Appendix: Nonstationary Nonseparable Random Fields

Kangrui Wang [1] **Oliver Hamelijnck** [1 2] **Theodoros Damoulas** [1 2] **Mark Steel** [2]

1. Proof of positive semi-definiteness for proposed kernels

In this section we prove positive semi-definiteness (PSD) for all our proposed kernels. We assume that all convolution kernels $K(\cdot)$ are positively bounded ($K(\cdot) > 0$) , $\int_{R^P} K(u)du < \infty$ and $\int_{R^P} K^2(u)du < \infty$ and the mixing distributions are valid probability distributions .
From the definition of a kernel we write: (Steinwart & Scovel, 2012):

$$\begin{aligned} C'(\mathbf{s}, \mathbf{s}'|\mathbf{a}) &= \langle \Phi(\mathbf{s}|\mathbf{a}), \Phi(\mathbf{s}'|\mathbf{a}) \rangle \\ C'(\mathbf{t}, \mathbf{t}'|\mathbf{b}) &= \langle \Phi(\mathbf{t}|\mathbf{b}), \Phi(\mathbf{t}'|\mathbf{b}) \rangle \end{aligned} \tag{1}$$

where $\mathbf{a} = \{a_1, ..., a_p\}$ and $\mathbf{b} = \{b_1, ..., b_p\}$ are the mixing variables. Thus, we know for any $\mathbf{x}$:

$$\begin{aligned} \mathbf{x}C'(\mathbf{s}, \mathbf{s}|\mathbf{a})\mathbf{x}^T &= \sum_i \sum_j \mathbf{x}_i C'(\mathbf{s}_i, \mathbf{s}_j|\mathbf{a})\mathbf{x}_j \\ &= \sum_i \sum_j \mathbf{x}_i \langle \Phi(\mathbf{s}_i|\mathbf{a}), \Phi(\mathbf{s}_j|\mathbf{a}) \rangle \mathbf{x}_j \\ &= \langle \sum_i \mathbf{x}_i \Phi(\mathbf{s}_i|\mathbf{a}), \sum_j \mathbf{x}_j \Phi(\mathbf{s}_j|\mathbf{a}) \rangle \\ &= || \sum_i \mathbf{x}_i \Phi(\mathbf{s}_i|\mathbf{a})||^2 \quad \geq 0 \\ \mathbf{x}C'(\mathbf{t}, \mathbf{t}|\mathbf{b})\mathbf{x}^T &= || \sum_i \mathbf{x}_i \Phi(\mathbf{t}_i|\mathbf{b})||^2 \quad \geq 0 \end{aligned} \tag{2}$$

which shows that both $C'(\mathbf{s}, \mathbf{s}'|\mathbf{a})$ and $C'(\mathbf{t}, \mathbf{t}'|\mathbf{b})$ are PSD. We next show that the the marginalized kernel:

$$\begin{aligned} &C(\mathbf{s}, \mathbf{s}', \mathbf{t}, \mathbf{s}') \\ &\quad = \int \int K(\mathbf{s} - \mathbf{u})K(\mathbf{s}' - \mathbf{u})K(\mathbf{t} - \mathbf{v})K'(\mathbf{t}' - \mathbf{v}) \\ &\qquad \int \int C'(\mathbf{s}, \mathbf{s}'|\mathbf{a})C'(\mathbf{t}, \mathbf{t}'|\mathbf{b})d\mu(\mathbf{a}, \mathbf{b})d\mathbf{u}d\mathbf{v} \end{aligned} \tag{3}$$

is also PSD. For any vector $\mathbf{x}$ we have:

$$\begin{aligned} &\mathbf{x}C(\mathbf{s}, \mathbf{s}, \mathbf{t}, \mathbf{t})\mathbf{x}^T \\ &\quad = \int \int K(\mathbf{s} - \mathbf{u})K(\mathbf{s} - \mathbf{u})K'(\mathbf{t} - \mathbf{v})K'(\mathbf{t} - \mathbf{v}) \\ &\qquad \int \int \mathbf{x}C'(\mathbf{s}, \mathbf{s}|\mathbf{a})C'(\mathbf{t}, \mathbf{t}|\mathbf{b})\mathbf{x}d\mu(\mathbf{a}, \mathbf{b})d\mathbf{u}d\mathbf{v} \\ &\quad = \int \int K(\mathbf{s} - \mathbf{u})K(\mathbf{s}' - \mathbf{u})K'(\mathbf{t} - \mathbf{v})K'(\mathbf{t} - \mathbf{v}) \\ &\qquad \mathbb{E}\left[|| \sum_i \mathbf{x}_i \Phi(\mathbf{s}_i|\mathbf{a})\Phi(\mathbf{t}_i|\mathbf{b})||^2 \right]_{p(\mathbf{a},\mathbf{b})} d\mathbf{u}d\mathbf{v} \quad \geq 0 \end{aligned} \tag{4}$$

where the last line follows from the fact that the expectation of positive elements is positive and the convolution function is bounded such that $K(\cdot) > 0$. This completes the PSD proof for our proposed kernels.

2. Weight-space view of stochastic processes

2.1. Weight-space view

In this section we provide an alternative presentation of stochastic process through the weight-space view (Rasmussen & Williams, 2005). This may be viewed as a discrete form to the continuous convolutions presented in the main paper and the same idea and methodology is used to achieve similar properties. The weight-space view describes a stochastic process as linear regression in a feature space such that the 'kernel trick' may be used to recover the function-space view.

In the weight-space view the function f is assumed to have a linear form

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{w}^T \Phi(\mathbf{x}) \\ &= \sum_{p=1}^{P} w_p \phi_p(\mathbf{x}) \end{aligned} \tag{5}$$

where $\Phi = \{\phi_1, ..., \phi_P\}$ is the feature map that maps from the D dimensional input into a P dimensional feature space and $\mathbf{w}$ is a zero mean Gaussian random variable with covariance Σ such that $\mathbf{w} \sim \mathcal{N}(0, \Sigma)$. The covariance of f is then given by

$$\begin{aligned} \text{Cov}(f(\mathbf{x}), f(\mathbf{x}')) &= \mathbf{E}[f(\mathbf{x})f(\mathbf{x}')] \\ &= \mathbf{E}\left[\sum_{i=1}^{P} w_i \phi_i(\mathbf{x}) \sum_{j=1}^{P} w_j \phi_j(\mathbf{x}')\right] \\ &= \sum_{i=1}^{P}\sum_{j=1}^{P} \phi_i(\mathbf{x}) \mathbf{E}[w_i w_j] \phi_j(\mathbf{x})') \\ &= \Phi(\mathbf{x}) \Sigma_w \Phi(\mathbf{x}')^T \end{aligned} \tag{6}$$

At this point we may define a kernel from the feature map $\Phi(\mathbf{x})\Sigma_w^{\frac{1}{2}}$ allowing us to apply the kernel trick and recover the function space view. Instead we define a kernel K using the feature map Φ. The kernel K is defined as an inner product between feature maps $\Phi : \mathcal{X} \rightarrow \mathcal{H}$, where $\mathcal{X}$ is the input space and $\mathcal{H}$ is an Hilbert space:

$$K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{H}}. \tag{7}$$

Rewriting Eq. 6 with this covariance:

$$\text{Cov}(f(\mathbf{x}), f(\mathbf{x}')) = K(\mathbf{x}) \Sigma_w K(\mathbf{x}')^T \tag{8}$$

where $K(\mathbf{x}) = \Phi(\mathbf{x})$. The rest of the sections follow by recognising Eq. 5 as a discrete convolution with a latent variable $\mathbf{x}_p$, and we will use $K_p(\mathbf{x})$ and $K(\mathbf{x}, \mathbf{x}_p)$ interchangeably.

2.2. Stationary processes

In this section we give the conditions for when the covariance in Eq. 6 is stationary. For stationarity to hold both it is necessary that both K and Σ be stationary kernels. A standard example is given by the squared exponential (SQE) kernel which we can derive by following the derivation given in (Rasmussen & Williams, 2005). Recall that f has a linear form:

$$f(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) \tag{9}$$

and when $w_i \sim iid$ with variance σ and $K(\mathbf{x}, \mathbf{x}_i) = exp(-\frac{(\mathbf{x}-\mathbf{x}_i)^2}{2\ell^2})$ and letting $I \rightarrow \infty$ the squared exponential covariance is recovered:

$$\begin{aligned} C(\mathbf{x}, \mathbf{x}') =& \mathbf{K}(\mathbf{x})[\sigma^2 \mathbf{I}] \mathbf{K}(\mathbf{x}')^T \\ =& \sum_{i=1}^{\infty} K(\mathbf{x}, \mathbf{x}_i) \sigma^2 K(\mathbf{x}', \mathbf{x}_i) \\ =& \sqrt{\pi} \ell \sigma^2 exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^2}{2(\sqrt{2}\ell)^2}\right) \end{aligned} \tag{10}$$

2.3. Nonstationarity processes

In this section we present the generalisation to a nonstationary process. A nonstationary covariance is one that cannot be written as a function of $\tau = \mathbf{x} - \mathbf{x}'$. This implies that the covariance of $f(\mathbf{x})$ depends on the input $\mathbf{x}$ in some way. Recall the definition of $f(\mathbf{x})$:

$$f(\mathbf{x}) = \sum_{i=1}^{P} w_i K_i(\mathbf{x}) \tag{11}$$

where $K_i(x) = \Phi_i(x)$. Introducing non-stationarity can be done in two ways:

- Define the random weights w_i to be parameterised by the input space, such that it is a function of $\mathbf{x}_i$:

  $$f(\mathbf{x}) = \sum_{i=1}^{P} w_i(\mathbf{x}_i) K_i(\mathbf{x}) \tag{12}$$

  where nonstationarity follows by recognising this is the same form as the the non-stationary kernel developed by (Fuentes & Smith, 2001) through a combination of locally stationary processes.

