Revealing the dependence structure of scenario-like inputs in numerical environmental simulations using Gaussian Process regression

Jeremy Rohmer, Olivier Roustant, Sophie Lecacheux, Jean-Charles Manceau

Revealing the dependence structure of scenario-like inputs in numerical environmental simulations using Gaussian Process regression

Jeremy Rohmer[1], Olivier Roustant[2], Sophie Lecacheux[3], Jean-Charles Manceau[1]

[1]{BRGM, 3 av. C. Guillemin - 45060 Orléans Cedex 2 - France}

[2]{INSA Toulouse, 135 avenue de Rangueil - 31077 Toulouse cedex 4 - France}

[3]{BRGM - Direction régionale Nouvelle-Aquitaine, Parc Technologique Europarc, 24 Avenue Léonard de Vinci - 33600 Pessac - France}

Correspondence to: J. Rohmer (j.rohmer@brgm.fr)

Abstract

Model uncertainties (related to the structure/form of the model or to the choice of "appropriate" physical laws) are generally integrated in environmental long running numerical simulators via scenario-like variables. By focusing on Gaussian Processes (GP), we show how different categorical covariance functions (exchangeable, ordinal, group, etc.) can bring valuable insights into the inter-dependencies of these scenarios. Supported by two real case applications (cyclone-induced waves and reservoir modelling), we have proposed a cross-validation approach to select the most appropriate covariance function by finding a trade-off between predictability, explainability, and stability of the covariance coefficients. This approach can be effectively used to support (or contradict) some physical assumptions regarding the scenario-like input. Through comparison to tree-based techniques, we show that GP models can be considered a satisfactory compromise when only a few model runs (~100) are available by presenting a high predictability and a concise and graphical way to map the dependence.

Keywords: Categorical variables; Computationally intensive simulator; Metamodel; Kriging; Model selection

1 Introduction

High-resolution numerical simulators are key ingredients of environmental science in order to get deeper insights into the natural systems' behavior. Some examples are Veeck et al. (2020) for hydrologic modelling; Zhao et al. (2013) for agricultural modelling; Vandromme et al. (2020) for landslide modelling; Abily et al. (2016) for urban flooding; Idier et al. (2020) for marine flooding, etc. To model the natural system, these simulators all have in common to involve a large spectrum of assumptions related to the system's geometry, to the loading/forcing conditions acting on the system, to the system's intrinsic physical law, to the properties' values, etc. While most of these modelling assumptions can mathematically be represented by means of continuous variables (like geotechnical properties of a given soil formation, or time series of rainfall conditions at a given location, etc.), some of them involve scenario-like variables.

This type of variable is assigned to different modelling scenarios like the choice in the structure/form of the model (e.g. 1D versus 2D modelling, Leandro et al., 2009), the selection of the physical processes regarded as "relevant and prominent" (e.g. account for spatial heterogeneity, Liu et al., 2017), the use of alternative physical laws (e.g. different soil water retention curves, Silva Ursulino et al., 2019), the system's future evolution (e.g., future gas emission scenarios, Le Cozannet et al., 2015; or land use change, Mishra et al., 2018), etc.

Depending on the modelling scenarios, the simulation results can differ, hence resulting in uncertainty. This category of uncertainty can be termed as *structural* since it is associated to the structure/form of the model or to the unambiguous choice of the "best" model to be used: this type of uncertainty should be understood with respect to parametric uncertainties, which are related to the difficulties in estimating the model input parameters (in a broad sense) due to the limited number, poor representativeness (caused by time, space and financial limitations), and imprecision of observations/data (e.g., Hill et al., 2013).

A pragmatic approach for mathematically representing a scenario-like variable is via a categorical variable: a multi-level indicator that takes up a finite number of discrete values; each discrete level being associated to a different scenario (e.g. level *a* is associated to scenario *a*); see some real case applications in the domain of: safety analysis of radioactive waste disposal by Storlie et al. (2013); earthquake risk assessments by Rohmer et al. (2014); marine

flooding induced by sea level rise by Le Cozannet et al. (2015); reservoir engineering for CO_2 geological storage by Manceau and Rohmer (2016); pollution risk analysis and management by Lauvernet & Helbert (2020), etc. Characterizing and quantifying structural uncertainty by means of categorical multi-level variables raises however several practical questions regarding:
(Q1) the dependence structure of the modelling scenarios: should each level (i.e. modelling scenario) of the considered categorical variable be treated equivalently with respect to the numerically simulated variable of interest? is there any dependence among the levels?, and can several levels be grouped?;
(Q2) the prediction using categorical variables: how to derive a predictive statistical model when only a few model runs (of the order of 100s) are available because of too high computation time cost? What is the performance of the derived model? (e.g. does it explain well the observations? what is its predictive capability?).
The objective of the present work is to explore how Gaussian Processes (Williams and Rasmussen, 2006), denoted GP, with mixed continuous/categorical inputs (e.g., Roustant et al., 2020; Qian et al., 2008; Zhang et al., 2020) can bring valuable insights into the inter-dependencies of the modelling assumptions, while fulfilling key characteristics like predictability (i.e. capability of the derived model of predicting "yet-unseen" input configurations) and explainability (i.e. capability of the derived model of representing the data). These characteristics are further detailed in Sect. 2.4.

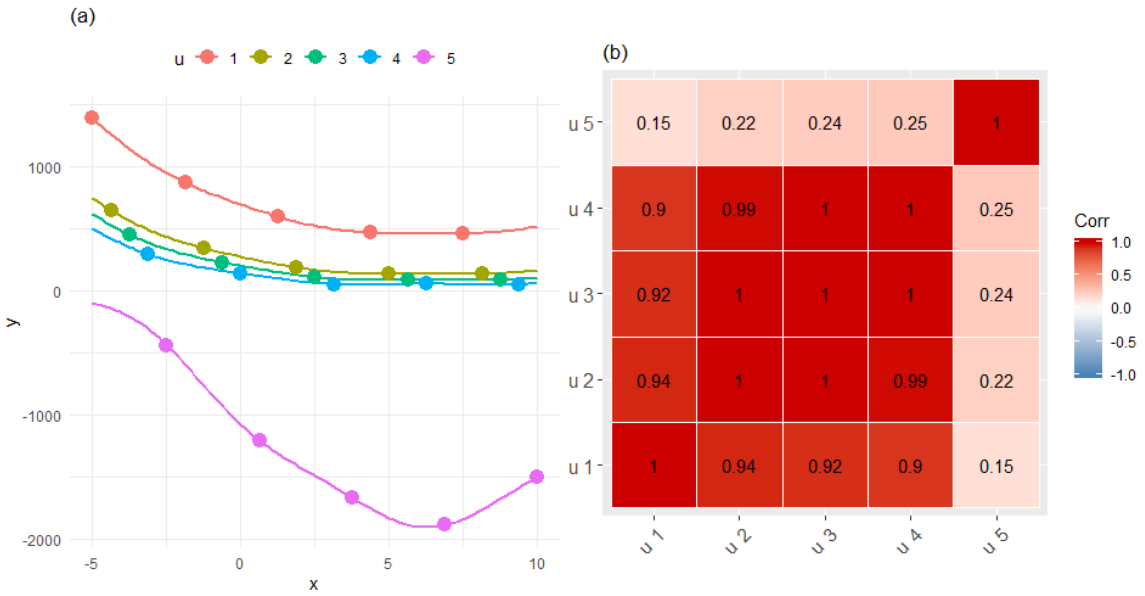

Figure 1 (a). Synthetic test function with continuous input variable *x*. Each color indicates a different level of the scenario-like (categorical) variable *u*. Each dot corresponds to a model run. (b) Correlation matrix for *u* derived from the GP-based analysis using 25 model runs.

The advantage of the GP-based method lies in the flexibility brought by the use of covariance functions (also known as kernels) which specify how similar (i.e. correlated) two instances of the variable of interest, e.g., y and y_0, are expected to be at two input values u and u_0 (i.e., in our case, at two levels of the scenario-like input). This "similarity" function can be encoded in different manners depending on the assumptions regarding the categorical variable (nominal/ ordinal, inter-dependence between the levels, interactions between given levels, etc.); see for instance Roustant et al. (2020); Lauvernet & Helbert (2020). Once fitted, the resulting correlation matrix provides the mapping of the dependence structure that can be used to support (or contradict) some physical assumptions regarding the scenario-like input's influence.

To illustrate the type of results that can be derived, Figure 1(a) depicts an unknown relation between a continuous and a categorical input variable with 5 levels (i.e. five scenarios). Figure 1(b) depicts the correlation matrix derived from the GP-based analysis given 25 model runs: this summarizes the interplay between the levels; a group of highly correlated levels are identified for $u_{1\text{-}4}$ as well as the decreasing correlation of u_5 with the others (from 25 to 15% considering u_4 to u_1). These observations are consistent with the test function. If the functional relation in Figure 1(a) had been known, this result would have been straightforward, but here the structure is unknown and can be learnt only with a limited number of numerical results (here with only 25 model runs). This result (that is further discussed in Sect. 3.1) depends on how the GP kernel is defined, which raises the question of covariance kernel model selection that is addressed in the present study by relying on a multi-criterion analysis.

The present paper is organized as follows. Section 2 describes the different steps of the proposed procedure as well as the statistical methods. In this section, a multi-criterion approach for selecting the categorical covariance kernel model is detailed. Section 3 presents the application to the synthetic case (described in Figure 1) and to two real cases, namely for cyclone-induced wave numerical modelling (Rohmer et al., 2016), and for reservoir modelling of CO_2 storage (Manceau and Rohmer, 2016). The results are then discussed in Section 4 based on which practical recommendations are defined.

2 Methods

2.1 Description of the procedure

The proposed procedure holds as follows:

- Step 1: a series of random computer experiments are performed by running the expensive-to-evaluate environmental simulator considering a limited number of randomly selected input variables' configurations;
- Step 2: using the set of random computer experiments (training dataset), different hypotheses regarding the structure of the considered input categorical variable are tested and modelled by means of different GP kernel (covariance) formulations (see further details in Sect. 2.2 and 2.3);
- Step 3: the question of selecting the most appropriate GP kernel is examined by analysing different aspects, i.e. by considering different criteria as described in Sect. 2.4. The objective is to select the resulting GP model which can achieve a trade-off between the different criteria;
- Step 4: since the practitioner is preferably interested in the dependence structure between the scenarios, the correlation matrix derived from the covariance matrix is analysed (see further explanation in Sect. 2.2): this summarizes the inter-dependencies between the levels (scenarios). This result can be confronted to some *a priori* physically-based interpretation of the scenario-like variable's influence that the practitioner may have before analysing the computer experiments.

