KNN-based ensembles for day-ahead solar power output forecasting

Yiannis Kamarianakis[1], Yannis Pantazis[1], Evangelia Kalligiannaki[1], Theodoros D. Katsaounis[2], Konstantinos Kotsovos[3], Issam Gereige[3], Marwan Abdullah[3], Aqil Jamal[3] and Athanasios Tzavaras[4]

[1] Data Science Group, Institute of Applied and Computational Mathematics, Foundation for Research and Technology-Hellas, Heraklion Crete (Greece)

[2] Dept. of Mathematics and Applied Mathematics, University of Crete, Heraklion (Greece)

[3] Saudi Aramco R&D Center, King Abdullah University of Science and Technology (KAUST), Thuwal (Saudi Arabia)

[4] Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal (Saudi Arabia)

Abstract

This manuscript evaluates a recently proposed times series KKN (KNN_{TS}) methodology, in terms of its accuracy in day-ahead forecasting of PV power outputs. Specifically, KNN_{TS} is evaluated against three widely applied benchmarks: seasonal ARIMA, spline-based daily profiles (SDP) and the persistence model (PRS). Two different solar cell architectures are examined under real operating conditions: Aluminum Back Surface Field (Al-BSF) and Back Contact (BC). Our analyses suggest that KNN_{TS} does not dominate the alternative, benchmark specifications. However, linear combinations of KNN_{TS} with SDP, may produce significantly improved forecasts. The outcomes of the model-building procedure are incorporated in a light online forecasting system for solar panels located in KAUST, Thuwal, Saudi Arabia.

Keywords: solar power output forecasting, time series KNN, ensemble modeling, shrinkage methods

1. Introduction

Accurate forecasts of photovoltaic (PV) outputs are essential to Distribution System Operators and Transportation System Operators (Antonanzas et al., 2017) for efficient solar energy trading and management of electricity grids (EPIA, 2012). The main aim of this work is to develop specifications for day-ahead forecasting of PV energy outputs, 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 (KNN_{TS}) (Martinez et al., 2019a), using observed energy yields (EY) from 2 panel technologies: Aluminum Back Surface Field (Al-BSF) and Back Contact (BC).

KNN_{TS} is straightforward to implement online as it is based solely on historical data and does not require next-day meteorological predictions as input. Of course, advanced predictive models (e.g. penalized regressions, random forests, neural networks; Nespoli et al., 2019) that use meteorological information should dominate such relatively simple schemes by a significant margin, to justify their substantially increased cost (both computational and in terms of the price to acquire continuously updated meteorological forecasts). Online forecasting systems for PV and wind-farm energy outputs are expected to employ such benchmark specifications while in "safe-mode" operation, when for instance, access to day-ahead meteorological forecasts has been interrupted.

In what follows, KNN_{TS} 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

doi:10.18086/eurosun.2022.16.09 Available at http://proceedings.ises.org

conventional RMSE; in addition, the relative accuracy metric proposed by Tofallis (2015) is monitored.

2. Methodology

The adopted KNN_{TS} methodology proposed recently by Martinez et al., (2019a), is implemented in `R` via the `tsfknn` package (Martinez et al., 2019b). Given a recently observed daily profile of PV outputs, conventional KNN identifies *k* profiles that are closest based on a chosen distance metric and reports the average of *k*, neighbor-specific subsequent values. The number of nearest neighbors is a decision variable that should be tuned. On the other hand, in KNN_{TS}, Martinez et al., (2019a) propose an ensemble scheme that instead of choosing *k*, averages forecasts corresponding to *k*=3,5,7. This scheme provided very satisfactory levels of accuracy in our preliminary analyses; hence all results depicted in the next sections are based on forecast combinations for *k*=3, 5, 7. In the application, a sensitivity analysis similar to the one presented in Martinez et al., (2019a) is performed, to evaluate alternative combination schemes (mean, median, weighted mean). Additional tuning decisions should be performed regarding a) the distance metric used to identify nearest neighbors (e.g. the widely adopted Euclidean, versus Manhattan distance; b) the multi-step ahead forecasting strategy and c) the time lags used as input variables. The first two choices are investigated in the application whereas for the third, we follow the recommendation in Martinez et al., (2019a) for periodic time series and set the number of time lags equal to the number of time-periods per day.

In contrast with KNN_{TS}, SDP reports daily profiles that utilize all measurements observed during a pre-specified time interval. Cowpertwait and Metcalfe (2009) present parametric specifications for SDP, based on a) dummy variables (SDP_a), or b) sine and cosine functions (SDP_b). Our preliminary experiments revealed that a non-parametric procedure based on regression splines can be more accurate relative to SDP_a and substantially faster to compute, relative to SDP_b without compromising forecasting performance. The procedure implemented herein, fitted piece-wise cubic polynomials with 15 knots (knots are breakpoints in the third derivative; the number of knots was chosen by performing a cross-validation experiment) arranged at equally-spaced time instants. The function `bs` in the `R` package `splines` is used to construct B-spline basis expansions. Daily profiles are then derived using conventional least squares estimation, which is compatible with the primary adopted accuracy criterion (RMSE). It is worth stressing that a simplified SDP-variant of the procedure applied here, typically computes static, month-specific daily profiles. This results in a straightforward online implementation which does not require computations on a daily basis; instead, reported forecasts are retrieved from a small database. On the contrary, here SDP is updated on a daily basis in terms of training data. Furthermore, the number of training weeks is a tuning parameter.

