DAY-AHEAD FORECASTING OF SOLAR IRRADIANCE: KNN-BASED ENSEMBLES

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

[1]FORTH, IACM, Heraklion, Greece, [2]ARAMCO, Renewable Energy, Carbon Management Division, Thuwal, Saudi Arabia, [3]KAUST, CEMSE, Thuwal, Saudi Arabia,
[4]University of Crete, Dept. of Mathematics & Applied Mathematics, Heraklion, Greece

ABSTRACT: This work presents a novel, outlier-robust forecasting scheme, which is derived by combining two specifications. Specifically, a recently proposed time series KNN (TSKNN) ensemble procedure is combined with spline-based daily profiles (SDP), estimated via median regression. An experiment that mimics real-world implementation of an online system is performed, to evaluate day-ahead solar irradiance forecasts of the proposed method, dubbed TSKKN$_{R0}$, versus TSKNN, SDP and the persistence model. Alternative specifications are evaluated with a new relative error metric which is outlier-robust. The analyzed measurements are collected at KAUST, Thuwal, Saudi Arabia for a full year, with 10-minute granularity.

1 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 [1], as they assist efficient solar energy trading and management of electricity grids [2]. This work evaluates a recently proposed, computationally-light time series KNN-regression scheme (TSKNN) [3], for day-ahead solar irradiance (SI) forecasting. SI is an essential predictor of photovoltaic (PV) energy output.

Time series KNN schemes (TSKNN) are straightforward to implement in online applications as they are based solely on historical data: next-day meteorological predictions are not required as additional input. Of course, advanced predictive models (e.g. penalized regressions, random forests, neural networks; [4] that use meteorological information, should dominate such 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 “cheap” specifications while in “safe-mode” operation, when for instance, access to day-ahead meteorological forecasts has been interrupted.

Outlying SI measurements may occur, due to extreme weather or malfunctioning equipment. This work emphasizes outlier-robust estimation methods and accuracy metrics, in contrast to the vast majority of published research. To the best our knowledge this is the first application of TSKNN to forecast SI; furthermore, it is the first study that constructs a forecast-combination-scheme, which aims at improving TSKNN performance.

2 DATA

SI measurements (*W/m²*) 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). Specifically, historical observations are collected at the New Energy Oasis (NEO) test field near the Red Sea coast (22.30 N, 39.10 E), during 2016 [364 consecutive days (52 weeks), with starting (ending) date, 01 January 2016 (29 December 2016)]. For each day, an observation is recorded every 10 minutes, starting at 8:00 and ending at 17:00 (earlier than 8 AM and later than 5 PM, SI levels are low, with negligible resulting PV energy yields; hence these measurements are omitted). Thus, the maximum number of daily observations, at times $t_n > 0, n = 1, \ldots, T$, is $T = 55$. Data cleaning and outlier-removal have been applied to eliminate statistical artifacts (measurements the lie outside a range of reasonable values in *W/m²*).

3 METHODOLOGY

3.1 TSKNN

The analyses presented herein are based on the methodology proposed in [3], which is implemented in `R` via the `tsfknn` package. Interestingly, [3] proposed an ensemble scheme that, instead of choosing the number of nearest neighbors *k* (typically a tuning parameter), it averages forecasts corresponding to *k* = 3, 5, 7. This is the default choice in `tsfknn` package, which provided very satisfactory levels of accuracy in the 111 time series of the NN3 competition [6]. Our preliminary analyses confirmed the above-mentioned finding; all results depicted in the next sections are based on forecast combinations for *k* = 3, 5, 7. In the application, TSKNN implementation is based on the default values regarding a) the distance metric (Euclidean is widely adopted), b) multi-step ahead forecasting strategy and c) the time-lags used as input variables. Regarding a) and b), preliminary investigations suggest that the default settings in `tsfknn` package provide superior levels of accuracy, whereas regarding c), we follow the recommendation in [3] for periodic time series and set the number of time lags equal to the number of time-periods per day. The effect of the length of the training period, N$_{TSKNN}$ is examined in the forecasting experiment, for N$_{TSKNN}$ = 1,…,5 weeks.

