Observation-Based Analog Ensemble Solar Forecast in Coastal California

Elynn Wu*, Mónica Zamora Zapata*, Luca Delle Monache†, Jan Kleissl*
*Center for Renewable Resources and Integration, Department of Mechanical and Aerospace Engineering,
University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
†Scripps Institution of Oceanography, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

***Abstract*—Historical observations from radiosondes, buoys, and satellite images are used to generate an analog ensemble (AnEn) solar forecast. In coastal California, Stratocumulus (Sc) clouds appear most frequently during late spring and summer months. Sc clouds form at night and begin to dissipate after sunrise, limiting solar energy generation in the morning hours. The AnEn method categorizes cloudy (as either well-mixed or decoupled) and clear events at the forecast initial time and uses several meteorological variables to find the closest analogs. The AnEn forecast is tested at the NKX weather station in San Diego, CA during May to September 2014-2017. The AnEn forecast has a lower root mean square error than a numerical weather prediction model and 24-hour persistence forecasts. The error is lowest for the clear cases and largest for the cloudy decoupled cases. The AnEn forecast is able to capture Sc dissipation for the well-mixed cases in the early morning, but decoupled cases display higher variability throughout the day and are much harder to predict as a result.**

***Index Terms*—Analog Ensemble, Solar Forecast, Coastal Stratocumulus**

I. Introduction

Solar generation can display high variability and an accurate solar forecast enables an improved operation of the electric grid. Despite steady improvements of forecasts from Numerical Weather Prediction (NWP) models, forecast skill is still limited by mean and conditional biases. In coastal California, solar production is often hindered by the formation and dissipation of Stratocumulus (Sc) clouds, which are most prominent in late spring and summer months. The majority of installed rooftop systems are located along the coast, posing a significant challenge. Sc clouds form in a shallow planetary boundary layer (PBL) at night and begin to dissipate after sunrise. Typical PBL height is often less than 1 km. Due to computational constraints, state-of-the-art NWP models have limited vertical resolution and parameterize the physical processes governing Sc clouds instead of directly solving them. This often leads to an erroneous representation of Sc clouds and thus an inaccurate solar forecast. While efforts have been made to improve the physical representation of Sc clouds in NWP models, imperfect initial conditions and numerical and physical approximations still affect the accuracy of the forecast.

Statistical post-processing of NWP models (e.g., support vector regression, gradient-boosted regression, random forest regression) have been shown to improve the skill of solar forecasts [1]–[3]. An analog ensemble (AnEn) for short-term solar forecast using historical runs from one NWP model (Regional Atmospheric Modeling System) outperformed quantile regression and persistence ensemble [4]. In this study, we aim to build upon [4] and generate an observation-based AnEn to forecast day-ahead (sunrise to sunset) solar irradiance. The observation-based AnEn benefits from knowledge of the initial state, albeit it is often only available at a single location. Satellite data is employed to expand knowledge of the initial state.

II. Methods

A. Data

1) Satellite: Coastal low cloud identification: Geostationary Operational Environmental Satellites (GOES) provide images at both visible and infrared wavelengths. A GOES-based long-term coastal low cloudiness (CLC) dataset along the California coast [5] identifies whether a 4 km pixel is covered with low clouds every 30 minutes. Each day in the CLC dataset is classified as Sc or non-Sc day, following criteria similar to [7]. The CLC cloud coverage percentage in three coastal and one ocean regions are used to generate the analogs.

2) Satellite: Solar irradiance: Another GOES derived product, SolarAnywhere [6], provides half-hourly solar irradiance data at 1 km spatial resolution. For Sc days, each low-cloud pixel is labeled based on its averaged normalized solar irradiance (k_t) between 0700 to 1000 Pacific Standard Time (PST), where k_t is defined as the ratio of global horizontal irradiance (GHI) to clear sky GHI. PST is 8 h behind Coordinated Universal Time (UTC) and will be used for the remainder of this study. Greater San Diego ($31.5° - 34.5°$N, $116.5° - 118.5°$W) with land elevation lower than 375 m is considered. Land pixels are classified as immediate (averaged $k_t < 0.6$), intermediate ($k_t < 0.7$), and extended coastal land ($k_t < 0.8$). One ocean region ($32.0° - 33.0°$N, $117.3° - 118.3°$W) is defined to capture Sc clouds over the ocean. Classifying three coastal land regions has two main advantages: (i) capturing more details of the Sc cloud location and (ii) determining whether Sc clouds exhibit dissipating signals during the early morning, which is commonly seen in the extended coast.