- Define the kernel K to be non-stationary:

  $$f(\mathbf{x}) = \sum_{i=1}^{P} w_i K_{\mathbf{x}}(\mathbf{x}, \mathbf{x}_i) \tag{13}$$

A general form that describes both approaches is:

$$f(\mathbf{x}) = \sum_{i=1}^{P} w_i(\mathbf{x}) K_{\mathbf{x}}(\mathbf{x}, \mathbf{x}_i) \tag{14}$$

where by Eq. 6 the covariance is given by:

$$C(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{P} \sum_{j=1}^{P} K_{\mathbf{x}}(\mathbf{x}, \mathbf{x}_i) \mathbf{E}[w_i(\mathbf{x}) w_j(\mathbf{x}')] K_{\mathbf{x}'}(\mathbf{x}', \mathbf{x}_j) \tag{15}$$

Thus, only if both the kernel function $K(\mathbf{x}, \mathbf{x}_i)$ and the random mapping w are stationary, then the function $f(\mathbf{x})$ will have a stationary covariance.

2.4. Separable Processes

In this section we present separable stochastic processes that are defined as having a separable covariance. Let $\mathbf{x} = \{s, t\}$ then a separable function can be written as $f(\mathbf{x}) = f(s)f(t)$ such that:

$$f(s,t) = f(s)f(t) = \left(\sum_{i=1}^{P} w_i^{(s)} K(s, s_i) \right) \cdot \left(\sum_{j=1}^{P} w_j^{(t)} K(t, t_j) \right) \tag{16}$$

where $(w^{(s)}, w^{(t)}) \sim \mathcal{N}(0, \Sigma)$ such that $\Sigma = \text{blkdiag}(\Sigma_1, \cdots, \Sigma_P)$ is block diagonal and each of the blocks Σ_p describes the covariance between $w_p^{(s)}$ and $w_p^{(t)}$. In general the number of basis function for $f(s)$ and $f(t)$ may be different (i.e different kernels may used for different dimensions) but for notational simplicity we will assume that they have the same number. We now show that for f to be separable then $w_p^{(s)}$ and $w_p^{(t)}$ must independent, implying Σ is diagonal. The covariance of f is given by

$$\begin{aligned} \text{Cov}(f(s,t), f(s',t')) &= \mathbf{E}[f(s,t)f(s't')] \\ &= \mathbf{E}\left[\left(\sum_{i=1}^{P} w_i^{(s)} K(s, s_i) \right) \cdot \left(\sum_{j=1}^{P} w_j^{(t)} K(t, t_j) \right) \left(\sum_{i=1}^{P} w_i^{(s)} K(s', s_i) \right) \cdot \left(\sum_{j=1}^{P} w_j^{(t)} K(t', t_j) \right) \right] \end{aligned} \tag{17}$$

which can be simplified through the block diagonal structure of Σ:

$$\begin{aligned} C(s,t,s',t') &= \mathbf{E}[f(s,t)f(s't')] \\ &= \sum_{i=1}^{P}\sum_{j=1}^{P} K(s,s_i)K(t,t_j)\mathbf{E}[w_i^{(s)}w_i^{(s)}w_j^{(t)}w_j^{(t)}]K(s',s_i)K(t',t_j) \end{aligned} \tag{18}$$

Thus for f to separable this covariance must simplify into a product of covariances between s and t. This only happens when $w^{(s)}$ and $w^{(t)}$ are independent, resulting in::

$$\begin{aligned} C(s,t,s',t') &= \sum_{i=1}^{P}\sum_{j=1}^{P} K(s,s_i)K(t,t_j)\mathbf{E}[w_i^{(s)}w_i^{(s)}]\mathbf{E}[w_j^{(t)}w_j^{(t)}]K(s',s_i)K(t',t_j) \\ &= \sum_{i=1}^{P} K(s,s_i)\mathbf{E}[w_i^{(s)}w_i^{(s)}]K(s',s_i)\sum_{j=1}^{P} K(t,t_j)\mathbf{E}[w_j^{(t)}w_j^{(t)}]K(t',t_j) \\ &= C(s,s')C(t,t') \end{aligned} \tag{19}$$

From eqn 19 we have shown that $f(s,t)$ is only separable if the random variables $w^{(s)}, w^{(t)}$ are independent because then the kernel function $K(s,t) = K(s)K(t)$ is separable.

3. Weight-space view of Stochastic mixture processes

3.1. Non-separable processes via a mixture of random variables

In this section we generalise from separable to non-separable processes. A non-separable process cannot be written as a product over input dimensions and we follow the mixing of variable construction shown in section 2 of the main paper to achieve this. Following section 2 we write a non-separable process as:

$$f(s,t) = \int f(s|a)f(t|b)p(a,b)dadb \tag{20}$$

where $f(s|a), f(t|b)$ are two (conditionally independent) stochastic functions and $p(a,b)$ induces correlation between the dimensions s,t. In general the conditional functions may be expressed in the same way as Eq. 14:

$$f(s|a) = \sum_{i=1}^{P} w_i(a)K_s(s,s_i|a) \tag{21}$$

where $w_i(a)$ is a random variable with parameter a and $K_s(s,s_i|a) = \langle \Phi_s(s|a), \Phi_s(s_i|a)\rangle$ represents a kernel with a parameter a. The function $f(s,t)$ may be written as an expectation:

$$\begin{aligned} f(s,t) &= \mathbf{E}_{p(a,b)}\left[f(s|a)f(t|b)\right] \\ &= \mathbf{E}_{p(a,b)}\left[\sum_{i=1}^{P} w_i^{(s)}(a)K_s(s,s_i|a)w_i^{(t)}(b)K_t(t,t_i|b)\right] \\ &= \sum_{i=1}^{P}\mathbf{E}_{p(a,b)}\left[w_i^{(s)}(a)K_s(s,s_i|a)w_i^{(t)}(b)K_t(t,t_i|b)\right] \end{aligned} \tag{22}$$

where $a = \{a_1, ..., a_P\}$ and $b_s = \{b_1, ..., b_P\}$. The covariance of $f(s,t), f(s',t')$ with zero mean and identical distributed w^s, w^t is:

$$\begin{aligned}
\text{Cov}(f(s,t), f(s't')) &= \mathbf{E}_f[f(s,t), f(s't')] \\
&= \mathbf{E}_f[\mathbf{E}_{p(a,b)}\left[f(s|a)f(t|b)f(s'|a)f(t'|b)\right]] \\
&= \mathbf{E}_f\left[\sum_{i=1}^{P} \mathbf{E}_{p(a,b)}\left[w_i^{(s)}(a)K_s(s,s_i|a)w_i^{(t)}(b)K_t(t,t_i|b)\right]\right. \\
&\cdot \left.\sum_{i=1}^{P} \mathbf{E}_f\left[\mathbf{E}_{p(a,b)}\left[w_i^{(s')}(a)K_{s'}(s',s_i|a)w_i^{(t')}(b)K_{t'}(t',t_i|b)\right]\right]\right] \\
&= \sum_{i=1}^{P} \mathbf{E}_{p(a,b)}\left[K_s(s,s_i|a)K_t(t,t_i|b)\mathbf{E}_f[w_i^{(s)}(a)w_i^{(s')}(a_{s_i})w_i^{(t)}(b)w_i^{(t')}(b)]K_{s'}(s',s_i|a)K_{t'}(t',t_i|b)\right]
\end{aligned} \tag{23}$$

This recovers the separable case when $p(a,b) = p(a)p(b)$ because the expectation over a will only affect terms with s and the same applies for b and t.

3.2. Nonstationary Nonseparable processes via SPM:Nonstationary Convolutions

In this section we present a weight-space view of the proposed SPM:Nonstationary Convolution. Under this framework all the nonstationarity comes from the convolution and the mixing only introduces nonseparability. Under the weight-space view Eqn. 5 is viewed a discretised convolution. As shown in §2.3 there is 2 ways to make this nonstationary but we restrict it to the case where it is expressed only in K. Following our SPM:Nonstationary Convolution framework we can construct nonstationary nonseparable processes by mixing these nonstationary separable processes. In general we have

$$\begin{aligned}
f(s,t) &= \mathbf{E}_{a,b}\left[f(s|a)f(t|b)\right] \\
&= \sum_{i=1}^{P} \mathbf{E}_{a,b}\left[w_{(s)}i(a)K_s(s,s_i|a)w_{(t)}i(b)K_t(t,t_i|b)\right]
\end{aligned} \tag{24}$$

As in §2.3 we split this general equation into two cases where the nonseparability is expressed either through the random weights or the kernel $K_\cdot(\cdot)$.

3.2.1. Nonseparability through the random weights

In this approach, we let the random weight depend on the mixing variable. We can now simplify eqn 24:

$$\begin{aligned}
f(s,t) &= \mathbf{E}_{a,b}\left[f(s|a)f(t|b)\right] \\
&= \sum_{i=1}^{P} \mathbf{E}_{a,b}\left[w_i^{(s)}(a)K_s(s,s_i)w^{(t)_i}(b)K_t(t,t_i)\right] \\
&= \sum_{i=1}^{P} K_s(s,s_i)K_t(t,t_i)\mathbf{E}_{a,b}\left[w_i^{(s)}(a)w_i^{(t)}(b)\right]
\end{aligned} \tag{25}$$

And the covariance function is given as:

$$\begin{aligned}
&\text{Cov}(f(s,t), f(s't')) \\
&= \sum_{i=1}^{P} K_s(s,s_i)K_t(t,t_i)\mathbf{E}_{a,b}\left[w_i^{(s)}(a)w_i^{(t)}(b)w_i^{(s)}(a)w_i^{(t)}(b)\right]K_{s'}(s',s_i)K_{t'}(t',t_i) \\
&= \sum_{i=1}^{P} K_s(s,s_i)K_t(t,t_i)|\Sigma_{w^{(s)},w^{(t)}}|K_{s'}(s',s_i)K_{t'}(t',t_i)
\end{aligned} \tag{26}$$

Although the kernel function is separable for s,t, the final process $f(s,t)$ is still nonseparable as the random weights $w_i^{(s)}(u), w_i^{(t)}(v)$ are correlated to each other.