The persistence model (PRS) is a typical benchmark in PV-output forecasting experiments. PRS implies that current conditions remain unaltered and that a future daily profile will be very close to the one most recently observed. Hence observations that correspond to the most recent, fully observed day, are used to forecast an unknown daily profile. ARIMA models constitute a second widely adopted benchmark specification. In fact, PV-outputs behave similar to the classic textbook example of periodic time series with a trend [e.g. the airline series used in Brockwell and Davis (2009)]. In the example application ARIMA model building is performed with the `auto.arima` function, which is available in the `forecast` package. A-priori, PV-outputs are expected to behave as non-stationary processes, as daily periodicities are coupled with a decreasing trend, due to declining PV efficiencies. Non-stationarity is accounted for, by treating the number of training weeks for KNN_{TS}, ARIMA and SDP, as a tuning parameter. This ensures that the computational scheme in the online implementation is computationally light, with low requirements in terms of training data.

The prioritized accuracy metric in the example application, is RMSE, which is compatible with conventional least squares estimates. An alternative, widely used accuracy metric is the mean absolute percentage error (MAPE). Tofallis (2015) showed that MAPE systematically favors specifications that produce low forecasts. Specifically, MAPE regression can be viewed as a weighted median regression with the observed measurements taking the role of weights (Tofallis, 2015); hence the lowest measurements are more influential and the predictive specification is pulled towards them. To overcome the shortcomings of MAPE, Tofallis (2015) proposed an alternative relative error metric, namely *LnQ*, which, is formulated as follows:

$$LnQ = \frac{1}{N}\sum_{t=1}^{N}\left[ln\left(\frac{\hat{Y}(t)}{Y(t)}\right)\right]^2 \qquad \text{(eq. 1)}$$

with $Y(t)$ denoting observed and $\hat{Y}(t)$ forecasted values.

3. Data

EY measurements (W/m^2) are collected at KAUST, Thuwal, Saudi Arabia: a challenging location for solar panels, which often need to operate in the presence of dust and/or at temperatures far beyond the Standard Test Conditions (STC). The analysed data are recorded every 10 minutes, from 8AM to 5PM (as reported EY is negligible before 8AM and later than 5PM), for 364 consecutive days (52 weeks) in 2016, with starting (ending) date, 01 January 2016 (29 December 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 TSKNN 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% (Fig.2). 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 [47 (52) days with missing percentage larger than 10% for Al-BSF (BC)] and b) available historical data included measurements of solar irradiance (W/m^2), an irradiance-based imputation scheme, which reveals intriguing characteristics of the two panel types, is adopted herein.

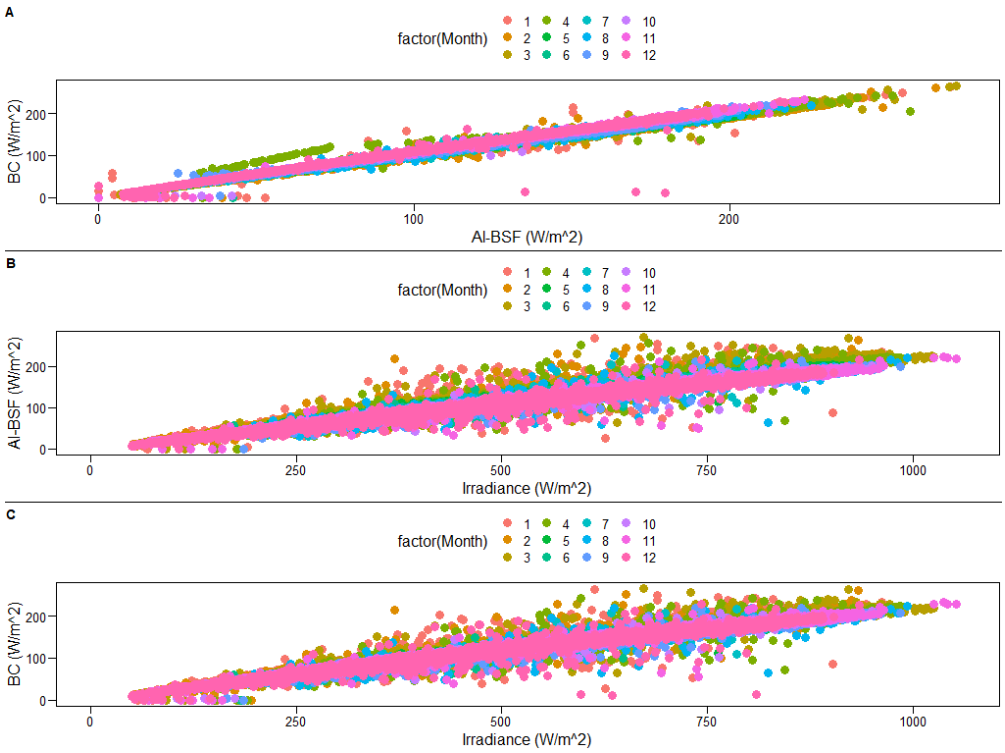

Fig. 1: Scatterplots for the associations between (A) Al-BSF versus BC energy outputs; (B) Al-BSF energy outputs versus irradiance; (C) BC energy outputs versus irradiance.