3) Radiosonde: Radiosonde data at the NKX Miramar Marine Corps Air Station in San Diego, CA (NKX: 32.85°N, 117.11°W) are reported at 0400 PST, describing the vertical structure of the PBL. When a single layer of Sc clouds is present, thermodynamic properties are nearly constant in height and the PBL is said to be well-mixed. On the other hand, decoupling describes a separated layer within the PBL and allows Cumulus clouds to form below the Sc cloud layer. Radiosonde profiles of temperature and moisture are used

to classify each day's PBL state into: (i) cloudy and well-mixed, (ii) cloudy and decoupled following [8], and (iii) clear. The following variables are available from the radiosonde: inversion base height (z_i), PBL averaged liquid water potential temperature ($\theta_{l_{\text{BL}}}$), inversion strength ($\Delta\theta_l$), tropospheric total water mixing ratio ($q_{t_{3km}}$), lifting condensation level (z_b), surface temperature (T_{sfc}), surface dew point temperature (T_{dew}), precipitable water for the entire sounding (PW), and geopotential height difference between 1000 and 500 hPa (ΔZ). A detailed description of the retrieval of each variable can be found in [9].

4) Surface weather observations: Surface weather observations at NKX (32.86°N, 117.14°W) are used to retrieve zonal and meridional wind speed (u, v) at 10 m. Finally, sea surface temperature (SST) is retrieved from one buoy (Torrey Pines Outer: 32.933°N, 117.391°W). For surface observations and buoy measurements, only data corresponding to the radiosonde launch time (0400 PST) are used.

B. Analog ensemble implementation

The AnEn forecast is issued at NKX daily at 0400 PST, the time when radiosonde data are available. For cloudy well-mixed or cloudy decoupled days, 12 variables including low cloud coverage CC (over ocean CC_{ocean}, immediate CC_{im}, intermediate CC_{int}, and extended coastal land CC_{ext}), z_i (which is also the cloud top height), $\theta_{l_{\text{BL}}}$, $\Delta\theta_l$, $q_{t_{3km}}$, z_b (also known as cloud base height), u, v, and SST are used to find the closest match. For clear days, the number of variables decreases to eight, with z_i, T_{sfc}, T_{dew}, PW, ΔZ, u, v, and SST. For each scenario, the top matching days are selected, and their corresponding observed GHI timeseries are used to constitute the mean and median ensemble forecast. The top matches are determined following [10], where the metric used to rank past observations' similarity is:

$$L_{O,A} = \sum_{i=1}^{N} \frac{w_i}{\sigma_i} \left(O_i - A_i\right)^2, \tag{1}$$

where O is the current observation, A is the analog observation in the past, N is the number of physical variables, w_i and σ_i are the weight and standard deviation of each physical variable, and i denotes different variables. The rank metric $L_{O,A}$ is calculated each day at 0400 PST. Sensitivity on the number of analogs and variable weights will be discussed in the next section.

During the testing stage, $L_{O,A}$ is computed from May to September 2014-2017, excluding the test day. While this study focuses on issuing an AnEn forecast at one location, it can be extended to other sites within the domain as specified in section II-A2 as a similar PBL state is expected.

Fig. 1 and Fig. 2 illustrate two sample AnEn forecasts where five analogs (gray lines) are used to generate the mean and median ensembles (blue and green lines). Fig. 1 shows an example of a cloudy and well-mixed case, where the cloud dissipation around 0800 PST is well captured by most of the analogs. Fig. 2 shows a cloudy and decoupled case that is highly variable throughout the day and that is very different from the previous day (orange line). The AnEn forecasts predict some of this variability as seen in the spread of the individual analogs.