3.2.2. Nonseparability through the convolution kernel

In this approach, we let the convolution kernel depend on the mixing variable. We can now simplify eqn 24 as:

$$\begin{aligned} f(s,t) &= \mathbf{E}_{a,b}\left[f(s|a)f(t|b)\right] \\ &= \sum_{i=1}^{P} \mathbf{E}_{a,b}\left[w_i^{(s)} K_s(s,s_i|a) w_i^{(t)} K_t(t,t_i|b)\right] \\ &= \sum_{i=1}^{P} \mathbf{E}_{a,b}\left[K_s(s,s_i|a)K_t(t,t_i|b)\right] w_i^{(s)} w_i^{(t)} \end{aligned} \tag{27}$$

And the covariance function of $f(s,t)$ is given by:

$$\begin{aligned} &\text{Cov}(f(s,t), f(s',t')) \\ &= \sum_{i=1}^{P} \mathbf{E}_{a,b}[K_s(s,s_i|a)K_t(t,t_i|b)K_{s'}(s',s_i|a)K_{t'}(t',t_i|b)]\mathbf{E}[w_i^{(s)} w_i^{(s')}]\mathbf{E}[w_i^{(t)} w_i^{(t')}] \end{aligned} \tag{28}$$

In both approaches, no matter if the mixing variable is part of the kernel or weights w, we can define conditional covariance function as :

$$\begin{aligned} C(s,s'|a) &= \sum_{i=1}^{P} K_{\mathbf{s}}(s,s_i|a)\mathbf{E}_w[w_i^{(s)}(a)w_i^{(s)}(a)]K_{\mathbf{s}'}(s',s_i|a) \\ C(t,t'|b) &= \sum_{i=1}^{P} K_{\mathbf{t}}(t,t_i|b)\mathbf{E}_w[w_i^{(t)}(b)w_i^{(t)}(b)]K_{\mathbf{t}'}(t',t_i|b) \end{aligned} \tag{29}$$

such that the full covariance function can be written as:

$$C(s,s',t,t') = \mathbf{E}_{u,v}\left[C(s,s'|a)C(t,t'|b)\right] \tag{30}$$

3.3. Nonstationary Nonseparable processes via SPM:Nonstationary Mixing

In this approach the convolution is assumed to stationary and all of the the nonstationarity is modelled through the mixing of variables. This is achieved by allowing the mixing variable to depend on the input space. Under the SPM: Nonstationary Mixing construction the nonstationary nonseparable process f is given by:

$$\begin{aligned} f(s,t) &= \mathbf{E}_{\mathbf{a},\mathbf{b}}\left[f(s|\mathbf{a})f(t|\mathbf{b})\right] \\ &= \sum_{i=1}^{P} \mathbf{E}_{a_{s_i},b_{t_i}}\left[w_i^{(s)}(a_{s_i})K(s,s_i|a_{s_i}) w_i^{(t)}(b_{t_i})K(t,t_i|b_{t_i})\right] \end{aligned} \tag{31}$$

We can recognise that this is a sum of locally stationary (nonseparable) processes and so is a generalisation of Eqn. 12 and Fuentes & Smith (2001) to the nonstationary nonseparable nonstationary case. Again following §3.2 the mixing parameters may either parameterize the kernel function K or the random weights w. These two cases are:

- kernel mixture:

$$\begin{aligned} f(s,t) &= \mathbf{E}_{a,b}\left[f(s|a)f(t|b)\right] \\ &= \sum_{i=1}^{P} \mathbf{E}_{a_{s_i},b_{t_i}}\left[w_i^{(s)} K(s,s_i|a_{s_i}) w_i^{(t)} K(t,t_i|b_{t_i})\right] \\ &= \sum_{i=1}^{P} \mathbf{E}_{a_{s_i},b_{t_i}}\left[K(s,s_i|a_{s_i})K(t,t_i|b_{t_i})\right] w_i^{(s)} w_i^{(t)} \end{aligned} \tag{32}$$

And the covariance function is given by:

$$\begin{aligned}\mathrm{Cov}(f(s,t), f(s',t')) &= \mathbf{E}_{a,b}\left[f(s|a)f(t|b)f(s'|a)f(t'|b)\right] \\ &= \sum_{i=1}^{P} \mathbf{E}_{a_{s_i},b_{t_i}}\left[K(s,s_i|a_{s_i})K(t,t_i|b_{t_i})\right]\mathbf{E}[w_i^{(s)}w_i^{(t)}w_i^{(s)}w_i^{(t)}] \\ &\mathbf{E}_{a_{s_i},b_{t_i}}\left[K(s',s_i|a_{s_i})K(t',t_i|b_{t_i})\right]\end{aligned} \tag{33}$$

- random weight mixture:

$$\begin{aligned}f(s,t) &= \mathbf{E}_{a,b}\left[f(s|a)f(t|b)\right] \\ &= \sum_{i=1}^{P} \mathbf{E}_{a_{s_i},b_{t_i}}\left[w_i^{(s)}(a_{s_i})K(s,s_i)w_i^{(t)}(b_{t_i})K(t,t_i)\right] \\ &= \sum_{i=1}^{P} K(s,s_i)K(t,t_i)\mathbf{E}_{a_{s_i},b_{t_i}}\left[w_i^{(s)}(a_{s_i})w_i^{(t)}(b_{t_i})\right]\end{aligned} \tag{34}$$

and the covariance is given by:

$$\begin{aligned}\mathrm{Cov}(f(s,t), f(s',t')) = &\sum_{i=1}^{P} K(s,s_i)K(t,t_i)\mathbf{E}_{a_{s_i},b_{t_i}}\left[w_i^{(s)}(a_{s_i})w_i^{(t)}(b_{t_i})w_i^{(s)}(a_{s_i})w_i^{(t)}(b_{t_i})\right] \\ &K(s',s_i)K(t',t_i)\end{aligned} \tag{35}$$

4. Additional Experimental results

In this section we provide further experimental details and results. We provide the full form of the covariance constructions and provide insights into the learnt parameters. To compare between models and baselines we use our covariance functions within a Gaussian process (GP) framework.

4.1. Nonseparable compound function

In our first toy example we are interested in recovering the following nonstationary nonseparable surface:

$$\begin{aligned}f(\mathbf{s},t) &= \sin\left(3 \cdot (s_1^2 + s_1 \cdot (2-s_2)^2 + t)\right) + 2 \\ y(\mathbf{s},t) &= f(g(\mathbf{s}),t) + \epsilon\end{aligned} \tag{36}$$

where s_1, s_2 are the first and second input dimensions of s respectively, $g(\mathbf{s}) = \Sigma^{\frac{1}{2}}\mathbf{s}$ is the input warping function that provides additional non-seperability, $\Sigma = \begin{bmatrix}1 & 0.3\\ 0.3 & 1\end{bmatrix}$ and $\epsilon \sim \mathcal{N}(0, 0.1)$. To provide emphasis on the learning of nonseparability we fix the time dimension to be $t = 0$ and study the function of $f(s_1, s_2) = f(s_1, s_2|t = 0)$. We assume the process $f(s_1, s_2)$ follows a GP run experiments with various kernel functions.

4.1.1. Stationary Separable Covariance

We assume the covariance function for s_1 and s_2 are separable, $C(s_1, s_1', s_2, s_2') = C(s_1, s_1')C(s_2, s_2')$. For each separable covariance term, we use the squared exponential function:

$$C(x,x') = \exp\left(-\frac{(x-x')^2}{\ell^2}\right) \tag{37}$$

4.1.2. Nonstationary Separable Covariance

For the nonstationary approach, we assume each component of the separable covariance function, $C(s_1, s_1', s_2, s_2') = C(s_1, s_1')C(s_2, s_2')$, follows a Gibbs covariance function (Gibbs, 1998):

$$C(x,x') = \sigma^2\sqrt{\frac{2\ell(x)\ell(x')}{\ell(x)^2 + \ell(x')^2}}\exp(-\frac{(x-x')^2}{\ell(x)^2+\ell(x')^2}) \tag{38}$$

Where $\ell(x)$ is the local lengthscale parameter at input location x. For both s_1 and s_2, we assume the local lengthscale follow linear functions of s_1 and s_2 : $\ell(s_1) = as_1 + b$ and $\ell(s_2) = cs_2 + d$.

4.1.3. Stationary Nonseparable Covariance

We use the nonseparable covariance constructed by Fonseca & Steel (2011). We assume the process $f(s_1, s_2)$ is a mixture of stationary separable process $f(s_1, s_2) = \mathbb{E}[f(s_1|a)f(s_2|b)]$. We assume the conditional process $f(x|a)$ follows a GP with covariance function:

$$C(x, x'|a) = \exp\left(-\frac{a(x-x')^2}{\ell_x^2}\right) \tag{39}$$

Thus, the covariance function of $C(s_1, s_1', s_2, s_2') = \mathbb{E}[C(s_1, s_1'|a)C(s_2, s_2'|b)]$. We assume a and b are random variables: $a = \lambda_0 + \lambda_1$ amd $b = \lambda_0 + \lambda_2$, where $\lambda_0, \lambda_1, \lambda_2$ are independent gamma variables ($\lambda_x \sim \text{Ga}(\beta_x, 1)$). The nonseparable convariance is then given by:

$$\begin{aligned}
C(s_1, s_1', s_2, s_2') &= \mathbb{E}_{a,b}\big[C(s_1, s_1'|a)C(s_2, s_2'|b)\big] \\
&= \int\int C(s_1, s_1'|a)C(s_2, s_2'|b)p(a,b)\,da\,db \\
&= \sigma^2 \int\int \exp\left(-a\frac{(s_1-s_1')^2}{\ell_{s_1}}\right) \exp\left((-b\frac{(s_2-s_2')^2}{\ell_{s_2}}\right) p(a,b)\,da\,db \\
&= \sigma^2 \int\int\int \exp\left(-(\lambda_0+\lambda_1)\frac{(s_1-s_1')^2}{\ell_{s_1}}\right) \exp\left((-(\lambda_0+\lambda_2)\frac{(s_2-s_2')^2}{\ell_{s_2}}\right) p(\lambda_0)p(\lambda_1)p(\lambda_2)\,d\lambda_0\,d\lambda_1\,d\lambda_2 \\
&= \sigma^2 \int \exp\left(-\lambda_0\left(\frac{(s_1-s_1')^2}{\ell_{s_1}} + \frac{(s_2-s_2')^2}{\ell_{s_2}}\right)\right) p(\lambda_0)\,d\lambda_0 \cdot \int \exp\left(-\lambda_1\frac{(s_1-s_1')^2}{\ell_{s_1}}\right) p(\lambda_1)\,d\lambda_1 \\
&\int \exp\left(-\lambda_2\frac{(s_2-s_2')^2}{\ell_{s_2}}\right) p(\lambda_2)\,d\lambda_2
\end{aligned} \tag{40}$$

which can be simplified to:

$$C(s_1, s_1', s_2, s_2') = \sigma^2\left(1 + \frac{(s_1-s_1')^2}{\ell_{s_1}^2} + \frac{(s_2-s_2')^2}{\ell_{s_2}^2}\right)^{\beta_0}\left(1 + \frac{(s_1-s_1')^2}{\ell_{s_1}^2}\right)^{\beta_1}\left(1 + \frac{(s_2-s_2')^2}{\ell_{s_2}^2}\right)^{\beta_2} \tag{41}$$

where ℓ_{s_1} and ℓ_{s_2} are the lengthscales for the inputs s_1 and s_2, and is final form of the covariance function used.

