Scalable Gaussian Process Models for Solar Power Forecasting

Astrid Dahl(✉) and Edwin Bonilla

University of New South Wales, Sydney, Australia
astrid.dahl@student.unsw.edu.au

Abstract. Distributed residential solar power forecasting is motivated by multiple applications including local grid and storage management. Forecasting challenges in this area include data nonstationarity, incomplete site information, and noisy or sparse site history. Gaussian process models provide a flexible, nonparametric approach that allows probabilistic forecasting. We develop fully scalable multi-site forecast models using recent advances in approximate Gaussian process methods to (probabilistically) forecast power at 37 residential sites in Adelaide (South Australia) using only historical power data. Our approach captures diurnal cycles in an integrated model without requiring prior data detrending. Further, multi-site methods show some advantage over single-site methods in variable weather conditions.

1 Introduction

Solar power forecasting is motivated by several areas of application, including grid management, load shifting (demand management) and energy storage management. As small scale residential solar penetration grows, challenges to forecasting power for multiple distributed small scale sites, in particular forecasting with incomplete site information and noisy power data, become of interest.

Challenges in this context include nonstationarity in the data,[1] and developing useful probabilistic forecasts. Many forecasting methods assume it is possible to detrend power data prior to stochastic modelling in order to 'flatten' the data and remove diurnal cyclical trends associated with cycles in solar radiation. However, methods to do so rely on comprehensive site information [7], or site history as in [5,17]. Overall, existing methods have high data demands, constraining their usefulness for new or unseen sites.

For certain applications it is desirable to work with a probabilistic distribution of forecasts that quantifies forecast uncertainty. Statistically-based methods, such as vector autoregressive (VAR) models, typically allow probabilistic forecasts however are constrained in their application to unflattened data and sites for which no training data is available. Machine learning methods such as

[1] Stationarity here refers to the property that distribution parameters remain stable (and finite) over time.

W.L. Woon et al. (Eds.): DARE 2017, LNAI 10691, pp. 94–106, 2017.
https://doi.org/10.1007/978-3-319-71643-5_9

neural networks (ANNs) are more widely applicable but do not generally allow probabilistic forecasts. Gaussian process models are advantageous in this regard, providing a flexible, nonparametric forecast approach that is also probabilistic in nature.

Transfer learning over distributed sites may assist in addressing site data limitations as well as improve prediction of weather-related power fluctuations. The literature suggests cross site data can be helpful in modelling cloud conditions to improve site level forecasts, as in [3,12,23], with evidence that cross site information in a dense network can be relevant from timescales of a few minutes, as in [27], to multi-hour horizons in a widely distributed network, as in [3].

A key constraint often associated with multisite approaches is scalability to large numbers of sites. Within the Gaussian process literature, several approximate methods have been developed that support stochastic parameter optimisation, thus maintaining scalability to large datasets and feasibility for real world application.

The current study considers the problem of short term (less than 30 min) power forecasting for large distributed networks of residential rooftop solar systems. We apply sparse variational Gaussian process (GP) approaches for probabilistic forecasting across multiple solar sites in Adelaide, Australia. Our aim is to test whether scalable GP methods can be applied to short term distributed forecasting to provide useful, probabilistic forecasts at the site level with limited site history and information.

1.1 Related Work

The literature around solar forecasting is extensive, including studies that investigate both solar irradiance and power forecasting over multiple forecast horizons (a few minutes to multiple days) using approaches that range from physics-based models to statistical and machine learning methods. Studies to date also examine multiple inputs including irradiance or power measurements, ground and satellite based weather data and meteorological forecasts. Several reviews [10,14,18,25] provide a thorough coverage of recent methods.

In the sphere of short term forecasting, stochastic models utilising only historical power data have been shown to perform relatively well in the past [18], although recent advances suggest highly accurate forecasts can be produced by including comprehensive climate data [16]. Predominant statistical methods include adaptively estimated VAR, autoregressive integrated moving average (ARIMA) and generalised autoregressive conditional heteroskedasticity (GARCH) models, for example as in [3,9]. These methods allow for weather-related nonstationarity and at shorter horizons (up to one hour) have been found to be competitive in forecasting clearness indices (i.e. flattened irradiance data) [9,15]. In a number of cases models are applied in a multisite setting as in [3,5,11,27].

The major machine learning methods explored for short term solar forecasting are neural networks and support vector machines (SVMs). Recent examples of ANNs for short term horizons include [12,20]. ANNs have been explored in a multisite setting in irradiance forecasting [25], although at time of writing no examples were identified of multivariate prediction at horizons less than one hour ahead.