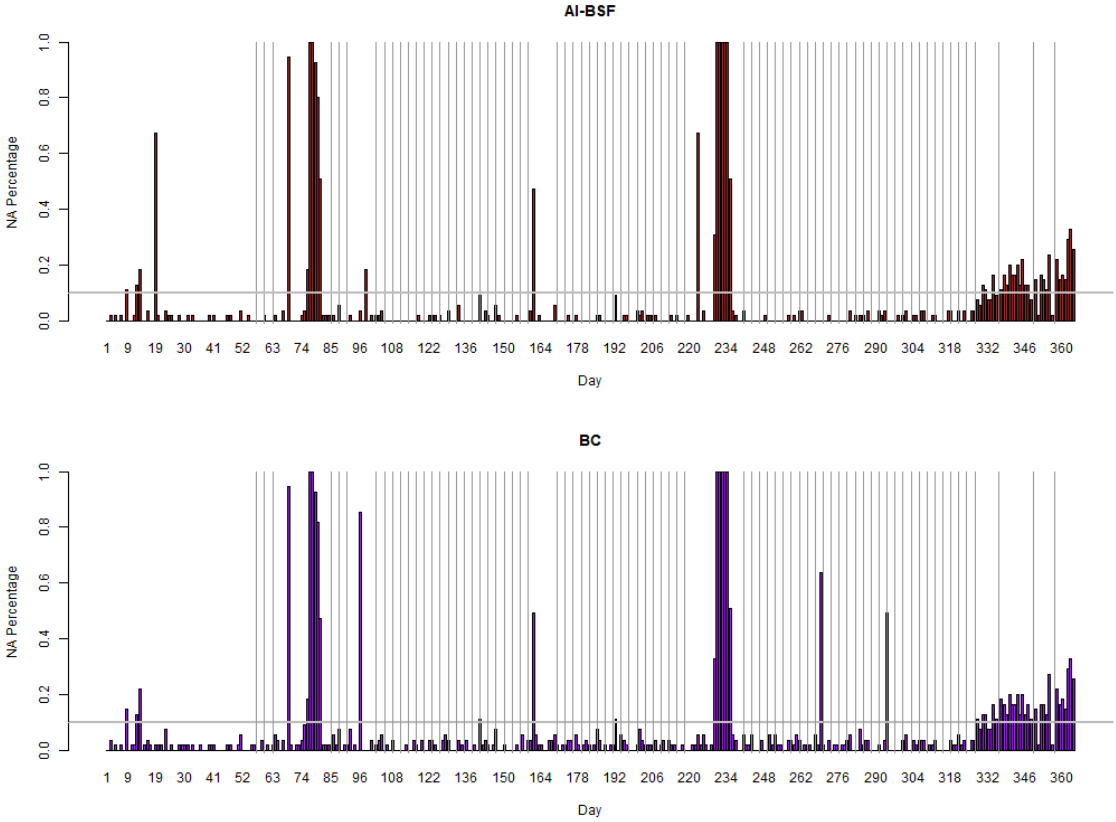

Fig. 2: Daily percentages of missing values for the two solar panel technologies. A horizontal line designates the 10% threshold, whereas vertical lines depict the days that comprise the first testing dataset.

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

$$\tilde{Y}_i(t) = c_i X(t) \qquad \text{(eq. 2)}$$

with $X(t)$ denoting observed irradiance at time t, c_1, c_2 designating Al-BSF and BC efficiencies and $\tilde{Y}_1(t), \tilde{Y}_2(t)$, imputed Al-BSF and BC power outputs, respectively. The overly parsimonious approximation in (2) should achieve subpar levels of accuracy (Fig. 1), given the expected decreasing trend for PV efficiencies. A more flexible scheme, which allows for monthly-varying efficiencies is formulated as:

$$\tilde{Y}_i(t) = c_i^m X(t) \qquad \text{(eq. 3)}$$

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.

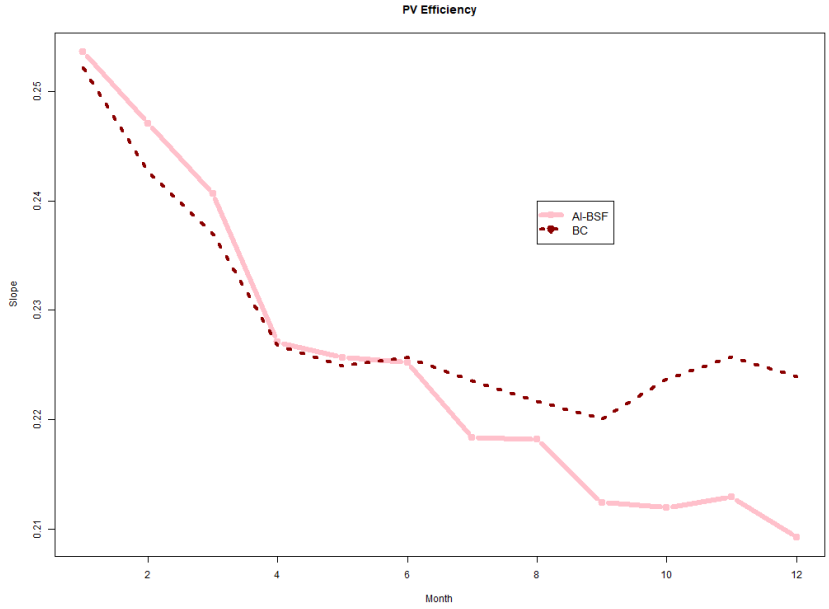

Fig. 3: LAD estimates for monthly efficiency rates corresponding to the specification shown in (3).

