

# Using input warping to improve the Bayesian optimisation of a complex epidemiological model of the sharka virus

Victor Picheny · Coralie Picard · Gael Thebaud

Received: date / Accepted: date

[Coralie's comment: les figures ne sont pas toutes de la même taille, mais je n'ai pas passé trop de temps dessus avant qu'on se mette d'accord sur celles qui seraient dans l'article ou pas]

**Abstract** On peut [Victor's comment: faire un commentaire] [Coralie's comment: chacun avec sa couleur], on peut aussi ~~enlever des trucs~~ ou bien ajouter d'autres trucs, et Gael aussi.

**Keywords** to do

## 1 Introduction

Mathematical models are increasingly used in many research fields to understand and optimize a process. For instance, they are useful in epidemiology to predict epidemics and to propose efficient control options [4, 5, 15, 28, 1, 12, 29, 9]. However, these epidemiological studies are mostly focused on improving one control option which generally depends on only one or two parameters in their model, although various control actions are usually applied simultaneously to manage an epidemic. All these actions could be jointly optimized but taking into account numerous management parameters in an optimization problem can be difficult, especially when the management efficiency depends on the interaction between these parameters.

In this study, we analyse a simulation model of sharka disease spread and management. This disease, caused by a virus transmitted by aphids through *Prunus* or-

---

V. Picheny  
MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France  
Tel.: +33561285551  
E-mail: victor.picheny@inra.fr

C. Picard  
BGPI, Montpellier SupAgro, INRA, Univ. Montpellier, Cirad, TA A-54/K, 34398, Montpellier

G. Thebaud  
to do

chard, is one of the most damaging diseases of stone fruit trees belonging to the genus *Prunus* (e.g. peach, apricot and plum) [3, 20]. Our model includes epidemiological parameters which vary between simulations, and various landscapes on which the virus can spread, which means that this model is stochastic. In addition, management parameters allow to simulate orchard surveillance. Here, we aim to optimize these management parameters using a efficient optimization algorithm.

Within the wide range of potential approaches to solve such optimization problems, black-box optimization methods have proven to be popular in this context [23], in particular because they are in essence non-intrusive: they only require pointwise evaluations of the model at hand (output value for a given set of inputs), as opposed to knowing the underlying mechanisms of the model, structural information, derivatives, etc. This greatly facilitates implementation and avoids developing taylored algorithms. In this work, we focus more particularly on the so-called *Bayesian optimization* (BO) approaches [14, 25], which are well-suited to tackle stochastic and expensive models.

In some cases, the user possesses relevant information regarding his model that could facilitate the optimization task. Accounting for this information within a black-box optimization framework (or rather: *grey box*) may be a challenging task as it is, in essence, unnatural. In this work, we focus on a particular type of information, which we refer to as *local invariance*: for some values of a subset of parameters, it is known that the model is insensitive to another subset of parameters. As an illustration, take a function  $y$  that depends on two discs, parameterized by  $x_1 = r_1 \in [0, r_{\max}]$  (radius of the first disc) and  $x_2 = \rho_{12} \in [0, 1]$  (ratio between  $r_1$  and the radius of the second disc,  $r_2$ ). An action  $A_1$  is conveyed on the first disc and another action  $A_2$  on the second. Setting  $x_1 = 0$ , we have  $r_2 = 0$  for any value of  $\rho_{12}$ , so  $y(0, x_2)$  is constant.

Intuitively, one may want to rework the definition of the parameters to optimize over in order to remove the invariances. However (as we show in 2), such a reformulation his is not always possible. Here, we propose to keep the optimisation problem unchanged, and convey the invariance information to the BO algorithm directly, by applying a *warping* [26, 27] to the parameter space.

The remainder of this paper is structured as follow. Section 2 describes the sharka model and its invariances. Section 3 presents the basics of Bayesian optimization and our warping strategy. Finally, section 3 analyses the efficiency of the warping on the sharka model.

## 2 Model description and problem set-up

The simulation model that we analyze in this work is a stochastic, spatially explicit, SEIR (susceptible-exposed-infectious-removed) model that simulates sharka spread and management actions [including surveillance, removals and replantations 18, 21, 22].

This model is orchard-based, with a discrete time step of one week. It allows to perform simulations on landscapes composed of uncultivated areas and patches on which peach trees are grown. The patches can be more or less aggregated in the landscape however, we only use in this work the 30 landscapes with a high level of patch

aggregation as described by [16]. During the simulation, the trees in the patches are characterized by different states. When the simulation begins, they are not infected: they are in the “susceptible” state. Then, the virus is introduced the first year of the simulation in one of the patches and spreads through orchards (new introductions can also occur during the entire simulation on all patches). The virus causes changes in tree status: from “susceptible”, they become “exposed” (infected but not yet infectious or symptomatic), “infectious hidden” (after the end of the latent period), “infectious detected” (when specific symptoms are detected on the tree during a survey), and “removed” (when the tree is removed from the patch). The model output is an economic criterion, the net present value (NPV), which accounts for the benefit generated by the cultivation of productive trees and the costs induced by fruit production and disease management [22].

In order to simulate wide range of epidemic and management scenarios, the model includes 6 epidemiological and 23 management parameters [22, 16]. In this work, we will use the 6 epidemiological parameters and only 10 management parameters (related to the surveillance of the orchards). They include distances of 3 zones for which the surveys are more or less frequent as well as their duration, the probability of the infected tree detection, and a contamination threshold which can request to increase the surveillance frequency in the focal zone. Details of epidemiological and management parameters used in this study are presented in Fig.1 and Table 1 (this table also includes the variation ranges of the parameters in the model).

Here, we aim to optimize the management strategy of the disease (i.e. to find the combination of management parameters allowing to obtain the best NPV), taking into account the epidemic stochasticity. However, we note that some combinations of management parameters can represent the same management, which may cause problems in the optimization process. Indeed, we observe that some management parameters are not useful when other parameters have a value of 0, which means that they can take any values without modifying the simulation. For example, when a zone radius is 0, the associated surveillance frequency have no impact on the NPV (regardless its value). The methodological developments that are proposed in this work address this issue by removing the parameter combinations which lead to the same management. The parameter invariances removed from the model are listed in Table 2.