Gaussian process and related models have been explored to a limited extent in solar forecasting. [9,15] include univariate GP models applied to clearness indices as comparative models. [4] also uses a GP model to forecast clearness index values over an irradiance field in 30 min increments. The authors apply probabilistic principal components dimension reduction to improve feasibility of real time adaptive GP modelling over multiple locations (by reducing the number of 'locations' for which a GP is estimated), and further assume independence between models, thus respecifying the multivariate problem as several univariate problems. However, even with these adaptations, scalability is still constrained by non-stochastic optimisation of the exact GP as described in Sect. 2.

Several studies use a closely related method, kriging, to predict clearness indices in a multisite setting [2,22,23,26]. Building on [22,26] develops one-hour-ahead clearness index forecasts using one month of hourly data from a group of 10 meteorological stations in Singapore. In [23], the authors 'nowcast' clearness index values for 25 sensor locations covering an approximately 30 km radius area in Osaka.

2 Theory

Gaussian process (GP) models provide a flexible nonparametric Bayesian approach to machine learning problems such as regression and classification [21] and have proved successful in various application areas involving spatio-temporal modeling [8]. Formally, a GP is a prior over functions for which every subset of function values $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)$ follows a Gaussian distribution. We denote a function drawn from a GP with mean function $\mu(\mathbf{x})$ and covariance function $\kappa(\cdot, \cdot)$ by $f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$.

One of the most widely used GP models is the standard regression setting with a zero-mean GP and i.i.d. Gaussian noise:

$$y_t \sim \mathcal{N}(f(\mathbf{x}_t), \sigma_y^2) \text{ with } f(\mathbf{x}_t) \sim \mathcal{GP}(\mathbf{0}, \kappa(\mathbf{x}_t, \mathbf{x}_{t'})), \tag{1}$$

where $\mathbf{x}_t$ denote features at time t and σ_y^2 is the noise variance.

Given a set of observations $\{(\mathbf{x}_t, y_t)\}_{t=1}^N$, we wish to learn a model in order to make predictions at a new datapoint $\mathbf{x}_*$. Given the likelihood and prior models in Eq. (1), the predictive distribution over $f(\mathbf{x}_*)$ is a Gaussian with mean and variance given by:

$$\mu_* = \kappa(\mathbf{x}_*, \mathbf{X})(\mathbf{K} + \sigma_y^2\mathbf{I})^{-1}\mathbf{y}, \quad \sigma_* = \kappa(\mathbf{x}_*, \mathbf{x}_*) - \kappa(\mathbf{x}_*, \mathbf{X})(\mathbf{K} + \sigma_y^2\mathbf{I})^{-1}\kappa(\mathbf{X}, \mathbf{x}_*),$$

where $\mathbf{X}$ and $\mathbf{y}$ denote all the training features and outputs, respectively; $\mathbf{K}$ is the covariance matrix induced by evaluating the covariance function at all training datapoints; and $\mathbf{I}$ is the identity matrix.

Although computing the exact predictive distribution above is appealing from the theoretical perspective and in a small-data regime, these computations become unfeasible for large datasets as their time and space complexity are $\mathcal{O}(N^3)$ and N^2 respectively.

Much of the research efforts in GP models have been devoted to this issue [19] with significant breakthroughs achieved over the last few years [13,24]. Indeed, here we study the *variational* approach to inference in GP models, which relies upon reformulating the prior via the so-called *inducing* variables [24].

2.1 Scalable Gaussian Process Regression via Variational Inference

Full details of the variational approach to scalable GP regression is out of the scope of this paper and we refer the reader to [6,13,24] for further reference. Here it suffices to explain that we introduce a set of M inducing variables $\mathbf{u} = (u_1, \ldots, u_M)$, which lie in the same space as the original function values and are drawn from the same GP prior. For these inducing variables we have their corresponding inputs $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_M)$, where each $\mathbf{z}_j$ is a D-dimensional vector in the same space as the original features $\mathbf{x}$.

The variational approach to GP inference involves a reformulation of the prior via the inducing variables and the proposal of an approximate posterior over these using $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$, which is estimated via the optimization of the so-called evidence lower bound (ELBO):

$$\mathcal{L}_{\text{elbo}}(\mathbf{m}, \mathbf{S}) = \text{KL}(q(\mathbf{u}) \| p(\mathbf{u})) - \mathbb{E}_{q(\mathbf{f})}[\log p(\mathbf{y}|\mathbf{f})], \tag{2}$$

where $\text{KL}(q\|p)$ denotes the Kullback-Leibler divergence between distributions q and p; $p(\mathbf{u}) = \mathcal{N}(\mathbf{0}, \kappa(\mathbf{Z}, \mathbf{Z}))$ is the Gaussian prior over the inducing variables; $\mathbb{E}_{q(\mathbf{f})}[\log p(\mathbf{y}|\mathbf{f})]$ is the expectation of the conditional likelihood (given in Eq. (1)) over $q(\mathbf{f}) = \int_{\mathbf{u}} q(\mathbf{u}) q(\mathbf{f}|\mathbf{u}) d\mathbf{u}$; and $q(\mathbf{u})$ the approximate posterior given above. Using simple properties of the Gaussian distribution it is possible to show that Eq. (2) can be solved analytically and, more importantly, $\mathcal{L}_{\text{elbo}}$ decomposes as a sum of objectives over the training data. This readily allows the application of stochastic optimization methods rendering the time and space complexity of the algorithm as $\mathcal{O}(M^3)$ and $\mathcal{O}(M^2)$, respectively, hence independent of N and applicable to very large datasets.

