Machine Learning (2019) 108:1287–1306
https://doi.org/10.1007/s10994-019-05808-z

Grouped Gaussian processes for solar power prediction

Astrid Dahl[1] · **Edwin V. Bonilla**[2]

Received: 26 November 2018 / Accepted: 3 May 2019 / Published online: 16 May 2019

Abstract
We consider multi-task regression models where the observations are assumed to be a linear combination of several latent node functions and weight functions, which are both drawn from Gaussian process priors. Driven by the problem of developing scalable methods for forecasting distributed solar and other renewable power generation, we propose coupled priors over groups of (node or weight) processes to exploit spatial dependence between functions. We estimate forecast models for solar power at multiple distributed sites and ground wind speed at multiple proximate weather stations. Our results show that our approach maintains or improves point-prediction accuracy relative to competing solar benchmarks and improves over wind forecast benchmark models on all measures. Our approach consistently dominates the equivalent model without coupled priors, achieving faster gains in forecast accuracy. At the same time our approach provides better quantification of predictive uncertainties.

Keywords Gaussian processes · multi-task learning · Bayesian nonparametric methods · scalable inference · solar power prediction

1 Introduction

The problem of forecasting local solar output in the short term is of significant interest for the purpose of distributed grid control and household energy management (Voyant et al. 2017; Widén et al. 2015). Variation in output is driven by two principal factors: diurnal cyclical effects (variation due to sun angle and distance) and variability due to weather effects, both inducing spatially-related dependence between proximate sites. In general, correlations across sites depend on many particulars relating to system configuration, local environment and so on. As such we wish to exploit spatial dependencies (and potentially other site-specific covariates) between sites in a flexible manner. More importantly, inherent to this application is the need for modeling uncertainty in a flexible and principled way (Antonanzas et al. 2016).

Editors: Karsten Borgwardt, Po-Ling Loh, Evimaria Terzi, Antti Ukkonen.

✉ Astrid Dahl
astridmdahl@gmail.com

[1] School of Computer Science and Engineering, University of New South Wales, Sydney, Australia

[2] Data61, Sydney, Australia

Gaussian process (GP) models are a flexible nonparametric Bayesian approach that can be applied to various problems such as regression and classification (Rasmussen and Williams 2006) and have been extended to numerous multivariate and multi-task problems including spatial and spatio-temporal contexts (Cressie and Wikle 2011). Multi-task GP methods have been developed along several lines (see e.g. Álvarez et al. 2012, for a review). Of relevance here are various mixing approaches that combine multiple latent univariate Gaussian processes via linear or nonlinear mixing to predict multiple related tasks (Wilson et al. 2012). The challenge in multi-task cases is maintaining scalability of the approach. To this end, both scalable inference methods and model constraints have been employed (Álvarez et al. 2010; Matthews et al. 2017; Krauth et al. 2017). In particular, latent Gaussian processes are generally constrained to be statistically independent (Wilson et al. 2012; Krauth et al. 2017).

In this paper we consider the case where the statistical independence constraint is relaxed for subsets of latent functions. We build on the scalable generic inference method of Krauth et al. (2017) to extend the model of Wilson et al. (2012) and allow nonzero covariance between arbitrary subsets, or 'groups' of latent functions. The grouping structure is flexible and can be tailored to applications of interest, and additional features can potentially be incorporated to govern input-dependent covariance across functions. By adopting separable kernel specifications, we maintain scalability of the approach whilst capturing latent dependence structures.

With this new multi-task GP model, we consider the specific challenge of forecasting power output of multiple, distributed residential solar sites. We apply our approach to capture spatial covariance between sites by explicitly incorporating spatial dependency between latent functions and test our method on three datasets comprised of solar output at distributed household sites in Australia.

For many of the same reasons, short term wind power forecasting represents significant challenges yet is critical to emerging energy technologies (Widén et al. 2015). Output variability is driven by wind speed, which (as for solar) is driven by multiple interacting environmental factors giving rise to spatial dependencies *a priori*. To demonstrate the broader applicability of the model, we also illustrate our approach on a wind speed dataset comprised of ground wind speed at distributed weather stations in Australia.

Our results show that, for solar models, introducing spatial covariance over groups of latent functions maintains or improves point-prediction forecast accuracy relative to competing benchmark methods and at the same time provides better quantification of predictive uncertainties. Further, wind forecast accuracy and uncertainty is improved on all measures by the introduction of spatial covariance. Timed experiments show that the new model dominates the equivalent model without spatial dependencies, achieving similar or superior forecast accuracy in a shorter time.

2 Related work

Gaussian Processes have been considered in the multi-task setting via a number of approaches. Several methods linearly combine multiple univariate GP models via coefficients that may be parameters (latent factor models as in Teh et al. 2005); linear coregional models (LCM; Goovaerts 1997), or themselves input dependent (Wilson et al. 2012).

Most mixing approaches focus on methods to combine multiple underlying independent latent functions. Recent developments in inference for multi-task GP models have improved scalability of mixing approaches, building upon the variational framework of Titsias (2009).

Nguyen and Bonilla (2014) develop a generic variational inference method that allows efficient optimization for multi-task models with arbitrary likelihoods, while the sparse, variational framework of Hensman et al. (2015); Matthews et al. (2017) supported significant gains in scalability of multi-task GP models. Dezfouli and Bonilla (2015) extend the approach in Nguyen and Bonilla (2014) to the sparse variational context, exploiting inducing points to improve scalability of the inference method using a general mixture-of-Gaussian sparse, variational posterior. More recently, the approach of Dezfouli and Bonilla (2015) was extended to integrate optimization that exploits leave-one-out objective learning in addition to the sparse, variational lower bound (Bonilla et al. 2016; Krauth et al. 2017).

Other multi-task GP approaches allow task-specific predictions through use of task-specific features or 'free-form' cross-task covariances (Bonilla et al. 2008), and more recently priors placed over cluster allocations allowing cluster-specific covariances (Hensman et al. 2014; Gardner et al. 2018). Combination via convolutions has also been developed and extended to sparse, variational settings (Álvarez and Lawrence 2009, 2011).

Coupling between Q node (but not weight) latent functions directly is considered by Remes et al. (2017), who build upon the Gaussian process regression network (GPRN) framework of (Wilson et al. 2012). The authors propose a rich, Generalized Wishart–Gibbs kernel that characterizes covariance for latent functions. The fully-coupled kernel is internally parameterized rather than utilizing feature-dependent cross-function covariance. The approach makes use of variational inference to approximate the model. Unlike our method, however, the natural disadvantage of such an approach is that it presents significant computational challenges in terms of scalability to a large number of observations and tasks. This is primarily due to the need for variational inference that requires batch optimization with $\mathcal{O}((NQ)^3)$ complexity, rendering it infeasible for larger scale applications. In fact, only small experiments were carried out in Remes et al. (2017) with NQ in the order of (approximately) 100 to 500, since the approach is primarily developed for small-data problems requiring a highly expressive latent covariance structure.

2.1 Multi-task solar power forecasting

A number of studies have confirmed that multi-task learning approaches can be useful for distributed solar irradiance or solar power forecasting, finding that cross-site information is relevant (Yang et al. 2013, 2015). Several studies build on the early work of Sampson and Guttorp (1992) and consider kriging methods for distributed solar irradiance forecasting or spatial prediction (notably Yang et al. 2013; Shinozaki et al. 2016). Other approaches include a range of linear statistical methods, shown to be competitive at shorter horizons, and neural network methods (Inman et al. 2013; Widén et al. 2015; Voyant et al. 2017). These approaches are generally constrained by data requirements, notably pre-flattening of data to remove diurnal cyclical trends, which requires knowledge of local system and environment variables. In the context of small scale, distributed residential sites, such information is often unavailable, motivating approaches that do not rely on rich data history or feature sets as are typically required by current approaches (Inman et al. 2013; Widén et al. 2015; Voyant et al. 2017; Antonanzas et al. 2016; Yang et al. 2018).

In addition to kriging studies, GP models have been considered for short term solar forecasting (Bilionis et al. 2014; Dahl and Bonilla 2017). Earlier approaches are generally constrained to small-data problems by poor scalability of exact GP models. More recently, Dahl and Bonilla (2017) use scalable sparse, variational inference to apply multi-task Gaussian (MTG) and linear coregional models (LCM) to forecast solar output at multiple, distributed residen-

tial sites. Multi-task approaches are found to improve model performance in mixed weather conditions, less so in sunny conditions. The specifications adopted, however, did not show strong improvement in overall forecast accuracy relative to the naive, univariate site GP benchmarks, with the LCM performing significantly worse than MTG and individual models in that setting. Moreover, the MTG presents scalability challenges since inducing inputs in the sparse, variational framework adopted are shared across all observations and tasks.

3 Multi-task Gaussian process models

A Gaussian process (GP, Rasmussen and Williams 2006) is formally defined as a distribution over functions such that $f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$ is a Gaussian process with mean function $\mu(\mathbf{x})$ and covariance function $\kappa(\mathbf{x}, \mathbf{x}')$ iff any subset of function values $f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)$ follows a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{K}$, which are obtained by evaluating the corresponding mean function and covariance function at the input points $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.

