Probabilistic Forecasting for Solar Energy

John Boland
Professor of Environmental Mathematics
University of South Australia

2023

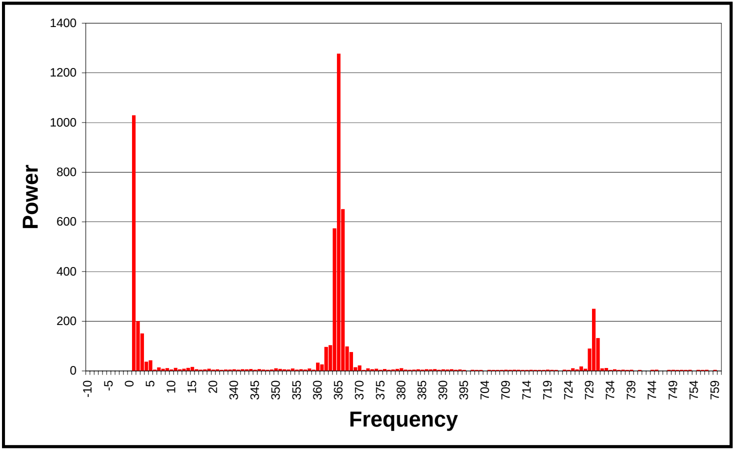
Data Sets

- I use two different sets - one is Global Horizontal Irradiation (GHI), and one is solar farm output
- GHI - Carpentras 15 minute power
- Solar Farm Output - Broken Hill 5 minute power

Point Forecast - GHI

- The first step is to model the seasonality.
- The assumption is that there are two significant frequencies, once a year and once a day.
- Spectral analysis shows we need more, one other being twice a day, probably showing asymmetry about solar noon.
- The frequencies of 364 and 366 cycles per year are significant.
- They are called beat frequencies or sidebands. They modulate the daily cycle to suit the time of year.

Power Spectrum - GHI



Fourier Series

$$
\begin{aligned}
F_t \quad = \quad & \alpha_0 + \alpha_1 \cos\frac{2\pi 365}{35040} + \beta_1 \sin\frac{2\pi 365}{35040} + \\
& \sum_{n=1}^{2}\sum_{m=-1}^{1} (\alpha_{nm} \cos\frac{2\pi(365n+m)t}{35040} + \\
& \beta_{nm} \sin\frac{2\pi(365n+m)t}{35040} + \beta_{nm}). \qquad (1)
\end{aligned}
$$

Modelling the Deseasoned Data

- The next step is to take the difference between the original data S_t and the seasonal model to form the residual series $R_t = S_t - F_t$.
- By analysing the sample autocorrelation function (SACF) and sample partial autocorrelation function (SPACF), we find that the best forecast model for the residuals is given by an autoregressive model with 5 lags $AR(5)$.
- The forecast at time $(t-1)$ for time t, $\hat{R}_t$ is given by

$$\hat{R}_t = \gamma_1 R_{t-1} + \gamma_2 R_{t-2} + \ldots + \gamma_5 R_{t-5}. \tag{2}$$

One Step Ahead Forecast

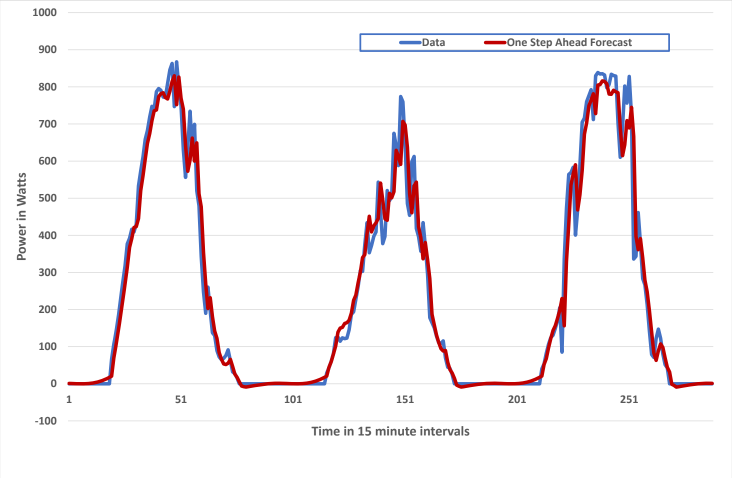


Figure : Forecasting one step ahead for Seattle

Putting Error Bounds on the Forecast

- Any one step ahead statistical forecasting method can be encapsulated by the structure

$$Y_t = f(F; Y_{t-1}, \ldots, Y_{t-p}; X_{i,t-1}, \ldots, X_{i,t-q}) + Z_t \qquad (3)$$