2.2 Gaussian Processes for Solar Power Forecasting

A key advantage of GP models is their flexibility to express potentially nonlinear relationships and nonstationary processes through various kernel forms. GP models have the capacity to account for nonstationarity associated with diurnal cycles through appropriate kernel functions. Further, their nonparametric nature allows models to flexibly reflect variable volatility i.e. nonstationarity associated with weather effects.

In the present study, we propose several Gaussian process model specifications for application to the residential solar forecasting problem where site information is unknown. Kernels are structured to capture both cyclical and autoregressive processes in the power data. We compare results under both 'site-independent' approaches, where Gaussian process models are applied to sites individually, and multi-site approaches, where forecasting for multiple sites is performed via collaborative Gaussian process models.

Site-Independent Models. Consider the timeseries of power observations for a single site p, denoted y_p, at times $t = 0, \dots, N$. As in (1), let

$$y_{pt} = f_p(\mathbf{x}_{pt}) + \epsilon_{pt}, \quad f_p(\mathbf{x}) \sim \mathcal{GP}(0, \kappa_p(\mathbf{x}_{pt}, \mathbf{x}_{ps})) \tag{3}$$

$\epsilon_{pt} \sim iid\mathcal{N}(0, \sigma^2_{y_p})$. Under the GP specification, observed power y_{pt} is a function of a latent Gaussian process, $f_p(\mathbf{x}_{pt})$, plus idiosyncratic noise ϵ_{pt}. The covariance between power at time t and time s, $s \neq t$ is thus given by the kernel function $\kappa_p(\mathbf{x}_{pt}, \mathbf{x}_{ps})$. The likelihood function is given by $y_{pt}|f_{pt} \sim \mathcal{N}(f_{pt}, \sigma^2_{y_p})$.

In the site-independent setting, the feature vector $\mathbf{x}_{pt}$ is comprised of two main elements: a time index t and a set of lagged power observations at prespecified five minute intervals denoted $\mathbf{g}$. In order to forecast power at $t + \delta$ for δ steps ahead, lag features are current observed power and past observed power at 5 and 10 min prior. Thus $\mathbf{x}_{pt} = (t, \mathbf{g}_{pt})$, $\mathbf{g}_{pt} = (y_{pt-\delta}, y_{pt-\delta-1}, y_{pt-\delta-2})$. Lags were selected in line with previous studies that find immediate lags are relevant for short term forecasting (see e.g. [26]).

Additional, 'extended site-independent' models are estimated using an augmented set of lag features. The feature vector is extended to include power observations of nearby sites, that is $\mathbf{g}_{pt} = (y_{pt-\delta}, y_{pt-\delta-1}, y_{pt-\delta-2}, y_{-pt-\delta}, y_{-pt-\delta-1}, y_{-pt-\delta-2})$, where y_{-p} denotes all sites near to site p. Utilising cross-site features in the form of lags allows separate site model estimation and has been applied in several studies including [3,12]. We define 'near' as being within a 10 km radius.[2]

Kernel functions for site-independent and extended site-independent models are comprised of several separable kernel elements. A periodic kernel is applied to the time index to capture daily cyclical trends in output and is defined as

$$\kappa_{Per.}(t, s) = \theta \exp\left[-0.5\left(\frac{\sin\left(\frac{\pi}{T}(t-s)\right)}{l}\right)^2\right] \tag{4}$$

where θ governs cycle amplitude, T denotes cycle period (fixed at one day), and lengthscale, l, governs rate of decay in covariance as the time-span between observations increases.

A linear kernel is applied to lag features $g_i \in g$ to capture short term variations from the regular diurnal trend:

$$\kappa_{Lin.}(\mathbf{g}_{pt}, \mathbf{g}_{ps}) = \sum_i \sigma_i g_{pti}, g_{psi} \tag{5}$$

[2] A fixed radius is applied to provide local regularisation, which has been found to reduce overfitting in multisite settings [11,27]. The 10 km threshold aims to limit 'neighbours' to sites most likely to be relevant given historic local windspeed.

where σ_i are in effect weight coefficients. The overall kernel structure for all site-independent models is:

$$\kappa_p(\mathbf{x}_{pt}, \mathbf{x}_{ps}) = \kappa_{Per.}(t, s)\kappa_{Lin.}(\mathbf{g}_{pt}, \mathbf{g}_{ps}). \tag{6}$$

Multi-site Models. Values for proximate sites would be expected to covary, due to both synchronous diurnal cycles in unflattened data and shared weather systems. Some efficiency would thus be expected from exploiting the shared covariance structure through collaborative learning.