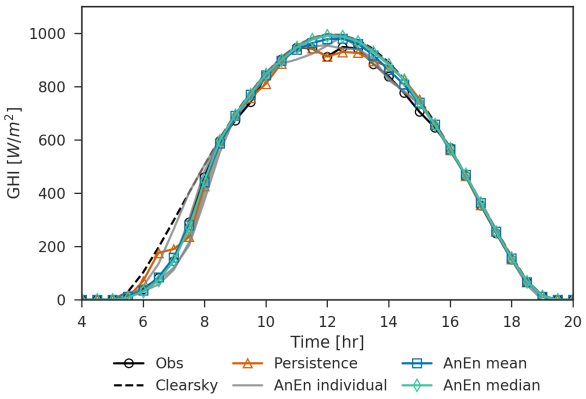

Fig. 1. Example of a cloudy and well-mixed AnEn forecast at NKX: July 21, 2017.

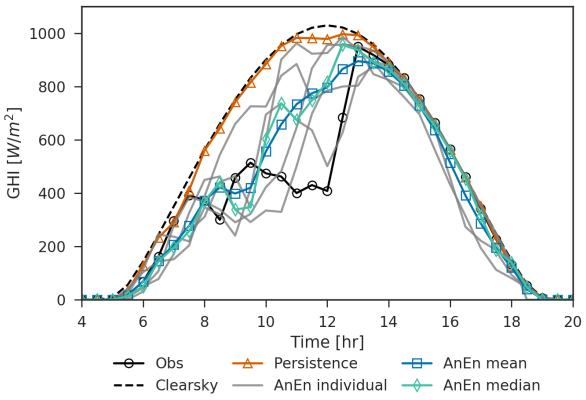

Fig. 2. Example of a cloudy and decoupled AnEn forecast at NKX: May 26, 2016.

C. Error metrics

The performance of the AnEn forecast is evaluated using the root mean square error (RMSE), which is decomposed into centered root mean square error (CRMSE) and bias (BIAS), as represented in [10], [11]:

$$\text{RMSE}^2 = \frac{1}{N_p} \sum_{i=1}^{N_p} \left(F_i - O_i\right)^2 = \text{CRMSE}^2 + \text{BIAS}^2, \tag{2}$$

where

$$\text{CRMSE}^2 = \frac{1}{N_p} \sum_{i=1}^{N_p} \left[\left(F_i - \overline{F}\right) - \left(O_i - \overline{O}\right)\right]^2, \tag{3}$$

and

$$\text{BIAS} = \overline{F} - \overline{O}. \tag{4}$$

The quantity N_p is the number of available (F_i, O_i) pairs, where F_i and O_i denote a forecast and an observation pair. $\overline{F}$

and $\overline{O}$ are the average of forecasts and observations among the N_p pairs. CRMSE is the random component of RMSE and can be interpreted as the intrinsic predictive skill of the forecast. BIAS is the systematic error.

III. Discussions and Results

A. Determining the best number of analogs

We study the sensitivity of the AnEn forecast to the number of analogs for a base setup of equal weights, and varying the number of analogs within [3, 5, 10, 15, 20, 25]. A small number of analogs is unlikely to be statistically consistent, while a large number of analogs will include irrelevant historical days and could be less accurate. Fig. 3 shows the error statistics for clear, cloudy decoupled, and cloudy well-mixed cases. The persistence forecast is defined as the observed GHI from the previous day and is thus insensitive to the number of analogs. For the AnEn mean and median, RMSE and CRMSE follow the same trend, while the trend for BIAS differs in clear and cloudy cases. In general, RMSE and CRMSE decrease with the number of analogs and plateau around 15 analogs. In some cases, the error increases slightly after 15 analogs. We therefore conclude that 15 analogs are sufficient to reach statistical consistency. In this sensitivity test, equal weights are used for the clear cases, and a best guess of weights is used for the cloudy cases where weights are doubled for z_i, $\Delta\theta_l$, and $\mathrm{CC}_{\mathrm{ocean}}$ while the other variables have equal weights.