Standard single-task GP regression assumes observations are iid versions of the latent function values corrupted by Gaussian noise. In this case, posterior inference can be carried analytically (Rasmussen and Williams 2006, chap. 2). In this paper we consider the more general case of multiple outputs, which sometimes is referred to in the literature as multi-task GP regression. In other words, we are given data of the form $\mathcal{D} = \{\mathbf{X} \in \mathbb{R}^{N \times D}, \mathbf{Y} \in \mathbb{R}^{N \times P}\}$ where each $\mathbf{x}_{(n)}$ in $\mathbf{X}$ is a D-dimensional vector of input features and each $\mathbf{y}_{(n)}$ in $\mathbf{Y}$ is a P-dimensional vector of task outputs. Furthermore, we are interested in the case of generally non-linear (non-Gaussian) likelihoods, for which there is no analytically tractable posterior.

3.1 Latent Gaussian process models with independent priors

Fortunately, advances in variational inference (Kingma and Welling 2014; Rezende et al. 2014) have allowed the development of efficient posterior inference methods with 'black-box' likelihoods. In the case of models with GP priors, Krauth et al. (2017) have extended these results to modeling multiple outputs under non-linear likelihoods and *independent* GP priors over multiple latent functions. In short, under such a modeling framework, correlations between the P outputs using Q independent latent functions $\{f_j(\mathbf{x})\}_{j=1}^{Q}$ each drawn from a zero-mean GP prior, i.e. $f_j(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, \kappa_j(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}_j))$, can be encoded via the likelihood. As we shall see in Sect. 5, an example of this is when the independent GPs are linearly combined via a set of weights, which can be deterministic as in the semi-parametric latent factor model of Teh et al. (2005) or stochastic and input-dependent as in the GPRN of Wilson et al. (2012).

Therefore, within the framework in Krauth et al. (2017) the prior over the latent function values corresponding to the N observations along with the likelihood model is given by:

$$p(\mathbf{F}|\boldsymbol{\theta}) = \prod_{j=1}^{Q} p(\mathbf{f}_j|\boldsymbol{\theta}_j) = \prod_{j=1}^{Q} \mathcal{N}(\mathbf{f}_j; \mathbf{0}, \mathbf{K}_{\mathbf{xx}}^{j}), \tag{1}$$

$$p(\mathbf{Y}|\mathbf{F}, \boldsymbol{\phi}) = \prod_{n=1}^{N} p(\mathbf{y}_{(n)}|\mathbf{f}_{(n)}, \boldsymbol{\phi}), \tag{2}$$

where $\mathbf{F}$ is the $N \times Q$ matrix of latent function variables; $\mathbf{f}_j$ is the N-dimensional vector for latent function j; and $\boldsymbol{\theta}_j$ the corresponding hyper-parameters; $\mathbf{f}_{(n)}$ is the Q-dimensional

vector of latent function values corresponding to observation n and $\boldsymbol{\phi}$ are the likelihood parameters.

Krauth et al. (2017) exploit the structure of the model in Eq. (1) to develop a scalable algorithm via variational inference. While the likelihood in this model is suitable for most unstructured machine-learning problems such as standard regression and classification, the prior can be too restrictive for problems where dependences across tasks can be incorporated explicitly. In this paper, driven by the solar power prediction problem where spatial relatedness can be leveraged to improve predictions across different sites, we lift this statistical independence (across latent functions) constraint in the prior to propose a new multi-task model where some of the functions are *coupled a priori*.

4 Grouped priors for multi-task GP models

To group latent functions *a priori*, we can define arbitrarily chosen subsets of latent functions in $\mathbf{F}$, $\mathbf{F}_r, r = 1, \ldots, R$, where R is the total number of groups. For each group the number of latent functions within is denoted Q_r, which we will also refer to as the group size, with $\sum_{r=1}^{R} Q_r = Q$. Each group is comprised of latent functions $\mathbf{F}_r = \{\mathbf{f}_j\}_{j \in \text{ group } r}$ and covariance between latent functions f_j and $f_{j'}$ is nonzero iff the functions f_j and $f_{j'}$ belong to the same group r.

Hence, the prior on $\mathbf{F}$ can be expressed similarly to the generic prior defined in Eq. (1):

$$p(\mathbf{F}|\boldsymbol{\theta}) = \prod_{r=1}^{R} p(\mathbf{F}_r|\boldsymbol{\theta}_r) = \prod_{r=1}^{R} \mathcal{N}(\mathbf{F}_r; \mathbf{0}, \mathbf{K}_{ff}^r), \tag{3}$$

where $\mathbf{K}_{ff}^r \in \mathbb{R}^{NQ_r \times NQ_r}$ is the covariance matrix generated by the group kernel function $\kappa_r(f_j(\mathbf{x}), f_{j'}(\mathbf{x}'))$, which evaluates the covariance of functions f_j and $f_{j'}$ at the locations $\mathbf{x}$ and $\mathbf{x}'$, respectively.

This structure allows arbitrary grouping of latent functions depending on the application (in our case, groups are structured for distributed forecasting, discussed below). However we emphasize that our inference method allows grouping between *any* latent functions in $\mathbf{F}$ and does not make any assumptions (beyond the standard iid assumption) on the conditional likelihood. Hence, since our model allows dependences between latent functions *a priori*, we refer to it as grouped Gaussian processes (GGP). Although we develop a generic and efficient method for GGP models in Sect. 6, our focus in this paper is on a particular class of flexible multi-task regression models referred in the literature to as Gaussian process regression networks (GPRN, Wilson et al. 2012).

4.1 Separable kernels

Before describing how GPRNs fit into the framework of Krauth et al. (2017) and how we generalize them to incorporate grouped priors, it is important to describe a simple yet efficient way of modeling correlations across groups. Once latent functions are coupled *a priori*, scalability becomes an important consideration. Thus, although $\kappa_r(f_j(\mathbf{x}), f_{j'}(\mathbf{x}'))$ is not constrained in terms of kernel choice, for the problem at hand we consider separable kernels of the form $\kappa_r(f_j(\mathbf{x}), f_{j'}(\mathbf{x}')) = \kappa_r(\mathbf{x}, \mathbf{x}')\kappa_r(\mathbf{h}_j, \mathbf{h}_{j'})$. $\mathbf{h}$ are defined as H-dimensioned feature vectors forming an additional feature matrix $\mathbf{H}_r \in \mathbb{R}^{Q_r \times H}$ that characterizes covariance

across functions $\mathbf{f}_j \in \mathbf{F}_r$. We describe in Sect. 7 below how $\mathbf{H}_r$ can be used to exploit spatial dependency between tasks.

This separable structure yields covariance matrices of the Kronecker form $\mathbf{K}^r_{ff} = \mathbf{K}^r_{\mathbf{hh}} \otimes \mathbf{K}^r_{\mathbf{xx}}$, where $\mathbf{K}^r_{\mathbf{xx}} \in \mathbb{R}^{N\times N}$ and $\mathbf{K}^r_{\mathbf{hh}} \in \mathbb{R}^{Q_r\times Q_r}$. By adopting the Kronecker-structured prior covariance over functions within a group, we reduce the maximum dimension of required matrix inversions, allowing scalable inference.

5 Grouped Gaussian process regression networks

Wilson et al. (2012) consider the case where the (noiseless) observations are a linear combination of Gaussian processes, $\{g_\ell(\mathbf{x})\}$, where the coefficients, $\{w_{p\ell}(\mathbf{x})\}$, are also input-dependent and drawn from Gaussian process priors. In other words, their conditional likelihood model for a single observation at input point $\mathbf{x}$ and task p is given by:

$$y_p(\mathbf{x}) = \sum_{\ell=1}^{Q_g} w_{p\ell}(\mathbf{x}) g_\ell(\mathbf{x}) + \epsilon_p,\ p = 1, \dots P, \tag{4}$$

where $\{w_{p\ell}(\mathbf{x}), g_\ell(\mathbf{x})\}$ are drawn from independent GP priors and ϵ_p is a task-dependent Gaussian noise variable. This model is termed the Gaussian process regression network (GPRN) by Wilson et al. (2012) and $\{w_{p\ell}\}$ and $\{g_\ell\}$ are referred to as weight functions and node functions, respectively. It is easy to see how GPRNs fit into the latent Gaussian process model formulation of Krauth et al. (2017), as described in Sect. 3.1. We simply make $\{w_{p\ell}\}$ and $\{g_\ell\}$ subsets of latent functions in $\{f_j\}_{j=1}^{Q}$ with PQ_g weight functions and Q_g node functions so that $Q_g(P+1) = Q$.