2.2 Gaussian Process for mixed continuous and categorical inputs

Let us consider the set of d continuous input variables $\mathbf{x}=(x_1,\ldots,x_d)$, and the set of J categorical inputs $\mathbf{u}=(u_1,\ldots u_J)$ with $L_1,\ldots, L_J$ levels that represent the scenario-like inputs. The output y is then computed using the numerical environmental simulator f(.) as $y = \mathrm{f}(\mathbf{x},\mathbf{u}) = \mathrm{f}(\mathbf{w})$.

In the context of Gaussian Process (GP) modelling (also named as kriging, Williams and Rasmussen, 2006), the function f(.) is assumed to be a realization of a GP ($Y(\mathbf{w})$) with a constant mean m and a covariance function $k(.,.)$, named kernel, that can be written as follows:

$$\forall \mathbf{w},\mathbf{w}', \mathrm{k}(\mathbf{w},\mathbf{w}') = \mathrm{cov}\big(Y(\mathbf{w}), Y(\mathbf{w}')\big) \quad (1)$$

Let denote $(\mathbf{w}^1,\ldots,\mathbf{w}^n)$ the training samples and $\boldsymbol{y} = (\mathrm{y}^1 = \mathrm{f}(\boldsymbol{w}^1),\ldots,\mathrm{y}^n = \mathrm{f}(\boldsymbol{w}^n))$ the corresponding results. The prediction at a new observation $\mathbf{w}^*$ is given by the kriging mean $\hat{Y}(\mathbf{w}^*)$ as follows:

$$\hat{Y}(\mathbf{w}^*) = \mathrm{E}(Y(\mathbf{w}^*)|Y(\mathbf{w}^1) = y^1,\ldots,Y(\mathbf{w}^n) = y^n) = m + \boldsymbol{c}_{\mathbf{w}^*}^T.\mathbf{C}^{-1}.(\boldsymbol{y} - m\mathbf{I}) \quad (2)$$

where **C** is the covariance matrix between the points $Y(\mathbf{w}^1),\ldots,Y(\mathbf{w}^n)$ whose element is $C[\mathrm{i},\mathrm{j}] = k(\mathbf{w}^i, \mathbf{w}^j)$; $\boldsymbol{c}_{\mathbf{w}^*}$ is the vector composed of the covariance between $\mathrm{Y}(\mathbf{w}^*)$ and the points $Y(\mathbf{w}^1),\ldots, Y(\mathbf{w}^n)$, and **I** is the vector of ones of length n.

The prediction at $\mathbf{w}^*$can be associated to an error estimate provided by the kriging variance $\hat{\sigma}^2$ given by:

$$\hat{\sigma}^2(\mathbf{w}^*) = \mathrm{Var}(Y(\mathbf{w}^*)|Y(\mathbf{w}^1) = y^1, \ldots, Y(\mathbf{w}^n) = y^n) = \mathrm{C}(\mathbf{w}^*, \mathbf{w}^*) - c_{\mathbf{w}^*}^T . \mathbf{C}^{-1} . c_{\mathbf{w}^*} \qquad (3)$$

Accounting for a mixture of input variables' types - continuous or categorical (ordinal or nominal) - is made via the covariance function $k(\mathbf{w}, \mathbf{w}')$. Here, it is here assumed to be the tensor product of the covariance function for the continuous inputs $k_{\mathrm{cont}}(\mathbf{x}, \mathbf{x}')$ and the one for the categorical inputs $k_{\mathrm{cat}}(\mathbf{u}, \mathbf{u}')$ as $k(\mathbf{w}, \mathbf{w}') = k_{\mathrm{cont}}(\mathbf{x}, \mathbf{x}') \prod_{i=1}^{I} k_{\mathrm{cat}}^i(u_i, u'_i)$. Hence, the covariance function k_{cont} can be described by kernel models that are commonly-used in the computer experiment community. In the present study, we restrict the analysis to the stationary two differentiable Matérn 5/2 model (Santner et al., 2003). The categorical covariance functions k_{cat}^i $(i = 1, \ldots, \mathrm{J})$ can be described in different manners depending on the assumption related to the scenario-like (categorical) input, as described in Sect. 2.3.

In practices, k_{cat}^i can be interpreted, under the homoscedastic assumption, as the kernel (up to a multiplicative constant) of the 1D section $u_i \to Y(\boldsymbol{x}, u_i, \boldsymbol{u}_{-i})$ where **x**, and $\boldsymbol{u}_{-i} = (u_1, \ldots, u_{i-1}, u_{i+1}, \ldots u_I)$ are fixed. In practices, we preferably use the scaled form of the covariance, i.e. the correlation to ease the interpretation on the dependencies between the levels of the categorical variable (i.e. of the scenarios). With the same notations, the correlation kernel (derived from k_{cat}^i) is interpreted as the correlation of the 1D section $u_i \to Y(\boldsymbol{x}, u_i, \boldsymbol{u}_{-i})$, whatever $\boldsymbol{x}$ and $\boldsymbol{u}_{-i}$. Thus the inspection of k_{cat}^i reveals the correlation of the simulator output explained by the i$^{\mathrm{th}}$ categorical input, the others being fixed. Note that such correlation does not depend on the other inputs, which is a result of constructing *k*(.,.) by tensor product.

2.3 Covariance kernel models for categorical inputs

The different options for defining a categorical covariance function are summarized in Table 1. For sake of clarity, we restrict the presentation to the case of a single input with L levels, which is hereafter denoted by *u*, because the kernel is constructed by tensor product of 1D categorical inputs (see Sect. 2.2).

Table 1. Different options for representing the scenario-like input variable using a kernel (covariance) model

Assumption	Interpretation of categorical variable	Type of covariance matrix	Symbol	Equation
No preference among the scenarios	Each level acts similarly	Compound Symmetry (also named Exchangeable)	CS	4
The variable of interest will act differently depending on the considered scenario, but without excluding inter-level dependencies	Each level has its own variance coefficient and a between-level structure exists	General	Gen	5
The variable of interest may act similarly depending on some subsets of scenarios	Some scenarios can be grouped based on expert information	Group	E	6
A grouping is a realistic option, but cannot be unambiguously defined	The grouping can be learnt from the data	Low rank approximation	LR	7
The scenarios can be ordered	The levels are discretized values of a latent ordinal variable	Ordinal	O	8

When the practitioner assumes that no preference can be given to the L levels (i.e. all considered scenarios are considered as having the same influence), k_{cat} can be described by an exchangeable covariance (Qian et al., 2008) - also named compound symmetry (denoted *CS*) - function as follows:

$$k_{\text{cat}}^{\text{CS}}(u,u') = \begin{cases} \sigma^2 & \text{if } u = u' \\ \rho.\sigma^2 & \text{if } u \neq u' \end{cases} \quad (4)$$

where ρ is a unique correlation coefficient satisfying $\rho \in \left]-\frac{1}{L-1}, 1\right[$.

When the practitioner assumes that the variable of interest will act differently depending on the considered scenario, but without excluding some dependencies between these different responses, k_{cat} can then be described by the most general (and complex) dependence structure where each pairwise coefficient can take a different value depending on the considered levels u, u'. The covariance function reads as follows:

$$k_{\text{cat}}^{\text{Gen}}(u,u') = \begin{cases} c_{u,u'} & \text{if } u \neq u' \\ v_u & \text{if } u = u' \end{cases} \quad (5)$$

The latter structure can be simplified by adding *a priori* information on the dependence between the levels; for instance by relying on expert-based information. A possible option is to assume that some scenarios perform similarly and that they can be grouped. Assume that the L levels of *u* are partitioned in G groups, and denote *g*(*u*) the group number corresponding to a given level *u*. Then, the covariance function can be written as (Roustant et al., 2020):

$$k_{\text{cat}}^{\text{E}}(u,u') = k_{\text{cat}}^{\text{Gen}}(g(u),g(u')) = \begin{cases} c_{g(u),g(u')} & \text{if } g(u) \neq g(u') \\ v_{g(u)} & \text{if } g(u) = g(u') \end{cases} \quad (6)$$

where for all $i, j \in \{1, \ldots, G\}$, the terms $\frac{c_{i,i}}{v_i}$ are within-group correlation, and $\frac{c_{i,j}}{\sqrt{v_i}\sqrt{v_j}} (i \neq j)$ are between-group correlations. The structure can be simplified by assuming that the correlation value for each pair of groups is unique by means of a compound symmetry matrix (Pinheiro & Bates, 2006).

Instead of deriving the groups based on expert information, a possible option is to derive the groups from the data using a low rank approximation (Roustant et al. 2020) so that the covariance matrix can be defined as:

$$k_{\text{cat}}^{\text{LR}}(u,u') = \mathbf{F}.\mathbf{F}^{\text{T}} \quad (7)$$

where the matrix **F** is of size $L \times q$ where q is low, with typical values of 2 or 3.

When the practitioner assumes that the levels can be ordered, this means that the categorical variable can be described by an ordinal continuous variable that is not directly observed (i.e. it is said to be "latent") and the levels are seen as discretized values of this ordinal variable (Qian

et al., 2008). The corresponding covariance function can be defined by taking advantage of the tools available for the continuous variables as follows:

$$k_{\text{cat}}^{\text{O}}(u, u') = \tilde{k}_{\text{cont}}(F(u), F(\text{u}')) \quad (8)$$

where $F(.)$ is a one-dimensional non-decreasing function (also called warping) so that $F: \{1, \dots, L\} \to \mathbb{R}$, and $\tilde{k}_{\text{cont}}$ is a one-dimensional continuous kernel.