3.2 Robust smooth daily SI profiles

A typical day-ahead forecasting benchmark is based on average daily profiles (DP). A significant advantage of DP estimators is that uncertainty quantification (forecasting intervals) is straightforward, via quantile regression [7]. The simplest method to calculate such profiles, utilizes time-specific dummy variables [8]. Such specifications are straightforward to implement using conventional least-squares-based estimators, or outlier-robust procedures [e.g. weighted least squares, least absolute deviations, a.k.a. median regression]. Unfortunately, the resulting number of unknown parameters is large in our case (55 coefficients) and requires an extensive historical period of available measurements to produce, reasonable, smooth daily profiles.

A second class of DP estimators is based on sine and cosine functions [8], which induce smooth variation into periodic models. Again, the resulting statistical models are straightforward to estimate with conventional or robust procedures. Their main disadvantage, is that model building (selection of statistically significant sine/cosine frequencies) is computationally demanding. On the other hand, our experiments indicated very similar levels of accuracy to the spline-based estimators that are presented next; hence performance of such periodic models will not be discussed further.

The specifications described thus far correspond to parametric models. An alternative, non-parametric modeling strategy, which is fast to compute without compromising accuracy, uses regression splines [9]. The procedure implemented here (dubbed SDP, for smooth daily profiles), fitted piece-wise cubic polynomials with 15 knots. Knots are breakpoints in the third derivative (the number of knots was chosen in a preliminary cross-validation experiment) arranged at equally-spaced time instants. The function `bs` in the `R` package `splines` was used to construct B-spline basis expansions.

Daily profiles can be again derived using least squares estimation or robust alternatives. An additional robust estimation procedure that wass evaluated in this case, is likelihood-based and adopts the heavy-tailed *t* distribution for the residuals (`R` package `mgcv`).

Similar to TSKNN, the effect of the length of the training period, N_{SDP} is examined in the forecasting experiment, for N_{SDP} = 1,…,8 weeks. The forecasting experiments that follow, focus on median-regression-based SDPs, which according to our experiments dominate alternatives for the adopted accuracy criterion that is discussed in Section 3.4. Furthermore, forecasting intervals are straightforward to compute by switching to quantiles other than the median.

3.3 Robust TSKNN

TSKNN is expected to perform well when environmental conditions are unstable: for instance, in a simplified situation with two daily regimes, namely clear sky and sandstorm, when a series of days with clear skies are followed by days with sandstorms, TSKNN will produce a forecast that is based on historical information from the most recent regime. On the other hand, spline-based average daily profiles are expected to perform well in time periods which are stable, since the profiles are computed using a substantially larger sample size (number of days) relative to TSKNN. Furthermore, robustly estimated (e.g. via median regression) smooth profiles essentially neglect measurements that correspond to extreme environmental conditions and exploit the vast majority data.

This work proposes a robust TSKNN variant, dubbed $TSKKN_{R0}$. $TSKKN_{R0}$ is motivated by shrinkage estimators, which are widely applied in regression problems via the LASSO [10]. Essentially, given that TSKNN exploits a relatively small number of days from the training data to compute forecasts for daily profiles, $TSKKN_{R0}$ ``shrinks" TSKNN towards a smooth, outlier-robust SDP profile, derived from a significantly larger sample. An interesting research question, relates to the mixing weights in $TSKKN_{R0}$; should they be kept fixed or do different periods within a year require different weights? Obviously, a scheme based on varying weights is expected to cope with periods with different characteristics throughout a year; unstable periods are expected to correspond to relatively large weights for TSKNN relative to SDP. In general, an estimator that combines TSKNN with robust SDP can also be viewed as a classic forecast combination scheme. In light of the forecast combination puzzle [11], the application that follows examines a simple scheme with equal weights, which does not require any additional tuning.