Given the observed data $\mathcal{D}$, for each latent process (over weights or node functions) we need to create as many latent variables as observations. Therefore, it is useful to conceptualize the weights as PQ_g latent variables of dimension $N \times 1$ arranged into a tensor $\mathbf{W} \in \mathbb{R}^{P\times Q_g\times N}$. Similarly, the node functions can be represented by Q_g latent variables of dimension $N \times 1$ arranged into a tensor $\mathbf{G} \in \mathbb{R}^{Q_g\times 1\times N}$. Therefore, the conditional likelihood for input $\mathbf{x}_{(n)}$ can be written in matrix form as

$$p(\mathbf{y}_{(n)}|\mathbf{f}_{(n)}, \boldsymbol{\phi}) = \mathcal{N}(\mathbf{y}_{(n)}; \mathbf{W}_{(n)}\mathbf{g}_{(n)}, \boldsymbol{\Sigma}_y), \tag{5}$$

where the latent functions are given by node and weight functions, i.e. $\mathbf{f}_{(n)} = \{\mathbf{W}_{(n)}, \mathbf{g}_{(n)}\}$; the conditional likelihood parameters $\boldsymbol{\phi} = \boldsymbol{\Sigma}_y$ and $\boldsymbol{\Sigma}_y$ is a diagonal matrix. P-dimensional outputs are constructed at $\mathbf{x}_{(n)}$ as the product of a $P \times Q_g$ matrix of weight functions, $\mathbf{W}_{(n)}$, and Q_g-dimensional vector of node functions $\mathbf{g}_{(n)}$.

5.1 Grouping structure

Although our modeling and inference framework allows for arbitrary grouping structure, we consider a correlated prior over the rows of the weight functions for the grouped GPRN, and give details of the exact setting for the solar and wind applications in Sect. 7. Figure 1 illustrates our GGP framework for the GPRN likelihood.

Naturally, the greater flexibility of our approach comes at the expense of a high time-and-memory complexity, which poses significant challenges for posterior estimation. In the following section, we develop an efficient variational inference algorithm for our GGP model that is not much more computationally expensive than the original GPRN's. In fact, we show

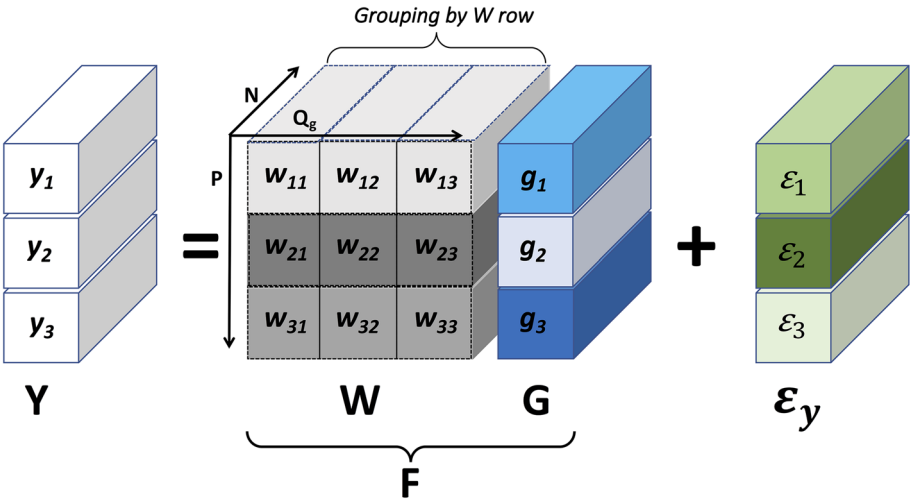

Fig. 1 Gaussian process regression network model where $\mathbf{Y}$ is a linear combination of node and weight latent functions comprising $\mathbf{F}$. In the grouped Gaussian process (GGP) framework, latent functions may be grouped arbitrarily. A grouping scheme is illustrated where weight functions in $\mathbf{W}$ are grouped by rows (grouped functions are shown in the same shade) and given a fully-coupled prior, while node functions in $\mathbf{G}$ are independent. Here N is the number of observations per task; P is the number of tasks; and Q_g is the group size

in our experiments in Sect. 9 that our inference method can converge faster than GPRN's while achieving similar or better performance.

6 Inference

Our inference method is based on the generic inference approach for latent variable Gaussian process models set out by Krauth et al. (2017). This is a sparse variational method that considers the case where latent functions are conditionally independent. We adapt that framework to the more general case where latent functions covary within groups, and for our case exploit the Kronecker structures noted at Sect. 5. Since our inference method does not exploit any of the specifics of the GPRN likelihood, we consistently use the general grouped prior notation defined in Sect. 4.

Under the sparse method, the prior at (3) is augmented with inducing variables, $\{\mathbf{u}_r\}_{r=1}^R$, drawn from the same GP priors as $\mathbf{F}_r$ at new inducing points $\mathbf{Z}_r$, where $\mathbf{Z}_r \in \mathbb{R}^{M\times D}$ lie in the same space as $\mathbf{X} \in \mathbb{R}^{N\times D}$, $M \ll N$. Since $\mathbf{u}_r$ are drawn from the same GP priors, inducing variables within a group r are similarly coupled via $\kappa_r(f_j(\mathbf{x}), f_{j'}(\mathbf{x}'))$ evaluated at points $\mathbf{Z}_r$. The prior in Eq. (3) is thus replaced by

$$p(\mathbf{u}|\boldsymbol{\theta}) = \prod_{r=1}^{R} p(\mathbf{u}_r|\boldsymbol{\theta}_r) = \prod_{r=1}^{R} \mathcal{N}(\mathbf{u}_r; \mathbf{0}, \mathbf{K}_{uu}^r), \tag{6}$$

$$p(\mathbf{F}|\mathbf{u}) = \prod_{r=1}^{R} \mathcal{N}(\mathbf{F}_r; \tilde{\boldsymbol{\mu}}_r, \widetilde{\mathbf{K}}_r), \tag{7}$$

Table 1 A summary of the prior covariance matrix structures for a given group r for scalable variational inference in the GGP model

Notation	Specification	Description
$\mathbf{K}^r_{ff}$	$\mathbf{K}^r_{\mathbf{hh}} \otimes \mathbf{K}^r_{\mathbf{xx}}$	Covariance between latent functions
$\mathbf{K}^r_{uu}$	$\mathbf{K}^r_{\mathbf{hh}} \otimes \mathbf{K}^r_{\mathbf{zz}}$	Covariance between inducing variables
$\mathbf{K}^r_{fu}$	$\mathbf{K}^r_{\mathbf{hh}} \otimes \mathbf{K}^r_{\mathbf{xz}}$	Covariance between latent functions and inducing variables
$\widetilde{\mathbf{K}}_r$	$\mathbf{K}^r_{ff} - \mathbf{A}_r \mathbf{K}^r_{uf}$	Covariance of conditional prior $p(\mathbf{F}\|\mathbf{u})$
$\mathbf{A}_r$	$\mathbf{I}_{Qr} \otimes \mathbf{K}^r_{\mathbf{xz}} (\mathbf{K}^r_{\mathbf{zz}})^{-1}$	Auxiliary matrix

where $\tilde{\boldsymbol{\mu}}_r = \mathbf{A}_r \mathbf{u}_r$, $\widetilde{\mathbf{K}}_r = \mathbf{K}^r_{ff} - \mathbf{A}_r \mathbf{K}^r_{uf}$ and $\mathbf{A}_r = \mathbf{K}^r_{fu} (\mathbf{K}^r_{uu})^{-1} = \mathbf{I}_{Qr} \otimes \mathbf{K}^r_{\mathbf{xz}} (\mathbf{K}^r_{\mathbf{zz}})^{-1}$. $\mathbf{K}^r_{uu} \in \mathbb{R}^{MQ_r \times MQ_r}$ is the covariance matrix induced by $\kappa_r(f_j(\mathbf{x}), f_{j'}(\mathbf{x}'))$ evaluated over $\mathbf{Z}_r$, $\mathbf{H}_r$, yielding the structure $\mathbf{K}^r_{uu} = \mathbf{K}^r_{\mathbf{hh}} \otimes \mathbf{K}^r_{\mathbf{zz}}$ and importantly the decomposition $(\mathbf{K}^r_{uu})^{-1} = (\mathbf{K}^r_{\mathbf{hh}})^{-1} \otimes (\mathbf{K}^r_{\mathbf{zz}})^{-1}$. We similarly define $\mathbf{K}^r_{fu}$ and $\mathbf{K}^r_{uf}$ (Table 1).