Two separate multi-site GP model structures are estimated for site-level power forecasting. The first is a pooled structure, where (standardised) site data are used in a joint specification with shared kernel parameter values. The second structure is the linear coregionalisation model or LCM. This structure assumes site observations covary through a lower dimension set of shared latent processes.

For each multi-site model structure, two alternative kernel specifications are explored. These four model specifications are detailed below.

Pooled Model. The pooled, or 'joint', model is a pooled Gaussian process model where all site observations share a common kernel that includes an additional kernel element defining a spatial covariance factor.

The first pooled model kernel ('Joint Model 1') is defined as a multiplicative, separable spatiotemporal kernel added to a shared linear kernel applied to lagged power values. Feature vector $\mathbf{x}$ is extended to include $\mathbf{h} = (latitude, longitude)$ i.e. $\mathbf{x}_{pt} = (t, \mathbf{g}_{pt}, \mathbf{h}_p)$. A radial basis function (RBF) kernel is applied to $h_i \in \mathbf{h}$ to capture spatial dependencies, thus for sites p and q,

$$\kappa(\mathbf{x}_{pt}, \mathbf{x}_{qs}) = \kappa_{Per.}(t, s)\kappa_{RBF}(\mathbf{h}_p, \mathbf{h}_q) + \kappa_{Lin.}(\mathbf{g}_{pt}, \mathbf{g}_{qs}). \tag{7}$$

where

$$\kappa_{RBF}(\mathbf{h}_p, \mathbf{h}_q) = \sigma^2 \exp\Big\{ -\frac{1}{2}\sum_{i=1}^{2}((h_{pi} - h_{qi})/l_i)^2 \Big\}. \tag{8}$$

In the RBF kernel, σ^2 governs maximum covariance between points h_p and h_q, and lengthscale, l_i, governs the rate of decay in covariance as distance between observations along the relevant axis increases.

The second joint model ('Joint Model 2') is similarly specified however replaces the shared linear kernel with separately parameterised linear kernels for each site. Specifically, $\kappa_{Lin.}(\mathbf{g}_{pt}, \mathbf{g}_{qs})$ becomes

$$\kappa_{Lin.,p}(\mathbf{g}_{pt}, \mathbf{g}_{qs}) = \sum_i \sigma_{pi} g_{pti}, g_{qsi}, \quad \kappa_{Lin.,p} = 0 \quad \text{for} \quad p \neq q \tag{9}$$

Coregional Model. The linear coregional model (LCM) assumes y_p is a function not of a single latent process $f_p(\mathbf{x_p})$ but a linear combination of several independent latent Gaussian processes. Covariance between sites arises from these shared latent processes. Weights defining the linear combination for a given site are site-specific,[3] $f_p(\mathbf{x}) = \sum_{j=1}^{Q} w_{pj} u_j(\mathbf{x})$.

We assume three latent processes $u(x)_j, j = 1, ..., 3$ in the first LCM model ('LCM Model 1') and two latent processes in the second model ('LCM Model 2'). Each latent process has an associated kernel, κ_j, giving rise to a shared covariance structure across sites driven by both kernel elements and weight matrices.

Let $\mathbf{B}_j = \mathbf{W}_j \mathbf{W}_j' + \kappa_j$ where $\mathbf{W}_j$ is a $p \times 1$ matrix of weights w_{pj}, and κ_j is a diagonal matrix of isotropic noise. We define $\kappa_1 = \kappa_{Per.}(t, s)$ and $\kappa_2 = \kappa_{RBF}(t, s)$ respectively as periodic and RBF kernels applied to time indices. The third latent process kernel is defined as $\kappa_3 = \kappa_{Lin.}(\mathbf{g}_t, \mathbf{g}_s)$. The shared kernel structure in LCM Model 1 is thus given by:

$$\mathbf{K}(f_p(\mathbf{x}_{pt}), f_q(\mathbf{x}_{qs})) = \sum_{j=1}^{3} [\mathbf{B}_j]_{pq} \, \kappa_j(\mathbf{x}_{pt}, \mathbf{x}_{qs}). \tag{10}$$

The second coregional model is similar to the above, however again the linear kernel component is treated slightly differently. In LCM Model 2, $Q = 2$ with κ_1 and κ_2 defined as above, and lag features are included in a separate kernel component defined as in (9). Thus

$$\mathbf{K}(f_p(\mathbf{x}_{pt}), f_q(\mathbf{x}_{qs})) = \sum_{j=1}^{2} [\mathbf{B}_j]_{pq} \, \kappa_j(\mathbf{x}_{pt}, \mathbf{x}_{qs}) + \kappa_{Lin.,p}(\mathbf{g}_{pt}, \mathbf{g}_{qs}). \tag{11}$$

The specification in (11) allows a slightly more expressive parameterisation of the linear kernel than (10).