2.4 Model selection

As aforementioned, different kernel modelling choices can be made to represent the categorical variable associated to the scenario-like variable. This raises the question of selecting the most appropriate kernel covariance model. Depending on the modelling objective, different approaches exist for selecting an optimal model with respect to a specific criterion (Burnham and Anderson, 2002); for instance, a model that satisfactorily represents (i.e. explains) the relationships between inputs and outputs might not necessarily perform as well as for prediction. Therefore, we propose to examine different viewpoints on the problem of kernel selection by examining different criteria. This multi-criterion approach shares similarities with the data science framework proposed by Yu & Kumbier (2020), who advocate analysing three core principles, predictability, computability, and stability. To select the most appropriate kernel model (given the training dataset), we investigate whether the considered GP model is capable of:

Predictability. It is related to whether the GP model is capable of predicting "yet-unseen" input configurations, i.e. samples that have not been used for training. This can be examined by using cross-validation approaches (e.g. Hastie et al., 2009). Two indicators are estimated.

The first indicator, denoted Q^2, measures the deviation from the true output value. Given a test set $\boldsymbol{T}$, Q^2 is defined as follows:

$$Q^2 = 1 - \frac{\sum_{i \in \boldsymbol{T}} (y_i - \hat{y}_i)^2}{\sum_{i \in \boldsymbol{T}} (y_i - \bar{y})^2} \quad (9)$$

where $\hat{y}_i$ is the i$^{\text{th}}$ GP-based prediction of the model output y_i, and $\bar{y} = \frac{1}{|\boldsymbol{T}|} \sum_{i \in \boldsymbol{T}} (y_i)$ is the average value for the test set. A coefficient Q^2 close to 1.0 indicates that the GP model is successful in matching the new observations that have not been used for the training.

The second criterion is the coverage of the prediction intervals (denoted PI_α) at the given confidence level α calculated on the test set $\boldsymbol{T}$ and holds as follows:

$$CA = \frac{1}{|\boldsymbol{T}|}\sum_{i\in\boldsymbol{T}} \mathbf{1}_{(y_i \in PI_\alpha(\boldsymbol{w}^i))} \quad (10)$$

where $\mathbf{1}_A$ is the indicator function. *CA* evaluates whether the model output y_i is within the bounds of the prediction interval computed at the input variables $\boldsymbol{w}^i$ using the GP mean and variance (Eqs. 2&3). The GP-derived PI_α is "optimal" when *CA* reaches the theoretical value of α.

Explainability and simplicity. The former concept relates to whether the considered GP model is capable of representing the data, for instance by analysing the likelihood *l*. However, adding more model parameters results in increasing the explainability. To counterbalance this tendency (related to overfitting), a penalty term is generally introduced (see e.g., Höge et al., 2018) to select simpler model. Here, simplicity is understood with respect to the number of GP model parameters. By assuming that the true GP model exists and that is among the set of candidate GP models, we propose to rely on the Bayesian Information Crierion BIC (Schwarz, 1978) described as follows:

$$BIC = 2\log(l) + k.\log(n) \quad (11)$$

where *k* is the number of parameters, *n* is the number of observations.

Stability. We explore to which extend the kernel correlation matrix (derived from the covariance matrix) is stable to the perturbations in the training dataset. By following a cross-validation procedure, we evaluate the following error term:

$$err = \frac{1}{n_p}\sum_{i=1}^{n_p}\left(\frac{1}{n_{\mathrm{cv}}}\sum_{j=1}^{n_{\mathrm{cv}}}(\hat{c}_{i,0} - \hat{c}_{i,j})^2\right) \quad (12)$$

where n_p is the number of terms in the correlation matrix (by restricting to the non-diagonal elements in the upper triangular part of the matrix), n_{cv} is the number of folds of the cross-validation procedure, $\hat{c}_{i,0}$ is the i$^{\mathrm{th}}$ coefficient of the correlation matrix (read in the column order for instance) derived from the GP model fitted using the whole training dataset, $\hat{c}_{i,j}$ is the i$^{\mathrm{th}}$ corresponding coefficient derived from the GP model fitted using the training dataset at the j$^{\mathrm{th}}$ iteration of the cross validation procedure (i.e. for a 5-fold cross validation, this training dataset corresponds to the whole training dataset, from which 20% of the observations have been randomly removed).

3 Application

In this section, we apply the GP-based procedure described in Sect. 2 to a synthetic test function (Sect. 3.1) and to two real cases, in the domain of cyclone-induced wave modelling (Sect. 3.2) and reservoir engineering (Sect. 3.3). For all GP models, we consider a Matérn 5/2 covariance matrix for the continuous variables. A GP model with constant trend is fitted using the R package *kergp* (Deville et al., 2018) by applying a pre- scaling and centering of the continuous input variables and of the variable of interest. The covariance parameters are estimated via a maximum likelihood approach using the derivative-free constrained optimiser by linear approximations named *COBYLA* developed by Powell (1994) with 250 randomly selected initial starts. The predictability and the stability are assessment via a 5-fold cross-validation procedure repeated 25 times.

3.1 Synthetic case

We first consider a synthetic test function using a modified version of the two-dimensional Branin function[1] where one continuous variable u is assumed to reach only discrete values as follows:

$$y = \begin{cases} h(x,-20), & if\ u = 1 \\ h(x,-10), & if\ u = 2 \\ h(x,-7.5), & if\ u = 3 \\ h(x,-5.0), & if\ u = 4 \\ -5.h(x,20), & if\ u = 5 \end{cases} \quad (13)$$

where $h(x,z) = \left(z - \frac{5}{4*\pi^2}x^2 + \frac{5}{\pi}x - 6\right)^2 + 10.\left(1 - \frac{1}{8\pi}\right).\cos(x) + 10$, with $x \in [-5,10]$. By construction, levels 1-4 are highly correlated (see Fig. 1a). Different categorical kernels are defined as follows:

- *Compound Symmetry* (denoted *CS*): no preference is given to the scenarios i.e. to the levels of the categorical variable *u*;
- *General* (denoted *Gen*): this is the most generic structure where the pairwise correlation coefficient differs from one level to another;

[1] http://www.sfu.ca/~ssurjano/branin.html

- *Expert-based* groups: levels are clustered based on expert information. We assume that two experts have given their opinions: the first one (denoted *E*) indicates a 'realistic' grouping of levels (i.e. consistent with the true function), and the second one (denoted *EW*) indicates a unrealistic grouping (i.e. (u_1,u_3), and (u_2,u_4,u_5));
- *Low rank* approximation: groups of levels are identified through the data-driven low rank approach. A matrix rank of 2 is here tested (denoted *LR2*);
- *Ordinal* variable: the levels are ordered by following the level index, and a continuous kernel is defined via a spline-based warping (kernel denoted *O*). As for the expert-based grouping, we assume that another expert does not know the intrinsic ordering and assumes an unrealistic ordering (denoted *OW*), i.e. $u_4<u_2<u_5<u_3<u_1$.

The training dataset is defined through random sampling by considering *m* points per level with *m*=4, 5 and 6, i.e. with different training dataset sizes of 20, 25 and 30 respectively.

Fig. 2 depicts the GP-derived correlation matrices for each of the afore-described kernel assumptions using the training dataset of intermediate size (*m*=5). The application of the expert-based and of the ordinal kernel assumptions (Fig. 2c,e) shows some consistent structures among the levels, namely the high-correlated group for u_1 to u_4 and the particular behaviour of u_5. The magnitude of the inter-dependencies between the identified group and u_5 slightly differs: assumption *E* indicates a low-to-moderate correlation (~10%), whereas assumption *O* indicates a decreasing correlation from 25 to 15% for $u_{4\text{-}1}$. The structure of inter-level dependencies is here richer for the general (Fig. 2a) and the LR2-based GP models (Fig. 2b). Though the former assumption is hardly exploitable, the latter still allows identifying some interesting features, namely the particular behavior of u_5, and a possible grouping of $u_{2\text{-}4}$. In this case, the exchangeable assumption (i.e. compound symmetry *CS*) leads to a low-to-moderate inter-correlation coefficient of ~40% (not shown).

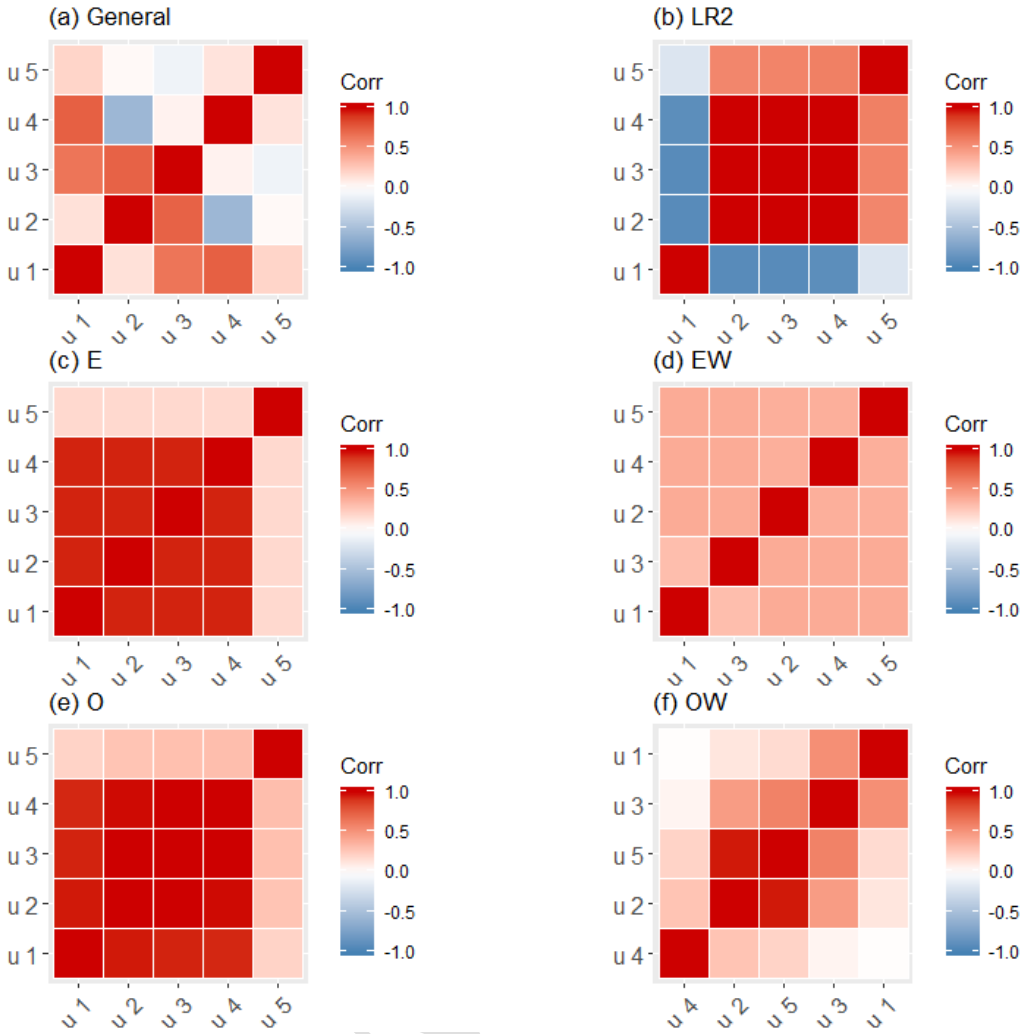

Figure 2. Correlation matrix for the synthetic test case (with m=5) considering different assumptions regarding the kernel associated to the categorical input variable denoted u. For the expert-based assumptions (c and d), the ordering of matrix coefficients follows the expert-based clustering, i.e. (u_1-u_4), and (u_5).