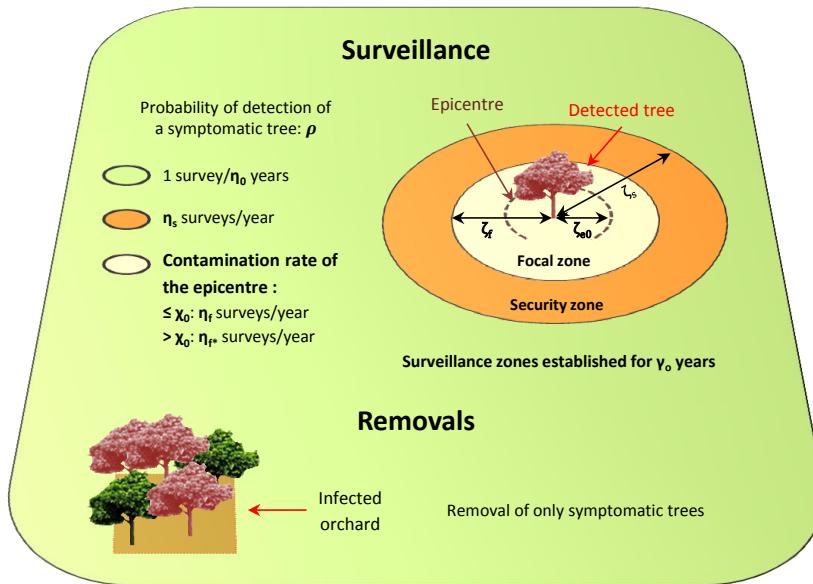
### 3 Methods: Bayesian optimization

#### 3.1 Overview

Bayesian optimization can be seen as a modernization of the statistical response surface methodology for sequential design [2], where the basic idea is to replace an expensive function by a cheap-to-evaluate surrogate one. In BO, Gaussian process (GP) regression, or kriging, is used to provide flexible response surface fits. GPs are attractive in particular for their tractability, since they are simply characterized by their mean  $m(\cdot)$  and covariance (or kernel)  $k(\cdot, \cdot)$  functions, see e.g., Rasmussen and

**Table 1** Management parameters implemented in the previously developed model with minimum and maximum values corresponding to the variation range of each parameter.

|              |   | Min | Max  |
|--------------|---|-----|------|
| $\rho$       | Probability of detection of a symptomatic tree  | 0   | 0.66 |
| $\gamma_O$   | Duration of observation zones (years)   | 0   | 10   |
| $\zeta_s$    | Radius-distance of security zones (m)   | 0   | 5800 |
| $\zeta_f$    | Radius-distance of focal zones (m)  | 0   | 1    |
| $\zeta_{eo}$ | Radius-distance of observation epicenter (m)  | 0   | 1    |
| $1/\eta_0$   | Maximal period between 2 observations (year)  | 1   | 15   |
| $\eta_s$     | Observation frequency in security zones (year-1)  | 0   | 8    |
| $\eta_f$     | Observation frequency in focal zones (year-1)   | 0   | 8    |
| $\eta_{f*}$  | Modified observation frequency in focal zones (year-1)  | 0   | 8    |
| $\chi_o$     | Contamination threshold in the observation epicenter, above which the observation frequency in focal zone is modified | 0   | 1    |



**Fig. 1** Management actions implemented in the model.

Williams [19]. In the following, we consider zero-mean processes ( $m = 0$ ) for the sake of conciseness.

Conditionally on  $n$  noisy observations  $\mathbf{f} = (f_1, \dots, f_n)$ , with independent, centered, Gaussian noise, that is,  $f_i = y(\mathbf{x}_i) + \varepsilon_i$  with  $\varepsilon_i \sim \mathcal{N}(0, \tau_i^2)$ , the predictive

**Table 2** Invariances of management parameters. For instance, when  $\gamma_O = 0$  or when  $\rho = 0$ ,  $\chi_o$  does not influence the model output.

| Management parameters | OR  | OR | OR |
|-----------------------|---|----|----|
| $\rho$                |   |    |    |
| $1/\eta_0$            |   |    |    |
| $\gamma_O$            |   |    |    |
| $\chi_o$              | $\gamma_O = 0$   $\rho = 0$                 |    |    |
| $\zeta_{eo}$          | $\gamma_O = 0$   $\zeta_s = 0$   $\rho = 0$ |    |    |
| $\zeta_f$             | $\gamma_O = 0$   $\zeta_s = 0$              |    |    |
| $\eta_{f*}$           | $\gamma_O = 0$   $\rho = 0$                 |    |    |
| $\zeta_s$             | $\gamma_O = 0$   $\eta_s = 0$               |    |    |
| $\eta_s$              | $\gamma_O = 0$                              |    |    |
| $\eta_f$              | $\gamma_O = 0$                              |    |    |

distribution of  $y$  is another GP, with mean and covariance functions given by:

$$\mu(\mathbf{x}) = \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1} \mathbf{f}, \quad (1)$$

$$\sigma^2(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1} \mathbf{k}(\mathbf{x}'), \quad (2)$$

where  $\mathbf{k}(\mathbf{x}) := (k(\mathbf{x}, \mathbf{x}_1), \dots, k(\mathbf{x}, \mathbf{x}_n))^\top$  and  $\mathbf{K} := (k(\mathbf{x}_i, \mathbf{x}_j) + \tau_i^2 \delta_{i=j})_{1 \leq i, j \leq n}$ ,  $\delta$  standing for the Kronecker function.

Commonly,  $k(., .)$  belongs to a parametric family of covariance functions such as the Gaussian and Matérn kernels, based on hypotheses about the smoothness of  $y$ . Corresponding hyperparameters are often obtained as maximum likelihood estimates, see e.g., Rasmussen and Williams [19] or Roustant et al [24] for the corresponding details.

BO typically tackles optimization problems of the form:

$$\begin{aligned} \min \quad & y(\mathbf{x}) \\ \text{s.t. } & \mathbf{x} \in \mathbb{X}, \end{aligned}$$

with  $\mathbb{X} \in \mathbb{R}^d$  is usually a bounded hyperrectangle and  $y : \mathbb{R}^d \rightarrow \mathbb{R}$  is a scalar-valued objective function.

Optimization amounts here to choosing a sequence of points  $\mathbf{x}_{n+1}, \dots, \mathbf{x}_{n+N}$  at which the function  $y$  is evaluated. Sequential design decisions, so-called *acquisitions*, are based on the GP model and judiciously balance exploration and exploitation in search for global optima. The GP model is updated after each new value is calculated.