4.1.4. SPM:nonstationary convolution

Following the construction presented in section 3, we have $f(s_1, s_2) = \mathbb{E}[f(s_1|a)f(s_2|b)]$, where $f(s_1|a)$ and $f(s_2|b)$ are nonstationary processes. We parameterize them with Gibbs covariances:

$$C(x, x'|a) = \sigma^2\sqrt{\frac{2\ell(x)\ell(x')}{\ell(x)^2 + \ell(x')^2}}\exp(-a\frac{(x-x')^2}{\ell(x)^2 + \ell(x')^2}) \tag{42}$$

To perform the mixing we use the same parameterisation as §4.1.3 which is results in the follow nonstationary nonseparable convariance:

$$\begin{aligned}
C(s_1, s_1', s_2, s_2') &= \sigma^2 A_{s_1,s_1'}A_{s_2,s_2'}(1 + Q_{s_1,s_1'} + Q_{s_2,s_2'})^{\beta_0}(1 + Q_{s_1,s_1'})^{\beta_1}(1 + Q_{s_2,s_2'})^{\beta_2} \\
Q_{x,x'} &= (x-x')\left(\frac{\ell(x)+\ell(x')}{2}\right)^{-1}(x-x') \\
A_{x,x'} &= \sqrt{\frac{2\ell(x)\ell(x')}{\ell(x)^2+\ell(x')^2}}
\end{aligned} \tag{43}$$

where the local lengthscales follow a linear function of s_1 and s_2: $\ell(s_1) = as_1 + b$ and $\ell(s_2) = cs_2 + d$, and all the nonstationarity arises from the convolution.

4.1.5. SPM:NONSTATIONARY MIXING

Following the construction presented in section 4 we create a nonstationary nonseparable function using a nonstationary mixing:

$$\begin{aligned} &C(s_1, s_1', s_2, s_2') \\ &= \int\int K(s_1 - u)K(s_1' - u)K(s_2 - v)K'(s_2' - v) \\ &\int\int C'(s_1, s_1'|a(u))C'(s_2, s_2'|b(v))\, d\mu(a, b)\, du\, dv \end{aligned} \tag{44}$$

We assume the convolution functions $K(s_1 - u), K(s_2 - v)$ are stationary and use an exponential convolution:

$$K(s_1 - u) = \exp\left(-\frac{(s_1 - u)^2}{\ell^2}\right) \tag{45}$$

and we assume the local stationary processes have the conditional covariance:

$$\begin{aligned} &C'(s_1, s_1'|a(u))C'(s_2, s_2'|b(v)) = \\ &\exp\left(-a(u)\frac{((s_1 - s_1')^2}{\ell_s}\right) \exp\left(b(v)\frac{((s_2 - s_2')^2}{\ell_t}\right)^s \end{aligned} \tag{46}$$

We construct the nonstationary mixing via a linear function of independent variables. Assume $\lambda_0 \sim$ Ga$(\beta_0, 1)$, $\lambda_0 \sim$ Ga$(\beta_1(u), 1)$ and $\lambda_2 \sim$ Ga$(\beta_2(v), 1)$, we have $a(u) = \lambda_0 + \lambda_1$ and $b(v) = \lambda_0 + \lambda_2$, where $\beta_1(u) = |au + b|$ and $\beta_2(v) = |cv + d|$ are the polynomial functions related to the location u and v . Thus, we have:

$$\begin{aligned} C_{u,v}(s_1, s_1', s_2, s_2') =&\mathbf{E}_{a,b}(C(s_1, s_1'|a(u))C(s_2, s_2'|b(v)) \\ =&\left(1 + \frac{(s_1 - s_1')^2}{\ell_{s_1}^2} + \frac{(s_2 - s_2')^2}{\ell_{s_2}^2}\right)^{\beta_0} \left(1 + \frac{(s_1 - s_1')^2}{\ell_{s_1}^2}\right)^{\beta_1(u)} \left(1 + \frac{(s_2 - s_2')^2}{\ell_{s_2}^2}\right)^{\beta_2(v)} \end{aligned} \tag{47}$$

which results in the final nonstationary nonseparable kernel.

4.2. Spatio-temporal experiments

In this section we provide additional details for the remaining spatio-temporal experiments. These experiments have two spatial inputs, s_1, s_2 and one time input t. For all models and baselines we run GP models with various kernels which we describe in full in the following sections. In the spatio-temporal heat equation, and two real world applications, London air quality and Irish wind data we construct use covariances of the same form; which we now present.

4.2.1. STATIONARY SEPARABLE COVARIANCE

We assume the covariance function for all three inputs are separable, $C(s_1, s_1', s_2, s_2', t, t') = C(s_1, s_1')C(s_2, s_2')C(t, t')$. For each separable covariance, we use the squared exponential covariance function:

$$C(x, x') = \exp\left(-\frac{(x - x')^2}{\ell^2}\right) \tag{48}$$

4.2.2. NONSTATIONARY SEPARABLE COVARIANCE

For the nonstationary approach, we assume each components of the separable covariance function,$C(s_1, s_1', s_2, s_2', t, t') = C(s_1, s_1')C(s_2, s_2')C(t, t')$, follows the Gibbs covariance function (Gibbs, 1998):

$$C(x, x') = \sigma^2 \sqrt{\frac{2\ell(x)\ell(x')}{\ell(x)^2 + \ell(x')^2}} \exp(-\frac{(x - x')^2}{\ell(x)^2 + \ell(x')^2}) \tag{49}$$

where $\ell(x)$ is the local lengthscale parameter at input location x. For all s_1, s_2 and t, we assume the local lengthscales follow linear function of s_1 , s_2 and t: $\ell(s_1) = as_1 + b$, $\ell(s_2) = cs_2 + d$ and $\ell(t) = et + f$.

4.2.3. Stationary nonseparable Covariance

We use the nonseparable covariance constructed by Fonseca & Steel (2011). We assume the process $f(s_1, s_2)$ is a mixture of stationary separable processes $f(s_1, s_2) = \mathbb{E}[f(s_1|a)f(s_2|b)]$. We assume the conditional process $f(x|a)$ follows a GP with covariance:

$$C(x, x'|a) = \exp\left(-\frac{a(x-x')^2}{\ell_x^2}\right) \tag{50}$$

Thus, the covariance function of $C(s_1, s_1', s_2, s_2') = \mathbb{E}[C(s_1, s_1'|a)C(s_2, s_2'|b)C(t, t'|h)]$. We assume the following mixing structure:

$$\begin{aligned} a &= \lambda_0 + \lambda_1 + \lambda_2 \\ b &= \lambda_0 + \lambda_1 + \lambda_3 \\ h &= \lambda_0 + \lambda_4 \end{aligned} \tag{51}$$

where $\lambda_0, ..., \lambda_4$ are independent gamma variables ($\lambda_x \sim \text{Ga}(\beta_x, 1)$). We can see that the time variable is assumed to have the same correlation across spatial locations, whereas the spatial inputs have extra dependency through λ_1. The resulting kernel is:

$$\begin{aligned} C(s_1, s_1', s_2, s_2', t, t') = \sigma^2 &\left(1 + \frac{(s_1 - s_1')^2}{\ell_{s_1}^2} + \frac{(s_2 - s_2')^2}{\ell_{s_2}^2} + \frac{(t-t')^2}{\ell_t^2}\right)^{\beta_0} \left(1 + \frac{(s_1 - s_1')^2}{\ell_{s_1}^2} + \frac{(s_2 - s_2')^2}{\ell_{s_2}^2}\right)^{\beta_1} \\ &\left(1 + \frac{(s_1 - s_1')^2}{\ell_{s_1}^2}\right)^{\beta_2} \left(1 + \frac{(s_2 - s_2')^2}{\ell_{s_2}^2}\right)^{\beta_3} \left(1 + \frac{(t-t')^2}{\ell_t^2}\right)^{\beta_4} \end{aligned} \tag{52}$$

where ℓ_{s_1} and ℓ_{s_2} are the lengthscales for the inputs.