B. Determining the best weights

In order to determine the weights that minimize the RMSE of ensemble mean and median forecasts, we explore the sensitivity of the method to different combinations of weights. In this section, we focus on cloudy decoupled and well-mixed cases. We begin with 12 variables as stated in Section II-B. To establish a baseline reference, equal weights for all 12 variables are used (Sens1 in Fig. 4). In the next 12 sensitivity tests ($\mathrm{Sens}2-13$), we double one variable weight while keeping the remaining 11 weights equal. Fig. 4 shows the RMSE of the baseline and sensitivity tests for cloudy decoupled and cloudy well-mixed days. We find that cloudy decoupled days have lower errors than the baseline (Sens1) when the weight for z_i or v is doubled ($\mathrm{Sens}2, 8$). For cloudy well-mixed days, lower errors are found when doubling the weight for z_i, $\Delta\theta_l$, $q_{t_{3km}}$, z_b, u, $\mathrm{CC}_{\mathrm{im}}$, or $\mathrm{CC}_{\mathrm{int}}$ ($\mathrm{Sens}2, 4-7, 11, 12$). The results of this test suggest that the decoupled days may not have a lot of similarity associated with the radiosonde nor satellite variables. This is expected as the well-mixed assumption does not represent decoupled days, and as a result the detection of variables such as $\Delta\theta_l$, $q_{t_{3km}}$, z_b is questionable. The exception is z_i, since it is independent of the well-mixed assumption, and strongly decoupled PBLs typically show larger z_i. For cloudy well-mixed days, both radiosonde and satellite variables play an important role in generating a good AnEn forecast. Overall, the AnEn median has a lower RMSE than the AnEn mean.

The weight sensitivity test reveals that some variables impact the results of the AnEn more than others. In order to find one set of optimal weights for all cloudy cases, we perform a sequential search on all possible weight combinations with some constraints based on the previous test (Fig. 4). We limit $w_i \geq 0.2$ for $i \in \{z_i, \Delta\theta_l, u\}$, and set increments of 0.1 for each weight, generating 715 possible weight combinations. The AnEn median is used to quantify the skill of AnEn forecast under each weight combination. Fig. 5 shows the distribution of RMSE for cloudy decoupled and cloudy well-mixed days. The AnEn for decoupled days is more sensitive to the different weight combinations as the spread is wider. Of the 715 cases, 105 (or 15%) runs have a smaller RMSE than the reference case (Sens1) for cloudy decoupled days. As for the cloudy well-mixed days, the number increases to 229 (or 32%). We select the optimal set of weights as the one that results in the lowest sum of RMSEs for cloudy decoupled and well-mixed days. The final weights are $w_{z_i} = 0.3$, $w_{\theta_{l_{BL}}} = 0.2$, $w_u = 0.2$, $w_{q_{t_{3km}}} = 0.1$, $w_{\mathrm{SST}} = 0.1$, and $w_{\mathrm{CC_{im}}} = 0.1$. The final configuration results in a reduction in RMSE of 3.0 W m^{-2} for decoupled days and 2.5 W m^{-2} for well-mixed days when compared to the reference case (Sens1).

C. Overall error statistics

We quantify the performance of the AnEn forecast with the optimal number of analogs (15) and weights. An additional NWP model– Weather Research and Forecasting (WRF)– is used as a baseline for comparison with the AnEn method. WRF is an in-house operational forecast issued for San Diego, with initial and boundary conditions derived from the 1600 PST North American Model (NAM) data. A more detailed description of the model setup can be found in [13]. Table I shows the half-hourly averaged error metrics calculated for the testing dataset during May to September from 2014-2017. Both AnEn mean and median have a lower overall RMSE and CRMSE than the persistence forecast. Errors are consistently lower for the clear cases, since clear conditions are expected to prevail for the whole day (equal weights were used for clear days). In contrast, decoupled days show the largest errors, indicating that their higher variability is more difficult to capture with both the persistence and AnEn forecast. The good performance of the AnEn forecast for the well-mixed cloudy cases indicates that the initial meteorological state is a good predictor of dissipation time. The differences between the mean and median forecast suggest that the distribution of the set of analogs is not Gaussian. AnEn forecasts have lower errors than WRF forecasts, especially for cloudy decoupled cases.