Benchmark Models. Without clear sky normalisation and under the assumption of short site history, there are few existing models for comparison. One feasible benchmark prevalent in the literature is the persistence model,[4] which forecasts the next observation as the current observation i.e. $y_{t+1} = y_t$. Persistence models are estimated for each site separately.

In addition, the site-independent Gaussian process models serve as a benchmark. These are approximately equivalent to linear Bayesian regression models assuming a standard Gaussian prior distribution over regression coefficients. As such they are closely related to VAR models as applied to flattened data.

3 Experiments

The analysis makes use of a sample of 37 residential photovoltaic systems installed within an approximately 10 by 15 km 'box' in the central Adelaide area.

[3] A useful exposition of coregional models can be found at [1].

[4] The persistence model in the present study is applied to unflattened data.

Most sites have an installed capacity of 2 to 5 kW. The dataset is comprised of 5-minute average power readings over a 30 day period in January 2017 (specifically 30 days ending 28 January 2017). Days were defined as 7 am to 7 pm, yielding a total of 144 observations for a site over a day. This accounts for a total of 159,840 observations, which is clearly unfeasible for standard (non-scalable) GP models.

The goal of the experiments is to test whether GP models estimated under a sparse variational framework can be applied to forecast distributed power output at the site level for multiple distributed sites. In particular, whether (a) combined kernel forms can be used to model nonstationary data characteristics, and (b) collaborative learning can improve forecast accuracy or reduce data requirements compared to independent site forecasts.

The four multisite models set out above are used to forecast output for each site. These are compared to results under the site-independent and persistence models. Models are trained for forecasting horizons from five to thirty minutes at five minute intervals. The forecast target in each case is five minute average power at that horizon. Models were trained using the first 60% of observations (18 days). Forecasts were then generated for a test set of the following 40% of observations (12 days) for each site.

All models are estimated via the sparse, variational approach described in Sect. 2. Inducing points are initialised at cluster centroids and optimised within the model. To illustrate the scalability of the approach, we use 2300 inducing points for joint models, or approximately 2.4% of the data dimension. Maintaining the same ratio, 63 inducing points per site were used for individual models.

3.1 Accuracy Metrics

Forecast accuracy is assessed for each site for each model using three measures: mean absolute error (MAE) in kilowatts, standardised mean squared error (SMSE) and standardised mean log loss (SMLL), as defined in [21]. Specifically,

$$SMSE = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} \left(\frac{y - \hat{y}}{\sigma_{y_{te}}} \right)^2 \tag{12}$$

$$MAE = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} |y - \hat{y}| \tag{13}$$

$$SMLL = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} (nlpd_i - nll_i), \quad \text{where} \tag{14}$$

$$nlpd_i = \frac{1}{2} \left[ln(2\pi) + ln\sigma^2_{\hat{y}_i} + \left(\frac{y_i - \hat{y}_i}{\sigma_{\hat{y}_i}} \right)^2 \right],$$

$$nll_i = \frac{1}{2} \left[ln(2\pi) + ln\sigma^2_{y_{tr}} + \left(\frac{y_i - \mu_{y_{tr}}}{\sigma_{y_{tr}}} \right)^2 \right].$$

Subscripts *te* and *tr* refer to training and test sets respectively and $\hat{y}$ denotes the predicted value of y. SMSE is standardised by reference to test set variance $\sigma^2_{y_{te}}$. Values less than one indicate the model improves on a simple mean forecast. SMLL measures the (negative log) likelihood of the test data under the model, denoted *nlpd*, relative to (negative log) likelihood under the trivial normal distribution with parameters ($\overline{y}_{tr}$, $\sigma^2_{y_{tr}}$), denoted *nll*. More negative metric values indicate better relative performance of the model.[5]

3.2 Results

Forecast Accuracy. Results at the site level suggest the site-independent model performs as well as or better than the joint (pooled) model in terms of average site accuracy (Fig. 1). SMSE for both the site independent and joint models ranges from 0.05–0.12 over 5–30 forecast horizons, however MAE and SMLL are consistently improved under the site-independent model over all forecast horizons e.g. MAE of 0.14–0.26 kW versus 0.17–0.29 kW under site and joint models respectively. The LCM specifications perform poorly on all measures relative to the joint and basic site-independent models. At all forecast horizons, the better performing models (joint and site-independent) are more accurate than the persistence benchmark.