- This contains the seasonality F and any autoregressive qualities. Knowledge of the statistical qualities of Z_t is necessary in order to construct the error bounds of the forecast.
- It is hoped, and sometimes assumed that Z_t is independent and identically distributed (i.i.d.). But, for solar irradiation, none of these assumptions hold.
- To estimate the error bounds, or the limits of the prediction intervals, we use quantile regression. We perform that on the Z_t, but only on data for which the solar elevation is greater than 10^o.

Methods for constructing prediction intervals

- Quantile regression. Take the noise and forecast various quantiles rather than the mean, and then add the quantiles of the noise to the point forecast.
- Autoregressive Conditional Heteroscedastic forecast of the variance, and then form the prediction intervals from that. One complication is that it assumes normal noise. So, transform the noise to normal, do the ARCH model, form the intervals, and back transform.
- The benchmark for comparison is assuming the variance does not change and the noise is normal, and then form the intervals using these naive assumptions.

Evaluation

- There are many measures to evaluate the performance of probabilistic forecasting.
- Philippe Lauret, Mathieu David, Pierre Pinson, Verification of solar irradiance probabilistic forecasts, Solar Energy, Volume 194, 2019, Pages 254-271, ISSN 0038-092X, $https://doi.org/10.1016/j.solener.2019.10.041$.
- In this study we will use two methods, coverage and interval width.

Metrics

- Coverage. If one is designing a 95% prediction interval, approximately 95% of the observations should fall within the interval.
- Width - as well as coverage, a smaller mean width of intervals is sharper and better.
- We compare our method with a benchmark. In this case, the naive model we use is that the errors are independent and identically normally distributed.

Coverage

	Naive	Quantile	ARCH
99	96.3	98.9	98.8
95	93.7	94.9	96.0
90	91.8	90.0	92.4
80	88.4	79.9	83.3

Table : The coverage for the methods with the prescribed probabilities

Width

	Naive	Quantile	ARCH
99	324.9	482.2	345.0
95	250.0	276.0	236.4
90	211.0	183.9	176.4
80	165.7	96.2	111.7

Table : The interval width for the methods with the prescribed probabilities

Results for 80% prediction intervals

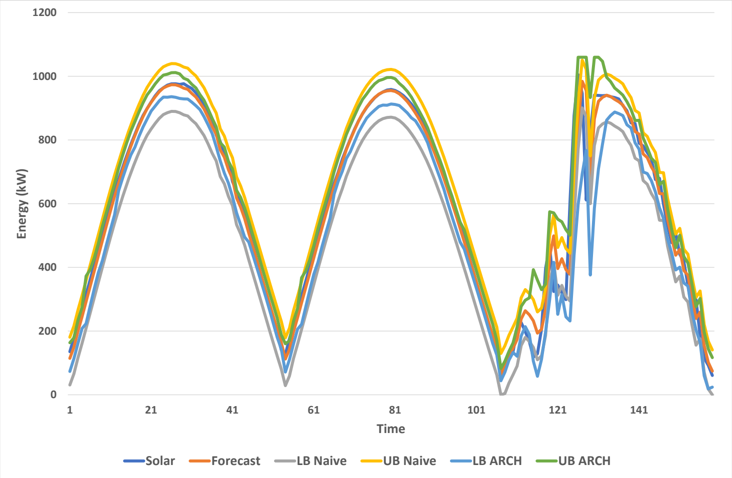


Figure : Comparing prediction intervals using an ARCH forecast, with the naive approach