The four criteria for kernel model selection are examined in Fig. 3. For the sake of comparability between the different assumptions for m, we preferably plot the difference of BIC with the minimum one (named ΔBIC). Several observations can be made:

- The criterion Q^2 is commonly used in the computer experiment community to rank different models with respect to their predictive capability. In our case, basing the analysis on this unique criterion is difficult: at low size of the training dataset (m=4), models *E*, *CS* and *O* all minimize 1-Q^2, and show similar performance (see median values in Fig. 3b): selecting one of them is here hardly achievable. Besides, the model

associated to the unrealistic expert-based grouping *EW* turns to have a satisfactory predictive capability at low *m* value (though it should be noted that the width of the confidence interval is larger than the others);

- For large enough *m* value - here *m*=6 - (i.e. with the largest size of the training dataset), the Q^2 criterion allows to identify model *O* as the most appropriate one with respect to the predictive capability ($\log_{10}(1-Q^2)$ is the lowest in Fig. 3b);
- The other facet of predictability related to the coverage of the prediction intervals turns to be informative to discard kernel assumptions *Gen* and *LR2* (because *CA* is here far larger than the level of the prediction interval), but hardly allows differentiating the other assumptions whatever *m* (Fig. 3c);
- Explainability measured by BIC (Fig. 3a) appears here efficient to exclude the unrealistic assumptions *OW* and *EW*: they present very large BIC differences whatever *m*. For instance, Burnham & Anderson (2002) suggested a difference of the considered information criterion (relative to the minimum value) of at least 10 to support the ranking between model candidates with confidence;
- BIC appears to be very informative to discriminate the kernel assumptions, and clearly selects the ordinal assumption as the most appropriate one.
- The stability criterion (Fig. 3d) tends to suffer from the same sensitivity than Q^2 to the size of the training dataset, but has a higher discriminative power: at low *m* value (*m*=4), both kernel models *CS* and *O* are selected as very stable (because of low $\log_{10}(err)$ values in Fig. 3d). It should however be noted that a high stability is also reached for *EW*. At a high enough *m* value, the ordinal assumption is then unambiguously selected as the one leading to the most stable correlation matrix.

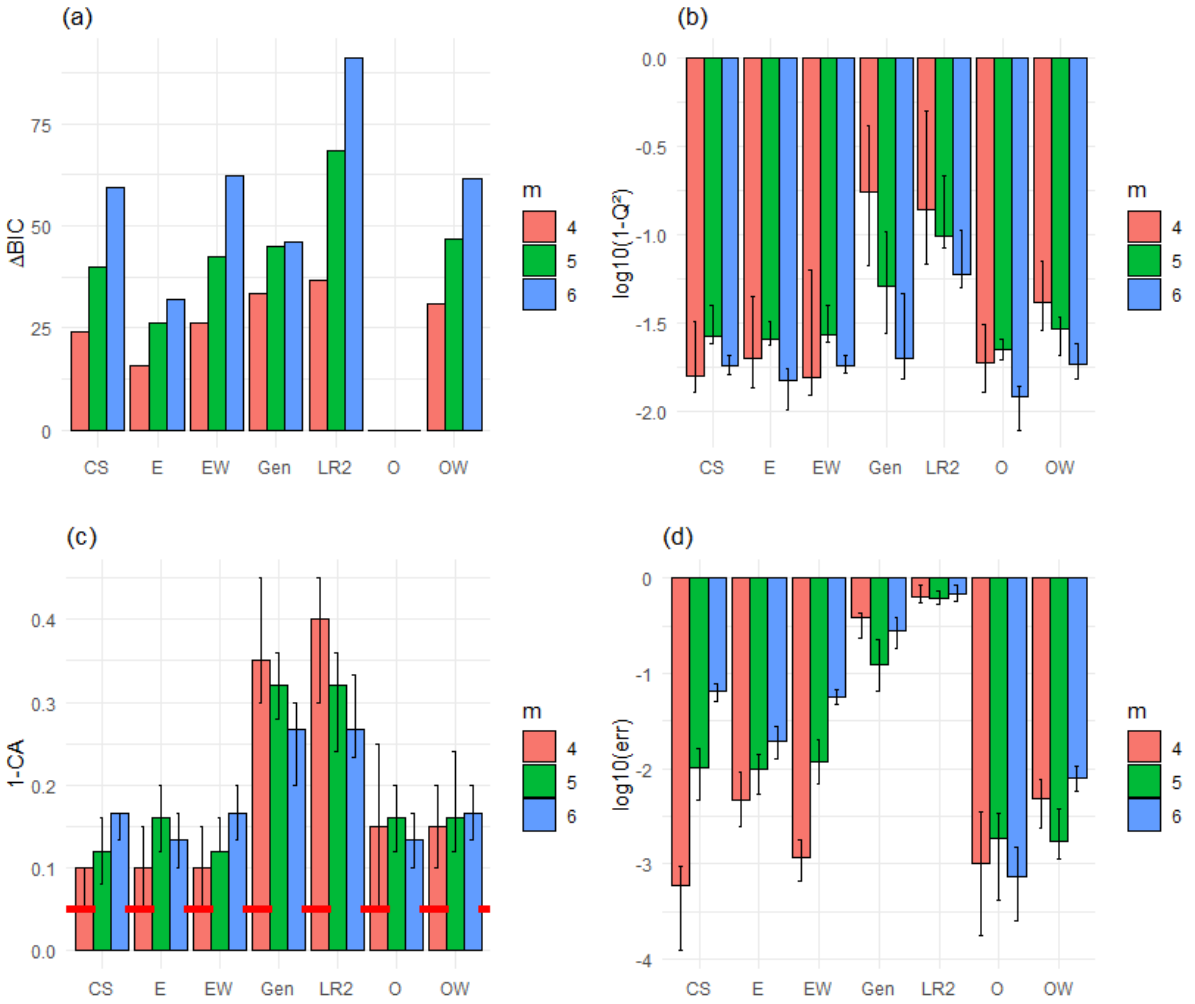

Figure 3. Selection criterion for the synthetic test case considering different numbers of points per level *m*: (a) BIC difference with respect to the minimum value; (b) Predictability measured by 1-Q^2 (log10 scale); (c) Coverage measured by 1-*CA*. The horizontal dashed line corresponds to 5%, i.e. the threshold consistent with the level of the 95% prediction interval; (d) Stability error *err* (log10 scale). The three latter criteria are derived from the 5-fold cross validation repeated 25 times: the height of the barplot is the median value and the lower and upper bounds are defined using the 25th and 75th percentiles.

From this analysis, we can conclude that the ordinal assumption allows to successfully fulfill three of the criteria (explainability, predictability, and stability) especially at high enough *m* value (m≥5). For coverage, the ordinal assumption is not ranked first, but *CA* appears of reasonable order of magnitude (median value of ~85%), i.e. with moderate deviation from the level of the 95% predictive interval. For low size of the training dataset (*m*=4), selecting

unambiguously *EW* and *CS* as valid assumptions turns to be difficult if the explainability criterion (BIC criterion) is not taken into account. This can partly be explained by the analysis of the *EW*-based correlation matrix (Fig. 2d), which reveals a quasi-homogeneous structure with correlation coefficients ranging from 44 to 56%, i.e. of the same order of magnitude than the *CS* assumption (of ~40%), hence indicating that both GP models should perform similarly. This suggests that at a low number of training points, the *CS* assumption remains the most reasonable assumption (when the goal is the joint maximization of predictability, stability and explainability).

3.2 Real case application 1: cyclone-induced waves

The first real case is based on Rohmer et al. (2016) and deals with the modelling of waves induced by cyclones at Sainte Suzanne city located in the North East of Reunion Island (Fig. 4b). The aim is to analyse the evolution of the significant wave height H_s (maximum value over time) as a function of the cyclone characteristics. These are modelled by means of five scalar continuous input parameters, namely the maximum wind speed, the radius of maximum winds (i.e. the distance from the cyclone eye at which the maximum wind intensity is reached); the shift around the central pressure; the forward speed defined as the translation speed of the cyclone eye, and the landfall position, that both characterize the minimum distance and the relative position of the track to the studied site. A set of seven historical cyclone tracks are considered: these are randomly shifted (via the continuous input variable modelling the relative position of the track) from their original track so that they cross the centre of Reunion Island (see Fig. 4b): a categorical variable is here defined; each level corresponding to a given track. A series of 100 computer experiments were performed by randomly sampling the five scalar inputs using a Latin Hypercube Sampling approach combined with a *maximin* criterion (Johnson et al., 1990). The sampling of the cyclone tracks is done by sampling with replacement.

