests, neural networks; [4]) that use meteorological information should dominate such relatively simple schemes by a significant margin, to justify their substantially increased cost (both computational and financially to acquire continuously updated meteorological forecasts). Online forecasting systems for PV and wind-farm energy outputs are expected to employ such benchmark specifications when for instance, access to day-ahead meteorological forecasts has been interrupted or their availability is limited.

In the following sections, KNNTS is evaluated against three frequently adopted benchmark specifications: seasonal ARIMA, spline-based daily profiles (SDP) and the persistence model (PRS). Evaluation is based primarily on conventional RMSE; in addition, the relative accuracy metric proposed by [5] is monitored.

II. DATA ANALYSIS - RESULTS

Power output measurements (W/m²) are collected at KAUST, Thuwal, Saudi Arabia. The analyzed data were recorded every 10 minutes, from 8AM to 5PM for 364 consecutive days (52 weeks) in 2016. Data cleaning (removal of extremely high, clearly erroneous measurements) has been applied to eliminate statistical artefacts; however, Fig. 1A suggests that some outliers still remain. In addition to outlying measurements the observed power output series contain gaps, which hamper application of the forecasting techniques that follow, especially KNNTS and ARIMA. Specifically, missing values constitute 5.9% (6.9%) of the Al-BSF (BC) measurements. Typically sequences of missing values are short as in the vast majority of examined days the percentage of missing values is clearly below 10%. A simple linear interpolation scheme would have been sufficient in the presence of occasional short gaps. However, given a) the relatively large proportions of missing values occasionally observed and b) available historical data included measurements of solar irradiance (W/m²), an irradiance-based imputation scheme, which reveals intriguing characteristics of the two panel types, is adopted herein.

A naïve imputation scheme would assume fixed performance rates for the two PV-technologies during the whole study period:

$$\hat{Y}_t(t) = c_1 \tilde{X}(t)$$  \hspace{1cm} (1)

with X(t) denoting observed irradiance at time t, c1, c2 designating Al-BSF and BC efficiencies and $$\hat{Y}_t(t)$$, $$\tilde{X}(t)$$, imputed Al-BSF and BC power outputs, respectively. The overly parsimonious approximation in (1) should achieve...
subpar levels of accuracy, given the expected decreasing trend for PV efficiencies. A more flexible scheme, which allows for monthly-varying efficiencies is formulated as:

$$\hat{Y}_i(t) = c_i X_i(t)$$

with $m=1,\ldots,12$ denoting a monthly index. Fig. 3, presents outlier-robust, least absolute deviation (LAD, a.k.a. median regression) estimates of monthly efficiencies, which indeed decrease with time. Interestingly, although Al-BSF appears to perform slightly better than BC in the first six months, BC clearly outperforms Al-BSF during the last six months of the study period.

The adopted imputation scheme extends the model shown in (3) by allowing for month-specific, daily efficiency-rate profiles:

$$\hat{Y}_i(t) = c_{i}^{m,t_d} X_i(t)$$

with $t_d=1,\ldots,55$ designating 10-minute intervals ($t_d=1$ to 55 corresponds to the interval 8AM to 5PM). Essentially the above specification takes (indirectly) into account dependence of PV efficiencies on operating temperatures [6].

Robust imputation is prioritized here, that is why alternative imputation procedures are evaluated via mean absolute error (MAE). Fig. 2 displays results of a leave-2-week-out cross-validation experiment, which evaluates imputation accuracy achieved from expressions (1), (2) and (3). In this experiment, measurements are divided in 16 consecutive 2-week periods. Each period is used as a testing subset, with the remaining periods utilized as training data. Fig. 2 clearly demonstrates the superiority of the daily profiles in (3); the average MAE achieved from (3) across all periods is close to 5 W/m², whereas for the parsimonious scheme in (2) is close to 10 W/m² for both technologies. Median regression estimates performed better than conventional least squares in terms of average MAE: that is why the adopted imputation scheme utilized in (3) with coefficients estimated via LAD. Fig. 2 presents observed and imputed data for the performance profiles that constitute the last week of the study period.

**Figure 2.** MAE of alternative imputation schemes for Al-BSF (a) and BC (b). LM denotes conventional least-squares whereas LAD corresponds to outlier-robust median regression estimates. MDS1 represents the static model in eq. (1), MDS2 stands for the specification shown in eq. (2) and MDS3 is the adopted imputation scheme shown in (3).

**Figure 3.** Observed (imputed) measurements for the last week of the study period are shown as colored (empty) triangles: Al-BSF (top), BC (bottom).

Fig. 3 presents observed and imputed data for the performance profiles that constitute the last week of the study period.