Broken Hill Solar Farm Output

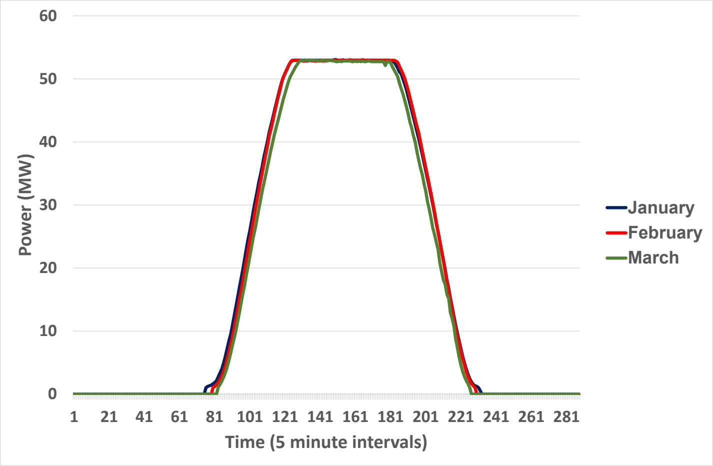


Figure : Capped Solar Farm Output - Summer

Broken Hill Solar Farm Output

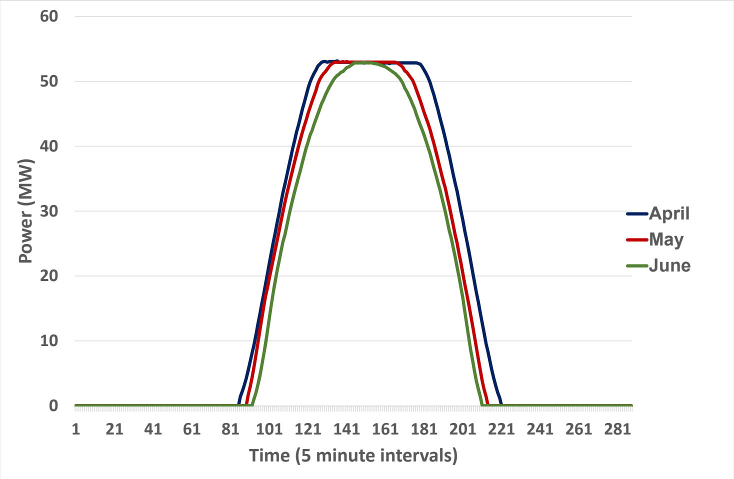
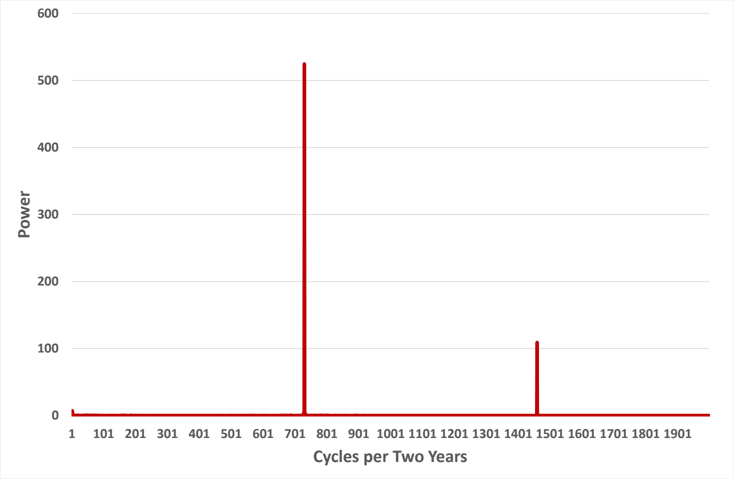


Figure : Capped Solar Farm Output - Winter

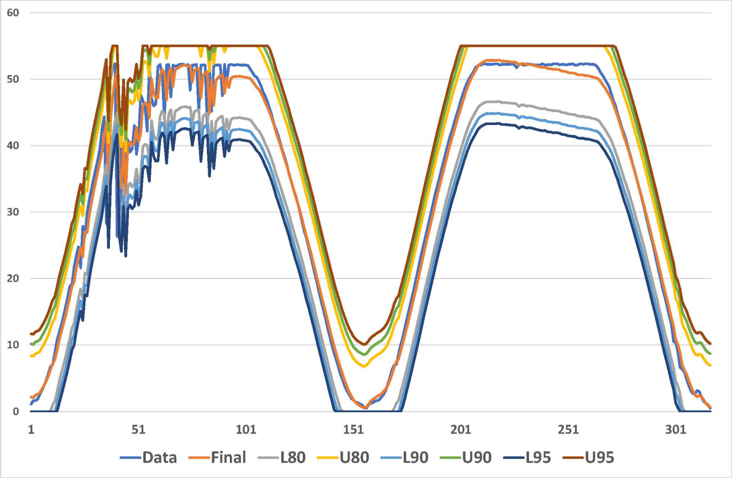
Power Spectrum - Solar Farm



Model

- From the power spectrum, we only need the Fourier series to contain the daily and twice daily frequencies.
- The residuals are modelled with an $ARMA(2,1)$ process.
- The combination of the two give the final one step ahead forecast.
- We then use quantile regression to obtain the prediction intervals.