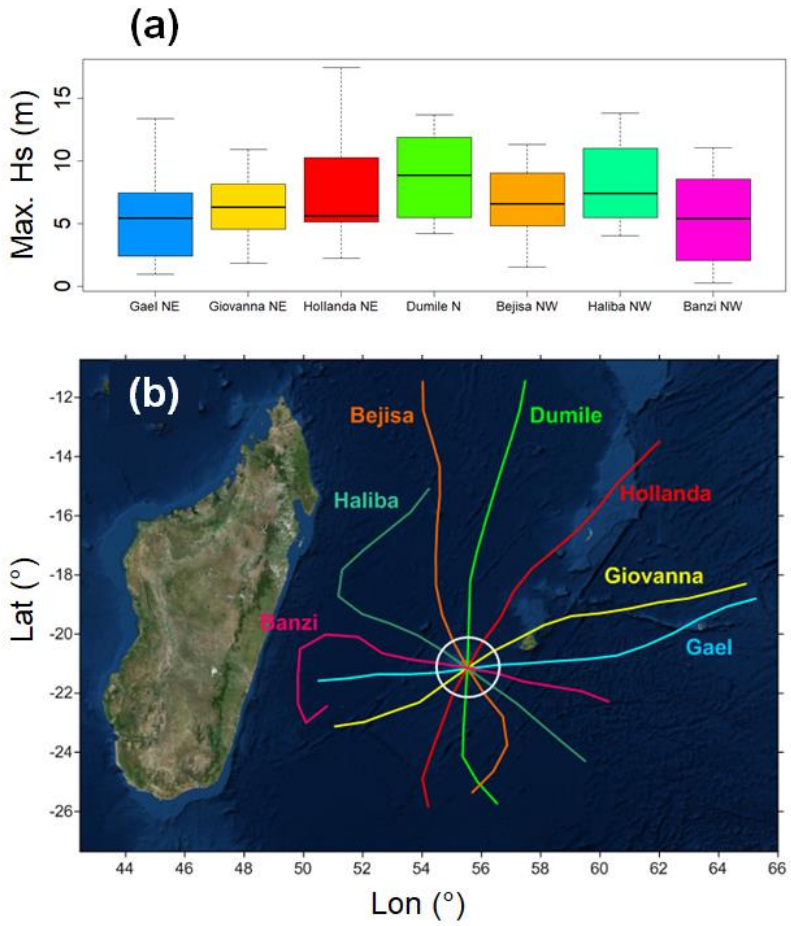

Figure 4. (a) Boxplot of the maximum significant wave height H_s (m) considering each cyclone track ordered according to the angle of approach from 0° to 180° (from the East – leftmost part, to the West – rightmost part); (b) Cyclone tracks used for modelling the waves at Saint Suzanne city (Reunion island).

An *a priori* physical interpretation of the track influence speculates that H_s is strongly related to the angle of approach of the cyclone in the vicinity of the studied site, which increases from 0° (East direction) to 180° (West direction); 90° being the North. This was confirmed by the sensitivity analysis of Rohmer et al. (2016). Despite the spatial variability of the track (as illustrated in Fig. 4b), the effect of the track scenarios could be summarized by a scalar continuous input, i.e. in relation to the angle of approach. The analysis of the boxplots in Fig. 4a seems to support this hypothesis; in particular, the median value of the maximum H_s appears to increase as the angle of approach increases from 0 to 90° (from *Gael* to *Dumile* cyclone). Yet, the tendency from 90° to 180° (from *Dumile* to *Banzi* cyclone) is less clear, especially for

Banzi; the difficulty in the interpretation may here be related to the complexity of this cyclone track compared to the quasi linear shape of the others (Fig. 4b). To support the evidence of the angle of approach's influence, a more rigorous analysis is here needed: the validity of this hypothesis is further investigated using a GP model with different categorical kernels as follows:

- *Compound Symmetry* (denoted *CS*): no preference is given to the tracks;
- *General* (denoted *Gen*): this is the most generic structure where the pairwise correlation coefficient differs from one level to another;
- *Expert-based* groups (denoted *E*): levels are clustered based on expert information. A first assumption relies on the selection of two groups: one composed of the 4 cyclones (Gael, Giovanna, Hollanda, and Dumile) coming from the north-eastern (NE) quadrant and another one composed of 3 tracks (Bejisa, Haliba, and Banzi) coming from the north-western (NW) quadrant (Fig. 4). A second assumption based on three groups is also tested by differentiating the track whose angle is quasi at 90° (North), namely *Dumile* cyclone (Fig. 4). In addition, two assumptions are made regarding the link between the groups by specifying a general (assumption *E2* and *E3*), or a compound symmetry covariance (assumption *E2cs* and *E3cs*);
- *Low rank* approximation: groups of levels are identified through the low rank approach with a assumed rank of 2 or of 3 (assumptions *LR2* and *LR3*);
- *Ordinal* variable: the levels are ordered following the angle of approach, and a continuous kernel is defined via a spline-based warping (kernel denoted *O*).

Fig. 5 depicts the GP-derived correlation matrices for each of the afore-described kernel assumptions. Fig. 5a reflects the hypothesis of a single group without any preference among the tracks. The derived correlation appears to be high (~80%). The application of alternative kernel assumptions show some consistent structures among the cyclone tracks. Two groups of cyclones appear to be highly correlated (coefficient >75%), namely the ones coming from NE, and the ones coming from NW as shown by Fig. 5b (General formulation), and Fig. 5e,f (*E2* and *E2cs*). The high correlation among these groups is also indicated by the other assumptions: (1) the low rank approximation, *LR2* and *LR3* (Fig.5c, d) - but these assumptions lead to a richer inter-dependency structure; (2) to a lesser extent,the expert-based assumption *E3cs* (Fig. 5h). However, we can note that there is an ambiguity for *Dumile* cyclone (see in particular, Fig. 5g),

which is either highly correlated with the NE cyclones (*E2*, *E2cs*) or with the NW cyclones (*LR3*, *E3cs*). The different assumptions all suggest a moderate correlation (of the order of 40-60%) between groups of cyclones coming from NE and NW. These observations are consistent with the ordinal assumption (Fig. 5h), which indicates a decreasing correlation from *Gael* (the track with the lowest angle of approach) to *Banzi* cyclone (the track with the largest angle of approach), and a central role of *Dumile*, with a decreasing correlation either with the NE or with the NW cyclones.

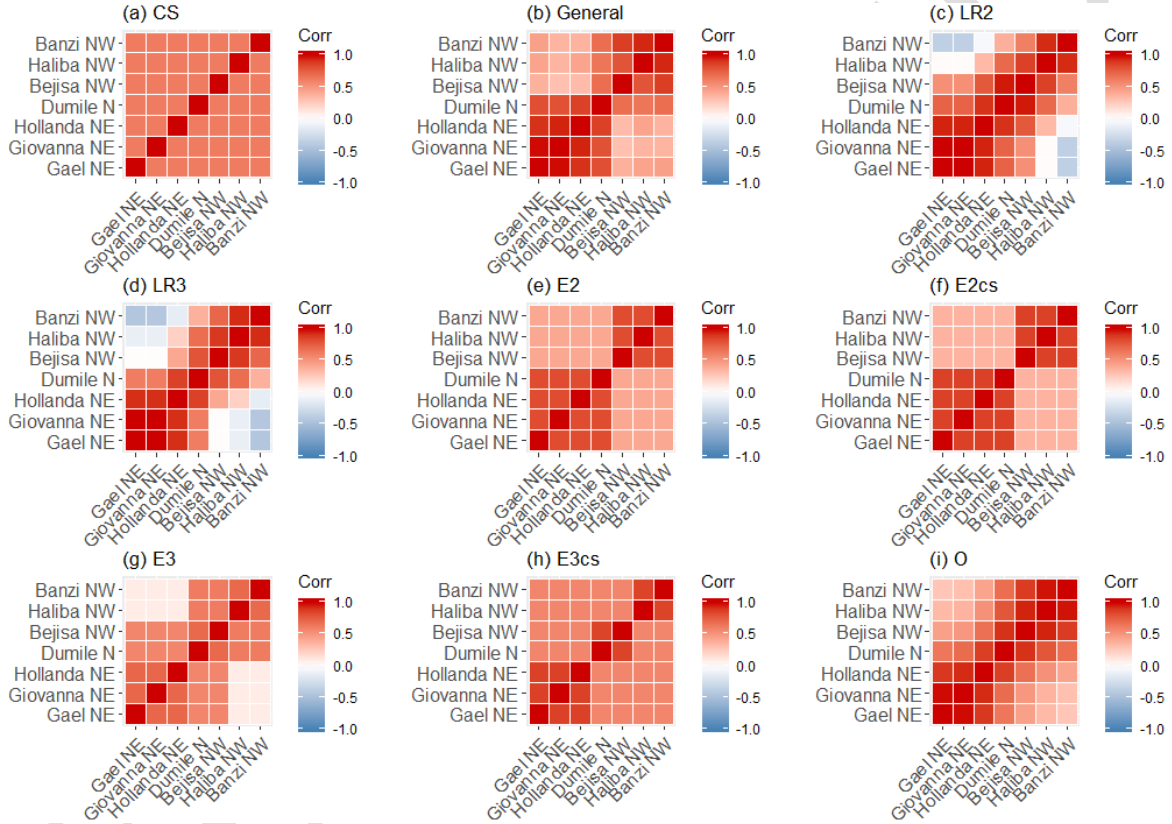

Figure 5. Correlation matrix for the cyclonic test case considering different assumptions regarding the kernel associated to the categorical input variable.

The four criteria of kernel model selection are examined in Fig. 6. Regarding explainability and simplicity, the ordinal assumption *O* appears to be the most appropriate: the derived GP model presents the minimum BIC value and the differences with the alternative models is large; here larger than 20. Regarding predictability, the ordinal assumption also leads to the best model regarding this criterion. Yet, it should be noted that the predictability of the expert-based model

candidates (*E2*, *E2cs*, *E3*, *E3cs*) remains of moderate-to-high degree (with median value of Q^2 around 90%). Regarding *CA*, the use of *E3cs* allows to reach the level of the 95% prediction interval, but alternative assumptions (*E2*, *E2*cs, *E3*cs, *CS* and *O*) lead to a satisfactory coverage of the prediction intervals as well. Finally, Fig. 6d indicates the poor stability of *Gen* and *LR3* kernel, i.e. the high sensitivity of the correlation coefficients' estimates, which may be related to the high number of coefficients to estimate (respectively 21 and 11). This also shows the satisfactory stability of two alternative assumptions, i.e. the expert-based ones and the ordinal one.