3.4 Benchmark Models and Accuracy Metrics

The persistence model (PRS) is a typical benchmark in renewable-energy-related forecasting experiments. This naive scheme 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 day, are used to forecast unobserved SI profiles, for one or more days ahead.

An accuracy metric that is widely used in engineering is the mean absolute percentage error (MAPE). As shown, among others in [5], MAPE systematically favors methodologies that produce low forecasts. Specifically, MAPE regression can be viewed as a weighted median regression with the observed measurements taking the role of weights; hence the lowest measurements are more influential and the predictive specification is pulled towards them. To overcome the above-mentioned shortcomings, [5] proposed an alternative relative error metric, namely *LnQ*, which is formulated as follows

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

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

Inspired by *LnQ*, the accuracy criterion that is reported in our forecasting experiments, is an outlier-robust' variant, dubbed LnQ_R

$$LnQ_R \ = \frac{1}{N}\sum_{t=1}^{N}\left|ln\left(\frac{\widehat{Y(t)}}{Y(t)}\right)\right| \qquad (2)$$

It is worth highlighting that *LnQ* complies with a multiplicative error model, which is appropriate for heteroscedastic processes [5]. *LnQ*-optimal regression estimates via least squares, coefficients of a linear model on log-transformed responses, whereas LnQ_R -optimal regression uses median regression instead.

4 RESULTS AND DISCUSSION

Figures 1 and 2 depict static SDP variants where differences between a) parametric versus nonparametric specifications and b) conventional versus outlier-robust estimation procedures are illustrated. The forecasting experiment that is discussed next evaluates four day-ahead forecasting methodologies in terms of their LnQ_R –performance: a) the naïve PRS, b) TSKNN, c) robust-SDP that combines smoothing splines with median regression and d) TSKKN$_{R0}$, which does not require additional tuning as it corresponds to simple averaging of TSKNN and SDP-based daily SI forecasts.

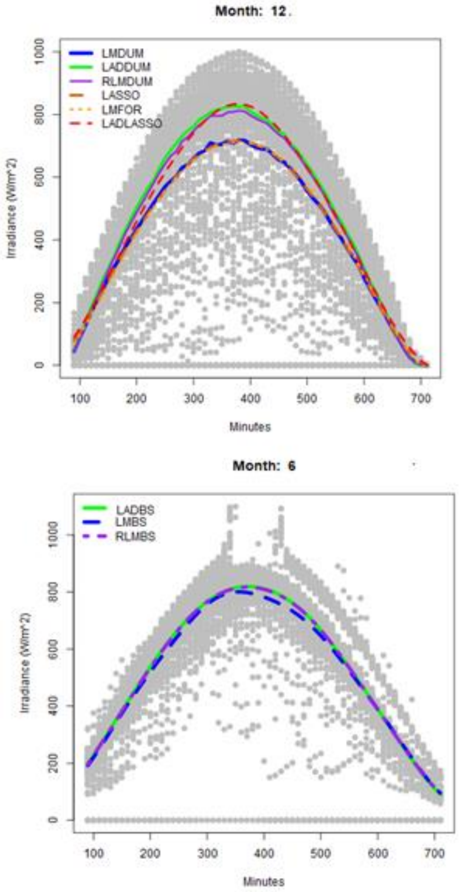

Figure 1: Irradiance SDP for December (top) based on 6 parametric procedures (three dummy-variable based and three based on periodic functions). The effect of extremely low measurements on outlier-sensitive least-squares-based daily profiles is clearly observed: ls-based SDPs lie substantially below their robust variants. SI SDP for June based on 3 non-parametric procedures (bottom); robust (LADBS, RLMBS) SDPs are very close to outlier-sensitive LMBS in this case. This example illustration uses more data relative to the ones utilized in our forecasting experiments.