Broken Hill Solar Farm Output with Prediction Intervals



Results

	Percentage	Present	Benchmark
Coverage			
	80	83.6	91.0
	90	92.2	93.3
	95	96.2	94.9
Width			
	80	6.10	11.24
	90	12.54	14.01
	95	18.91	16.37

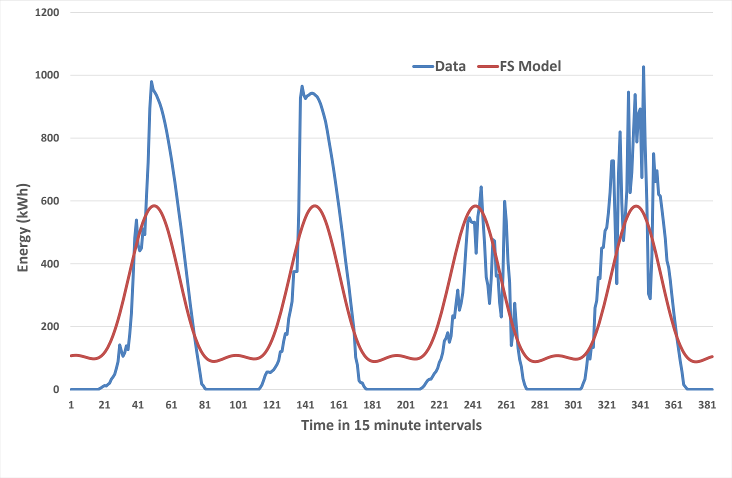
Conclusion

- We have shown the difference in point forecasting for solar energy versus solar farms.
- We used quantile regression to form prediction intervals for one step ahead forecasting.
- The results show that this approach behaves much better than the benchmark except for the 95% prediction interval.
- Presently we are also using a more sophisticated approach by transforming the noise, then using an ARCH model to forecast the variance, setting the prediction interval bounds and back transforming them. This appears to add further improvement.

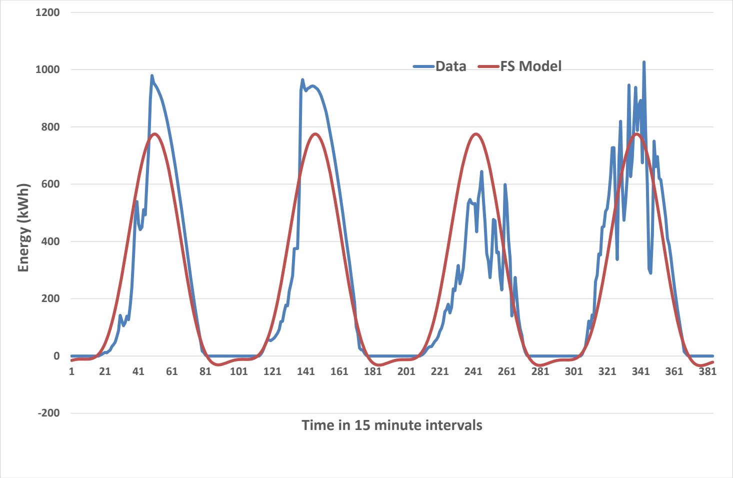
TIES 2024

We are excited that TIES 2024 will be hosted by John Boland and his team at University of South Australia in Adelaide, Australia from 2nd to 5th Dec 2024. This is the annual conference of The International Environmetrics Society

Without Beat Frequencies



Including Beat Frequencies



Quantile Regression

For quantile level τ of the response, the goal is to

$$\min_{\beta_0(\tau),\beta_1(\tau),\ldots,\beta_p(\tau)} \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0(\tau) - \sum_{j=1}^{p} z_{ij}\beta_j(\tau))^2 \tag{4}$$

$$\rho = \tau \max(r, 0) + (1-\tau)\max(-r, 0) \tag{5}$$

is the check function. If the error in the regression in a single period, r, is positive, then the check function multiplies the error by τ and by $1-\tau$ if negative. In the study performed on Seattle data, the predictor variables are the previous 5 lagged values of the noise.