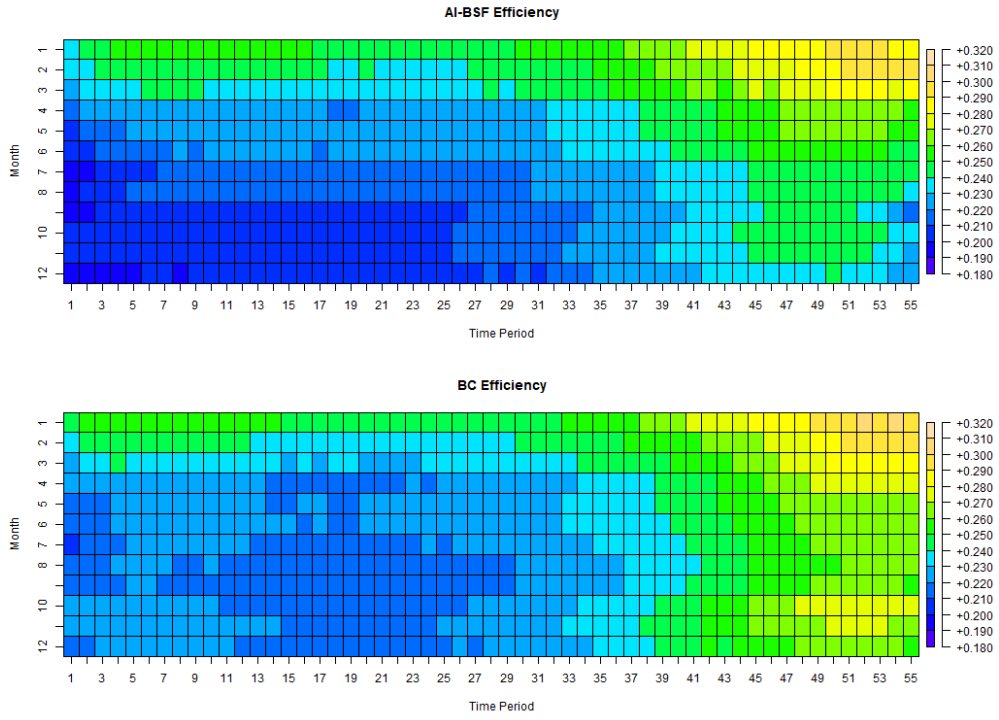

Fig. 4: LAD estimates of month-specific, daily efficiency rate profiles, captured by the slopes in (4). Row numbers correspond to months whereas columns represent time periods within a day: $t_d = 1$ ($t_d = 55$) denotes 8AM (5PM).

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

$$\tilde{Y}_i(t) = c_i^{m,t_d} X(t) \qquad \text{(eq. 4)}$$

with $t_d = 1, \ldots, 55$ designating 10-minute intervals [$t_d = 1$ ($t_d = 55$) corresponds to 8AM (5PM)]. Essentially the above specification takes (indirectly) into account dependence of PV efficiencies on operating temperatures (Skoplaki and Palyvos, 2009). Figures 4 and 5 depict curvilinear, increasing daily profiles, which are derived with outlier robust, median regression (Koenker, 2005). During the first months of the study period, differences in PV efficiencies are mainly observed when the levels of irradiance are low (around 8AM and 5PM). These differences increase with time; as shown in Fig. 5, BC achieves superior efficiency rates relative to Al-BSF towards the end of the study period, during the whole day.

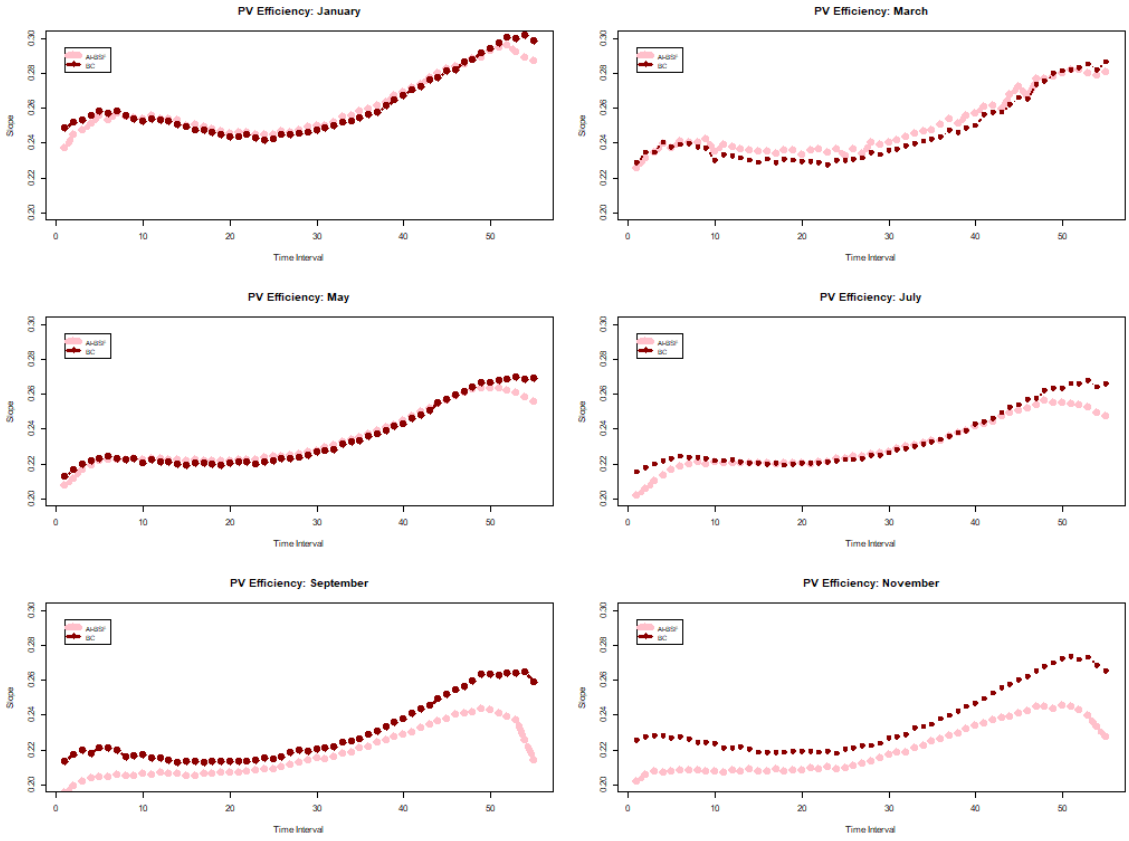

Fig. 5: LAD estimates of month-specific, daily profiles for the PV efficiency rates in (4).