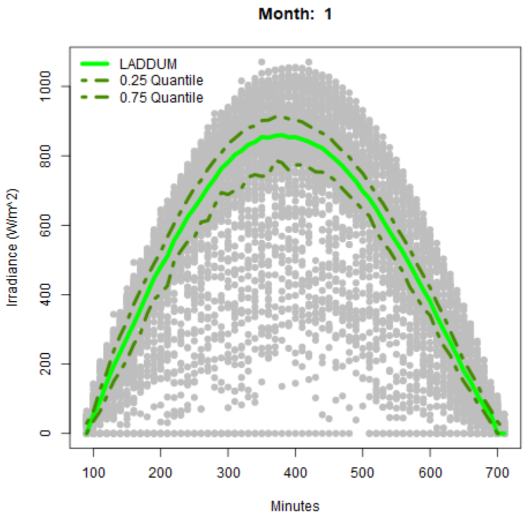

Figure 2: Irradiance SDP for January, based on median regression on dummy variables; estimated forecasting intervals correspond to 0.25 and 0.75 quantiles. This example illustration uses more data relative to the ones utilized in our forecasting experiments. Still, the curve that corresponds to the lower forecasting boundary is not smooth enough.

The forecasting experiment comprises $D = 76$ testing days (Figure 3) dispersed in the training data and mimics a realistic online implementation: sometime during each testing day $d = 1,\ldots,D$, a forecast is computed for the SI profiles of the next day. Results are summarized in Table I; a subset of observed and forecasted values is shown in Figure 4. Figure 5 depicts SDP forecasts for the last day of the experiment, coupled with quantile-regression-derived intervals that quantify uncertainty on point forecasts. In accordance with prior expectations, uncertainty increases towards mid-day.

Table I: LnQ_R performance with increasing amount of training data for the four alternative, day-ahead forecasting specifications.

# of tr. weeks	**PRS**	**TSKNN**	**SDP**	**TSKNN$_{R0}$**
1	0.1323	0.2030	0.1117	0.1351
2	0.1323	0.1527	0.1114	0.1180
3	0.1323	0.1432	0.1052	0.1104
4	0.1323	0.1413	0.1070	0.1105
5	0.1323	0.1409	0.1092	0.1118

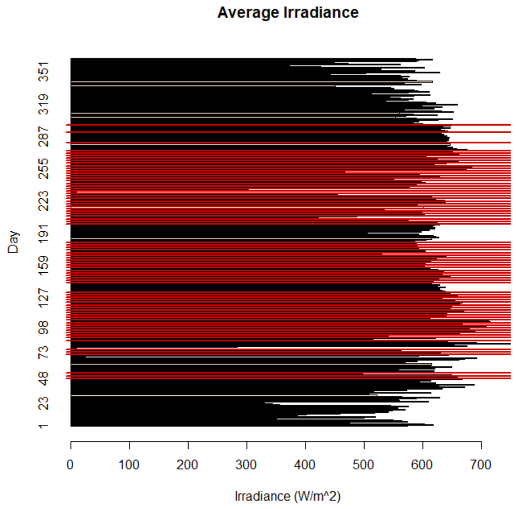

Figure 3: Average levels of daily solar irradiance across the examined period. The testing days of the forecasting experiment are highlighted in red.

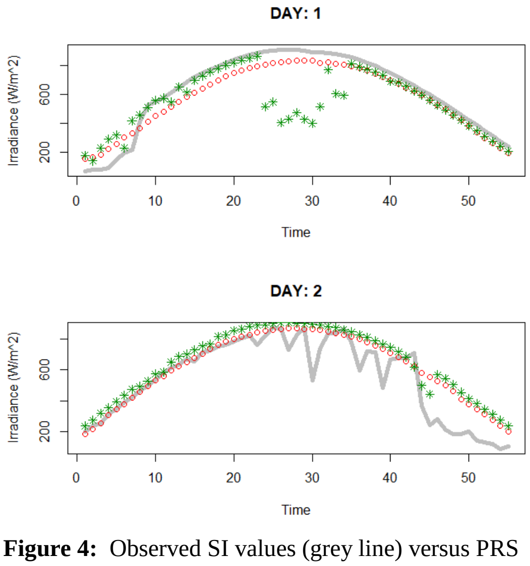

Figure 4: Observed SI values (grey line) versus PRS (green stars) and TSKKN$_{R0}$ (red circles) forecasts, for the first 2 days of the forecasting experiment.