4.2.4. SPM:nonstationary convolution

Following the construction presented in section 3, we have $f(s_1, s_2) = \mathbb{E}[f(s_1|a)f(s_2|b)]$, where $f(s_1|a)$ and $f(s_2|b)$ are nonstationary processes. We parameterize them with Gibbs covariances:

$$C(x, x'|a) = \sigma^2 \sqrt{\frac{2\ell(x)\ell(x')}{\ell(x)^2 + \ell(x')^2}} \exp(-a\frac{(x-x')^2}{\ell(x)^2 + \ell(x')^2}) \tag{53}$$

To perform the mixing we use the same parameterisation as §4.1.3 which is results in the following nonstationary nonseparable convariance:

$$\begin{aligned} C(s_1, s_1', s_2, s_2') &= \sigma^2 A_{s_1,s_1'} A_{s_2,s_2'} A_{t,t'} (1 + Q_{s_1,s_1'} + Q_{s_2,s_2'} + Q_{t,t'})^{\beta_0} (1 + Q_{s_1,s_1'} + Q_{s_2,s_2'})^{\beta_1} \\ &\quad (1 + Q_{s_1,s_1'})^{\beta_2} (1 + Q_{s_2,s_2'})^{\beta_3} (1 + Q_{t,t'})^{\beta_4} \\ Q_{x,x'} &= (x - x') \left(\frac{\ell(x) + \ell(x')}{2}\right)^{-1} (x - x') \\ A_{x,x'} &= \sqrt{\frac{2\ell(x)\ell(x')}{\ell(x)^2 + \ell(x')^2}} \end{aligned} \tag{54}$$

We assume the local lengthscales follow a linear function of s_1 and $s_2 : \ell(s_1) = as_1 + b$ and $\ell(s_2) = cs_2 + d$.

4.2.5. SPM:nonstationary mixing

Following the construction presented in section 4 we create a nonstationary nonseparable function using a nonstationary mixing:

$$\begin{aligned} &C(s_1, s_1', s_2, s_2', t, t') \\ &\quad = \int\int K_{s_1}(s_1 - u_1)K_{s_1}(s_1' - u_1)K_{s_2}(s_2 - u_2)K_{s_2}(s_2' - u_2)K_t(t - v)K_t(t - v) \\ &\quad\quad \int\int C'(s_1, s_1'|a(u_1))C'(s_2, s_2'|b(u_2))C'(t, t'|h(v))\, d\mu(a, b, h)\, du_1\, du_2\, dv \end{aligned} \tag{55}$$

We assume the convolution functions $K(s_1-u_1), K(s_2-u_2), K(t-v)$ are stationary and are parameterised as an exponential convolution:

$$K(s_1-u)=\exp\left(-\frac{(s_1-u)^2}{\ell^2}\right) \tag{56}$$

We assume the local stationary processes follow the conditional covariance:

$$\begin{aligned}&C'(s_1,s_1'|a(u_1))C'(s_2,s_2'|b(u_2))C'(t,t'|h(v))=\\&\quad\exp\left(-a(u_1)\frac{(s_1-s_1')^2}{\ell_{s_1}}\right)\exp\left(-b(u_2)\frac{(s_2-s_2')^2}{\ell_{s_2}}\right)\exp\left(-h(v)\frac{(t-t')^2}{\ell_t}\right)\end{aligned} \tag{57}$$

Thus, we can construct the nonstationary mixing via a linear function of independent variables:

$$\begin{aligned}a&=\lambda_0+\lambda_1+\lambda_2\\b&=\lambda_0+\lambda_1+\lambda_3\\h&=\lambda_0+\lambda_4\end{aligned} \tag{58}$$

Assume $\lambda_0\sim\text{Ga}(\beta_0,1)$, $\lambda_1\sim\text{Ga}(\beta_1,1)$ and $\lambda_2\sim\text{Ga}(\beta_2(u_1),1)$,$\lambda_3\sim\text{Ga}(\beta_3(u_2),1)$,$\lambda_4\sim\text{Ga}(\beta_4(t),1)$, then we have:

$$\begin{aligned}C_{u_1,u_2,v}(s_1,s_1',s_2,s_2',t,t')=&\mathbf{E}_{a,b}(C(s_1,s_1'|a(u))C(s_2,s_2'|b(v))\\=&\sigma^2\left(1+\frac{(s_1-s_1')^2}{\ell_{s_1}^2}+\frac{(s_2-s_2')^2}{\ell_{s_2}^2}+\frac{(t-t')^2}{\ell_t^2}\right)^{\beta_0}\left(1+\frac{(s_1-s_1')^2}{\ell_{s_1}^2}+\frac{(s_2-s_2')^2}{\ell_{s_2}^2}\right)^{\beta_1}\\&\left(1+\frac{(s_1-s_1')^2}{\ell_{s_1}^2}\right)^{\beta_2(u_1)}\left(1+\frac{(s_2-s_2')^2}{\ell_{s_2}^2}\right)^{\beta_3(u_2)}\left(1+\frac{(t-t')^2}{\ell_t^2}\right)^{\beta_4(v)}\end{aligned} \tag{59}$$

which results in the final nonstationary nonseparable kernel.

4.2.6. SPM:NONSTATIONARY MIXING VIA HEAT EQUATION

The spatio-temporal heat equation is an SDE that defines the dispersion of heat through an object:

$$\frac{df(s_1,s_2,t)}{dt}-D(\frac{d^2f(s_1,s_2,t)}{ds_1^2}+\frac{d^2f(s_1,s_2,t)}{ds_2^2})=\phi(s_1,s_2,t) \tag{60}$$

The fundamental solution is given by:

$$G(s_1,s_2,t)=\frac{1}{(4\pi Dt)}exp\left(-\frac{s_1^2+s_2^2}{4Dt}\right). \tag{61}$$

Construction the covariance of $f(s_1,s_2,t)$ we have:

$$\begin{aligned}&C(s_1,s_2,t,s_1',s_2',t')\\&=\int\int G(s_1-u_{s_1},s_2-u_{s_2},t-v)G(s_1'-u_{s_1},s_2'-u_{s_2},t'-v)C_{u_{s_1},u_{s_2},v}(s_1,s_2,t,s_1',s_2',t')du_{s_2}du_{s_1}dv\end{aligned} \tag{62}$$

For computational simplicity , we assume only spatial dependency exists in the latent structure, i.e. $C_{u_{s_1},u_{s_2},v}(s_1,s_2,t,s_1',s_2',t')=C_{u_{s_1},u_{s_2}}(s_1,s_2,s_1',s_2')C(t,t')$. We use the construction of Eq. 41 as our space dimension mixture.

4.3. Inference methods

To ensure the consistency for all experiments, we apply uniform priors to all parameters in kernels in §4. We optimize all hyperparameters through the variational ELBO set the number of inducing variables to be same as the observations.

4.3.1. VARIATIONAL APPROXIMATIONS

To make the experiments with the DGP fair we perform variational inference on all models for all experiments. We use the variational lowerbound presented in (Hensman et al., 2013) for the single GP models and for the DGP we use the doubly stochastic framework proposed by (Salimbeni & Deisenroth, 2017).

4.3.2. Monte Carlo integration

Because the SPM:NM kernel does not have a closed form, we require Monte Carlo integration to approximate the required integrals. Assuming $\mathbf{m}_{u_1} = \{m_{u_1}^{(1)}, ..., m_{u_1}^{(k)}\}$, $\mathbf{m}_{u_2} = \{\{m_{u_2}^{(1)}, ..., m_{u_2}^{(k)}\}\}$, $\mathbf{m}_v = \{m_v^{(1)}, ..., m_v^{(k)}\}$ are uniformly distributed in $[-1, 1]$, we have:

$$\begin{aligned}
&C(s_1, s_1', s_2, s_2', t, t') \\
&\quad= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} K_{s_1}(s_1 - u_1) K_{s_1}(s_1' - u_1) K_{s_2}(s_2 - u_2) K_{s_2}(s_2' - u_2) K_t(t - v) K_t(t - v) \\
&\quad\quad \mathbf{E}_{p(a,b,h)}[C'(s_1, s_1'|a(u_1)) C'(s_2, s_2'|b(u_2)) C'(t, t'|h(v))] \, du_1 \, du_2 \, dv \\
&\quad= \int_{-1}^{1} \int_{-1}^{1} \int_{-1}^{1} \frac{1 + m_{u_1}^2}{(1 - m_{u_1}^2)^2} \frac{1 + m_{u_2}^2}{(1 - m_{u_2}^2)^2} \frac{1 + m_v^2}{(1 - m_v^2)^2} K_{s_1}(s_1 - u_1) K_{s_1}(s_1' - u_1) K_{s_2}(s_2 - u_2) K_{s_2}(s_2' - u_2) \\
&\quad\quad K_t(t - v) K_t(t - v) \mathbf{E}_{p(a,b,h)}[C'(s_1, s_1'|a(u_1)) C'(s_2, s_2'|b(u_2)) C'(t, t'|h(v))] \, dm_{u_1} \, dm_{u_2} \, dm_v \\
&\quad\approx V \frac{1}{k^3} \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{q=1}^{k} \frac{1 + m_{u_1}^{(i)2}}{(1 - m_{u_1}^{(i)}{}^2)^2} \frac{1 + m_{u_2}^{(j)2}}{(1 - m_{u_2}^{(j)}{}^2)^2} \frac{1 + m_v^{(q)2}}{(1 - m_v^{(q)}{}^2)^2} K_{s_1}(s_1 - u_1^{(i)}) K_{s_1}(s_1' - u_1^{(i)}) K_{s_2}(s_2 - u_2^{(j)}) K_{s_2}(s_2' - u_2^{(j)}) \\
&\quad\quad K_t(t - v^{(q)}) K_t(t - v^{(q)}) \mathbf{E}_{p(a,b,h)}[C'(s_1, s_1'|a(u_1^{(i)})) C'(s_2, s_2'|b(u_2^{(j)})) C'(t, t'|h(v^{(q)}))]
\end{aligned} \tag{63}$$

where $u_1^{(i)} = \frac{m_{u_1}^{(i)}}{1 - m_{u_1}^{(i)2}}$, $u_2^{(j)} = \frac{m_{u_2}^{(j)}}{1 - m_{u_2}^{(j)2}}$ and $v^{(q)} = \frac{m_v^{(q)}}{1 - m_v^{(q)2}}$. Since we are using a radial basis function as our convolution function, most of the region falls into the tail part of the basis functions $K(\cdot)$ and returns very small values, i.e when far from data the convolution will be close to zero. Thus, we can approximate the infinite integral with a good understanding of training+prediction region:

$$\begin{aligned}
&C(s_1, s_1', s_2, s_2', t, t') \\
&\quad\approx V \frac{1}{k^3} \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{q=1}^{k} K_{s_1}(s_1 - u_1^{(i)}) K_{s_1}(s_1' - u_1^{(i)}) K_{s_2}(s_2 - u_2^{(j)}) K_{s_2}(s_2' - u_2^{(j)}) \\
&\quad\quad K_t(t - v^{(q)}) K_t(t - v^{(q)}) \mathbf{E}_{p(a,b,h)}[C'(s_1, s_1'|a(u_1^{(i)})) C'(s_2, s_2'|b(u_2^{(j)})) C'(t, t'|h(v^{(q)}))]
\end{aligned} \tag{64}$$

where $\mathbf{u}_1 = \{u_1^{(1)}, ..., u_1^{(k)}\}$,$\mathbf{u}_2 = \{u_1^{(2)}, ..., u_2^{(k)}\}$ and $\mathbf{v} = \{v^{(2)}, ..., v^{(k)}\}$ are in the same region as s_1, s_2,t