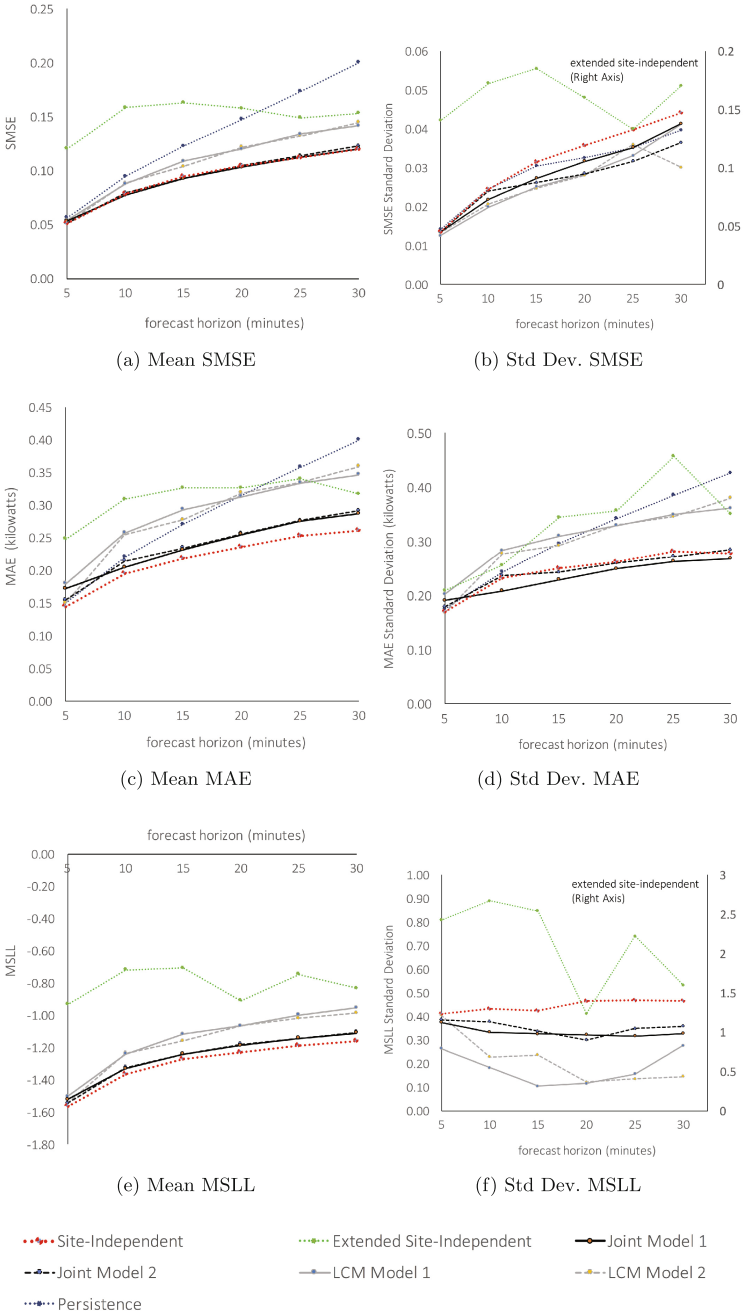
Additional expressiveness in the kernel due to the more flexible linear lag kernel structure does not significantly improve forecast accuracy in the joint or LCM models, and in some cases tends to contribute to higher forecast variability across sites (Fig. 1). Interestingly, the extended site-independent model performs very poorly relative to other models, however forecast error remains fairly stable over 10–30 min horizons. This result may indicate overspecification of this (very flexible) kernel structure.

Estimation of Daily Power Curve. It is difficult to evaluate the current approach as an alternative to those that require flattening the data without a direct (flattened) benchmark for the given dataset. However, examining forecast accuracy on clear[6] days provides some insight into how the approach accounts for clear sky curves. Table 1 summarises forecast accuracy under the site-independent and joint model 1 specifications for clear (or mostly clear) and cloudy days in the test set, which each represent 50% of the test data.

Forecast accuracy appears competitive on clear days, with mean MAE across sites of 50 Watts on clear days at the five minute horizon, rising to 130 Watts at the 30 min horizon. Given mean power for the full dataset set across clear and cloudy days of 2.1 kW, MAE represents around 2.4 (6.2)% of mean power at

[5] Note that SMLL does not apply to the non-probabilistic persistence model.

[6] Clear days are defined as those where daily global horizontal irradiance (GHI) was more than 90% of mean maximum daily GHI for the month of January. Measurements are from the Adelaide (West Terrace) Australian Bureau of Meteorology weather station. GHI for clear (cloudy) days ranges from 93–97 (36–90)% of the mean January maximum.