In the noiseless setting ( $\tau = 0$ ), the canonical acquisition function is *expected improvement* (EI) [11]. Define  $f_{\min} = \min_{i=1, \dots, n} y_i$ , the smallest  $y$ -value seen so far, and let  $I(\mathbf{x}) = \max\{0, f_{\min} - Y(\mathbf{x})\}$  be the *improvement* at  $x$ .  $I(x)$  is largest when  $Y(\mathbf{x})$  has substantial distribution below  $f_{\min}$ . The expectation of  $I(x)$  over  $Y(x)$  has a convenient closed form, revealing balance between exploitation ( $\mu(x)$ )

under  $f_{\min}$ ) and exploration (large  $\sigma^n(x)$ ):

$$\mathbb{E}\{I(x)\} = (f_{\min} - \mu(x))\Phi\left(\frac{f_{\min} - \mu(x)}{\sigma(x)}\right) + \sigma(x)\phi\left(\frac{f_{\min} - \mu(x)}{\sigma(x)}\right), \quad (3)$$

where  $\Phi$  ( $\phi$ ) is the standard normal cdf (pdf).

### 3.2 Bayesian optimization of stochastic simulators

When  $y$  is only available through noisy evaluations, the EI acquisition cannot be used directly. Several authors have tackled this issue; we refer to [17] for a review on the topic. We chose here to focus on the *reinterpolation method* proposed in [10], which is based on the use of an instrumental noiseless kriging model, built from the original one. First, the (noisy) kriging predictions at the DOE points  $\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_n)$  are computed. Then, a reinterpolating model is built, by using the same covariance kernel and parameters and the same experimental design, but the observation vector is replaced by  $\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_n)$  and the noise variance is set to zero. Since this latter model is noise-free, the classical EI can be used as the infill criterion. Once the new design is chosen and the evaluation is performed, both kriging models are updated. One particular characteristic of this strategy is that it does not allow repetitions, which is desirable in our case.

### 3.3 Bayesian optimization with invariances

#### 3.3.1 Definition of local invariances

From Section 2, we observe that there are cases of invariances involving a single or several variables. Here, we formalize the different situations and generalize some cases.

We first introduce the following notation (this is purely notation, no actual permutation is performed):

$$y(\mathbf{x}) = y(x_i, \mathbf{x}_J, \mathbf{x}_{-iJ}) \quad (4)$$

$$\mathbb{X} = \{\mathbb{X}_i, \mathbb{X}_J, \mathbb{X}_{-iJ}\} \quad (5)$$

**Definition 1 (Simple)** We call *simple invariance* the following case:  $y$  is invariant with respect to  $\mathbf{x}_J$  ( $J$  a subset of  $\{1, \dots, d\} \setminus i$ ) if  $x_i = c_i$  ( $i \in \{1, \dots, d\}$ ):

$$y(c_i, \mathbf{x}_J, \mathbf{x}_{-iJ}) = y(c_i, \mathbf{x}'_J, \mathbf{x}_{-iJ}), \quad \forall \mathbf{x}_J, \mathbf{x}'_J \in \mathbb{X}_J, \mathbf{x}_{-iJ} \in \mathbb{X}_{-iJ}.$$

This corresponds for instance to the last line of Table 2: the observation frequency  $\eta_f$  does not have an effect on the model if the duration of observation  $\gamma_O$  is set to zero.

**Definition 2 (Or)** We call “or” *invariance* the following case:  $y$  is invariant with respect to  $\mathbf{x}_J$  ( $J$  a subset of  $\{1, \dots, d\} \setminus I$ ) if there exists at least one  $i \in I$  such that  $x_i = c_i$  ( $I$  a subset of  $\{1, \dots, d\} \setminus J$ ):

$$y(c_i, \mathbf{x}_{I \setminus i}, \mathbf{x}_J, \mathbf{x}_{-IJ}) = y(c_i, \mathbf{x}_{I \setminus i}, \mathbf{x}'_J, \mathbf{x}_{-IJ}), \quad \forall \mathbf{x}_J, \mathbf{x}'_J \in \mathbb{X}_J, \mathbf{x}_{I \setminus i} \in \mathbb{X}_{I \setminus i}.$$

This corresponds for instance to the fourth line of Table 2: the contamination threshold in the observation zone  $\xi_0$  does not have an effect on the model if the duration of observation  $\gamma_O$  is set to zero or if the probability of detection  $\rho$  is set to zero.

**Definition 3 (Linear)** We call *linear invariance* the following case:  $y$  is invariant with respect to  $\mathbf{x}_J$  ( $J$  a subset of  $\{1, \dots, d\} \setminus I$ ) if  $\mathbf{A}\mathbf{x}_I = \mathbf{b}$ , with  $I$  a subset of  $\{1, \dots, d\} \setminus J$ ,  $\mathbf{A}$  a matrix of size  $p \times \text{Card}(I)$  and  $\mathbf{b}$  a vector of size  $p$ :

$$y(\mathbf{x}_I, \mathbf{x}_J) = y(\mathbf{x}_I, \mathbf{x}'_J), \quad \forall \mathbf{x}_J, \mathbf{x}'_J \in \mathbb{X}_J, \text{ if } \mathbf{A}\mathbf{x}_I = \mathbf{b}.$$

There are two particular cases worth noting:

- setting  $p = \text{Card}(I)$ ,  $\mathbf{A} = \mathbb{I}_p$  and  $\mathbf{b} = \mathbf{c}_I$  results in an “AND” condition:  $y$  is invariant with respect to  $\mathbf{x}_j$  if,  $\forall i \in I, x_i = c_i$ ;
- setting  $p = 1$ ,  $\mathbf{A} = [1, -1]$  results in an invariance under the condition  $x_{i1} = x_{i2}$ .

### 3.3.2 Principle of input warping

The standard use of GPs implies a hypothesis of stationarity (the unconditional joint probability distribution of the process does not change when shifted in the  $\mathbb{X}$  space), which is in contradiction with the notion of local invariance.

There are several ways of incorporating structural information into Gaussian processes. One is to work on the kernel function  $k$  [8, 6]. Another, which is the one we use here, is to transform the original input space  $\mathbb{X}$  into a *warped* one  $\tilde{\mathbb{X}}$  and index the GP on  $\tilde{\mathbb{X}}$ , so that the new topology directly reflects the structural information [27, 13].