4.4. Results

4.4.1. Nonseparable compound function

In this section we provide further results and analysis for the nonseparable compound function introduced in §4.1. We provide further figures for the experiment described in the main paper. In this experiment we generate training sets based on 20-200 random observations and predict on the remaining. The results across these training sets and across all baselines is shown in Fig. 1. As expected all models perform well with a large number of observations but this is reduced the benefit of modeling nonseparability and nonstationarity becomes apparent because only models with nonstationarity are able to retain structure of the data. We plot the MAE on the left hand side of Fig 3 and provide a table of learnt parameters in Table. 1.

We have run an additional experiment where we split the data into two parts: the training set $\{f(x_1, x_2)|x_1 < 0.25\}$ and the testing set$\{f(x_1, x_2)|x_1 > 0.25\}$. The results are shown in Fig. 2. As shown our two non-stationary and non-separable kernels are able to better predict in the testing region because they can model the underlying non-stationary and non-separability of the data. The mean absolute error (MAE) is shown on the right hand side of 3, where the size of the testing region on the bottom axis. As expected the MAE for all models increases as the size testing region increases but our proposed models clearly outperform the competing ones.

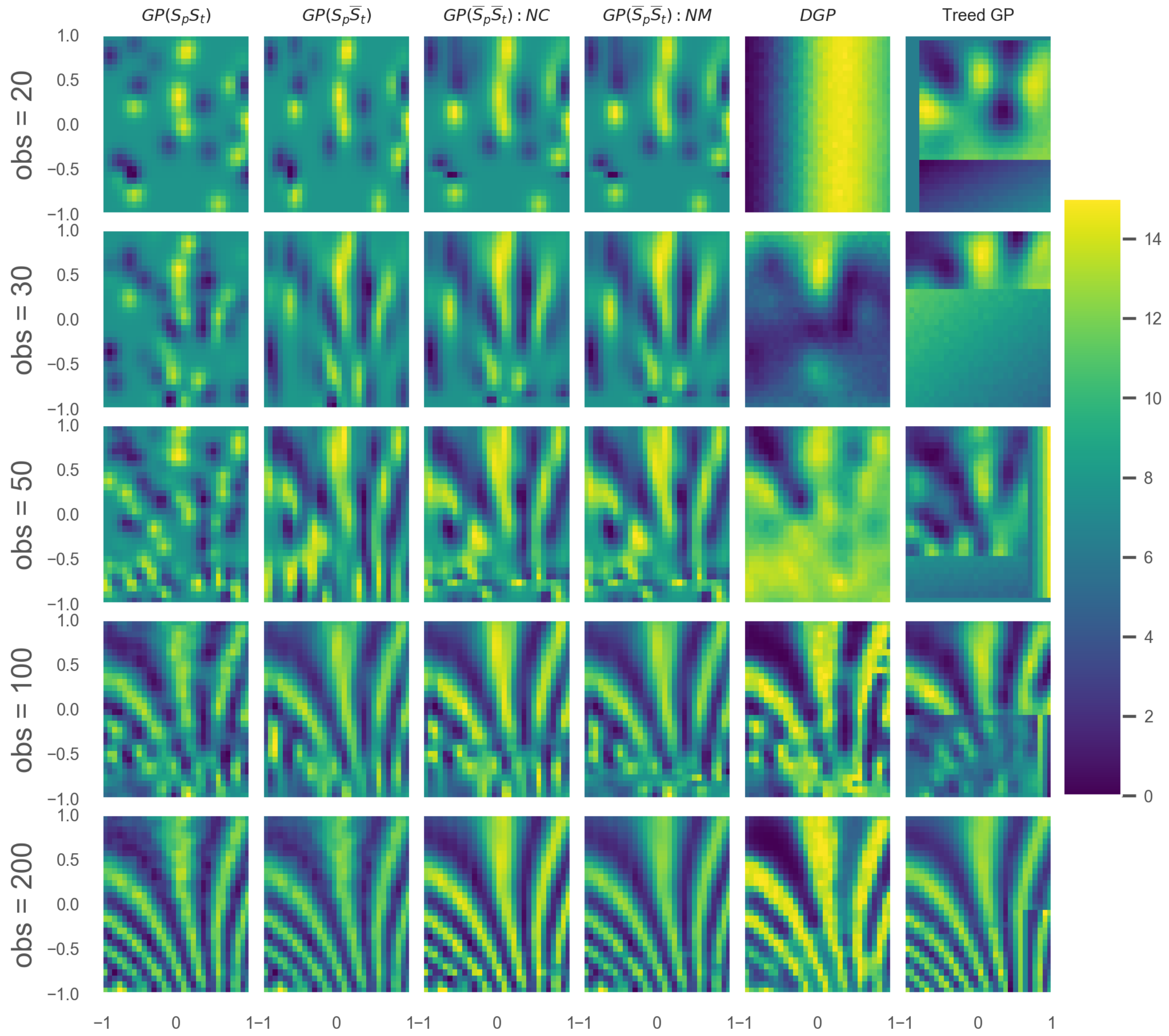

Figure 1. Prediction surface plots from the nonseparable compound function experiment across an increasing number of training observations. Even with low number of observations the two nonstationary and nonseparable kernel are able to maintain more structure. The treed GP works well when we have enough observations to create partitions.

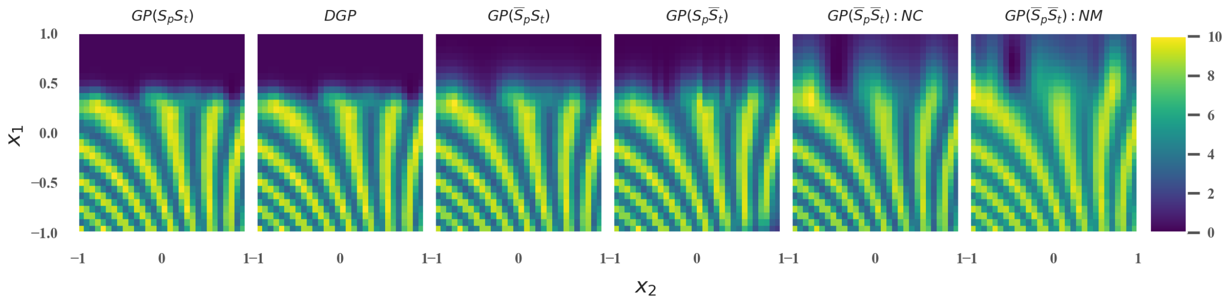

Figure 2. Prediction surface plots from the nonseparable compound function experiment with split training and testing data where the training data is only available at $x_1 < 0.25$. The two nonstationary and nonseparable kernel are able to better predict into the testing region. This is because they are able to learn the varying lengthscales and dependency structure of the data, and so unlike than other kernels do not return the mean as quickly. The treed GP is not applicable in this setting as the partition is only on the training data and it predicts zero when the prediction inputs are outside these partitions.

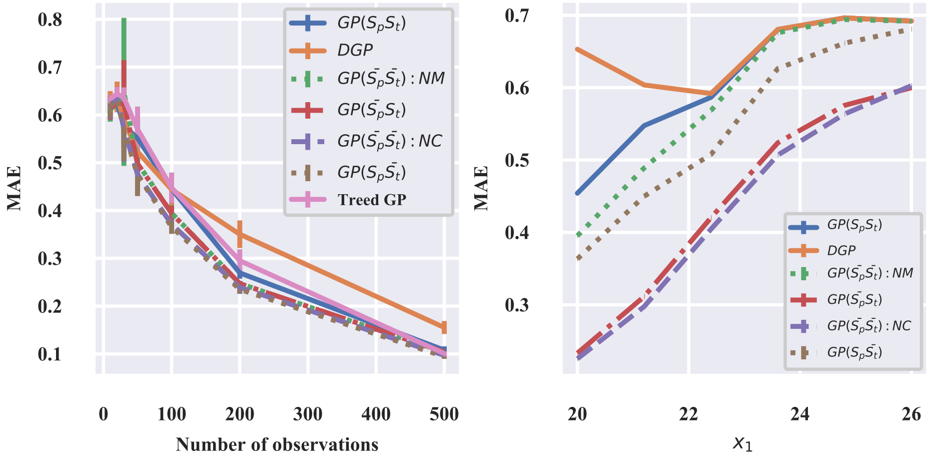

Figure 3. We plot MAE metrics for the two nonseparable compound function experiments. The left panel relates to the random dataset and the right panel related to the split dataset. The left figure shows the MAE value for different number of observations across all kernel constructions in Fig 1. The right figure illustrates the MAE value for the split dataset in Fig 2. The MAE value is calculated for the predictions with $x_1 = 20, .., 26$. We can see the nonstationary and nonseparable kernels achieve smaller MAE values across all experiments.