(a) Mean SMSE

(b) Std Dev. SMSE

(c) Mean MAE

(d) Std Dev. MAE

(e) Mean MSLL

(f) Std Dev. MSLL

Fig. 1. Site forecast mean error and error variability

Table 1. Mean site forecast accuracy on clear versus cloudy days

Model	Horizon	Clear days			Cloudy days		
		MAE (kW)	SMSE	MSLL	MAE (kW)	SMSE	MSLL
joint model 1	5	0.091	0.003	−2.01	0.253	0.051	−1.03
site-independent	5	0.050	0.002	−2.10	0.238	0.052	−1.04
joint model 1	10	0.100	0.005	−1.77	0.308	0.066	−0.89
site-independent	10	0.078	0.005	−1.88	0.311	0.077	−0.85
joint model 1	15	0.122	0.007	−1.65	0.340	0.076	−0.83
site-independent	15	0.096	0.006	−1.77	0.341	0.085	−0.78
joint model 1	20	0.150	0.009	−1.56	0.359	0.080	−0.81
site-independent	20	0.109	0.007	−1.69	0.361	0.093	−0.77
joint model 1	25	0.171	0.011	−1.49	0.378	0.085	−0.79
site-independent	25	0.124	0.008	−1.63	0.380	0.100	−0.75
joint model 1	30	0.184	0.012	−1.44	0.389	0.087	−0.78
site-independent	30	0.133	0.009	−1.59	0.390	0.103	−0.73

the 5 (30) min horizon. Similarly, SMSE of 0.002 to 0.009 (for site-independent results) over 5 to 30 min horizons implies that average mean squared error is less than one percent of total power variation on clear days.

Considering transfer learning more generally, it is relevant to note the better performance of the joint model on cloudy days, which contrasts with the better performance of the site-independent models on clear days (Table 1). On all measures, the best performing joint model performs consistently better during variable weather, while the opposite is true for sunny weather periods (noting accuracy is significantly diminished for both models on cloudy days). This suggests 'negative' transfer effects with respect to forecasting diurnal cycles, while forecast errors are somewhat moderated during cloudy periods by the joint model.

4 Discussion

The scalable, approximate Gaussian process methods appear to have significant potential in the distributed forecasting setting. We are able to produce probabilistic site level forecasts using a flexible, nonparametric method in a large scale setting. Further, the approach seems to incorporate diurnal cycles within an integrated model successfully without exogenous site information.

Gaussian process based models produce a strong level of accuracy on sunny days for forecasts out to the 30 min horizon. Overall, however, accuracy of models during cloudy conditions appears low. Given the absence of feature data beyond location, time and output, however, it is possible accuracy can be substantially improved (as in [16]) via inclusion of weather or other external data, including site features where available.

Overall accuracy of forecasts is not improved by jointly estimated models (pooled and coregional) compared to site-independent models. Performance in cloudy versus clear weather, however, illustrates that there may be potential for transfer learning benefits during more variable weather.

One possible factor affecting model performance is the spatial covariance kernel, which is a stationary function resulting in sites equally distant along a fixed axis being assigned an equal covariance regardless of current weather direction. Ideally, a spatial kernel would more specifically reflect current cloud velocity. Further, the stationary kernel assigns higher weight to closer sites, which may not be optimal as forecast horizons increase. A more refined kernel or adaptive model structure may thus assist in identifying relevant cloud-related data features for transfer learning in a forecast setting.

Acknowledgements. This work was supported by Solar Analytics Pty Ltd. and performed on behalf of the Cooperative Research Centre for Low-Carbon Living (University of New South Wales and Solar Analytics Pty Ltd.).