6.1 Posterior estimation

The (analytically intractable) joint posterior distribution of the latent functions and inducing variables under the prior and likelihood models in Eqs. (1) and (6) is approximated via variational inference (Jordan et al. 1998). Specifically, $p(\mathbf{F}, \mathbf{u}|\mathbf{Y}) = p(\mathbf{F}|\mathbf{u}, \mathbf{Y}) p(\mathbf{u}|\mathbf{Y}) \approx q(\mathbf{F}, \mathbf{u}|\boldsymbol{\lambda}) \stackrel{\text{def}}{=} p(\mathbf{F}|\mathbf{u}) q(\mathbf{u}|\boldsymbol{\lambda})$. The variational posterior $q(\mathbf{u}|\boldsymbol{\lambda})$ is defined as a mixture of K Gaussians (MoG) with mixture proportions π_k. We assume that $q(\mathbf{u}|\boldsymbol{\lambda})$ also factorizes over groups (and in the diagonal case over individual latent functions). The variational posterior is thus defined as

$$q(\mathbf{u}|\boldsymbol{\lambda}) = \sum_{k=1}^{K} \pi_k \prod_{r=1}^{r} q_k(\mathbf{u}_r|\boldsymbol{\lambda}_{kr}) \tag{8}$$

where $q_k(\mathbf{u}_r|\boldsymbol{\lambda}_{kr}) = \mathcal{N}(\mathbf{u}_r; \mathbf{m}_{kr}, \mathbf{S}_{kr})$ and $\boldsymbol{\lambda}_{kr} = \{\mathbf{m}_{kr}, \mathbf{S}_{kr}, \pi_k\}$. We then estimate the model by maximizing the so-called evidence lower bound (ELBO), which de-constructs to $\mathcal{L}_{\text{elbo}}(\boldsymbol{\lambda}) \stackrel{\text{def}}{=} \mathcal{L}_{\text{ent}}(\boldsymbol{\lambda}) + \mathcal{L}_{\text{cross}}(\boldsymbol{\lambda}) + \mathcal{L}_{\text{ell}}(\boldsymbol{\lambda})$, which are the entropy, cross-entropy and expected log likelihood terms, respectively. The explicit expression required for $\mathcal{L}_{\text{elbo}}$ is a generalization of the results in Krauth et al. (2017). For the entropy term we have that (using Jensen's inequality):

$$\begin{aligned} \mathcal{L}_{\text{ent}}(\boldsymbol{\lambda}) &= \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\lambda})}\big[\log q(\mathbf{u}|\boldsymbol{\lambda})\big] \\ &\geq -\sum_{k=1}^{K} \pi_k \log \sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{m}_k; \mathbf{m}_l, \mathbf{S}_k + \mathbf{S}_l), \end{aligned} \tag{9}$$

where $\mathbf{m}_k$ is the vector $\{\mathbf{m}_{kr}\}_{r=1}^{R} = \{\mathbf{m}_{kj}\}_{j=1}^{Q}$ and $\mathbf{S}_k$ is the block diagonal matrix with diagonal elements $\{\mathbf{S}_{kr}\}_{r=1}^{R}$ (and equivalent for $\mathbf{m}_l, \mathbf{S}_l$). For the cross-entropy and the expected log likelihood terms:

$$\begin{aligned}\mathcal{L}_{\text{cross}}(\boldsymbol{\lambda}) &= \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\lambda})}\big[\log p(\mathbf{u}|\boldsymbol{\theta})\big] \\ &= -\frac{1}{2}\sum_{k=1}^{K}\pi_k\sum_{r=1}^{R}[M_r\log(2\pi)+\log\left|\mathbf{K}_{uu}^r\right|+\mathbf{m}_{kr}'(\mathbf{K}_{uu}^r)^{-1}\mathbf{m}_{kr}+\text{tr}\,((\mathbf{K}_{uu}^r)^{-1}\mathbf{S}_{kr})]\end{aligned} \tag{10}$$

$$\mathcal{L}_{\text{ell}}(\boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{F}|\boldsymbol{\lambda})}\big[\log p(\mathbf{Y}|\mathbf{F},\boldsymbol{\phi}))\big] = \sum_{n=1}^{N}\mathbb{E}_{q_{(n)}(\mathbf{f}_{(n)}|\boldsymbol{\lambda})}\big[\log p(\mathbf{y}_{(n)}|\mathbf{f}_{(n)},\boldsymbol{\phi})\big] \tag{11}$$

where $q(\mathbf{F}|\boldsymbol{\lambda})$ results from integration of the joint approximate posterior over inducing variables $\mathbf{u}$. Note that $\text{tr}\,((\mathbf{K}_{uu}^r)^{-1})$ factorizes as $\text{tr}\,((\mathbf{K}_{\mathbf{zz}}^r)^{-1})\,\text{tr}\,((\mathbf{K}_{\mathbf{hh}}^r)^{-1})$ and $\ln\left|\mathbf{K}_{uu}^r\right|$ factorizes as $Q_r\ln\left|\mathbf{K}_{\mathbf{zz}}^r\right| + M\ln\left|\mathbf{K}_{\mathbf{hh}}^r\right|$. Given factorization of the joint and variational posteriors over k and r and standard conjugacy results, we have

$$\begin{aligned}q(\mathbf{F}|\boldsymbol{\lambda}) &= \sum_{k=1}^{K}\pi_k\prod_{r=1}^{R}\mathcal{N}(\mathbf{b}_{kr},\boldsymbol{\Sigma}_{kr}), \\ \mathbf{b}_{kr} &= \mathbf{A}_r\mathbf{m}_{kr}, \quad \text{and} \quad \boldsymbol{\Sigma}_{kr} = \widetilde{\mathbf{K}}_r + \mathbf{A}_r\mathbf{S}_{kr}\mathbf{A}_r'\end{aligned} \tag{12}$$

The distribution $q_{(n)}(\mathbf{f}_{(n)}|\boldsymbol{\lambda})$ similarly factorizes as

$$\begin{aligned}q_{(n)}(\mathbf{f}_{(n)}|\boldsymbol{\lambda}) &= \sum_{k=1}^{K}\pi_k q_{k(n)}(\mathbf{f}_{(n)}|\boldsymbol{\lambda}_k) \\ &= \sum_{k=1}^{K}\pi_k\prod_{r=1}^{R}\mathcal{N}(\mathbf{b}_{kr(n)},\boldsymbol{\Sigma}_{kr(n)}).\end{aligned} \tag{13}$$

$\mathcal{L}_{\text{ell}}$ may be estimated by Monte Carlo, requiring sampling only from Q_r-dimensional multivariate Gaussians $\mathcal{N}(\mathbf{f}_{r(n)};\mathbf{b}_{kr(n)},\boldsymbol{\Sigma}_{kr(n)})$ where $\mathbf{b}_{kr(n)}$ is the vector comprised of every nth element of $\mathbf{b}_{kr}$, and $\boldsymbol{\Sigma}_{kr}$ is the (full) $Q_r\times Q_r$ matrix comprised of nth diagonal elements of posterior covariance $\boldsymbol{\Sigma}_{kjj'}$ sub-matrices of $\boldsymbol{\Sigma}_{kr}$. Thus, we estimate

$$\widehat{\mathcal{L}}_{\text{ell}} = \frac{1}{S}\sum_{n=1}^{N}\sum_{k=1}^{K}\pi_k\sum_{s=1}^{S}\ln p(\mathbf{y}_{(n)}|\mathbf{f}_{(n)}^{(k,s)}). \tag{14}$$

Under the separable structure adopted, each mixture component covariance for $\mathbf{f}_{r(n)}$, $\boldsymbol{\Sigma}_{kr(n)}$ can be seen to consist of structure arising from the grouped prior plus a term arising from the variational posterior:

$$\begin{aligned}\boldsymbol{\Sigma}_{kr(n)} &= \widetilde{\mathbf{K}}_{r(n)} + \mathbf{A}_{r(n)}\mathbf{S}_{kr}\mathbf{A}_{r(n)}', \quad \text{where} \\ \widetilde{\mathbf{K}}_{r(n)} &= \mathbf{K}_{\mathbf{hh}}^r \times \left[\kappa_r(\mathbf{x}_{(n)},\mathbf{x}_{(n)} - \kappa_r(\mathbf{x}_{(n)},\mathbf{Z}_r)\mathbf{K}_{\mathbf{zz}}^r\kappa_r(\mathbf{Z}_r,\mathbf{x}_{(n)})\right] \quad \text{and} \\ \mathbf{A}_{r(n)} &= \left[\mathbf{I}_{Qr}\otimes\kappa_r(\mathbf{x}_{(n)},\mathbf{Z}_r)(\mathbf{K}_{\mathbf{zz}}^r)^{-1}\right]\end{aligned} \tag{15}$$

Thus cross-function covariance within a group will be either driven by the prior, where $\mathbf{S}_{kr}$ is diagonal, or more flexible in form where $\mathbf{S}_{kr}$ is non-diagonal.

6.2 Prediction

Prediction for a new point $\mathbf{y}_\star$ given $\mathbf{x}_\star$ is taken as the expectation over the general posterior distribution for the new point:

$$p(\mathbf{y}_\star|\mathbf{x}_\star) = \int p(\mathbf{y}_\star|\mathbf{f}_\star)q(\mathbf{f}_\star|\boldsymbol{\lambda})\mathrm{d}\mathbf{F}$$
$$= \sum_{k=1}^{K} \pi_k \int p(\mathbf{y}_\star|\mathbf{f}_\star)q_k(\mathbf{f}_\star|\boldsymbol{\lambda}_k)\mathrm{d}\mathbf{F}_\star, \tag{16}$$

where $q_k(\mathbf{f}_\star|\boldsymbol{\lambda}_k)$ is defined as for $q_{k(n)}(\mathbf{f}_{(n)}|\boldsymbol{\lambda}_k)$ in (13). Given the explicit expression for the posterior distribution, the expectation in Eq. (16) is estimated by sampling:

$$\mathbb{E}_{p(\mathbf{y}_\star|\mathbf{x}_\star)}\left[\mathbf{y}_\star\right] \approx \frac{1}{S}\sum_{s=1}^{S} \mathbf{W}_\star^s \mathbf{g}_\star^s, \tag{17}$$

where $\{\mathbf{W}_\star^s, \mathbf{g}_\star^s\} = \mathbf{f}_\star^s$ are samples from $q_k(\mathbf{f}_\star|\boldsymbol{\lambda}_k)$.