Table 1. In this table we present the learnt parameters for the random split nonseparable compound function (§6.1). We provide parameters at two spatial locations across various kernels. In the nonstationary kernels the lengthscales are parameterised using linear functions of the input space. In $\mathcal{GP}(\bar{S}_t\bar{S}_p)$: NC we can see that the lengthscale values change between the two locations but the mixing parameter β_0 in constant, whereas for $\mathcal{GP}(\bar{S}_t\bar{S}_p)$: NM it is the reverse, and so highlights where the nonstationarity of these models is coming from.

PARAMETER	LOCATION (X_1, X_2)	$\mathcal{GP}(S_tS_p)$	$\mathcal{GP}(\bar{S}_tS_p)$	$\mathcal{GP}(S_t\bar{S}_p)$	$\mathcal{GP}(\bar{S}_t\bar{S}_p)$: NC	$\mathcal{GP}(\bar{S}_t\bar{S}_p)$: NM
LENGTHSCALES(ℓ_1, ℓ_2)	$(0.7, -0.7)$	$0.13, 0.09$	$0.12, 0.08$	$0.05, 0.23$	$0.09, 0.13$	$0.13, 0.21$
	$(0.5, 0.3)$	$0.13, 0.09$	$0.11, 0.2$	$0.05, 0.23$	$0.12, 0.25$	$0.13, 0.21$
AMPLITUDE	$(0.7, -0.7)$	0.41	0.41	0.39	0.45	0.53
	$(0.5, 0.3)$	0.41	0.41	0.39	0.45	0.53
MXING PARAMETER (β_0)	$(-0.7, -0.7)$	-	-	2.06	1.02	0.6
	$(0.5, 0.3)$	-	-	2.06	1.02	3.12
TRANSFERRED CORRELATION	$(0.7, -0.7)$	-	-	0.27	0.18	0.12
	$(0.5, 0.3)$	-	-	0.27	0.18	0.33
NOISE		0.034	0.006	0.021	0.002	0.003

Computational Complexity:

In this experiment we empirically demonstrate the computational complexity of our covariance functions and baselines by recording the training time across a varying number of observations. We find that SPM:NC has the same computational complexity as all closed form kernels. As expected both the SMP:NM kernel and the Deep GP take longer due to not having closed form solutions and so require monte-carlo simulations. As shown in Fig. 4, the Deep GP the running time increases faster when the number of observations increases.

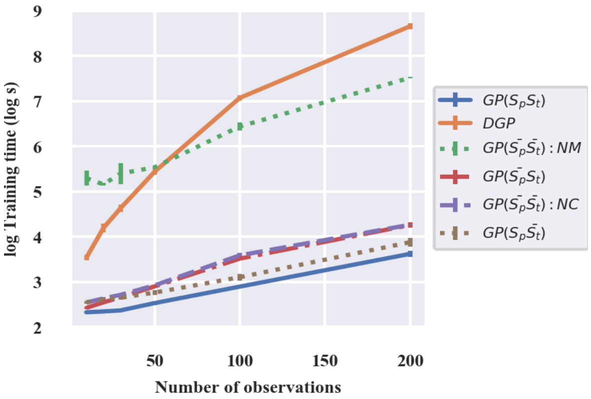

Figure 4. We plot the log running time (log seconds) required for 10000 training iterations across all kernels against the number of observations. As expected, both the Deep GP and SPM:NM take longer to run due to requiring monte-carlo approximation. However, compared to SPM:NM, we find that the training time of the Deep GP increases much faster with the size of the data. To keep a fair comparison we did not include the Treed GP in the above plot. This is due to its the implementation not being in Tensorflow.

4.4.2. Spatio-temporal heat equation

In this section we provide additional details on the spatio-temporal heat equation experiment. Synthetic observations are generated using one solution of the heat equation. For the SPM:nonstationary mixing kernel we use the Green's function of the SDE as the convolution kernel §4.2.6. As shown in the Fig. 5 all the kernels predict well in the training region, due to sufficient number of observations. In the testing region all models provide intervals that cover the test data but only the nonstationary mixing kernel with SDE convolution can still fit the data well. Table 3 provides the learnt parameters for different kernel functions listed in §4.2. For the nonstationary kernels, the lengthscale parameters are constrained using linear functions and varying across input locations. For a fair comparison, we calculate the corresponding hyper-parameters in different location. The SDE kernel has different parameters and so the comparison between it and the rest of kernels is not applicable. However, the learnt dependency structure is listed in the table.

Table 2. Learnt parameters across various kernels on the Spatio-temporal heat experiment.

PARAMETER	T	$\mathcal{GP}(S_tS_p)$	$\mathcal{GP}(\bar{S}_tS_p)$	$\mathcal{GP}(S_t\bar{S}_p)$	$\mathcal{GP}(\bar{S}_t\bar{S}_p)$: NC	$\mathcal{GP}(\bar{S}_t\bar{S}_p)$: SDE
LENGTHSCALES(ℓ_t)	$t=0$	0.25	0.22	0.5	0.21	-
	$t=0.25$	0.25	0.37	0.5	0.77	-
MIXING PARAMETER (β_1)	$t=0$	-	-	1.26	1.73	2.15
	$t=0.25$	-	-	1.26	1.73	2.15
MIXING PARAMETER (β_0)	$t=0$	-	-	0.72	0.2	-
	$t=0.25$	-	-	0.72	0.2	-
CORRELATION(s_1, s_2)	$t=0$	-	-	0.31	0.27	0.38
	$t=0.25$	-	-	0.31	0.27	0.38
CORRELATION(s_1, t)	$t=0$	-	-	0.12	0.05	-
	$t=0.25$	-	-	0.12	0.05	-
NOISE		0.048	0.032	0.027	0.015	0.007

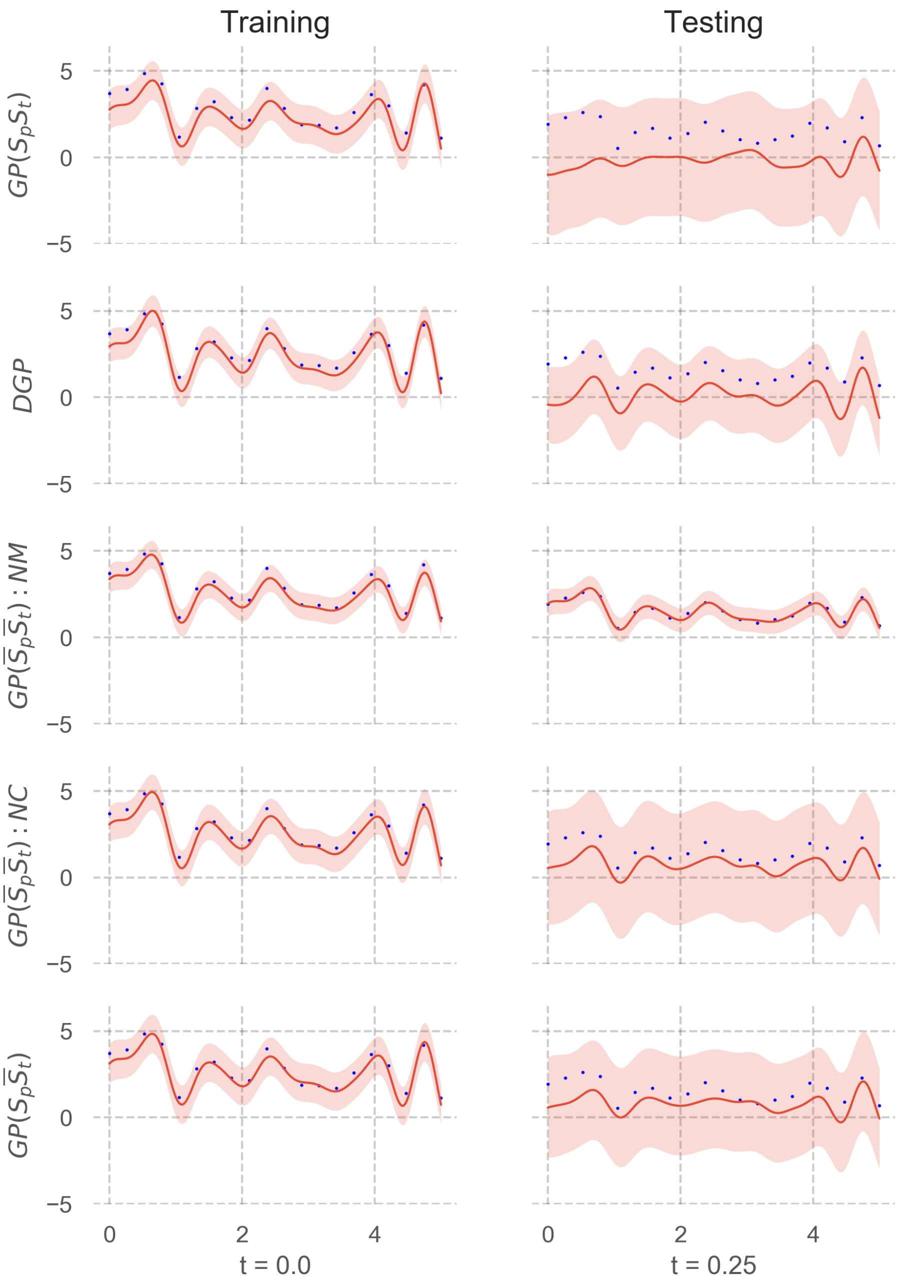

Figure 5. We plot timeseries predictions on the spatio-temporal heat equation experiment in the training region (left hand side) and testing region (right hand side). The training data is only available from $t = 0$ to $t = 0.1$. With enough training data at $t = 0$, all the covariance functions predict well. Whereas at $t = 0.25$ only the nonstationary mixing with SDE convolution is able to retain the correct structure.

4.4.3. London air quality

We model NO2 across London using observations from the London air quality network (www.londonair.org.uk) for 48 hours at 34 sensors recording dataevery hour. We use latitude, longitude, time as our spatio-temporal inputs and remove observations from Sensor No. $9, 25, 33$ for testing. We plot the time series for all models at a single observation station in Fig. 6.