The skill of the AnEn method relies on historical similarity for local cloudy conditions. For Sc clouds, we expect similarity in meteorological conditions at 0400 PST to be a strong predictor of dissipation time. If clouds form again in the afternoon, the timing of these events might be harder to capture with 0400 PST conditions due to greater changes in synoptic conditions.

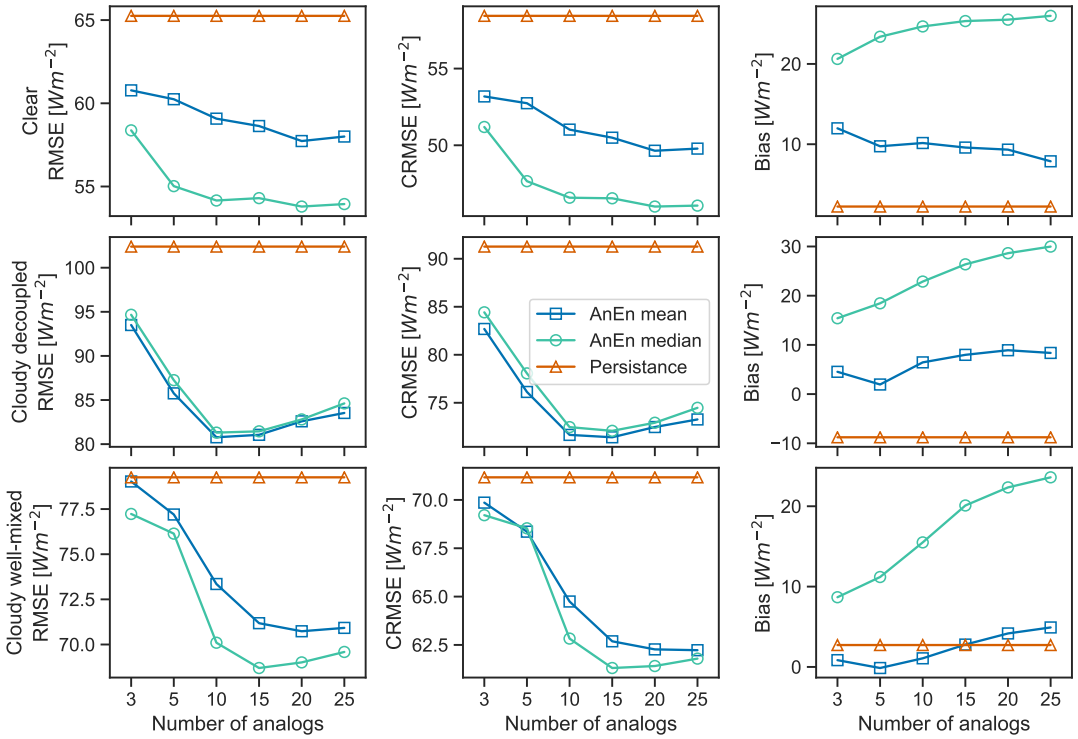

Fig. 3. Error statistics with different number of analogs under clear, cloudy decoupled, and cloudy well-mixed days. AnEn mean, median, and persistence forecast are shown.