6.3 Complexity

Under the GGP with a Kronecker-structured prior the time complexity per iteration changes slightly from the independent function case. For the same P, Q_g and M, fewer M-dimensioned inversions are required for GGP versus GPRN, without any increase in maximum dimension under the Kronecker specification assuming $M \geq Q_r$. This represents a substantial reduction in M-dimensioned inversions, depending on the grouping scheme.

The cost of calculating $\mathcal{L}_{\text{cross}}$ is dominated by the cost of inversions, being $\mathcal{O}\left(\sum_{r=1}^{R}(M^3 + Q_r^3)\right)$ for the grouped case and $\mathcal{O}\left(QM^3\right)$ for the independent case. Under the diagonal posterior specification, $\mathbf{S}_k$ in Eq. (9) reduces to the same form as the independent case of Krauth et al. (2017). Lastly, $\mathcal{L}_{\text{ell}}$ under the grouped structure requires sampling from low-dimensional $Q_r \times Q_r$ multivariate Gaussians with non-diagonal posterior covariance matrices, whereas this is avoided under the independent framework. However, the low dimensionality (number of tasks in our empirical evaluation) involved yields minimal additional cost.

7 Grouped Gaussian processes for spatially dependent tasks

It is natural to consider a multi-task framework in a spatio-temporal setting such as distributed solar forecasting, where power output at solar sites in a region would *a priori* be expected to covary over time and space. Given the expectation of spatially-driven covariance across sites, i.e. tasks, we seek to exploit this structure to increase both efficiency and accuracy of multi-task forecasts. Our approach does this by incorporating explicit spatial dependencies between latent functions in the model.

Latent functions in the general framework do not necessarily map to a particular task. The question therefore arises as to how to use spatial information relating to *tasks* to structure covariance between *latent functions*. We solve this by setting $Q_g = P$ and grouping latent functions within rows of $\mathbf{W}$ i.e. $\mathbf{f}_j \in \mathbf{W}_{i:}, i = 1, \ldots P$. We then define a feature matrix $\mathbf{H}_r$ that governs covariance across the P functions in each row (Fig. 1). With $Q_g \geq P$ it is

possible to obtain a very general representation of the multi-task process with full mixing between tasks via $\mathbf{G}$, which now contains P node functions. This grouping structure allows parameters to vary across tasks, and at the same time, the coupled prior can act to regularize latent function predictions.

Model settings In our setting, we consider each latent process in $\mathbf{G}$ to be an independent GP, i.e., $\langle \mathbf{g}_j, \mathbf{g}_{j'} \rangle = 0$ for $j \neq j'$. Furthermore, input features of $\mathbf{g}_j$, $j = 1, \ldots, P$ are defined to be task features i.e. features for $\mathbf{g}_j$ relate to task j, specifically lagged-target values for j.

We define spatial features $\mathbf{h}_j = (latitude_j, longitude_j)$ governing weightings applied to node functions. For a given task i, $\mathbb{E}_{p(\mathbf{Y}|\mathbf{F},\boldsymbol{\phi})}\left[\mathbf{y}_{(n)i}\right] = \mathbf{w}_{(n)i}\mathbf{g}_{(n)}$ where $\mathbf{w}_{(n)i}$ denotes the ith row vector of $\mathbf{W}_{(n)}$. It can be seen that, in addition to depending on input features $\mathbf{x}_{(n)}$, relative weights placed on node functions are now smoothed by spatial covariance imposed over the weights in $\mathbf{w}_{(n)i}$. This allows site-by-site optimization of spatial decay in (cross-task) weights in addition to site-specific parameterization and features in $\mathbf{w}_{(n)i}$. In total, this grouping structure yields $2P$ groups: P groups of size P (corresponding to $\mathbf{W}$) and P groups of size 1 (corresponding to $\mathbf{G}$).

Kernels and features for $\kappa_r(\mathbf{x}, \mathbf{x}')$ and $\kappa_r(\mathbf{h}_j, \mathbf{h}_{j'})$ are selected in line with previous studies relating to multi-task distributed forecasting (Inman et al. 2013; Dahl and Bonilla 2017). In particular, for our task of forecasting distributed solar output over time, for $\mathbf{g}_j$, $j = 1, \ldots, P$, we define $\kappa_{\mathbf{g}_j}(\mathbf{x}_t, \mathbf{x}_s) = \kappa_{\mathbf{g}_j}(\boldsymbol{l}_t, \boldsymbol{l}_s)$ as a radial basis function kernel (κ_{RBF}) applied to a feature vector of recent lagged observed power at site j, i.e. for site j at time t, $\boldsymbol{l}_{j,t} = (y_{j,t}, y_{j,t-1}, y_{j,t-2})$.

For row-group r, we define a separable, multiplicative kernel structure as discussed above, i.e. $\kappa_r(f_j(\mathbf{x}), f_{j'}(\mathbf{x}')) = \kappa_r(\mathbf{x}, \mathbf{x}')\kappa_r(\mathbf{h}_j, \mathbf{h}_{j'})$. We set the kernel over the inputs as $\kappa_r(\mathbf{x}, \mathbf{x}') = \kappa_{Per.}(t, s)\kappa_{RBF}(\boldsymbol{l}_{rt}, \boldsymbol{l}_{rs})$, where $\kappa_{Per.}(t, s)$ is a periodic kernel on a time index t capturing diurnal cyclical trends in solar output.

We adopt a compact kernel over functions, specifically a separable RBF- Epanechnikov structure, i.e., $\kappa_r(\mathbf{h}_j, \mathbf{h}_{j'}) = \kappa_{RBF}(\mathbf{h}_j, \mathbf{h}_{j'})\kappa_{Ep.}(\mathbf{h}_j, \mathbf{h}_{j'})$, $j, j' = 1 \ldots P$, where $\mathbf{h}_j = (latitude, longitude)$ for site j. By using a more flexible compact kernel, we aim to allow beneficial shared learning across tasks while reducing negative transfer by allowing cross-function weights to reduce to zero at an optimal distance.

8 Experiments

We evaluate our approach on forecasting problems for distributed residential solar installations and wind speed measured at proximate weather stations.

8.1 Solar forecasting

The task for solar is to forecast power production 15 min ahead at multiple distributed sites in a region. Data consist of 5 min average power readings from groups of proximate sites in the Adelaide and Sydney regions in Australia. We present results for three datasets: ten Adelaide sites (ADEL- AUTM) and twelve Sydney sites (SYD- AUTM), both over 60 days during Autumn 2016, and ten Adelaide sites (ADEL- SUMM) over 60 days in Spring–Summer 2016. We train all models on 36 days of data, and test forecast accuracy for 24 subsequent days (days are defined as 7 am to 7pm). In all, for each site, we have 5000 datapoints for training and 3636 datapoints for testing.

Datasets have varying spatial dispersions. ADEL- AUTM (ADEL- SUMM) sites are spread over an approximately 30 × 40 (20 × 20) km area, while SYD- AUTM sites are evenly dispersed over an approximately 15 × 20 km area.

8.1.1 Benchmark models

We compare forecast performance of our GGP method to the fully independent GPRN and several other benchmark models. We estimate (1) separate independent GP forecast models for each site (IGP), (2) pooled multi-task models with task-specific (spatial) features for sites (MTG), and (3) multi-task linear coregional models (LCM). The final benchmark model (4) is the GPRN with independent latent functions (Wilson et al. 2012).

These models can be expressed in terms of the general latent function framework with differing values of P, Q and R, and different likelihood functions. As discussed in Sect. 5, where latent functions are independent, group size is equal to 1 and $R = Q$. Key model constants are presented at Table 3.

Both IGP and MTG models have a standard, single task Gaussian likelihood functions, while the LCM model is comprised of P node functions mapped to outputs via a $P \times Q_g$ matrix of deterministic weights, i.e. $p(\mathbf{y}_{(n)}|\mathbf{f}_{(n)}, \boldsymbol{\phi}) = \mathcal{N}(\mathbf{y}_{(n)}; \mathbf{W}_{(n)}\mathbf{g}_{(n)}, \boldsymbol{\Sigma}_y)$ where $\mathbf{W}_{(n)ij} = w_{ij} \quad \forall n = 1, \ldots, N$ and $Q_g = P$. Kernels for all models are presented at Table 2. We maintain similar kernel specification across models. All kernels are based on the specification described at Sect. 7.

Models are presented for diagonal and full MoG posterior specifications, with $K = 1$. In the case of the GGP, to maintain the scalable specification, we adopt a Kronecker construction of the full posterior for each group R in line with the prior specification.