As noted by a reviewer, Figures 4 and 5 indicate higher panel efficiencies in the afternoon, which is not in accordance with what one would expect, based for instance in Skoplaki and Palyvos (2009). Typically, cell temperature is lower in the morning and lower in the afternoon; on the other hand, an evaluation of the dataset analyzed in Katsaounis et al., (2019), which relates strongly to the data analyzed here, suggests that both air and cell temperatures drop substantially in the afternoon (< 30 °C) to reach similar magnitudes. Furthermore, Skoplaki and Palyvos (2009) assume that the open circuit voltage, one of the critical factors that influence electrical efficiency of a PV cell/module, decreases with temperature. Such a decreasing association is not clearly manifested in the dataset analyzed in Katsaounis et al., (2019).

Robust imputation is prioritized here, that is why alternative imputation procedures are evaluated via mean absolute error (MAE). Fig. 6 displays results of a leave-2-week-out cross-validation experiment, which evaluates imputation accuracy achieved from specifications (2), (3) and (4). 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. 6 clearly demonstrates the superiority of the daily profiles in (4): the average MAE achieved from (4) across all periods is close to 5 W/m^2, whereas for the parsimonious scheme in (3) it is close to 10 W/m^2 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 (4) with coefficients estimated via LAD. Fig.7 presents observed and imputed data for the performance profiles that constitute the last week of the study period.

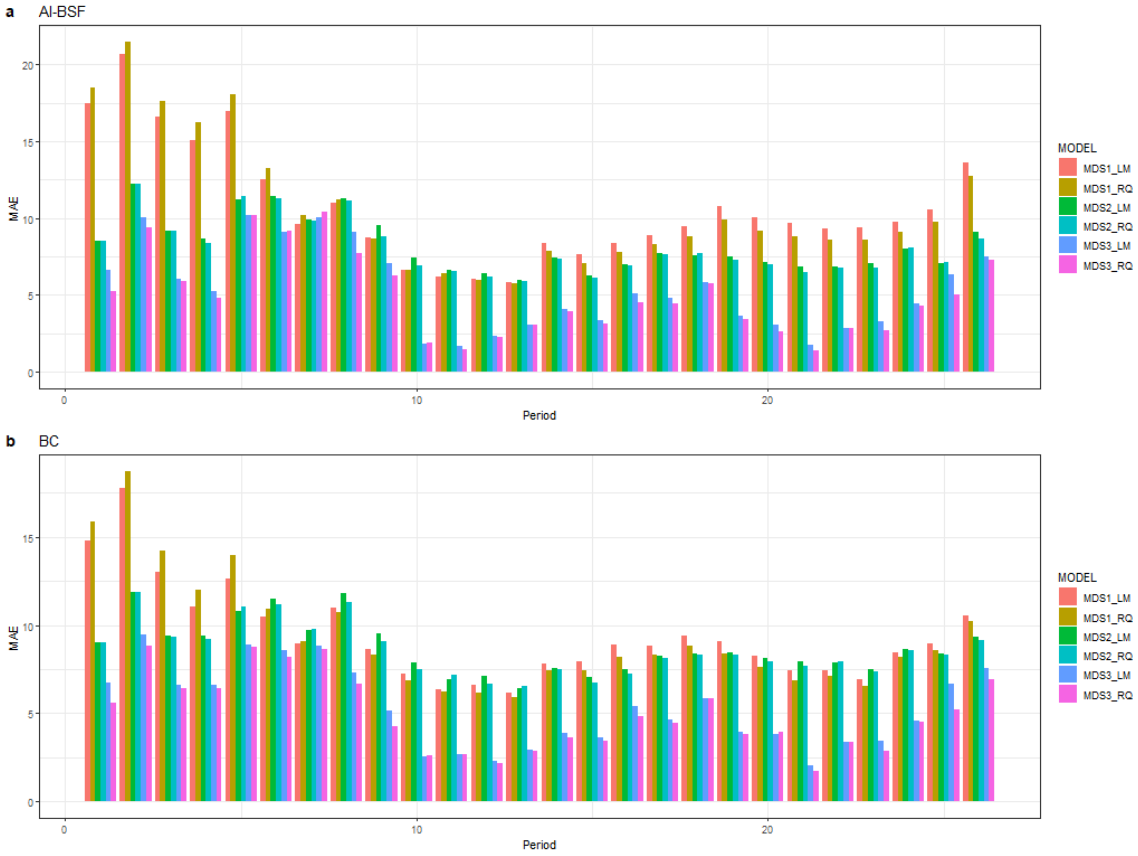

Fig. 6: MAE of alternative imputation schemes for Al-BSF (a) and BC (b). LM denotes conventional least-squares whereas RQ corresponds to outlier-robust median regression estimates (LAD). MDS1 represents the static model in eq. (2), MDS2 stands for the specification shown in eq. (3) and MDS3 is the adopted imputation scheme shown in (4).

4. Forecasting Experiment

This section presents a forecasting experiment which mimics real world, online implementation of the examined specifications. 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 (Fig. 2). Training data consist of days before d and forecasts correspond to the day d+1. Computational times were less than 5 seconds for both SDP and KNN$_{TS}$, 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 Figures 8, 9 and Table 1. Interestingly, focusing on RMSE, SDP outperforms both KNN$_{TS}$ and ARIMA; in accordance with prior expectations, PRS is the worst performing method by a significant margin. Regarding KNN$_{TS}$ 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 (KNN$_{TS}$) 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.