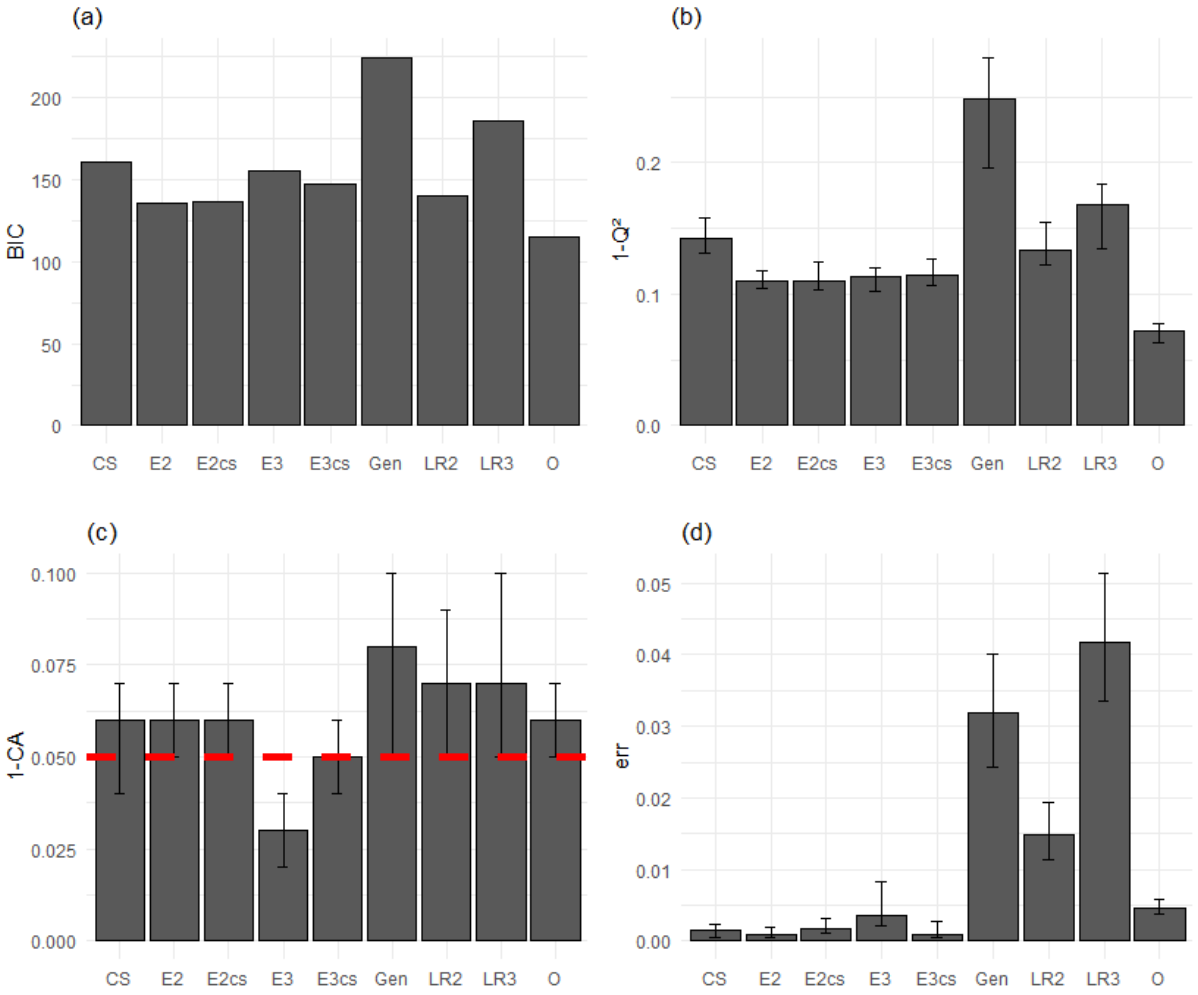

Figure 6. Selection criterion for the application case 1: (a) BIC; (b) Predictability measured by 1-Q^2; (c) Coverage measured by 1-*CA*. The horizontal dashed line corresponds to 5%, i.e. the threshold consistent with the level of the 95% prediction interval; (d) Stability error *err*. Criteria (b-d) are derived from a 5-fold cross validation repeated 25 times: the height of the barplot is the median value and the lower and upper bounds are defined using the 25th and 75th percentiles.

On this basis, it can be concluded that the ordinal assumption allows to reach a satisfactory trade-off between the four criteria. This appears to be consistent with the afore-described physical intuition; the "latent" continuous variable being here related to the angle of approach. Figure 7a summarizes this dependence via the spline-based warping function $F(.)$ used to set-up the ordinal covariance kernel model $k_{\text{cat}}^{\text{O}}$ (see Eq. 8): this shows a strong link between NE cyclones *Gael*-*Giovanna* (and to a lesser extent, for *Hollanda* as well), and a quasi-linear increasing influence to *Banzi*. This warping is the basis for computing the correlation matrix (Fig. 5i). To ease the interpretation, let us focus on a single row of Fig. 5i, i.e. the pairwise correlation between *Gael* track with the others. Fig. 7b provides a clear indication of a highly correlated group of cyclones coming from NE (within the red-coloured envelope), with correlation coefficient exceeding 80%, with decreasing correlation with those coming from NW. The analysis of the correlations derived from the cross-validation iterations (for each repetitions) - grey lines in Fig. 7b - confirms this result: more than 75% of the cross-validation-derived results show high correlation (>75%) of *Dumile* with the NE cyclones, hence in agreement with *E2* assumption (Fig. 5e).

Compared to *O*, the analysis of *E2* performance criteria shows that this assumption can also be considered a reasonable one with very satisfactory stability of the correlation coefficients (Fig. 6d), though the criterion values appear to be higher. The stability criterion appears to be lower than for *O*, which may be related to the lower number of correlation coefficients to be estimated (of 2 for *E2* assumption, and of 7 for *O* assumption).

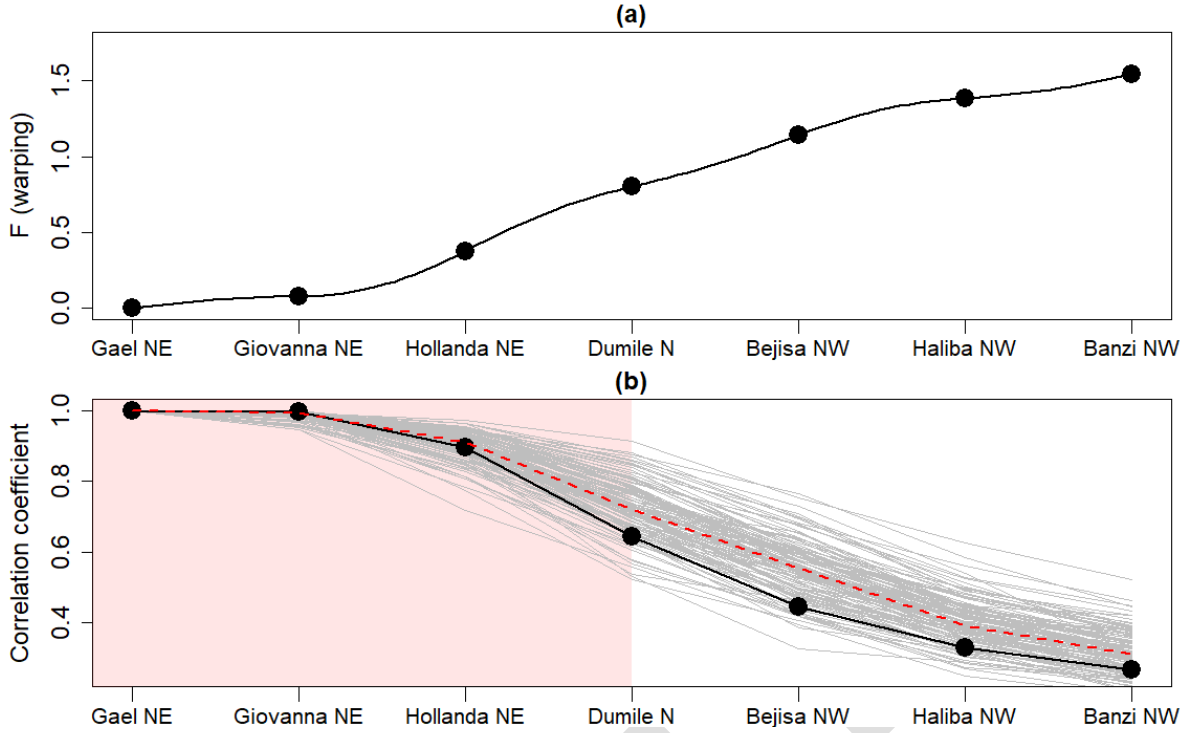

Figure 7. (a) Spline-based (un-normalized) warping function *F* (see Eq. 8) used in the ordinal assumption for the cyclone case. (b) Pairwise correlation between *Gael* track with the other ones. The red-coloured envelope indicates cyclones coming from North-East (denoted NE). The grey-coloured curves are the correlations derived at each iteration of the 5-fold cross-validation procedure (repeated 25 times). The red dashed line indicates the median value.

3.3 Real case application 2: CO_2 geological storage

The second real case application corresponds to the modelling of the long term fate of stored CO_2 in a deep aquifer on a potential project in the Paris basin (France) as described by Manceau and Rohmer (2016). The injection of 30 Mt of CO_2 during 30 years in the lower Triassic sandstone formation at approximately 1,000m depth was numerically simulated and the evolution of the quantity of mobile CO_2 for a time period of 150 years after the injection stops was investigated as a function of:

- two continuous input variables, namely the porosity and the intrinsic permeability of the aquifer rock formation;
- four categorical input variables related to: the scenarios of permeability anisotropy (“minor”, “medium”, and “large”), the scenarios of regional hydraulic gradient

("absence" and "activated"), the scenarios related to the capillary effect ("absence" and "activated"), and the physical laws used to model the relative permeability as a function of CO_2 saturation (ten choices) as depicted in Fig. 8. Due to the importance of the latter parameter (as shown by Manceau and Rohmer, 2016), the following analysis focuses on this variable.