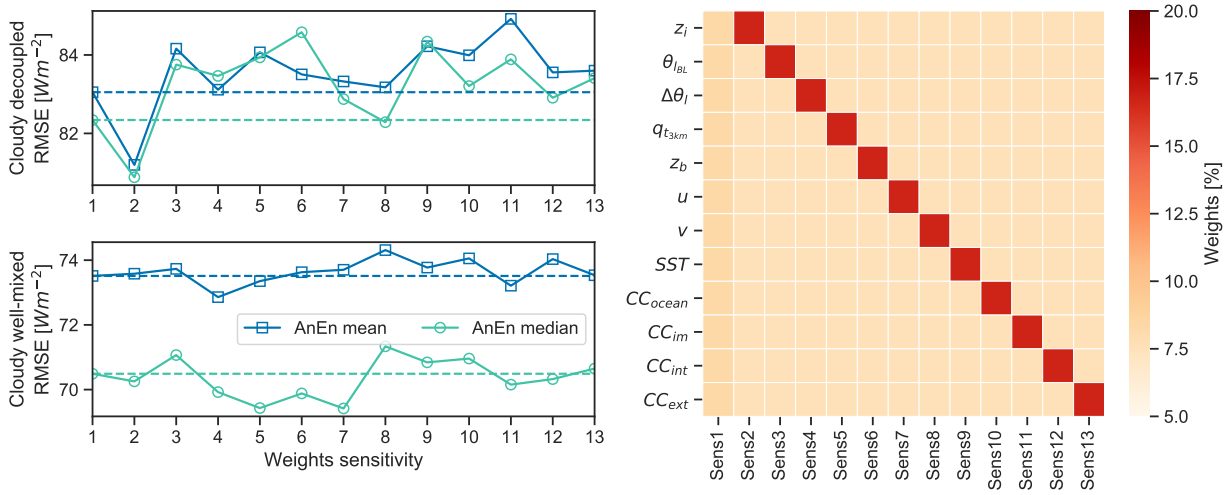

Fig. 4. Sensitivity to different weights for cloudy decoupled and cloudy well-mixed days. Left panel: RMSE of AnEn mean and median. Right panel: the corresponding weights used in each sensitivity test. The dashed line indicates the RMSE from the reference case ($Sens1$).

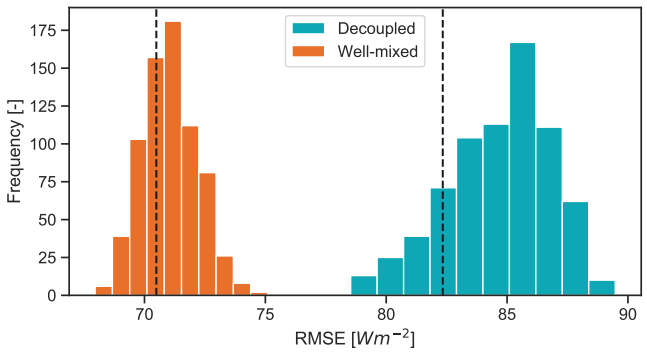

Fig. 5. Histogram of RMSE for 715 combinations of weights for cloudy decoupled and cloudy well-mixed days using AnEn median. The dashed lines are the reference case (Sens1 in Fig. 4) RMSE.

Higher GHI variability is observed for decoupled days, because Cumulus clouds can form underneath the Sc layer. There is a lack of knowledge on the dissipation of decoupled PBL clouds over coastal regions, but we expect them to dissipate later, as they usually have greater cloudiness over marine regions in comparison to Sc layers.

IV. Conclusions

We developed and tested an observation-based AnEn in San Diego, CA. The AnEn takes observations from satellite, radiosonde, and surface stations to find the closest historical days (analogs) and blends them to generate a forecast of solar radiation for the whole day. Sensitivity to the number of analogs and weights were performed in order to minimize the errors. We found 15 to be the ideal number of analogs, and that the optimal weights highlight the importance of the following variables: z_i, $\theta_{l_{BL}}$, u, $q_{t_{3km}}$, SST, and $\mathrm{CC_{im}}$ for cloudy days. Both AnEn mean and median forecasts have a lower RMSE and CRMSE compared to both persistence and WRF forecasts. The AnEn median method has the higher forecast skill of the two. The AnEn forecast is able to capture the Sc dissipation for the well-mixed cases in the early morning, but the variability in the decoupled cases is harder to capture, which leads to higher errors.

Future work should combine the AnEn forecast with WRF, to generate a hybrid statistical-dynamical method with forecast horizons that could exceed 24 h ahead. Future work should also explore the dynamics of coastal decoupled clouds in more depth to better understand which meteorological variables are of importance for decoupled Sc dissipation and its influence on solar variability, allowing for an implementation of the AnEn method based on physical insights. Finally, potential applications involve translating the AnEn irradiance forecast into an aggregated solar power production forecast at the University of California, San Diego to optimize the usage of the battery system and reduce peak demand costs.