Table 2 Latent function kernel specifications for GGP and benchmark models

Model	$\kappa_j(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}_j)$ (benchmark models)
IGP	$\kappa_{Per.}(t,s)\kappa_{RBF}(\boldsymbol{l}_{it}, \boldsymbol{l}_{is})$
MTG	$\kappa_{Per.}(t,s)\kappa_{RBF}(\boldsymbol{l}_{it}, \boldsymbol{l}_{js})\kappa_{RBF-Ep.(2)}(\mathbf{h}_i, \mathbf{h}_j)$
LCM	$\kappa_{Per.}(t,s)\kappa_{RBF}(\boldsymbol{l}_{it}, \boldsymbol{l}_{is}), \quad i = 1$ to P.
GPRN	
$\mathbf{W}_{i,j}$	$\kappa_{Per.}(t,s)\kappa_{RBF}(\boldsymbol{l}_{it}, \boldsymbol{l}_{is})$
$\mathbf{g}_j$	$\kappa_{RBF}(\boldsymbol{l}_{it}, \boldsymbol{l}_{is}), \quad i = 1$ to P
GGP	$\kappa_r(f_j(\mathbf{x}), f_{j'}(\mathbf{x}'))$ (GGP solar)
$\mathbf{W}_{i,:}$	$\kappa_{Per.}(t,s)\kappa_{RBF}(\boldsymbol{l}_{it}, \boldsymbol{l}_{is})\kappa_{RBF-Ep.(2)}(\mathbf{h}_i, \mathbf{h}_j)$
$\mathbf{g}_j$	$\kappa_{RBF}(\boldsymbol{l}_{it}, \boldsymbol{l}_{is}), \quad i = 1$ to P
GGP	$\kappa_r(f_j(\mathbf{x}), f_{j'}(\mathbf{x}'))$ (GGP wind)
$\mathbf{W}_{i,j\neq i}$	$\kappa_{Per.}(t,s)\kappa_{RBF}(\boldsymbol{l}_{it}, \boldsymbol{l}_{is})\kappa_{RBF-Ep.(2)}(\mathbf{h}_i, \mathbf{h}_j)$
$\mathbf{W}_{i,i}$	$\kappa_{Per.}(t,s)\kappa_{RBF}(\boldsymbol{l}_{it}, \boldsymbol{l}_{is})$
$\mathbf{g}_j$	$\kappa_{RBF}(\boldsymbol{l}_{it}, \boldsymbol{l}_{is}), \quad i = 1$ to P

$\kappa_{Per.}(t,s)$ is a periodic kernel applied to a time index; $\kappa_{RBF}(\boldsymbol{l}_{it}, \boldsymbol{l}_{is})$ is a radial basis function kernel applied to recent lagged power output; $\kappa_{RBF-Ep.(2)}(\mathbf{h}_i, \mathbf{h}_j)$ is a separable, multiplicative radial basis-Epanechnikov function kernel applied to cross-site spatial features (latitude and longitude)

Table 3 Key constants for GGP and benchmark models

Model	P	Q_g	Q	R	M	Agg. Ind.
ADEL- AUTM, ADEL- SUMM						
IGP	1	1		1	543	543
MTG	1	1		1	543	543
LCM	10	10	10	10	252	2520
GPRN	10	10	110	110	113	12,463
GGP	10	10	110	20	200	4000
SYD- AUTM						
IGP	1	1		1	577	577
MTG	1	1		1	577	577
LCM	12	12	12	12	252	3024
GPRN	12	12	156	156	107	16,718
GGP	12	12	156	24	200	4800
WIND						
IGP	1	1		1	524	524
MTG	1	1		1	524	524
LCM	6	6	6	6	288	1730
GPRN	6	6	42	156	107	6300
GGP	6	6	42	18	200	3600

Output dimension (P); number of node functions (Q_g, LCM, GPRN and GGP only); number of latent functions (Q); number of latent function groups (R); dimension of inducing point matrices $\mathbf{K}_{\mathbf{zz}}^{r}$ (M); and total inducing points (Agg. Ind.) M has been set to obtain roughly the same computational cost per iteration across all models

To compare model performance under equivalent settings, we consider the complexity of the different approaches and standardize model settings by reference to a consistent target computational complexity per iteration. In our variational framework, the time complexity is dominated by algebraic operations with cubic complexity on the number of inducing inputs M. We therefore set $QM^3 = RM^3 = 20 \times (200)^3$ for ADEL- AUTM and ADEL- SUMM models, $QM^3 = RM^3 = 24 \times (200)^3$ for SYD- AUTM, and adjust the number of inducing points, M, accordingly (Table 3).

8.1.2 Experiment settings and performance measures

All models are estimated based on the variational framework explained in Sect. 6. We optimize the ELBO iteratively until its relative change over successive epochs is less than 10^{-5} up to a maximum of 200 epochs. Optimization is performed using ADAM (Kingma and Ba 2014) with settings $\{LR = 0.005;\ \beta_1 = 0.09;\ \beta_2 = 0.99\}$. All data except time index features are normalized prior to optimization. Reported forecast accuracy measures are root mean squared error (RMSE) and negative log predictive density (NLPD). The non-Gaussian likelihood of GPRN models makes the usual analytical expression for NLPD intractable. We therefore estimate it using Monte Carlo:

$$\text{NLPD} = -\mathbb{E}_{q(\mathbf{f}_\star|\boldsymbol{\lambda})}\big[\ln p(\mathbf{y}_\star|\mathbf{f}_\star)\big] \approx -\frac{1}{S}\sum_{s=1}^{S} \ln \mathcal{N}(\mathbf{y}_\star; \mathbf{W}_\star^s \mathbf{g}_\star^s, \boldsymbol{\phi}),$$

where $\mathbf{W}_\star^s \mathbf{g}_\star^s$ are draws from their corresponding posterior over $\mathbf{f}_\star$. In addition, we compute average ranking (M- RANK) over both accuracy measures (RMSE and NLPD), and mean forecast variance (F- VAR), which is critical to the use of short term forecasts as inputs to system or market management algorithms.

8.1.3 Results

Results for solar models are presented at Table 4 with diagonal and full-Gaussian posterior specifications. GGP maintains or improves point accuracy when compared to best performing benchmarks on both RMSE and NLPD individually. For RMSE, accuracy under GGP differs by less than 1% relative to GPRN, and similarly matches or improves on NLPD relative to LCM and other benchmarks. GGP performs strongly in terms of overall accuracy across both measures, consistently achieving the highest average rank across both measures (M- RANK). In contrast, competing baselines either perform well on RMSE at the expense of poor performance under NLPD or *vice versa*.

The benefit of regularization under the GGP is clear when considering mean forecast variance, which is lower under GGP than all benchmark models for all experiments. Compared to the un-grouped GPRN (LCM), variance for solar forecasts is reduced by 18 to 24 (13 to 40)% under the most accurate GGP model.

We test statistical significance of differences in performance discussed above via 95% intervals estimated by Monte Carlo.[1] Results of the analysis show that RMSE under GPRN is statistically significantly lower than under GGP for solar datasets. In all other cases, RMSE is either not significantly different or significantly higher than under GGP. Results are similar for NLPD, which is statistically significantly lower under LCM for two of three datasets, and otherwise higher or not significantly different.

With the exception of the MTG model, all multi-task models consistently improve on the naive independent forecast models. Figure 3 illustrates the benefit observed under the GGP (and other multi-task models) in reducing large forecast errors associated with variable weather conditions.

8.2 Wind speed forecasting

Wind variability shares characteristics with solar variability, as discussed in Sect. 1, with similar approaches applied to the problem of short term forecasting (Widén et al. 2015). We test our GGP method forecasting ground wind speed 30 min ahead at six weather stations in Victoria, Australia, within an approximately 30 × 40 km area. Data are half-hourly wind speed readings collected over an eight month period. The WIND data present an interesting challenge, with frequent missing and noisy observations (Fig. 2). After filtering, we have 4000 training points and 1024 test points per station.

We adopt the same kernel and feature definitions as for solar (Table 2) however use a different grouping structure for the GGP. We allow functions on the diagonal of $\mathbf{W}$ to be independent and group off-diagonal functions within each row. This structure for each task j allows weight placed on its 'own' univariate node function $\mathbf{g}_j$ to be independent of weights placed over remaining sites, which are still spatially smoothed. Wea similarly adjust the

[1] The Monte Carlo procedure tests for significance of differences between model performance metrics by repeatedly resampling from the test data and recalculating the difference in performance metrics across models for each sample (sample size is N_{test}). Differences are deemed statistically significantly different from zero where the null value falls outside the interval defined by percentiles (0.025, 0.975) of the constructed empirical distribution.