A series of 100 computer experiments were performed by randomly sampling both scalar inputs using a Latin Hypercube Sampling approach combined with a *maximin* criterion. The sampling of the categorical variables is done by sampling with replacement.

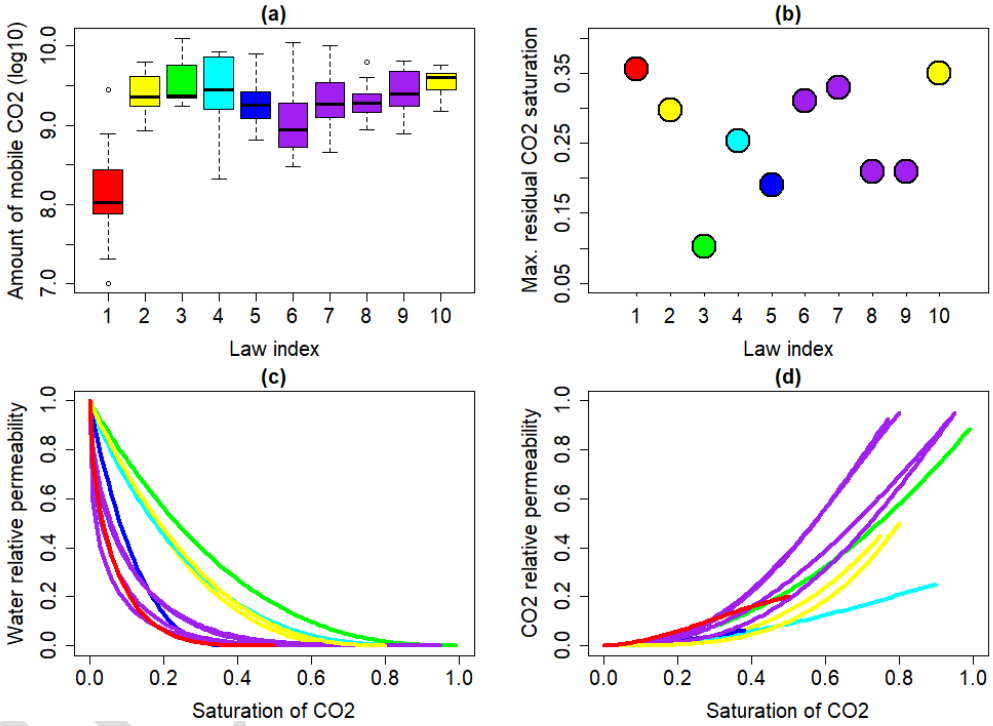

Figure 8. (a) Boxplot of the amount of mobile CO_2 considering each physical law index (1-10). Colors indicate the expert-based grouping of the laws. (b) Maximum residual saturation of CO_2 for each law. Relative permeability law used in the reservoir test case: (c) water; (d) CO_2.

Contrarily to the case described in Sect. 3.2, the physical intuition on an *a priori* influence of scenario-like variable related to the relative permeability laws is harder to give: this is related to the richness of the information associated to the process of residual trapping that is related to different aspects: (1) the capacity of the porous medium to allow the flow of the gaseous phase in the presence of another phase: this is represented as a function of the saturation (see

examples in Fig. 8c). This capacity is associated to a potential hysteretic effect resulting in a non-unique dependence over time; (2) the capacity of the alternative phase (water) that is represented by another function of the saturation (Fig. 8d); (3) the considered phase (gaseous or aqueous) progressively becomes isolated when its saturation decreases in a porous medium. This leads to a saturation that cannot be reduced: these specific saturations are called residual situation for the gaseous phase (Fig. 8b) and irreducible saturation for the aqueous phase.

To help in the formulation of a physically-based assumption about the inter-dependencies, boxplots in Fig. 8a allow to identify some specific law behaviours especially for law 1; some tendency can also be noticed when ordering the laws in a specific order with respect to the median values of the variable of interest. To get a clearer picture, we test the validity of these observations via the proposed GP-based approach by defining different categorical kernels as follows:

- *Compound Symmetry* (denoted *CS*): no preference is given to the physical laws;
- *General* (denoted *Gen*): the most generic dependence structure;
- *Expert-based* groups: the grouping should account for the three facets of the CO_2 flow in porous media, i.e. by integrating the three pieces of information depicted in Fig. 8b-d: dissimilarities in both the imbibition, the drainage curve's shape and in the residual trapping model. On this basis, the following grouping is proposed: (law 2; law 10); (law 6-9), (law 1); (law 3); (law 4); (law 5). In addition, an assumption is made regarding the link between the groups by specifying a general or a compound symmetry covariance (assumption respectively denoted *E* and *Ecs*);
- *Low rank* approximation: groups of levels that are identified through the low rank approach. A rank of 2 and of 6 (i.e. of the same number of the expert-based groups) are tested (kernel respectively denoted *LR2* and *LR6*);
- *Ordinal* variable: the levels are ordered with respect to the value of the maximum CO_2 residual saturation (Fig. 8b), and a continuous kernel is defined via a spline-based warping (assumption denoted *O*).

The other categorical variables are respectively assigned a Compound Symmetric kernel for the regional hydraulic gradient and for the capillary effect. Owing to the "natural" ordering of the levels of the categorical variable for the permeability anisotropy, an ordinal kernel with spline-

based warping is defined. A logarithm (base 10) transformation of the variable of interest is applied due to high asymmetry in the variable histogram.

Fig. 9 depicts the GP-derived correlation matrices for each of the afore-described kernel assumptions. Some consistent structures can be noticed regarding law 1, which appears to be anti-correlated with the others as shown in Fig. 9a (general assumption), and in Fig. 9b,c (low rank approximation *LR2* and *LR6*). The specificity of law 1 is also outlined by the expert-based grouping in Fig. 9d, which indicates here a moderate positive correlation of 35-40%. This is in agreement with the ordinal assumption (Fig.9f) which indicates that the pairwise correlation coefficients of law 1 with the others (see last row of Fig. 9f) rapidly decreases. Though disagreeing on the correlation magnitude, *LR2*, *Gen*, and *E* models all suggest a moderate correlation among all laws except for law 1. The assumption *LR6* also suggests the particular behaviour of law 5, which goes in the same direction as the expert-based clustering of considering it as a belonging to a single group. The assumption *Ecs* (i.e. using a *CS* between-group covariance assumption) leads to a less complex correlation structure and clearly highlights the grouping of law 6-9 (as suggested by the experts, see also the purple-coloured laws in Fig. 8), which is in agreement with the group of highly correlated laws as outlined by Fig. 9f (though the size of the group is larger and includes law 2 and 4 as well).

To summarize, the inspection of the correlation matrices is here more difficult than for the cyclonic application case (Sect. 3.2), where all assumptions more or less agree regarding the information supplied by the modellers. Still, this inspection highlights the specificities of law 1 and law 5 (LR6 assumption), that both strongly differ from the other laws: this is suggested by the irreducible water saturation (red curve in Fig. 8c,d): law 1 is even more different with an irreducible water saturation associated to a large maximum gas residual saturation, which seems to explain why the model behaves so specifically when this law is accounted for.

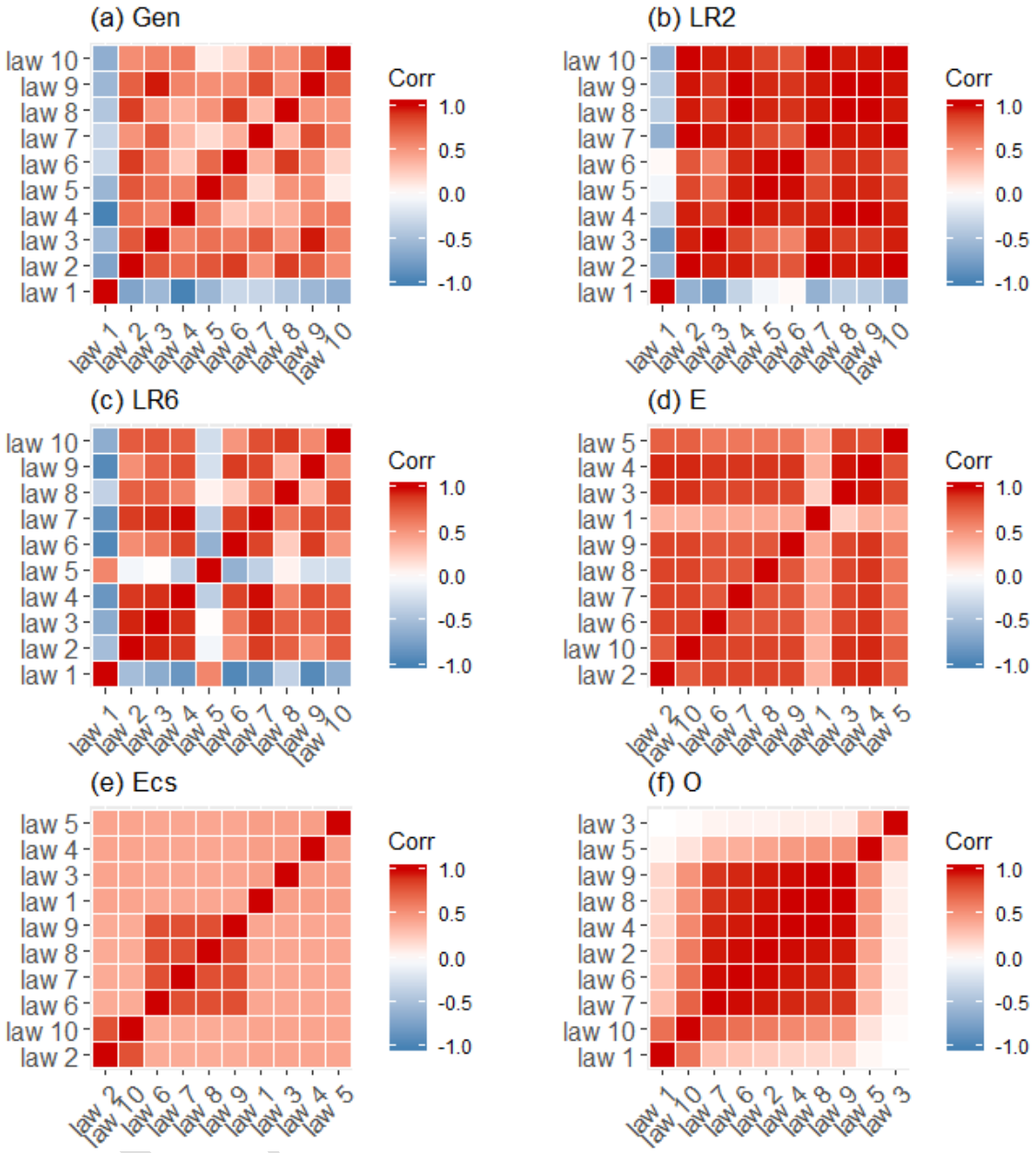

Figure 9. Correlation matrix for the reservoir test case considering different assumptions regarding the kernel associated to the categorical input variable. For the expert-based assumptions (d and e), the ordering of matrix coefficients follows the expert-based clustering, i.e. (law 2; law 10); (law 6-9), (law 1); (law 3); (law 4); (law 5). For the ordinal assumption (f), ordering of matrix coefficients follows the ordering of the maximum residual saturation of CO_2.