Acknowledgment

E. Wu and M. Zamora Zapata contributed equally to this work. M. Zamora Zapata is funded by CONICYT Becas-Chile. We acknowledge (i) Rachel E.S. Clemesha for providing the Coastal low cloudiness dataset, (ii) Clean Power Research for providing high resolution satellite-derived irradiance data [6]. Radiosonde data are obtained from weather.uwyo.edu/upperair/sounding.html. Buoy data can be accessed directly from ndbc.noaa.gov. Surface weather observation data can be accessed directly from mesonet.agron.iastate.edu/request/download.phtml

TABLE I
Half-hourly averaged error statistics for three PBL states from 2014-2017: clear (122 days), cloudy and decoupled (106 days), cloudy and well-mixed (382 days).

	RMSE/CRMSE/Bias [Wm^{-2}]			
PBL state	AnEn mean	AnEn median	Persistence	WRF†
Clear	59/50/10	54/47/25	65/58/2	58/52/21
Cloudy decoupled	79/70/6	79/70/23	102/91/-9	96/86/11
Cloudy well-mixed	72/63/2	68/61/19	79/71/3	76/66/28

†Note that WRF forecast is only available in 2016 and 2017.

References

[1] A. Gensler, J. Henze, B. Sick, and N. Raabe, Deep Learning for solar power forecasting An approach using AutoEncoder and LSTM Neural Networks, in 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2016, pp. 2858-2865.

[2] Y. Gala, Á. Fernández, J. Díaz, and J. R. Dorronsoro, Hybrid machine learning forecasting of solar radiation values, Neurocomputing, vol. 176, pp. 48-59, Feb. 2016.

[3] S. E. Haupt and B. Kosovic, Big Data and Machine Learning for Applied Weather Forecasts: Forecasting Solar Power for Utility Operations, in 2015 IEEE Symposium Series on Computational Intelligence, 2015, pp. 496-501.

[4] S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone, An analog ensemble for short-term probabilistic solar power forecast, Applied Energy, vol. 157, pp. 95-110, Nov. 2015.

[5] R. E. S. Clemesha, A. Gershunov, S. F. Iacobellis, A. P. Williams, and D. R. Cayan, The northward March of summer low cloudiness along the California coast, Geophysical Research Letters, vol. 43, no. 3, pp. 1287-1295, 2016.

[6] SolarAnywhere - The Trusted Leader in Solar Data & Intelligence, SolarAnywhere. [Online]. Available: https://www.solaranywhere.com/. [Accessed: 07-Jan-2019].

[7] E. Wu, R. E. S. Clemesha, and J. Kleissl, Coastal Stratocumulus cloud edge forecasts, Solar Energy, vol. 164, no. September 2017, pp. 355-369, 2018.

[8] V. P. Ghate, M. A. Miller, B. A. Albrecht, C. W. Fairall, "Thermodynamic and radiative structure of stratocumulus-topped boundary layers," Journal of the Atmospheric Sciences, vol. 72, no. 1, pp. 430-451, 2015.

[9] M. Zamora Zapata, J. R. Norris, and J. Kleissl, Coastal stratocumulus dissipation dependence on initial and boundary conditions in a mixed layer model, (In preparation).

[10] L. Delle Monache, T. Nipen, Y. Liu, G. Roux, and R. Stull, Kalman Filter and Analog Schemes to Postprocess Numerical Weather Predictions, Mon. Wea. Rev., vol. 139, no. 11, pp. 3554-3570, Mar. 2011.

[11] K. E. Taylor, Summarizing multiple aspects of model performance in a single diagram, Journal of Geophysical Research: Atmospheres, vol. 106, no. D7, pp. 7183-7192, 2001.

[12] P. N.Blossey, C. S. Bretherton, M. Zhang, A. Cheng, S. Endo, T. Heus, Y. Liu, A. P. Lock, S. R. de Roode, and K.M. Xu (2013), "Marine low cloud sensitivity to an idealized climate change: The CGILS LES intercomparison", J. Adv. Model. Earth Syst., vol. 5, no. 2, pp. 234-258, 2013.

[13] D. K. Sahu, H. Yang, and J. Kleissl, Assimilating observations to simulate marine layer stratocumulus for solar forecasting, Solar Energy, vol. 162, pp. 454471, Mar. 2018.