Table 4 Forecast accuracy and variance of GGP and benchmark models using diagonal (D) and full (F) Gaussian posteriors

	ADEL- AUTM							ADEL- SUMM						
	RMSE		NLPD		M- RANK	F- VAR		RMSE		NLPD		M- RANK	F- VAR	
GGP (D)	0.282		0.243		**2.5**	0.140		0.318		0.323		**2.5**	0.118	
GGP (F)	0.288	*	0.265	*	4.0	**0.136**	*	0.321	*	0.352	*	4.0	**0.113**	*
LCM (D)	0.294	*	0.240		4.0	0.162	*	0.325	*	0.332	*	4.5	0.165	*
LCM (F)	0.293	*	**0.240**	*	3.0	0.160	*	0.323	*	**0.323**		3.0	0.158	*
GPRN (D)	**0.278**	*	0.311	*	3.0	0.173	*	**0.315**	*	0.376	*	3.0	0.152	*
GPRN (F)	0.283		0.320	*	4.5	0.174	*	0.316	*	0.382	*	4.0	0.152	*
MTG (D)	0.301	*	0.337	*	7.0	0.174	*	0.444	*	0.675	*	10.0	0.256	*
MTG (F)	0.304	*	0.376	*	9.0	0.206	*	0.441	*	0.674	*	9.0	0.267	*
IGP (D)	0.315	*	0.368	*	9.0	0.177	*	0.341	*	0.415	*	7.5	0.153	*
IGP (F)	0.314	*	0.370	*	9.0	0.183	*	0.343	*	0.414	*	7.5	0.156	*
	SYD- AUTM							WIND						
GGP (D)	0.284		**0.257**		**2.5**	0.157		0.454	*	0.670	*	2.5	0.282	*
GGP (F)	0.298	*	0.286	*	6.0	**0.142**	*	**0.450**		**0.661**		**1.0**	**0.281**	
LCM (D)	0.310	*	0.273	*	6.5	0.180	*	0.465	*	0.675	*	4.0	0.305	*
LCM (F)	0.302	*	0.257		5.5	0.178	*	0.465	*	0.677	*	5.0	0.308	*
GPRN (D)	0.281	*	0.323	*	3.5	0.185	*	0.460	*	0.702	*	3.5	0.308	*
GPRN (F)	0.284		0.326	*	5.5	0.187	*	0.455	*	0.693	*	5.0	0.306	*
MTG (D)	**0.280**		0.342	*	5.0	0.207	*	0.474	*	0.735	*	10.0	0.348	*
MTG (F)	0.283		0.360	*	6.5	0.219	*	0.472	*	0.728	*	8.0	0.336	*
IGP (D)	0.286		0.340	*	7.5	0.204	*	0.472	*	0.721	*	7.0	0.335	*
IGP (F)	0.286		0.335	*	6.5	0.202	*	0.473	*	0.728	*	9.0	0.346	*

M- RANK is mean of RMSE and NLPD ranks and F- VAR is mean forecast variance. Lower values indicate better performance for all measures. * indicates significant difference from best GGP model ((D) or (F)) based on 95% credible interval

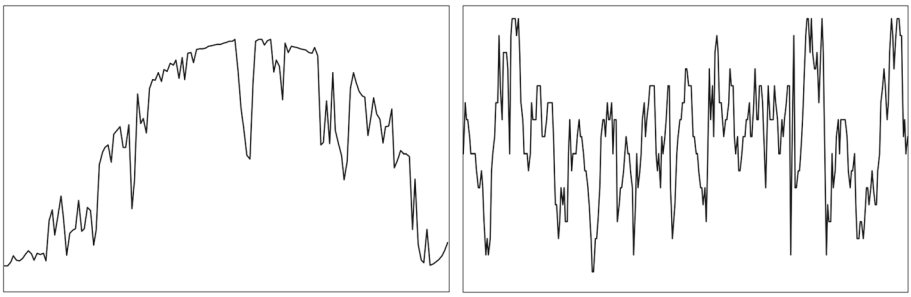

Fig. 2 Example data for normalized solar power (ADEL- SUMM—left hand side) and normalized wind speed (WIND—right hand side). Wind data exhibit strong noise relative to Summer time solar data

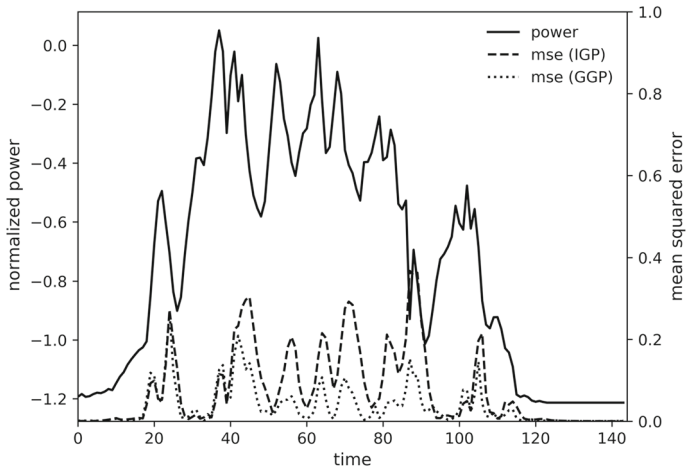

Fig. 3 Mean squared error under the GGP and IGP approaches for ADEL- AUTM. Results shown for a single day with variable cloud cover causing high variability in power output

number of inducing points, M to test models under equivalent settings, specifically setting $QM^3 = RM^3 = 18 \times (200)^3$.

Results for WIND are presented at Table 4 with diagonal and full-Gaussian posterior specifications. On this dataset GGP outperforms all other models on all measures including point accuracy (NLPD and RMSE), overall accuracy as measured by mean model ranking across both RMSE and NLPD, and forecast variance. Consistent reductions in variance are observed for the WIND dataset, ranging from 7 to 25% improvements over competing models. As for solar, confidence intervals are constructed via Monte Carlo. For WIND, all differences in model performance are confirmed to be statistically significant.

Comparison to approach of Remes et al. (2017) In addition to the above benchmarks, estimated using the generic sparse, variational inference framework, we also consider the approach of Remes et al. (2017). Since this method is a variational approach with complexity of $\mathcal{O}\left((QN)^3\right)$, which does not use inducing points, in order to fit a model under equivalent complexity conditions, we take a subset of the training data such that $(QN)^3$ approximates the settings above. We estimate a model for the WIND dataset, which has a manageable number

of tasks. We set $Q = 2$. Equivalent complexity would imply $N = 66$ for WIND, however we limit the minimum data size at $N = 200$.

We utilize the model implementation made available by the authors and allow all parameters to be optimized[2] The model gradient was optimized over 50 iterations, repeated ten times using different random parameter initial values. The model with the best performance (lowest objective function) was used to generate predictions.

The estimated value for RMSE was 0.88 for the WIND dataset, significantly higher than results under the GGP.

9 Timed experiments

To further examine the properties of the GGP model in relation to existing scalable multi-task methods, we conduct a series of timed experiments. We re-estimate models for the same forecasting problems as presented at Sect. 8 and, for each epoch completed, capture time and performance measures at that point. The goal of the analysis is to evaluate time taken for the GGP approach to achieves gains in forecast accuracy relative to the independent GPRN and other benchmarks, as well as final forecast performance attained upon completion.

We reiterate that, as for experiments presented at Sect. 8, the number of inducing points for each model is set to approximately standardize computational complexity per iteration.

All models are estimated on a multi-GPU machine with four NVIDIA TITAN Xp graphics cards (memory per card 12 GB; clock rate 1.58 GHz). Experiments were run until either convergence criteria were reached (see Sect. 8), or to a maximum of 500 epochs or 300 min runtime (these constraints were set conservatively based on previous experimental results). Starting values for common components were set to be equal across all models. Optimization settings were as for all other experiments.

9.1 Results of timed experiments

Representative rates of improvement in performance measures over time are presented at Fig. 4 for two datasets, WIND and ADEL- SUMM. These datasets were selected since results for ADEL- AUTM and SYD- AUTM were similar to those for WIND. Results are shown for all multi-task benchmark approaches with full variational posterior specifications (similar results were obtained for the diagonal posterior setting). Performance metric values shown are recorded at the end of each epoch (hence the first value for each model is recorded at different times, being the time taken to estimate the initial epoch) and adjusted for calculation time for performance capture.

For performance at a given point in time, results suggest a consistent ranking across models tested. We observed that GGP achieves higher forecast accuracy significantly faster than GPRN in the majority of cases, with a few cases performing similarly to GPRN. Specifically, for all datasets except ADEL- SUMM, RMSE reduces significantly faster under the GGP method relative to the GPRN, and NLPD for the GGP surpasses GPRN relatively early in the optimization. Relative rates of improvement in RMSE and NLPD as shown for WIND at Fig. 4 provide a typical example of the performance difference between the two models.

[2] The Matlab implementation of the model presented in Remes et al. (2017) is available at https://github.com/sremes/wishart-gibbs-kernel. Models presented in Remes et al. (2017) for the GPRN likelihood were estimated using fixed parameter values for weight latent functions.