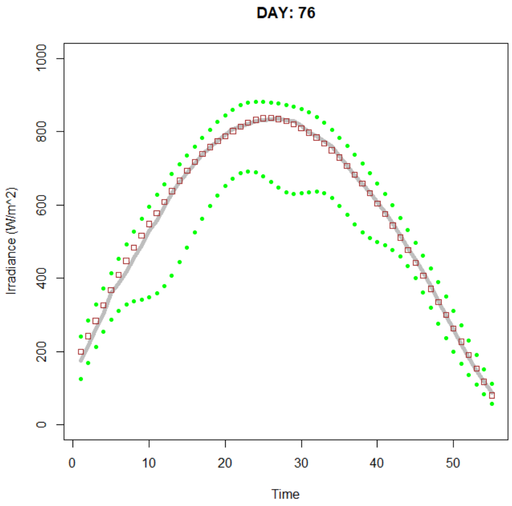

Figure 5: Observed SI values (grey line) versus SDP forecasts (brown squares) and SDP forecasting intervals (0.1 and 0.9 quantiles) shown as green points.

A striking result in Table I is the clear dominance of SDP relative to the recently proposed TSKNN scheme. SDP performs well even with a small training dataset. On the other hand, TSKNN performance improves very fast initially, when the available historical data increase from one to two weeks and slowly afterwards, as the number of training weeks increases from three to five. It is surprising that TSKNN is less accurate relative to the persistence model. On the other hand, the proposed combined scheme, TSKKN$_{R0}$ performs significantly better relative to TSKNN, with performance close to SDP.

5 CONCLUSIONS

This work evaluated an easy-to-implement recently proposed, times-series KNN, ensemble scheme for day-ahead SI forecasting. Although its performance is not satisfactory, especially versus spline-based smooth daily profiles estimated with median-regression, our investigation suggests that a novel scheme, which shrinks TSKNN towards SDP may save practitioners from extremely low levels of accuracy. The new scheme, namely TSKKN$_{R0}$ does not require additional tuning as it weighs equally TSKNN and SDP. A research question that is worth examining is related to the forecast combination puzzle: is it worth trying to optimize the forecast combination scheme?

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.

REFERENCES

[1] J. Antonanzas, et al., Solar Energy (2017) **158**, 140-146.

[2] EPIA (2012) Available online: http://pvtrin.eu/assets/media/PDF/Publications/other_publications/263.pdf.

[3] F. Martínez, et al., Artificial Intelligence Review (2019) **52**, 2019-2037.

[4] A. Nespoli et al., Energies (2019) **12**, 1621. https://doi.org/10.3390/en12091621

[5] C. Tofallis, Journal of the Operational Research Society (2015) **66**, 1352-1362.

[6] S.F. Crone at al., International Journal of Forecasting (2011) **27**, 635-660.

[7] R. Koenker, Quantile Regression, Cambridge University Press (2005).

[8] P.S.P. Cowpertwait and A.V. Metcalfe, Introductory Time Series with R, Springer (2009).

[9] S.N. Wood, Generalized Additive Models: An Introduction with R, Chapman and Hall/CRC (2006).

[10] T. Hastie et al., The elements of Statistical Learning: data mining, inference and prediction. Springer (2009).

[11] G. Claeskens et al., International Journal of Forecasting (2016) **32**, 754-762.