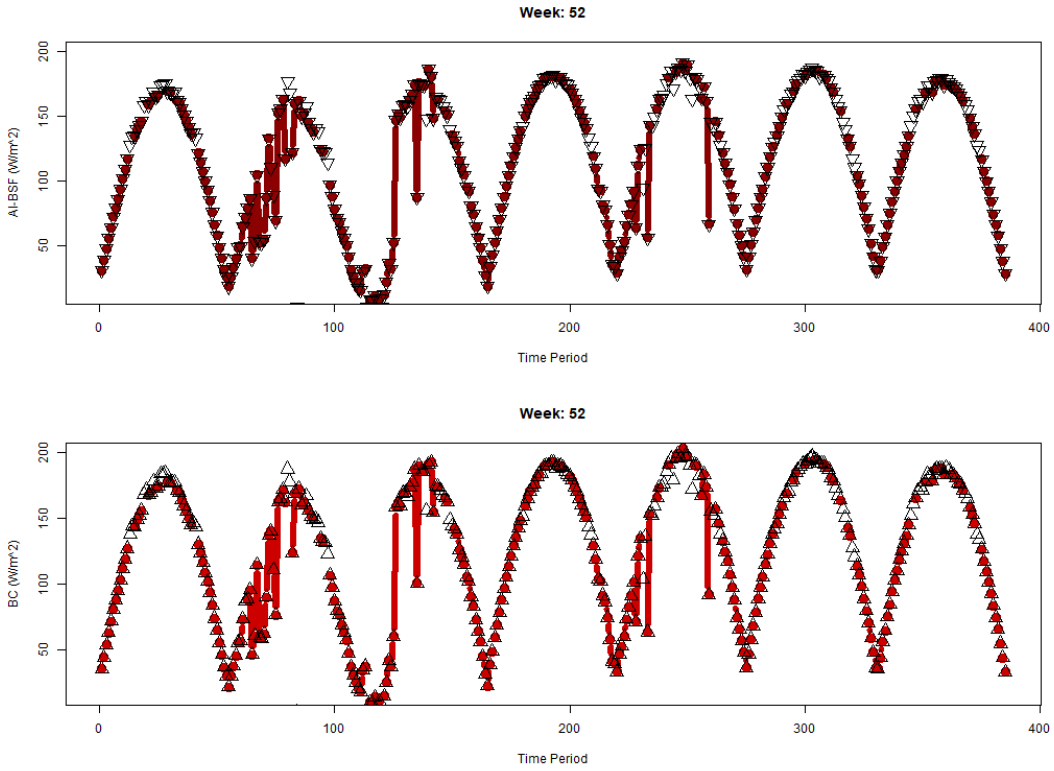

Fig. 7: Observed (imputed) measurements for the last week of the study period are shown as colored (empty) triangles: Al-BSF (top), BC (bottom).

The above evaluation is not insensitive to the chosen accuracy metric. A practitioner that focuses on relative accuracy and adopts LnQ to compare models, would not discard PRS, which despite its simplicity achieves levels of accuracy very close to ARIMA and SDP. Interestingly, KNN_{TS} is by far the worst performing methodology in terms of LnQ. The optimal combination function used to aggregate the targets associated with the nearest neighbors is again the median. SDP appears to dominate in terms of LnQ as it did for RMSE. It is worth emphasizing though, that its performance degrades significantly when $N_W \neq 2$ and that SDP is far from the best performing method when N_W is chosen with a priority on RMSE performance. Fig. 10 shows that forecasts from different methodologies are very strongly correlated. Hence the expected gains from a forecast combination scheme are not dramatic in this application.