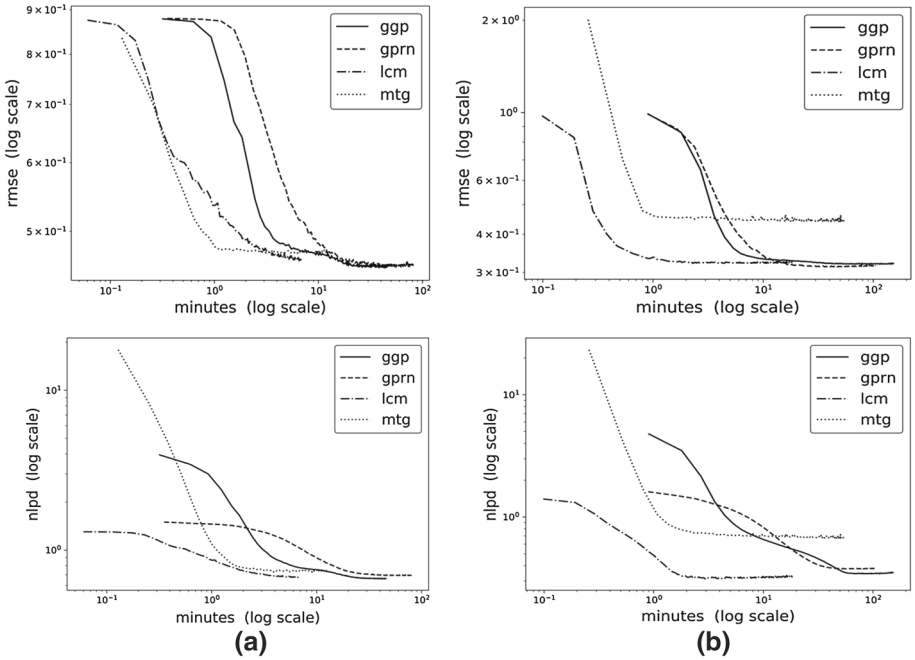

Fig. 4 Forecast accuracy for WIND (**a**) and ADEL- SUMM (**b**) datasets over optimization time in minutes for all multi-task benchmark models. RMSE and NLPD are recorded after each epoch during optimization

In terms of final accuracy, over the four datasets considered, results confirm the general rankings of final model accuracy as shown in Table 4. Specifically, GPRN achieves the best accuracy in terms of RMSE in two of four cases (ADEL- AUTM and ADEL- SUMM), while GGP improves on GPRN in two cases (SYD- AUTM and WIND, with MTG performing best on SYD- AUTM). Considering both speed and final accuracy together, GGP dominates the GPRN, achieving lower RMSE and NLPD in a significantly shorter time than GPRN in the majority of cases. In some cases, GPRN after some time will overtake GGP to achieve a slightly better result on RMSE, however in no case achieves a better result on NLPD. The results of timed experiments are therefore consistent with an improvement over the GPRN in terms of speed of convergence without loss of accuracy in terms of NLPD, and minor loss of accuracy in terms of RMSE.

With respect to the LCM and MTG models, these methods achieve improvements in forecast accuracy significantly more quickly than the GGP and GPRN. Consistent with results shown at Table 4, LCM achieves lower or similar NLPD to the GGP, with GGP outperforming LCM in two of four cases (SYD- AUTM and WIND). However, we note that as show in Fig. 4, the LCM converges relatively prematurely, and never achieves GGP or GPRN performance on RMSE. A similar phenomenon was observed to a greater degree for the MTG, which converges quickly but achieves poor accuracy relative to other models on both RMSE and NLPD, the exception being RMSE for the SYD- AUTM dataset.

Across the datasets considered, the GGP approach tends to achieve better forecast accuracy than the LCM where data are noisier, consistent with improved accuracy where the data require a more expressive model than the (fixed-weight) LCM approach. For example, the ADEL-SUMM dataset has significantly less noise relative to Autumn and wind datasets (Fig. 2).

Consequently, GPRN, LCM and GGP all perform similarly for the ADEL- SUMM dataset in terms of final accuracy, but LCM is significantly faster, suggesting there is little advantage from a more costly, expressive model such as GPRN or GGP. In contrast, for noisier datasets, GGP and GPRN continue to improve over LCM, and GGP does so at a faster rate than the GPRN. Figure 4 illustrates the typical relative accuracy over time of multitask models.

10 Discussion

We have proposed a general multi-task GP model, where groups of functions are coupled *a priori*. Our approach allows for input-varying covariance across tasks governed by kernels and features and, by building upon sparse variational methods and exploiting Kronecker structures, our inference method is inherently scalable to a large number of observations.

We have shown the applicability of our approach to forecasting short term distributed solar power and wind speed at multiple locations, where it matches or improves point forecast performance of single-task learning approaches and other multi-task baselines under similar computational constraints while improving quantification of predictive variance. We have also demonstrated that our approach can yield important reductions in time taken to achieve the same accuracy relative to the equivalent model without coupled priors. In general, the GGP strikes a balance between flexible, task-specific parameterization and effective regularization via structure imposed in the prior.

While we focus on *a priori* spatial dependencies, we emphasize that other grouping structures and kernels, likelihood functions or applications are possible. For example, non-spatial covariates in other domains, or grouping of functions according to clusters of tasks, could be adopted.

Acknowledgements This research was conducted with support from the Cooperative Research Centre for Low-Carbon Living in collaboration with the University of New South Wales and Solar Analytics Pty Ltd.

References

Álvarez, M., & Lawrence, N. D. (2009). Sparse convolved Gaussian processes for multi-output regression. In *Neural Information Processing Systems*.

Álvarez, M. A., & Lawrence, N. D. (2011). Computationally efficient convolved multiple output Gaussian processes. *Journal of Machine Learning Research*, *12*(5), 1459–1500.

Álvarez, M. A., Luengo, D., Titsias, M. K., & Lawrence, N. D. (2010). Efficient multioutput Gaussian processes through variational inducing kernels. In *Artificial Intelligence and Statistics*.

Álvarez, M. A., Rosasco, L., & Lawrence, N. D. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, *4*(3), 195–266.

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., de Pison, F. M., & Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. *Solar Energy*, *136*, 78–111.

Bilionis, I., Constantinescu, E. M., & Anitescu, M. (2014). Data-driven model for solar irradiation based on satellite observations. *Solar Energy*, *110*, 22–38.

Bonilla, E. V., Chai, K. M. A., & Williams, C. K. I. (2008). Multi-task Gaussian process prediction. In *Neural Information Processing Systems*.

Bonilla, E. V., Krauth, K., & Dezfouli, A. (2016). Generic inference in latent Gaussian process models. arxiv:1609.00577.

Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Hoboken: Wiley.

Dahl, A. & Bonilla, E. (2017). Scalable Gaussian process models for solar power forecasting. In W. L. Woon, Z. Aung, O. Kramer, & S. Madnick (Eds.), *Data analytics for renewable energy integration: Informing the generation and distribution of renewable energy* (pp. 94–106). Springer International Publishing.

Dezfouli, A., & Bonilla, E. V. (2015). Scalable inference for Gaussian process models with black-box likelihoods. In *Neural Information Processing Systems*.
Gardner, J. R., Pleiss, G., Wu, R., Weinberger, K. Q., & Wilson, A. G. (2018). Product kernel interpolation for scalable Gaussian processes. arXiv:1802.08903.
Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford: Oxford University Press.
Hensman, J., Matthews, A., & Ghahramani, Z. (2015). Scalable variational Gaussian process classification. In *AISTATS*.
Hensman, J., Rattray, M., & Lawrence, N. D. (2014). Fast nonparametric clustering of structured time-series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(2), 383–393.
Inman, R. H., Pedro, H. T., & Coimbra, C. F. (2013). Solar forecasting methods for renewable energy integration. *Progress in Energy and Combustion Science*, *39*(6), 535–576.
Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1998). *An introduction to variational methods for graphical models*. Berlin: Springer.
Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*. arXiv:1412.6980.
Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. *International Conference on Learning Representations*.
Krauth, K., Bonilla, E. V., Cutajar, K., & Filippone, M. (2017). AutoGP: Exploring the capabilities and limitations of Gaussian process models. In *Uncertainty in Artificial Intelligence*.
Matthews, A. G. de G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., et al. (2017). GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, *18*(40), 1–6.
Nguyen, T. V. & Bonilla, E. V. (2014). Automated variational inference for Gaussian process models. In *Neural Information Processing Systems*.
Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge: The MIT Press.
Remes, S., Heinonen, M., & Kaski, S. (2017). A mutually-dependent Hadamard kernel for modelling latent variable couplings. In M.-L. Zhang & Y.-K. Noh (Eds.), *Proceedings of the Ninth Asian Conference on Machine Learning*, vol. 77 of *Proceedings of Machine Learning Research*, (pp. 455–470).
Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*.
Sampson, P. D., & Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, *87*(417), 108–119.
Shinozaki, K., Yamakawa, N., Sasaki, T., & Inoue, T. (2016). Areal solar irradiance estimated by sparsely distributed observations of solar radiation. *IEEE Transactions on Power Systems*, *31*(1), 35–42.
Teh, Y. W., Seeger, M., & Jordan, M. I. (2005). Semiparametric latent factor models. In *Artificial Intelligence and Statistics*.
Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *AISTATS*.
Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., et al. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, *105*, 569–582.
Widén, J., Carpman, N., Castellucci, V., Lingfors, D., Olauson, J., Remouit, F., et al. (2015). Variability assessment and forecasting of renewables: A review for solar, wind, wave and tidal resources. *Renewable and Sustainable Energy Reviews*, *44*, 356–375.
Wilson, A. G., Knowles, D. A., & Ghahramani, Z. (2012). Gaussian process regression networks. In *International Conference on Machine Learning*.
Yang, D., Gu, C., Dong, Z., Jirutitijaroen, P., Chen, N., & Walsh, W. M. (2013). Solar irradiance forecasting using spatial-temporal covariance structures and time-forward kriging. *Renewable Energy*, *60*, 235–245.
Yang, D., Kleissl, J., Gueymard, C. A., Pedro, H. T., & Coimbra, C. F. (2018). History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. *Solar Energy*, *168*, 60–101.
Yang, D., Ye, Z., Lim, L. H. I., & Dong, Z. (2015). Very short term irradiance forecasting using the lasso. *Solar Energy*, *114*, 314–326.