References

1. Alvarez, M.A., Rosasco, L., Lawrence, N.D., et al.: Kernels for vector-valued functions: a review. Found. Trends® Mach. Learn. **4**(3), 195–266 (2012)
2. Aryaputera, A.W., Yang, D., Zhao, L., Walsh, W.M.: Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging. Sol. Energy **122**, 1266–1278 (2015). http://www.sciencedirect.com/science/article/pii/S0038092X15005745
3. Bessa, R., Trindade, A., Silva, C.S., Miranda, V.: Probabilistic solar power forecasting in smart grids using distributed information. Int. J. Electr. Power Energy Syst. **72**, 16–23 (2015). http://www.sciencedirect.com/science/article/pii/S0142061515000897, the Special Issue for 18th Power Systems Computation Conference
4. Bilionis, I., Constantinescu, E.M., Anitescu, M.: Data-driven model for solar irradiation based on satellite observations. Sol. Energy **110**, 22–38 (2014). http://www.sciencedirect.com/science/article/pii/S0038092X14004393
5. Boland, J.: Spatial-temporal forecasting of solar radiation. Renew. Energy **75**, 607–616 (2015). http://www.sciencedirect.com/science/article/pii/S0960148114006624
6. Bonilla, E.V., Krauth, K., Dezfouli, A.: Generic inference in latent Gaussian process models (2016). arXiv preprint: arXiv:1609.00577
7. Copper, J., Sproul, A., Jarnason, S.: Photovoltaic (pv) performance modelling in the absence of onsite measured plane of array irradiance (poa) and module temperature. Renew. Energy **86**, 760–769 (2016)
8. Cressie, N., Wikle, C.K.: Statistics for Spatio-Temporal Data. John Wiley & Sons, Hoboken (2011)
9. David, M., Ramahatana, F., Trombe, P., Lauret, P.: Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models. Sol. Energy **133**, 55–72 (2016). http://www.sciencedirect.com/science/article/pii/S0038092X16300172
10. Diagne, M., David, M., Lauret, P., Boland, J., Schmutz, N.: Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. Renew. Sustain. Energy Rev. **27**, 65–76 (2013). http://www.sciencedirect.com/science/article/pii/S1364032113004334

11. Domke, J., Engerer, N., Menon, A., Webers, C.: Distributed solar prediction with wind velocity (2016)
12. Gutierrez-Corea, F.V., Manso-Callejo, M.A., Moreno-Regidor, M.P., Manrique-Sancho, M.T.: Forecasting short-term solar irradiance based on artificial neural networks and data from neighboring meteorological stations. Sol. Energy **134**, 119–131 (2016). http://www.sciencedirect.com/science/article/pii/S0038092X16300536
13. Hensman, J., Fusi, N., Lawrence, N.D.: Gaussian processes for big data. In: Uncertainty in Artificial Intelligence (2013)
14. Inman, R.H., Pedro, H.T., Coimbra, C.F.: Solar forecasting methods for renewable energy integration. Prog. Energy Combust. Sci. **39**(6), 535–576 (2013)
15. Lauret, P., Voyant, C., Soubdhan, T., David, M., Poggi, P.: A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. Sol. Energy **112**, 446–457 (2015)
16. Li, J., Ward, J.K., Tong, J., Collins, L., Platt, G.: Machine learning for solar irradiance forecasting of photovoltaic system. Renew. Energy **90**, 542–553 (2016). http://www.sciencedirect.com/science/article/pii/S0960148115305747
17. Lonij, V.P., Brooks, A.E., Cronin, A.D., Leuthold, M., Koch, K.: Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors. Sol. Energy **97**, 58–66 (2013)
18. Pelland, S., Remund, J., Kleissl, J., Oozeki, T., De Brabandere, K.: Photovoltaic and solar forecasting: state of the art. iea pvps task 14, subtask 3.1. report iea-pvps t14–01: 2013. Technical report (2013). ISBN: 978-3-906042-13-8
19. Quiñonero-Candela, J., Rasmussen, C.E.: A unifying view of sparse approximate Gaussian process regression. J. Mach. Learn. Res. **6**, 1939–1959 (2005)
20. Rana, M., Koprinska, I., Agelidis, V.G.: Univariate and multivariate methods for very short-term solar photovoltaic power forecasting. Energy Convers. Manag. **121**, 380–390 (2016)
21. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press, Cambridge (2006)
22. Sampson, P.D., Guttorp, P.: Nonparametric estimation of nonstationary spatial covariance structure. J. Am. Stat. Assoc. **87**(417), 108–119 (1992)
23. Shinozaki, K., Yamakawa, N., Sasaki, T., Inoue, T.: Areal solar irradiance estimated by sparsely distributed observations of solar radiation. IEEE Trans. Power Syst. **31**(1), 35–42 (2016)
24. Titsias, M.: Variational learning of inducing variables in sparse Gaussian processes. In: Artificial Intelligence and Statistics (2009)
25. Voyant, C., Notton, G., Kalogirou, S., Nivet, M.L., Paoli, C., Motte, F., Fouilloy, A.: Machine learning methods for solar radiation forecasting: a review. Renew. Energy **105**, 569–582 (2017). http://www.sciencedirect.com/science/article/pii/S0960148116311648
26. Yang, D., Gu, C., Dong, Z., Jirutitijaroen, P., Chen, N., Walsh, W.M.: Solar irradiance forecasting using spatial-temporal covariance structures and time-forward kriging. Renew. Energy **60**, 235–245 (2013). http://www.sciencedirect.com/science/article/pii/S0960148113002759
27. Yang, D., Ye, Z., Lim, L.H.I., Dong, Z.: Very short term irradiance forecasting using the lasso. Sol. Energy **114**, 314–326 (2015). http://www.sciencedirect.com/science/article/pii/S0038092X15000304