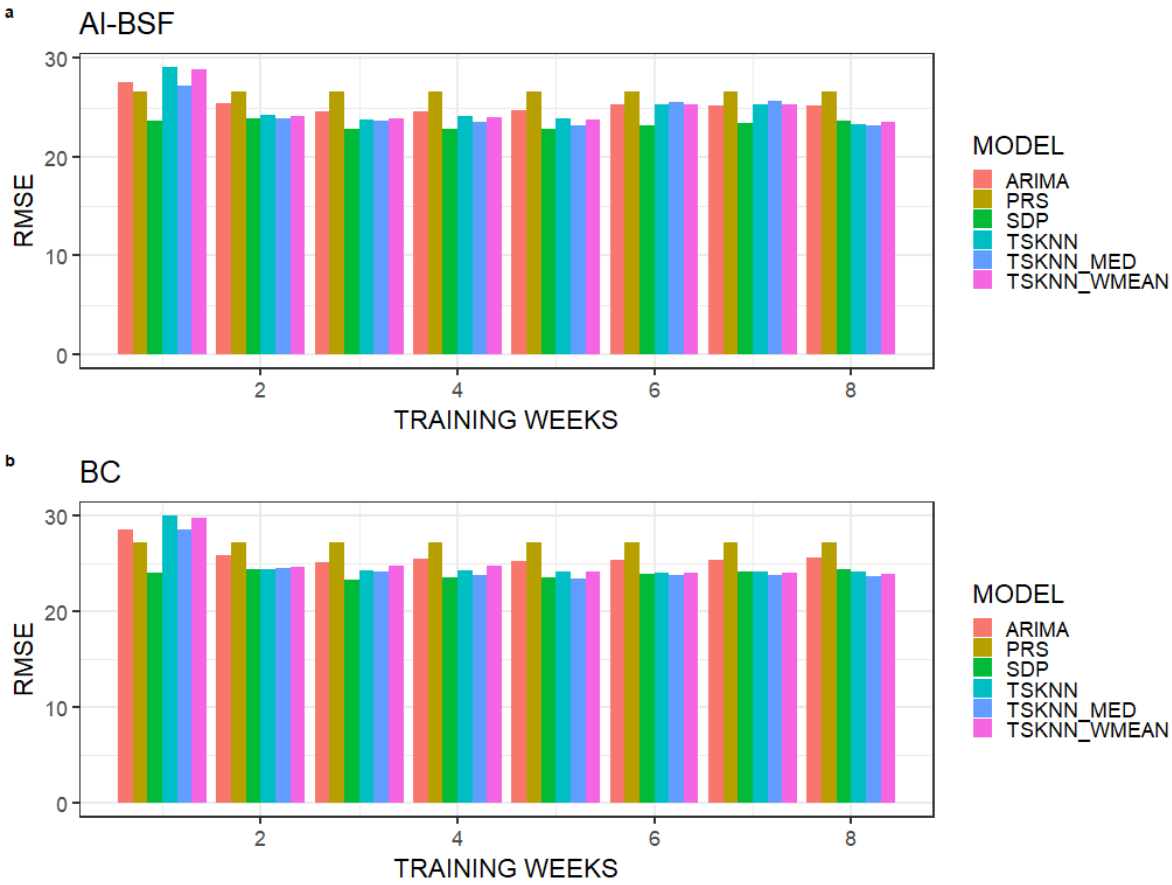

Fig. 8: RMSE performance of the persistence model, versus ARIMA, SDP, and 3 KNN_{TS} variants, with increasing training data. In TSKNN the combination function used to aggregate the targets associated with the nearest neighbors is the mean, in TSKNN_MED it is the median whereas in TS_WMEAN a weighted average is computed, with distance-based weights.

Tab. 1: RMSE and LnQ performance of the persistence model, versus ARIMA, SDP and KNN_{TS}. Reported accuracies depend on the number of training weeks N_W, which is shown in parentheses.

PV Type	Model	RMSE	LnQ
Al-BSF	PRS	26.619	42.462
Al-BSF	SDP	**22.809 (5)**	**41.441 (2)**
Al-BSF	ARIMA	24.582 (3)	43.601 (4)
Al-BSF	KNN_{TS}	23.120 (8)	56.394 (8)
BC	PRS	27.125	43.447
BC	SDP	**23.314 (3)**	**42.072 (2)**
BC	ARIMA	25.072 (3)	42.620 (3)
BC	KNN_{TS}	23.624 (8)	56.395 (8)

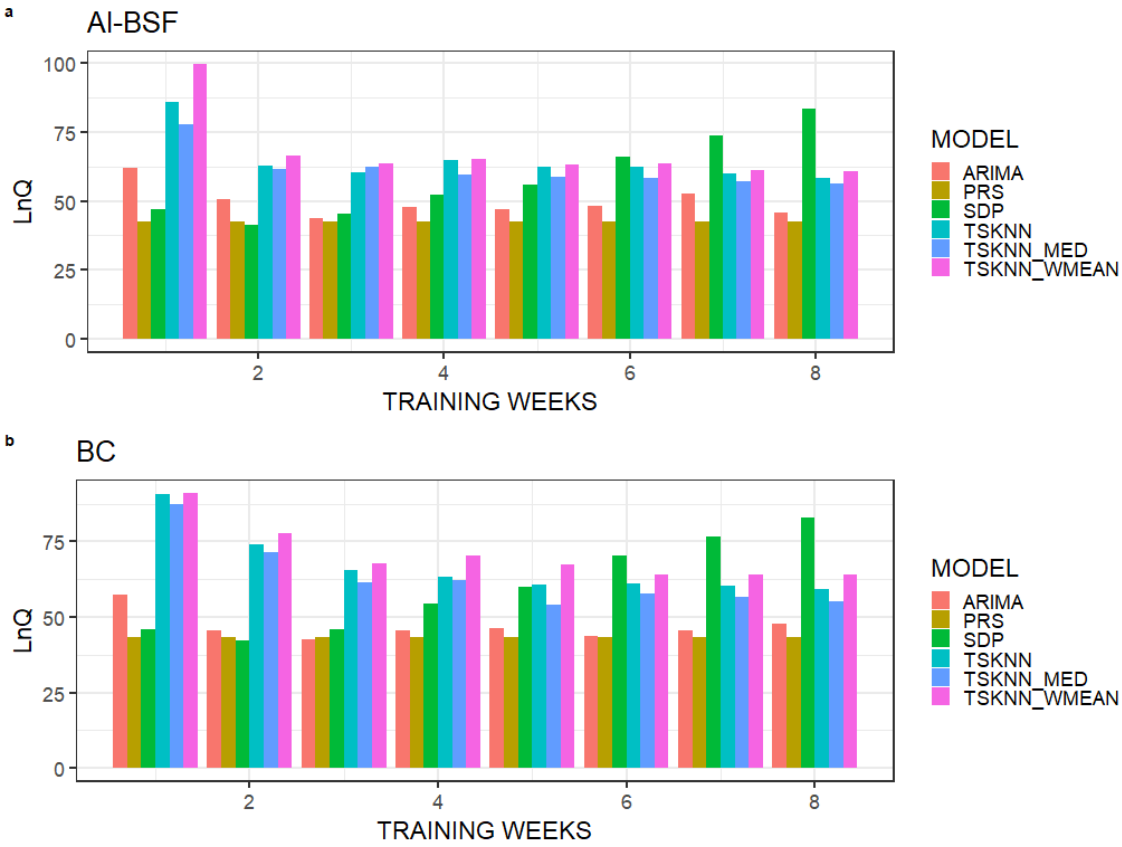

Fig. 9: LnQ performance of the persistence model, versus ARIMA, SDP, and 3 KNN_{TS} variants, with increasing training data. In TSKNN the combination function used to aggregate the targets associated with the nearest neighbors is the mean, in TSKNN_MED it is the median whereas in TS_WMEAN a weighted average (with distance-based weights) is computed.