The predicted profiles match very well the observed data indicating the improved performance of the adopted imputation scheme of expression (3).
III. FORECASTING ANALYSIS

This section presents a forecasting experiment based on actual measured real data. Specifically, the experiment comprises $D = 76$ testing days: sometime during each testing day $d = 1,...,D$ (in reality around noon), a forecast is computed for the PV outputs of the next day. Days with significant percentages of missing data are not included in the testing set to avoid issues in applying the statistical method. Training data consist of days before $d$ and forecasts correspond to the day after $d$. Computational times were less than 5 seconds for both SDP and KNNTS, even when the number of training weeks, $N_w=8$. ARIMA model building is substantially slower, with computational times close to 5 minutes on average, when $N_w=8$. Results of the experiment are depicted in Figure 4 and Table 1.

Interestingly, focusing on RMSE, SDP outperforms both KNNTS and ARIMA; in accordance with prior expectations, PRS is the worst performing method by a significant margin. Regarding KNNTS the optimal combination function used to aggregate the targets associated with the nearest neighbors is the median, by a small margin relative to the mean. ARIMA (KNNTS) performance is optimal when the training data comprise 3 (8) weeks for both PV technologies. On the other hand, SDP achieves minimum RMSE with 5 (3) training weeks for BC (Al-BSF). It should be stressed however that RMSE performances by SDP are very close when the number of training weeks, $N_w=1$. $N_w=3$ can be viewed as a choice that results in satisfactory performance for all examined models, while resulting in computationally fast implementations.

<table>
<thead>
<tr>
<th>PV Technology</th>
<th>Model</th>
<th>RMSE (W/m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al-BSF</td>
<td>PRS</td>
<td>26.619</td>
</tr>
<tr>
<td>Al-BSF</td>
<td>SDP</td>
<td><strong>22.809 (5)</strong></td>
</tr>
<tr>
<td>Al-BSF</td>
<td>ARIMA</td>
<td>24.582 (3)</td>
</tr>
<tr>
<td>Al-BSF</td>
<td>KNNTS</td>
<td>23.120 (8)</td>
</tr>
<tr>
<td>BC</td>
<td>PRS</td>
<td>27.125</td>
</tr>
<tr>
<td>BC</td>
<td>SDP</td>
<td><strong>23.314 (3)</strong></td>
</tr>
<tr>
<td>BC</td>
<td>ARIMA</td>
<td>25.072 (3)</td>
</tr>
<tr>
<td>BC</td>
<td>KNNTS</td>
<td>23.624 (8)</td>
</tr>
</tbody>
</table>

Table 1. RMSE performance of the persistence model, versus ARIMA, SDP, and KNNTS. Reported accuracies depend on the number of training weeks $N_w$, which is shown in parentheses.

The results of figure 4 and table 1 demonstrate that the RMSE performance of each model is similar, with SDP showing the lowest values, closely followed by KNNTS & ARIMA. This indicates that KNNTS is performing well under this experimental scheme provided the training period is sufficiently long, however there is no significant RMSE difference over the other methods.

IV. CONCLUDING REMARKS

This work evaluated a recently proposed time series KNN procedure against spline-based daily profiles, seasonal ARIMA and the persistence model, for day-ahead forecasting of PV outputs. Two technologies of solar panels are examined: Al-BSF and BC. Contrary to what one may expect, an extensive forecasting experiment revealed that KNN-based ensembles are not superior relative to the examined alternatives when performance is evaluated in terms of the widely adopted RMSE criterion. Despite this fact on the example application considered, KNNTS forecasts are expected to perform well when environmental conditions are highly variable, with regime-specific variability. The alternative specifications examined here can be combined to result in a forecast combination scheme that is expected to perform as well as the best performing method. Such ensembles can be easily incorporated in a light (in terms of computations and data requirements) forecasting system. Construction of such ensembles via weighted combination schemes, is a research topic that we plan to examine in the near future.

V. ACKNOWLEDGEMENTS

The authors acknowledge the support of KAUST and Saudi Aramco R&D Center - Carbon Management Division for their financial support in developing this work. This work was partially supported by grant #OSR–2020–4433.02 from KAUST and Saudi Aramco.
VI. REFERENCES


VII. BIOGRAPHIES

Dr. Yiannis Kamarianakis received his B.Sc in Mathematics in 1998 and his Ph.D in Mathematical Finance/Economics in 2007 both from the Univ. of Crete, Greece. He continued as a Postdoctoral Researcher in Cornell, USA and IBM Th. Watson Research Center, USA. In 2013 he joined the School of Mathematical and Statistical Sciences of Arizona State Univ. USA as an assistant professor. From 2018 he is a Principal Researcher in IACM-FORTH, Greece. His research interests are in the area of Statistical Modelling and Machine Learning and in particular on Spatial time-series modelling, Parametric nonlinear autoregressive models, Robust estimation and quantile regression. Yiannis during his career his received several awards and scholarships while he supervised many Ph.D and M.Sc theses. He has authored more than 30 refereed journal publications and more than 30 conference proceedings papers while he holds three patents.