Consider for simplicity a single invariance over  $x_J$  when  $x_i = c_i$ . A simple way to handle this problem is to distort locally the space so that the subspace  $\{(x_i, \mathbf{x}_J) | x_i = c_i\}$  collapses to a single point, for instance with  $\mathbf{x}_J$  at its average value:  $(c_i, \bar{\mathbf{x}}_J)$ .

Hence, we are seeking warping functions of the form:

$$\begin{aligned} \psi : \mathbb{X} &\rightarrow \tilde{\mathbb{X}} \\ \mathbf{x} &\mapsto \tilde{\mathbf{x}} = \psi(\mathbf{x}) \end{aligned}$$

such that:

1.  $\psi(x_i, \mathbf{x}_J, \mathbf{x}_{-iJ}) = (c_i, \bar{\mathbf{x}}_J, \mathbf{x}_{-iJ})$  if and only if  $x_i = c_i$ ;
2.  $\psi$  restricted to  $\mathbb{X} \setminus (c_i, \dots)$  and  $\tilde{\mathbb{X}} \setminus (c_i, \bar{\mathbf{x}}_J, \dots)$  is a diffeomorphism.

In addition, we will search for deformations that decrease monotonically when  $|x_i - c_i|$  increases, that is:

$$\begin{aligned} ((x_i, \mathbf{x}_J, \mathbf{x}_{-iJ}), \psi[(x_i, \mathbf{x}_J, \mathbf{x}_{-iJ})]) &\leq d((x'_i, \mathbf{x}_J, \mathbf{x}_{-iJ}), \psi[(x'_i, \mathbf{x}_J, \mathbf{x}_{-iJ})]) \\ &\text{if } |x_i - c_i| \leq |x'_i - c_i|, \end{aligned}$$

for some distance  $d(., .)$ .

Since the  $\mathbf{x}_J$  dimension collapses to  $\overline{\mathbf{x}_J}$  at  $x_i = c_i$ , we write:

$$\forall j \in J, \quad \tilde{x}_j = \overline{x_j} + (x_j - \overline{x_j}) \alpha(x_i, c_i), \quad (6)$$

with  $\alpha(x_i, c_i)$  an attenuation function such that:

1.  $\alpha(c_i, c_i) = 0$ ;
2.  $\alpha$  increases monotonically with  $|x_i - c_i|$ ;
3.  $0 < \alpha \leq 1, \forall x_i \neq c_i$ .

Condition 1 ensures that  $\widetilde{\mathbf{x}}_j = \overline{\mathbf{x}_j}$  when  $x_i = c_i$  (the  $J$ -th dimensions collapse).

### 3.3.3 Warping for a simple invariance

In the simple invariance case, we propose linear and correlation-based attenuation functions:

$$\alpha_{\text{lin}}(x_i, c_i) = \frac{|x_i - c_i|}{\delta_i}, \quad (7)$$

$$\alpha_{\text{cor}}(x_i, c_i) = 1 - r(x_i, c_i), \quad (8)$$

where  $r$  is a  $\mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  correlation function. Typically,  $\delta_i$  may be set to the range of variation of  $x_i$ , so that the condition  $\alpha \leq 1$  is ensured. Choosing  $r$  as the generalized exponential correlation, we have:

$$\alpha_{\text{exp}}(x_i, c_i) = 1 - \exp \left[ - \left( \frac{|x_i - c_i|}{\theta_i} \right)^d \right], \quad (9)$$

with  $\theta_i$  and  $d$  positive parameters to be tuned.

Figure 2 shows a 2D rectangular space distorted by three warpings, when the invariance is on a boundary of  $x_1$ . Figure 3 shows (unconditional) realizations of GPs with a Gaussian kernel applied on the warped space. We see that the invariance at  $x_1$  maximum is ensured. The linear warping induces a strong anisotropy, while with the two other warpings, the process seems stationary far from the critical value.

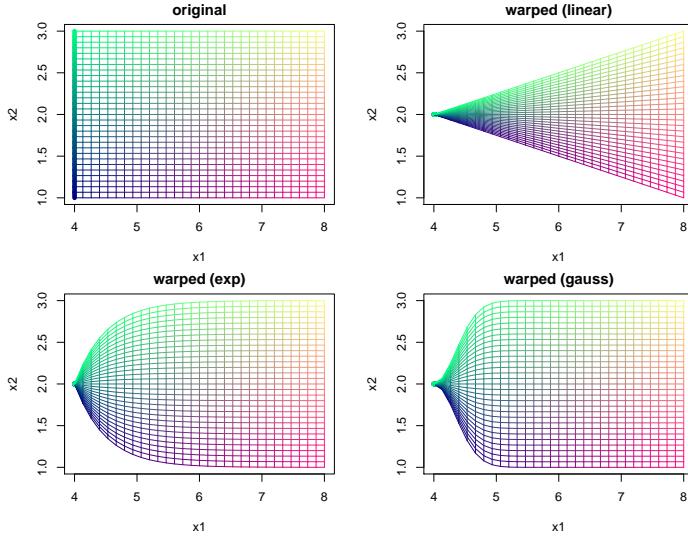
### 3.3.4 Warping for linear invariances

For simplicity, we consider first the particular linear case where  $\mathbf{A} = \mathbb{I}_p$  and  $\mathbf{b} = \mathbf{c}_I$ , that is where invariances occur when a set of variables takes simultaneously a set of critical values:  $\mathbf{x}_I = \mathbf{c}_I$ . In that case, a possible warping is:

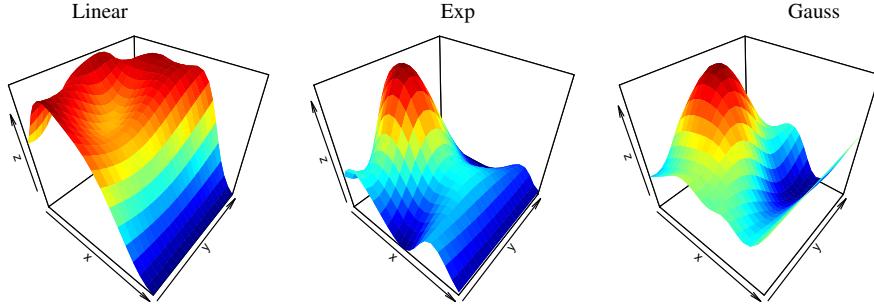
$$\forall j \in J, \quad \tilde{x}_j = \overline{x_j} + (x_j - \overline{x_j}) \alpha_I(\mathbf{x}_I, \mathbf{c}_I). \quad (10)$$

with  $\alpha_I$  now a multivariate attenuation function ( $\mathbb{R}^{\text{Card}(I)} \times \mathbb{R}^{\text{Card}(I)} \rightarrow \mathbb{R}$ ), so that, similarly to the simple case:

1.  $\alpha_I(\mathbf{c}_I, \mathbf{c}_I) = 0$ ;
2.  $\alpha_I$  increases monotonically with  $d(\mathbf{x}_I, \mathbf{c}_I)$  (for some distance  $d(., .)$ );
3.  $0 < \alpha_I \leq 1, \forall \mathbf{x}_I \neq \mathbf{c}_I$ .