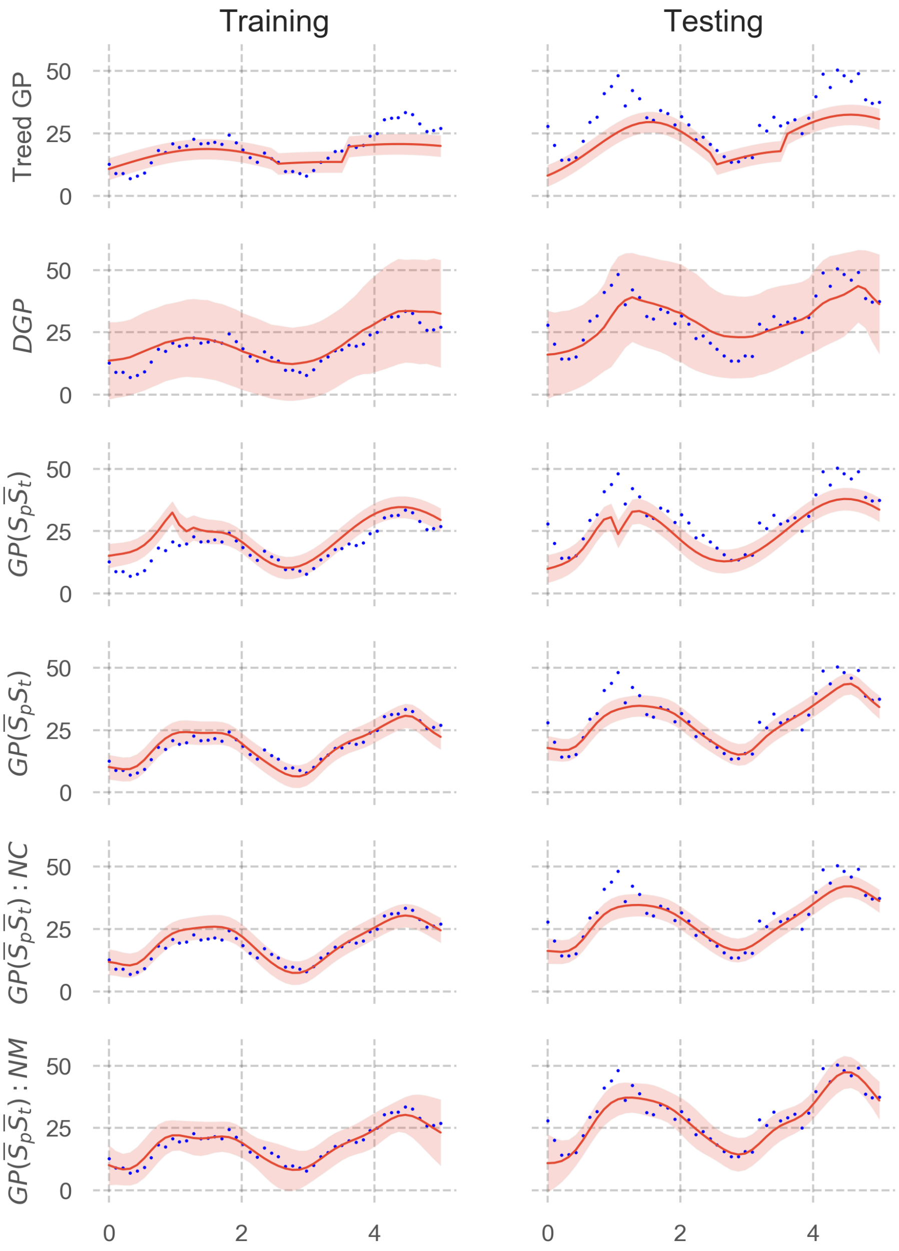

Figure 6. Predictive mean and 95% credible intervals for our proposed kernels and baselines. The left hand side are predictions at sensor No. 35 where observations are present. The right hand side are at sensor No. 9 where we have removed observations for testing. The deep GP obtains a relatively large variance and is the only model that covers all the testing points in the confidence interval. Both GP($S_t\bar{S}_p$) and SPM:NC fit the training observations, indicating that the model is gaining from modelling nonseparability. The SPM:NM has larger intervals than GP($S_t\bar{S}_p$) and SPM:NC, and has testing intervals that cover more training points. This indicates that GP($S_t\bar{S}_p$) and SPM:NC might be overfitting slightly.

4.4.4. Irish wind data

In this experiment we consider the classic Irish wind data set. The data is collected daily at 12 observation stations across Ireland. We use wind speed observed over 80 days. We plot spatial slices for two time points for all models in Fig. 7 and plot time series of the fitted models in Fig 8.

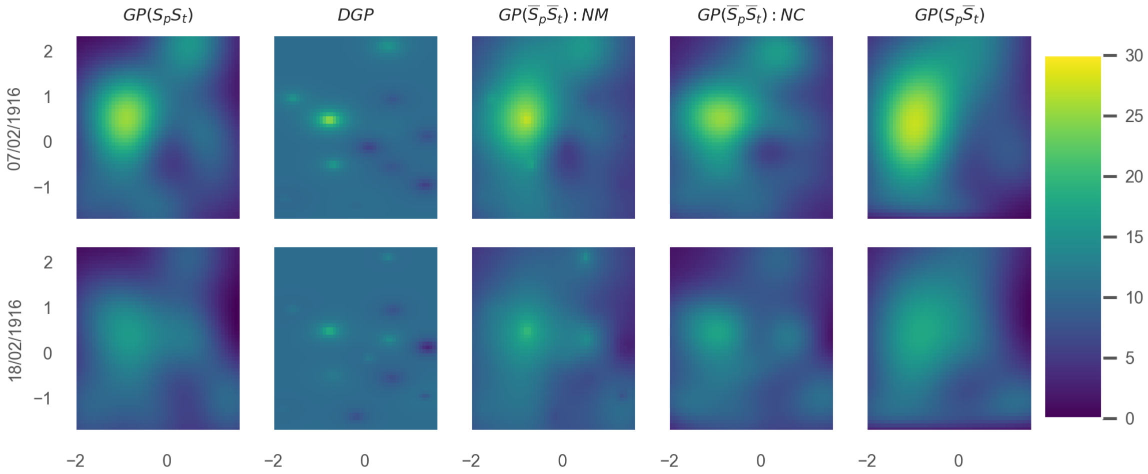

Figure 7. Prediction surface of Irish Wind data at two time points. We can see that for both SPM constructions the reconstructed surfaces are similar because of the constant nonseparability of the data.

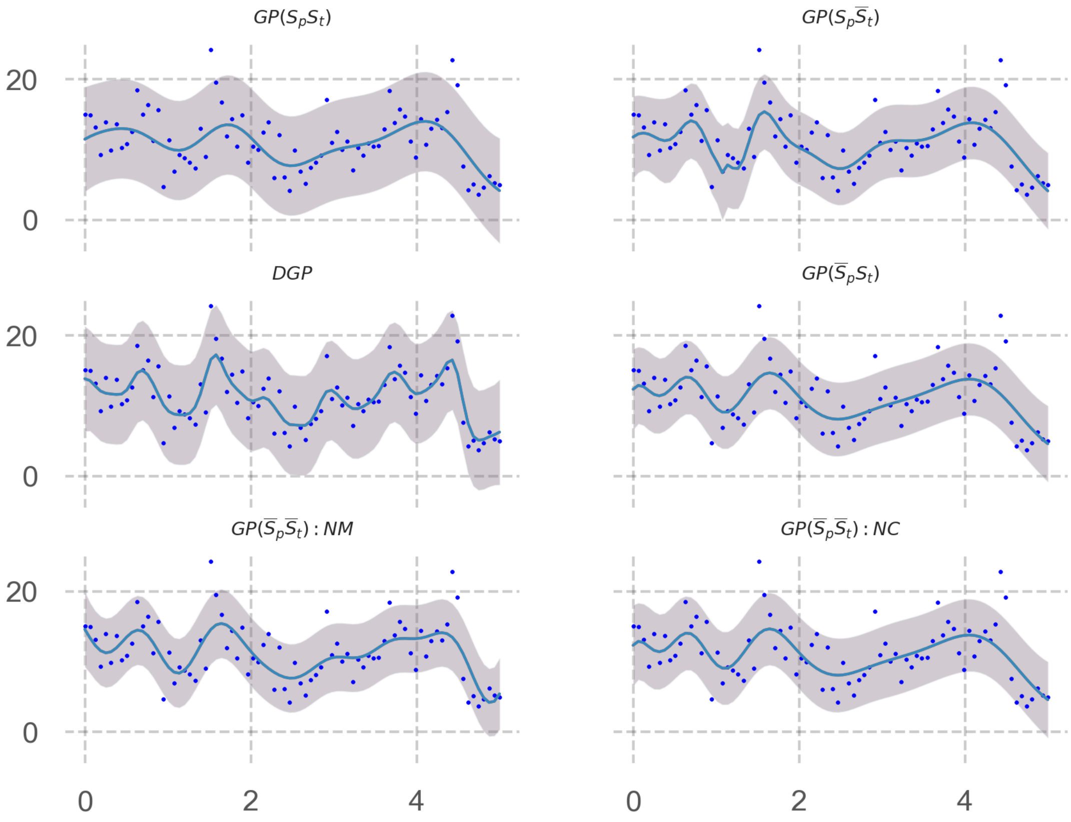

Figure 8. Prediction on Irish Wind data across time at the RochesPoint station.

References

Fonseca, T. C. and Steel, M. F. A general class of nonseparable space–time covariance models. *Environmetrics*, 22(2): 224–242, 2011.

Fuentes, M. and Smith, R. L. A new class of nonstationary spatial models. Technical report, Technical report, North Carolina State University, Raleigh, NC, 2001.

Gibbs, M. N. *Bayesian Gaussian processes for regression and classification*. PhD thesis, Citeseer, 1998.

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

Salimbeni, H. and Deisenroth, M. P. Deeply non-stationary gaussian processes. In *Proc. NIPS Workshop Bayesian Deep Learn*, 2017.

Steinwart, I. and Scovel, C. Mercer's theorem on general domains: on the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35(3):363–417, 2012.