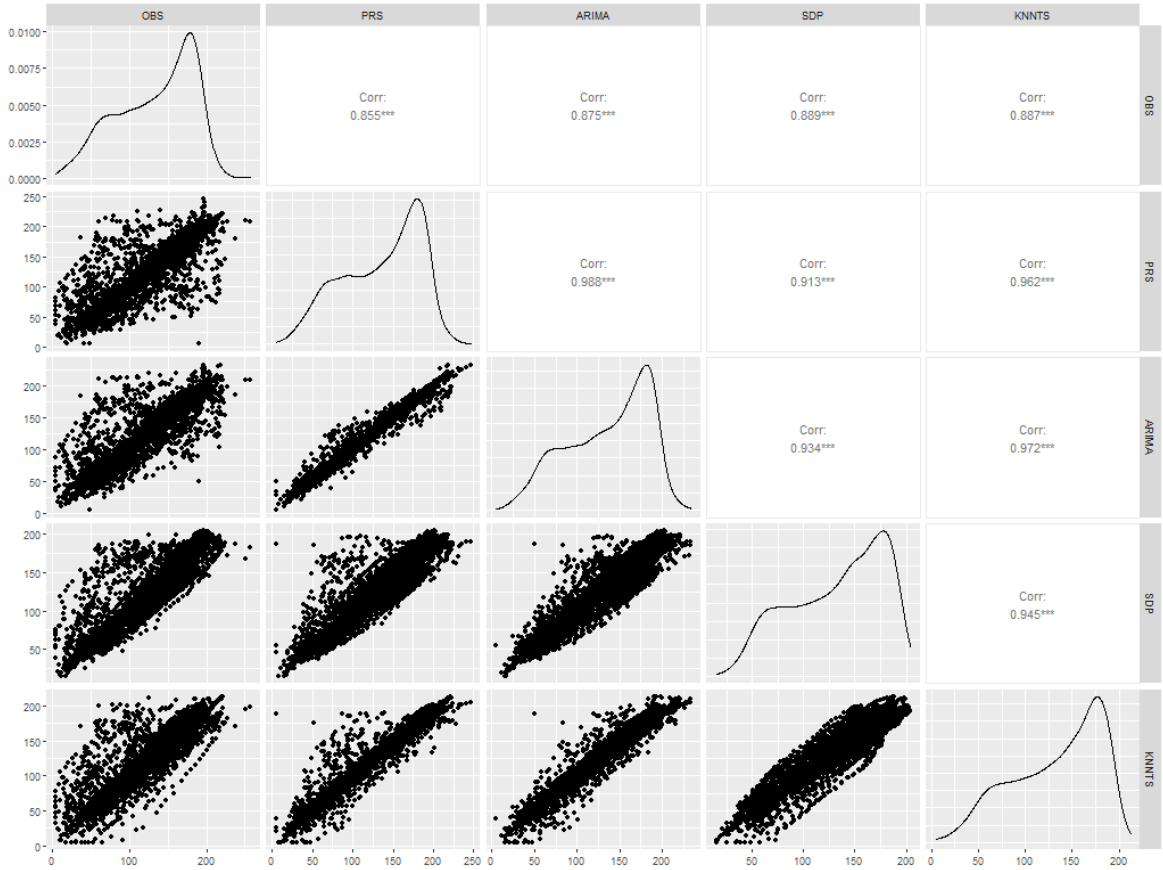

Fig. 10: Scatterplots and Pearson's bivariate correlation metrics for Al-BSF. Stars designate strong evidence against the null hypothesis that assumes zero correlation. Observed values (OBS), PRS, ARIMA with N_W=3, TSKNN with N_W=8 and SDP with NW=5 forecasts, are shown.

5. 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 types 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. In fact, if one adopts a relative error metric, KNN_{TS} is by far the worst performing method. Despite the poor performance on the example application KNN_{TS} forecasts are expected to perform well when environmental conditions are highly variable, with regime-specific variability. The 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 next.

6. Acknowledgments

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.

7. References

Antonanzas, J., Pozo-Vázquez, D., Fernandez-Jimenez, L.A., Martinez-de-Pison, F.J., 2017. The value of day-ahead forecasting for photovoltaics in the Spanish electricity market. Solar Energy 158, 140-146. https://doi.org/10.1016/j.solener.2017.09.043.

Brockwell, P.J., Davis, R.A., 1991. Time Series: theory and methods (2nd edition). Springer science & business media. http://dx.doi.org/10.1007/978-1-4419-0320-4

Cowpertwait, PSP, Metcalfe, A.V., 2009. Introductory Time Series with R, Springer. https://doi.org/10.1007/978-0-387-88698-5

EPIA, 2012. Available online: http://pvtrin.eu/assets/media/PDF/Publications/other_publications/263.pdf

Katsaounis, Th., Kotsovos, K, Gereige, I., Basaheeh, M., Abdullah, A., Khayat, A., Al-Habshi, E., Al-Saggaf, A., Tzavaras, A.E., 2019, Performance assessment of bifacial c-Si PV modules through device simulations and outdoor measurements, Renewable Energy, 143, 1285-1298. https://doi.org/10.1016/j.renene.2019.05.057

Koenker, R., 2005. Quantile Regression, Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511754098

Martínez, F., Frías, M.P., Pérez, M.D., Rivera, A.J, 2019a. A methodology for applying k-nearest neighbor to time series forecasting. Artif Intell Rev 52, 2019-2037. http://dx.doi.org/10.1007/s10462-017-9593-z

Martínez, F., Frías, M.P., Charte, F., Rivera, A.J, 2019b. Time series forecasting with knn in R: the tsfknn package. R Journal, 11, 229. http://dx.doi.org/10.32614/RJ-2019-004

Nespoli A, Ogliari E, Leva S, Massi Pavan A, Mellit A, Lughi V, Dolara A., 2019. Day-Ahead Photovoltaic Forecasting: A Comparison of the Most Effective Techniques. Energies 12(9):1621. http://dx.doi.org/10.3390/en12091621

Skoplaki, E., Palyvos, J.A., 2009. On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations. Solar Energy 83, 614-624. http://dx.doi.org/10.1016/j.solener.2008.10.008

Tofallis, C., 2015. A better measure of relative prediction accuracy for model selection and model estimation. Journal of the Operational Research Society, 66, 1352-1362. http://dx.doi.org/10.1057/jors.2014.103