**Fig. 2** Three deformations of a 2D space. The local invariance is at  $x_1 = 0$ , highlighted with larger lines.



**Fig. 3** Three GP realizations using warping functions as shown previously.

As in the simple case, linear and correlation-based warpings can be defined as:

$$\alpha_{\text{lin}}(\mathbf{x}_I, \mathbf{c}_I) = \frac{1}{\text{Card}(I)} \sum_{i \in I} \frac{|x_i - c_i|}{\delta_i}, \quad (11)$$

$$\alpha_{\text{cor}}(\mathbf{x}_I, \mathbf{c}_I) = 1 - r_I(\mathbf{x}_I, \mathbf{c}_I), \quad (12)$$

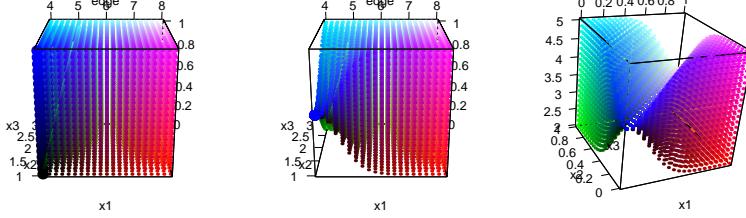
with  $r_I$  a  $\mathbb{R}^{\text{Card}(I)} \times \mathbb{R}^{\text{Card}(I)} \rightarrow \mathbb{R}$  correlation function, for instance:

$$r_I(\mathbf{x}_I, \mathbf{c}_I) = \exp \left[ - \sum_{i \in I} \left( \frac{|x_i - c_i|}{\theta_i} \right)^d \right]$$

Generalizing to the affine case  $\mathbf{Ax}_I = \mathbf{b}$ , the warping function is the same as in Equation 10, with now:

$$\alpha(\mathbf{x}_I, \mathbf{c}_I) = 1 - r_A(\mathbf{Ax}_I, \mathbf{b}). \quad (13)$$

Figure 4 shows two deformations of the unit cubic space when  $y$  is invariant w.r.t.  $x_3$  when  $1-x_1 = x_2 = 0$ , and 2- $x_1 = x_2$ . On both cases a Gaussian warping (exponential with  $d = 2$ ) is applied.



**Fig. 4** Left: original space (with the critical edge highlighted). Centered: warping for  $c_1 = 0$  AND  $c_2 = 1$ . Right: warping for  $x_1 = x_2$ .

### 3.3.5 Combining warpings

*Independent conditions* Now, we consider that we have a series of invariance conditions, defined with respect to sets  $I_1, \dots, I_n$  and corresponding  $J_1, \dots, J_n$ . If  $J_k \cap I_l = \emptyset$ ,  $1 \leq j \neq k \leq n$  and  $I_i \cap J_k = \emptyset$ ,  $1 \leq j, k \leq n$ , the set of warped variables are distinct from the set on which the conditions are written, the invariance conditions are written only once for each variable. In that case, the warpings can be applied independently.

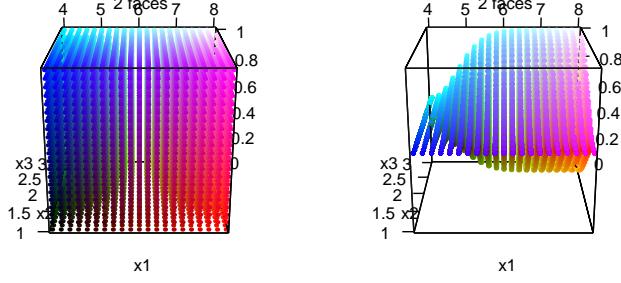
*Combinations of simple conditions: “OR” invariance* Now, we consider the case when  $y$  is invariant w.r.t. a set  $\mathbf{x}_J$  for different conditions on sets  $I_1, \dots, I_n$  (that, for  $\mathbf{x}_{I_1} = \mathbf{c}_{I_1}$  OR  $\mathbf{x}_{I_2} = \mathbf{c}_{I_2}$  OR ...). If  $J \cap I_i = \emptyset$ ,  $1 \leq i \leq n$ , the warping function we propose is:

$$\forall j \in J, \quad \tilde{x}_j = \bar{x}_j + (x_j - \bar{x}_j) \prod_{I \in \{I_1, \dots, I_n\}} \alpha_I(\mathbf{x}_I, \mathbf{c}_I). \quad (14)$$

We see directly that the product of  $\alpha$ 's ensure that  $\tilde{x}_j = \bar{x}_j$  if any  $x_i = c_i$ , and the distortion reduces only when *all* the  $x_i$ 's are far from the  $c_i$ 's. Figure 5 shows a deformation of a cubic space when  $x_3$  is not influent when  $x_1$  or  $x_2$  are minimal, when a Gaussian warping (exponential with  $d = 2$ ) is applied.