Dr. Ioannis Pantazis received his B.Sc and Ph.D in Computer Science from Univ. of Crete, Greece in 2004 and 2010 respectively. He continued as a Postdoctoral Researcher in Univ. of Massachusetts USA. In 2018 he was elected as an Assistant Researcher in IACM-FORTH, Greece. His research interests are in the area Applied Probability, Stochastic Modelling Uncertainty Quantification and Machine Learning. During his career he received several awards and scholarships. He has authored more than 15 refereed journal publications and more than 20 conference proceedings papers.

Dr. Evangelia Kalligianaki received her B.Sc in Mathematics from Univ. of Crete in 1999 and her Ph.D in applied mathematics in 2007 from the same university. Evangelia held various post-doctoral positions in Oak Ridge National Lab, Univ. of Southern California, Univ. of Massachusetts, USA and in KAUST, Saudi Arabia. In 2018 she started as a research scientist in IACM-FORTH leading a project investigating Complex Stochastic systems, funded by Hellenic Foundation for Research and Innovation. Her research interests include, Uncertainty quantification, Statistical inference, Stochastic modelling and simulations, Machine learning. She has authored more than 15 journal publications while she has numerous conference publications.

Dr. Theodoros Katsaounis received his B.Sc in Mathematics from Univ. of Crete, Greece in 1997 and in his Ph.D in 1994 in applied and computational mathematics from the Univ. of Tennessee, USA. He held various postdoc positions in Univ. of Crete, Greece, Univ. Tennessee, USA and Ecole Normale Superieure, France. In 2002 he joined as an assistant professor the Dept. Of Applied Mathematics of Univ. of Crete, Greece where he is currently a full professor. His research interests are in the area of Applied and Computational Mathematics and in particular in Mathematical Modelling of various physical processes such as fluid flows, water waves and solar cells and the numerical solution and simulation of such models via finite element/volume methods. Theodoros has many publications in journals and conference proceedings while he has been invited/participated in numerous conferences and workshops.

Dr. Konstantinos Kotsovos holds a diploma in Electrical Engineering from the University of Patras, Greece and a PhD degree from the National Technical University of Athens (NTUA). He joined Aramco R&D since 2013 and currently is based at KAUST. His research interests include high efficiency solar cell & module design, simulation & characterization and renewable energy storage. He is the main author of coauthor of more than 20 technical papers and 2 patents.

Mr. Marwan Abdullah joined Saudi Aramco on 2011. He holds a diploma in Mechanical Engineering from KFUPM, and MS degree from the King Abdulaziz University (2022). He is the main author and co-author of more than 11 international journal and conference publications. Since 2015 he moved to Saudi Aramco R&D center at KAUST to support Sustainable
Team’s projects and activities, and established a state-of-the-art system for steam generation using solar thermal collectors.

**Dr. Issam Gereige** is a specialist in renewable energy, with over 10 years of experience in academic and industrial research. He joined Saudi Aramco in June 2014 and is currently a senior scientist in the Research and Development Center, Carbon Management division. He received his BSc degree in Physics from the Lebanese University in 2004, MSc and Ph.D. in Optics and Photonics from University of Saint Etienne, France in 2005 and 2008, respectively. He has co-authored 30 research publications in various high impact journals and presented his work in many international conferences. His research interests include high-efficiency crystalline silicon technologies for hot climate, novel energy storage systems, nanostructured materials for CO2 reduction, solar-driven electrochemical processes, and renewable energy market analysis.

**Dr. Aqil Jamal** is the Chief Technologist leading the Carbon Management Research Division of Saudi Aramco’s R&D Center in Dhahran. Dr. Jamal is responsible for directing research in Carbon Capture and Utilization, Energy Efficiency and Renewables. Dr. Jamal holds a Ph.D degree in Chemical Engineering from the University of British Columbia and did his post-doc research with National Energy Technology Lab (NETL), U.S. DOE, in Pittsburgh PA. Dr. Jamal has over 30 years of experience in sustainability related research including hydrogen production technologies, renewables and CO2 capture and utilization.

**Dr. Athanasios Tzavaras** is Professor in the Applied Mathematics and Computational Science program (AMCS) and Principal Investigator of the Applied Partial Differential Equations research group at the King Abdullah University of Science and Technology, Saudi Arabia. Professor Tzavaras obtained a Diploma in Naval Architecture and Marine Engineering from the National Technical University of Athens, Greece, and subsequently moved to USA, where he obtained M.Sc. and Ph.D degrees in Applied Mathematics from Brown University. He held academic positions at the University of Wisconsin-Madison (1987-2005), University of Maryland (2005-2009) and University of Crete, Greece (2002-04 and 2010-2014) and visiting positions in Purdue University, Stanford University, Ecole Polytechnique and University Paris VI. His research interests include mathematical modelling, and the analysis and computation of partial differential equations arising in fluid mechanics and material modelling across multiple scales. He is a member of the American Mathematical Society (AMS), the Society of Industrial and Applied Mathematics (SIAM), and of the International Society for the Interaction of Mechanics and Mathematics as well as corresponding Editor of the SIAM Journal for Mathematical Analysis, and an Associate Editor of the Journal of Analysis and its Applications (ZAA).