*“Circular” conditions* Difficulty only arises when some variables appear in both  $I_l$ 's and  $J_m$ 's sets. Take for instance a “reciprocal” condition, e.g.,  $y$  is invariant w.r.t.  $\mathbf{x}_J$  when  $\mathbf{x}_I = \mathbf{c}_I$ , and invariant w.r.t.  $\mathbf{x}_I$  when  $\mathbf{x}_J = \mathbf{c}_J$ . In that case, applying independently warping functions would lead to:

$$\begin{aligned} \psi(\mathbf{c}_I, \mathbf{x}_J, \mathbf{x}_{-IJ}) &= (\mathbf{c}_I, \bar{\mathbf{x}}_J, \mathbf{x}_{-IJ}), \\ \psi(\mathbf{x}_I, \mathbf{c}_J, \mathbf{x}_{-IJ}) &= (\bar{\mathbf{x}}_I, \mathbf{c}_J, \mathbf{x}_{-IJ}), \\ \text{but: } \psi(\mathbf{c}_I, \mathbf{c}_J, \mathbf{x}_{-IJ}) &= (\mathbf{c}_I, \mathbf{c}_J, \mathbf{x}_{-IJ}), \end{aligned}$$



**Fig. 5** Warping for  $c_1 = 4$  (left face of the cube) OR  $c_2 = 1$  (front face). Left: original space, right: distorted space.

which induces a discontinuity.

In that case, a simple solution is to fix the non influent variable to its critical value instead of its average, hence applying:

$$\forall k \in K = (\cup_{1 \leq l \leq n} I_l) \cap (\cup_{1 \leq m \leq n} J_m), \tilde{x_k} = c_k + (x_k - c_k) \prod_{i \in I_k} \alpha(x_i, c_i) \quad (15)$$

*Remark* This formula does not apply in the affine case (Equation 13).

We first show the deformations on a 2D space on Figure 6, where the two critical values are on the boundaries of  $x_1$  and  $x_2$ . Here, the warping of Equation 15 is applied on each variable ( $K = \{1, 2\}$ ). Again, except for the linear warping, the local topology is preserved far from the critical edges.

As another illustrative example, we consider a cubic space with the following circular conditions:

- $y$  is invariant w.r.t.  $x_2$  if  $x_1 = 4$ ;
- $y$  is invariant w.r.t.  $x_3$  if  $x_2 = 1$ ;
- $y$  is invariant w.r.t.  $x_1$  if  $x_3 = 0$ .

All critical values correspond to the lower bounds of the variables. Equation 15 is applied to each variable, hence with  $K = \{1, 2, 3\}$ ,  $C = [4, 1, 0]$  and  $I_1 = 3$ ,  $I_2 = 1$ , and  $I_3 = 2$ . The original and distorted space is shown in Figure 7.

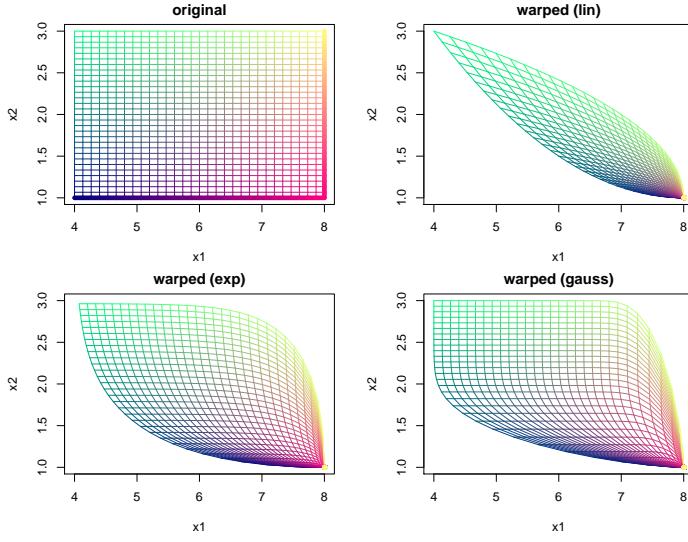
### 3.3.6 Warping parameters tuning

[Victor's comment: a faire...]

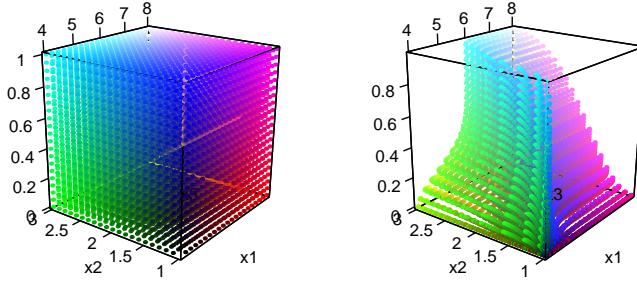
Note that choosing the range of the correlation  $r_A$  is non-trivial. A possible solution is  $\theta_A = \mathbf{A}^T \theta_I$ .

### 3.3.7 BO on warped spaces

[Victor's comment: a faire...]



**Fig. 6** Three deformations of a 2D space, with invariance at  $x_1 = 8$  OR  $x_2 = 1$ , highlighted with larger lines.



**Fig. 7** Warping with circular conditions. Left: original space, right: distorted space.

#### 4 A warping-based Bayesian optimization of the Sharka model

##### 4.1 Numerical setup

###### 4.1.1 Experiments description

To evaluate the benefits of including the warping step in the optimization process (i.e. reducing the parameter space removing the combinations which lead to the same management), we conducted 50 independent optimizations of sharka management parameters with and without the warping step. The criterion to optimize was the mean of the NPV ( $\overline{NPV}$ ). For this to happen, we randomly selected 50 times 200 management strategies using a maximin Latin hypercube sampling design [7]. Then, for each

sampling design of 200 strategies, we performed 2 optimizations in parallel: with and without the warping step. For one optimization, we performed sequentially 200 iterations allowing to choose 200 new strategies, resulting in a total of 400 evaluated strategies. These 200 new strategies were selected each time among 100,000 randomly generated candidate points over the parameter space and 10,000 more locally around the best point found. In addition, for each evaluated strategy, 1000 simulations were carried out (with different random seed) to take into account the variability due to the epidemic and landscape characteristics.

#### 4.1.2 Comparison with standard BO

Description of comparison metrics

[Victor's comment: Idem juste pour les méthodes de comparaison, je me charge du paragraphe pour dire quoi on se compare et je m'occupe de la partie krigage et warping.]

[Coralie's comment: on compare ici les résultats obtenus sans le problème d'identification : avec ton script denoise.results.v6.R. Mais je ne sais pas trop comment l'expliquer ici]

We firstly compared the optimization results by subtracting the  $\overline{NPV}$  achieved using the optimization with the warping step and the optimization without the warping step (obtained from the same sampling design).

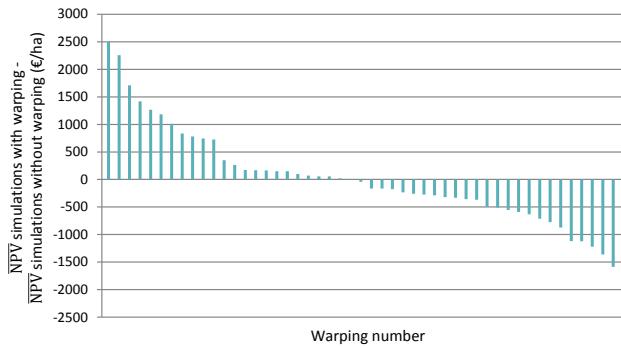
In addition, we compared the optimization speed between the optimizations with or without warping. To this end, we used two different ways. Firstly, we performed a nonlinear regression of  $\overline{NPV}$  obtained for all the selected strategies during the optimization process with and without the warping step, and we compared the growth parameter  $c$  of the following regression:  $a + b \times \exp^{-c \times x}$ .

Secondly, we used a specific algorithm developed by [Coralie's comment: référence???]. Briefly, we uniformly defined 100  $\alpha$  values between a minimum and a maximum values. Then, for each iteration performed in the optimization process (i.e. for each of the 200 evaluated strategies), we add: the number of optimizations (under 50) which exceed  $\alpha$  1, the number of optimizations which exceed  $\alpha$  2, ..., the number of optimizations which exceed  $\alpha$  100. We used  $\alpha \in [0;18,012.12]$ , and then  $\alpha \in [10,000;18,012.12]$ . The value 18,012.12 corresponds to the maximal value of  $\overline{NPV}$  identified in all the optimizations.

## 4.2 Results and insights into the Sharka model

We firstly subtracting the  $\overline{NPV}$  obtained with optimizations with and without the warping step. In 24 out of the 50 optimization cases, we obtained better  $\overline{NPV}$  with the warping step than without (Fig.8). This result means that with 200 iterations in the optimization, the final optimization result is not impacted by the use of a warping step.

However, we showed that the warping can impact the optimization speed (Fig.9). Indeed, the parameter  $c$  corresponding to the growth parameter of a nonlinear regression was higher with (0.26) than without (0.18) warping (Fig.10). In addition, we can



**Fig. 8** Comparison of  $\overline{NPV}$  obtained at the end of the optimization with and without warping.

visually observe that the warping step allow to improve the optimization speed on the Fig.11 and 12 which present the results of the algorithm developped by [Coralie's comment: reference???

## 5 Conclusion

What we did (the problem we solved)

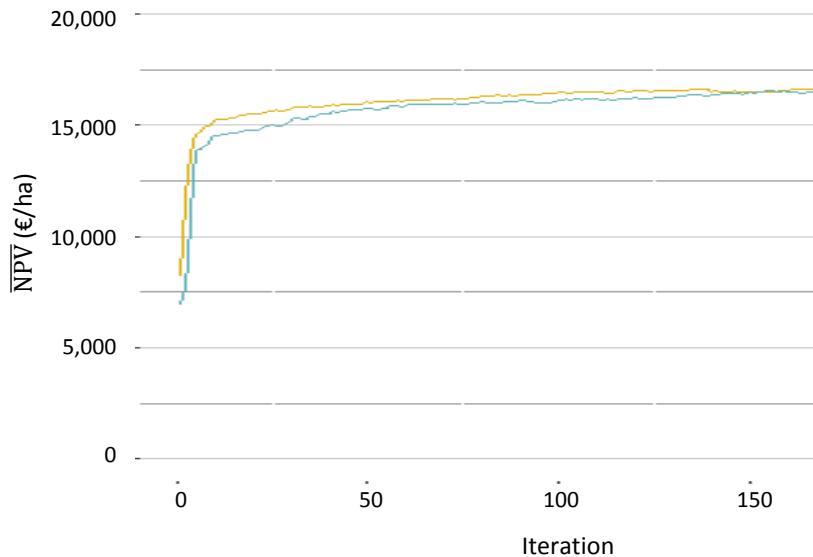
What we proposed: warping to tackle invariances. Proof of concept

Possible extensions

## References

### References

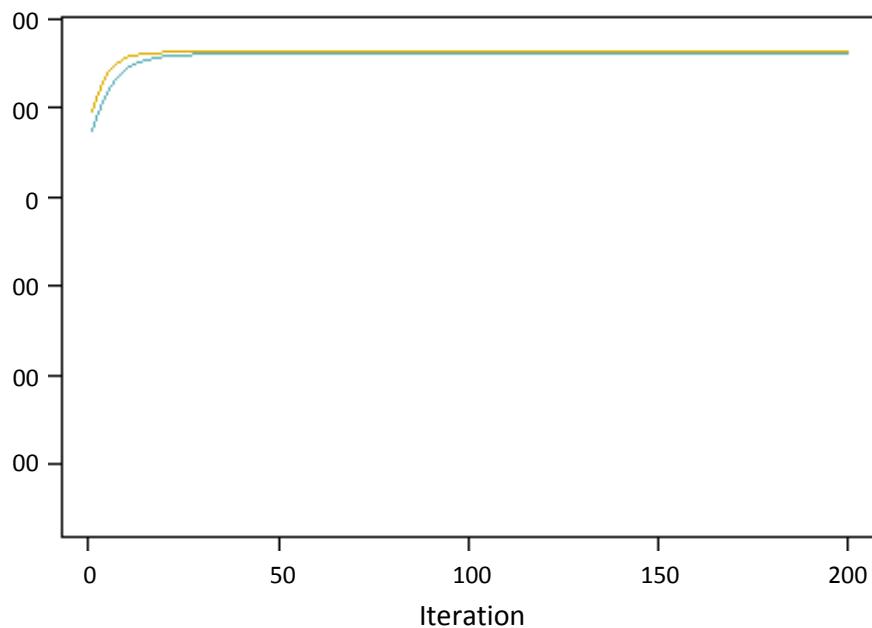
1. Bajardi P, Barrat A, Savini L, Colizza V (2012) Optimizing surveillance for live-stock disease spreading through animal movements. *Journal of the Royal Society Interface* 9(76):2814–2825
2. Box GE, Draper NR (1987) Empirical model-building and response surfaces. John Wiley & Sons
3. Cambra M, Capote N, Myrta A, Llácer G (2006) Plum pox virus and the estimated costs associated with sharka disease. *EPPO Bulletin* 36(2):202–204
4. Cunniffe NJ, Koskella B, Metcalf CJF, Parnell S, Gottwald TR, Gilligan CA (2015) Thirteen challenges in modelling plant diseases. *Epidemics* 10:6–10
5. Cunniffe NJ, Cobb RC, Meentemeyer RK, Rizzo DM, Gilligan CA (2016) Modeling when, where, and how to manage a forest epidemic, motivated by sud-



**Fig. 9** Comparison of  $\overline{NPV}$  obtained during optimizations with and without warping. Yellow and blue lines represent the mean of the  $\overline{NPV}$  selected at each iteration for the 50 optimizations respectively performed with and without the warping step.

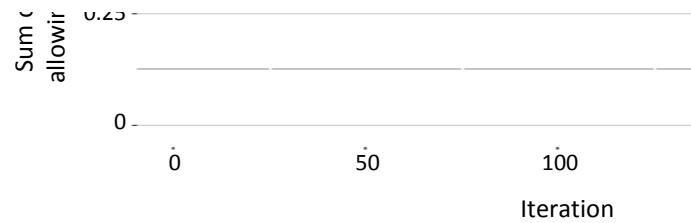
den oak death in California. *Proceedings of the National Academy of Sciences* 113(20):5640–5645

6. Duvenaud D (2014) Automatic model construction with gaussian processes. PhD thesis, University of Cambridge
7. Fang KT, Li R, Sudjianto A (2005) Design and modeling for computer experiments. Chapman and Hall/CRC
8. Ginsbourger D, Durrande N, Roustant O (2013) Kernels and designs for modelling invariant functions: From group invariance to additivity. In: mODa 10—Advances in Model-Oriented Design and Analysis, Springer, pp 107–115
9. Grechi I, Ould-Sidi MM, Hilgert N, Senoussi R, Sauphanor B, Lescourret F (2012) Designing integrated management scenarios using simulation-based and multi-objective optimization: Application to the peach tree–myzus persicae aphid system. *Ecological Modelling* 246:47–59



**Fig. 10** Non linear regression on  $\overline{NPV}$  obtained at each iteration of the optimizations with (yellow) and without (blue) warping.

10. J Forrester AI, Keane AJ, Bressloff NW (2006) Design and analysis of "noisy" computer experiments. *AIAA journal* 44(10):2331–2339
11. Jones DR, Schonlau M, Welch WJ (1998) Efficient global optimization of expensive black-box functions. *Journal of Global optimization* 13(4):455–492
12. Kompas T, Ha PV, Nguyen HTM, East I, Roche S, Garner G (2017) Optimal surveillance against foot-and-mouth disease: the case of bulk milk testing in australia. *Australian Journal of Agricultural and Resource Economics* 61(4):515–538
13. Marmin S, Ginsbourger D, Baccou J, Liandrat J (2018) Warped gaussian processes and derivative-based sequential designs for functions with heterogeneous variations. *SIAM/ASA Journal on Uncertainty Quantification* 6(3):991–1018
14. Mockus J (2012) Bayesian approach to global optimization: theory and applications, vol 37. Springer Science & Business Media
15. Mushayabasa S, Tapedzesa G (2015) Modeling the effects of multiple intervention strategies on controlling foot-and-mouth disease. *BioMed research international* 2015
16. Picard C, Samuel S, Jacquot E, Thbaud G (2018) Analyzing the influence of landscape aggregation on disease spread to improve management strategies. (in



**Fig. 11** Results of the Coco algorithm with (yellow) and without (blue) warping ( $\alpha \in [0;18,012.12]$ ).

- prep)
17. Picheny V, Wagner T, Ginsbourger D (2013) A benchmark of kriging-based infill criteria for noisy optimization. Structural and Multidisciplinary Optimization 48(3):607–626
  18. Pleydell DR, Soubeyrand S, Dalloz S, Labonne G, Chadœuf J, Jacquot E, Thébaud G (2018) Estimation of the dispersal distances of an aphid-borne virus in a patchy landscape. PLoS computational biology 14(4):e1006,085
  19. Rasmussen CE, Williams C (2006) Gaussian Processes for Machine Learning. MIT Press, URL <http://www.gaussianprocess.org/gpml/>
  20. Rimbaud L, Dalloz S, Gottwald T, Decroocq V, Jacquot E, Soubeyrand S, Thébaud G (2015) Sharka epidemiology and worldwide management strategies: learning lessons to optimize disease control in perennial plants. Annual review of phytopathology 53:357–378

Iteration

**Fig. 12** Results of the Coco algorithm with (yellow) and without (blue) warping ( $\alpha \in [10,000;18,012.12]$ ).

21. Rimbaud L, Bruchou C, Dallot S, Pleydell DR, Jacquot E, Soubeyrand S, Thébaud G (2018) Using sensitivity analysis to identify key factors for the propagation of a plant epidemic. Royal Society open science 5(1):171,435
22. Rimbaud L, Dallot S, Bruchou C, Thoyer S, Jacquot E, Soubeyrand S, Thébaud G (2018) Heuristic optimisation of the management strategy of a plant epidemic using sequential sensitivity analyses. bioRxiv p 315747
23. Rios LM, Sahinidis NV (2013) Derivative-free optimization: a review of algorithms and comparison of software implementations. Journal of Global Optimization 56(3):1247–1293
24. Roustant O, Ginsbourger D, Deville Y (2012) DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based meta-modeling and optimization. Journal of Statistical Software 51(1):1–55, URL <http://www.jstatsoft.org/v51/i01/>
25. Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N (2016) Taking the human out of the loop: A review of bayesian optimization. Proceedings of the IEEE 104(1):148–175

26. Snelson E, Ghahramani Z, Rasmussen CE (2004) Warped gaussian processes. In: Advances in neural information processing systems, pp 337–344
27. Snoek J, Swersky K, Zemel R, Adams R (2014) Input warping for bayesian optimization of non-stationary functions. In: International Conference on Machine Learning, pp 1674–1682
28. Tildesley MJ, Savill NJ, Shaw DJ, Deardon R, Brooks SP, Woolhouse ME, Grenfell BT, Keeling MJ (2006) Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the uk. *Nature* 440(7080):83
29. VanderWaal K, Enns EA, Picasso C, Alvarez J, Perez A, Fernandez F, Gil A, Craft M, Wells S (2017) Optimal surveillance strategies for bovine tuberculosis in a low-prevalence country. *Scientific reports* 7(1):4140