The four criteria for kernel model selection are examined in Fig. 10. For the considered case, we show that the expert-based (*Ecs*, i.e. with the simplified correlation structure between the groups) and the *CS* assumption both lead to GP models that satisfactorily fulfill the four criteria,

with a slightly higher performance for the simpler structure of *CS*. This means that the expert-based grouping of laws, in particular of laws 6-9 (in purple in Fig. 8), is informative (in the sense that it leads to a competitive GP model), but only makes a slight difference with the simpler structure especially regarding explainability (with BIC difference ~10) and regarding stability (due to the lowest number of CS correlation coefficients, namely of 1). From this analysis, it can be concluded that, given the 100 simulation results, there is only a mild evidence supporting the assumption of the structure associated to the permeability laws that was intuited from the analysis of the boxplots in Fig. 8a and of the curve similarities (Fig. 8b-d). Similarly as for the synthetic case, additional simulation results should here be performed in order to discriminate unambiguously the most appropriate kernel assumption.

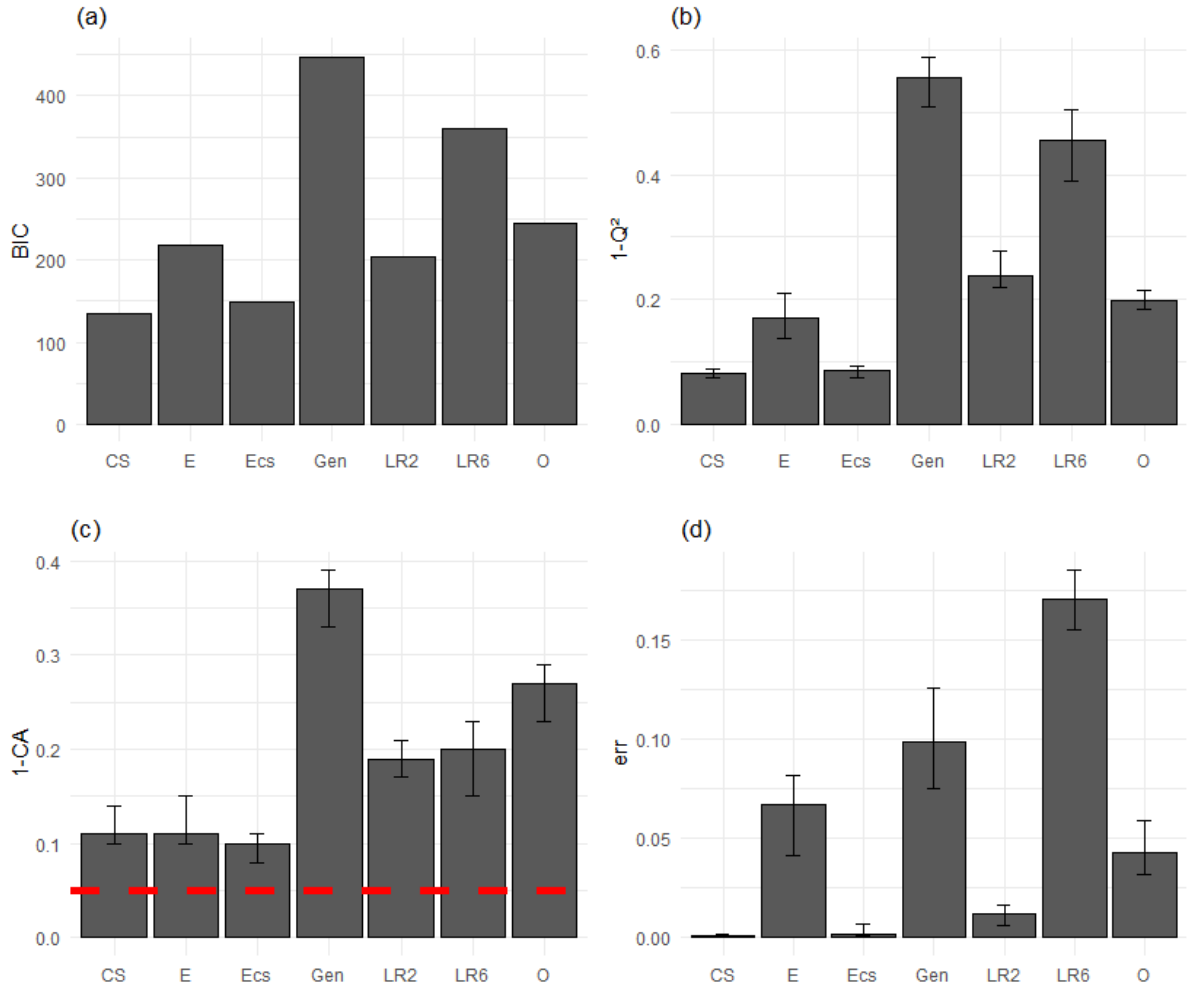

Figure 10. Selection criterion for the application case 2: (a) BIC; (b) Predictability measured by 1-Q^2; (c) Coverage measured by 1-*CA*. The horizontal dashed line corresponds to 5%, i.e. the threshold consistent with the level of the 95% prediction interval; (d) Stability error *err*.

Criteria (b-d) are derived from a 5-fold cross validation repeated 25 times: the height of the barplot is the median value and the lower and upper bounds are defined using the 25^{th} and 75^{th} percentiles.

4 Summary, Discussion and Further works

Model uncertainties (related to the structure/form of the model or to the unambiguous choice of "appropriate" physical laws) are generally integrated in environmental models via scenario-like variables, i.e. categorical variables. In the present study, we have explored the applicability of GP meta-modelling in order to:

- Inform on the structure of dependence for scenario-like inputs (research question Q1) using different formulations of the categorical covariance function (exchangeable, ordinal, group, etc.);
- Derive a predictive statistical model using a limited number of computer experiments, i.e. 100-200 (research question Q2) that satisfactory fulfil predictive capability, explainability and stability of parameters' estimates.

Regarding Q1, we discuss the added value of the approach in Sect. 4.1, as well as its limitations. Regarding Q2, we compare in Sect. 4.2 the GP-based procedure with popular modelling alternatives, namely tree-based methods, with respect to predictability as well as from a practical viewpoint. Different lines for further research works are also outlined.

4.1 Selecting the dependence structure

We have proposed to rely on a multiple-criterion selection approach to provide indications on which kernel model is the most appropriate to represent the dependence structure of the scenario-like inputs. By construction, the proposed framework should provide a level of sufficient flexibility to critically analyse the different physically-based assumptions: the advantage of testing and confronting each assumption to one another is to keep the procedure transparent to the practitioner. This flexibility is well shown in the cyclone application case (Sect. 3.2), where the proposed procedure allows us to confront an *a priori* physically-based assumption (i.e. the cyclone track effect can be summarized by a scalar ordinal variable) to alternative views on the dependence structure, and to support the evidence of the *a priori* assumption.

Yet, the procedure does not ensure that a unique model is selected: multiple assumptions may eventually turn to be valid (with respect to the four selection criteria) or a trade-off may be difficult to find. In the reservoir case, the application (Sect. 3.3) only moderately supports the evidence of some dependence structure; the compound symmetric and the expert-based kernel model perform similarly with respect to the four selection criteria. Though this result is informative per se, in particular in situations where the practitioner is preferably interested in explaining the numerical results, additional investigations are necessary to confirm this conclusion. This is shown in the synthetic case (Sect. 3.1), where the ordinal assumption was also successfully identified provided that a minimum number of training samples are available.

On the one hand, if additional model runs are computationally affordable, a possible option is to rely on an adaptive sampling strategy. This question deserves however further investigation in the presence of a mixture of continuous and categorical variables, and could be based on recent advances in the context of optimization by Pelamatti et al. (2019) and Munoz Zuniga and Sinoquet (2020). On the other hand, if additional model runs are not possible, an option is to aggregate the information provided by the "plausible" GP models (i.e. the ones that satisfactorily fulfil the criteria) while accounting for some weight reflecting their "plausibility" (with respect to the selection criteria). This option can take advantage of adequate averaging techniques developed, for instance, within the Bayesian framework as proposed by Zhang & Taflanidis (2019) for uncertainty quantification, and by Ginsbourger et al. (2008) for optimization problems.

4.2 Comparison to a popular alternative

It should be acknowledged that GPs are not the first statistical modelling option that comes to mind when addressing the problem of scenario-like variables. A popular approach relies on tree-based methods like regression decision trees, denoted DT (Breiman, 1984), and random forest regression, denoted RF (Breiman, 2001), which both natively handle categorical predictors without having to first transform them (e.g., by using feature engineering techniques) because they are based on binary recursive partitioning. Examples of real case applications are provided by Jaxa-Rozen and Kwakkel (2018) and Rohmer et al. (2018).

From a practical viewpoint, the advantage of DT is to provide the structure of dependence (as well as the interactions) with a graphical presentation of the results in the form of a tree, which

greatly eases the interpretation. For instance, in the cyclone real case application, Fig. 11a gives the tree structure derived from the analysis of the cyclone real case.

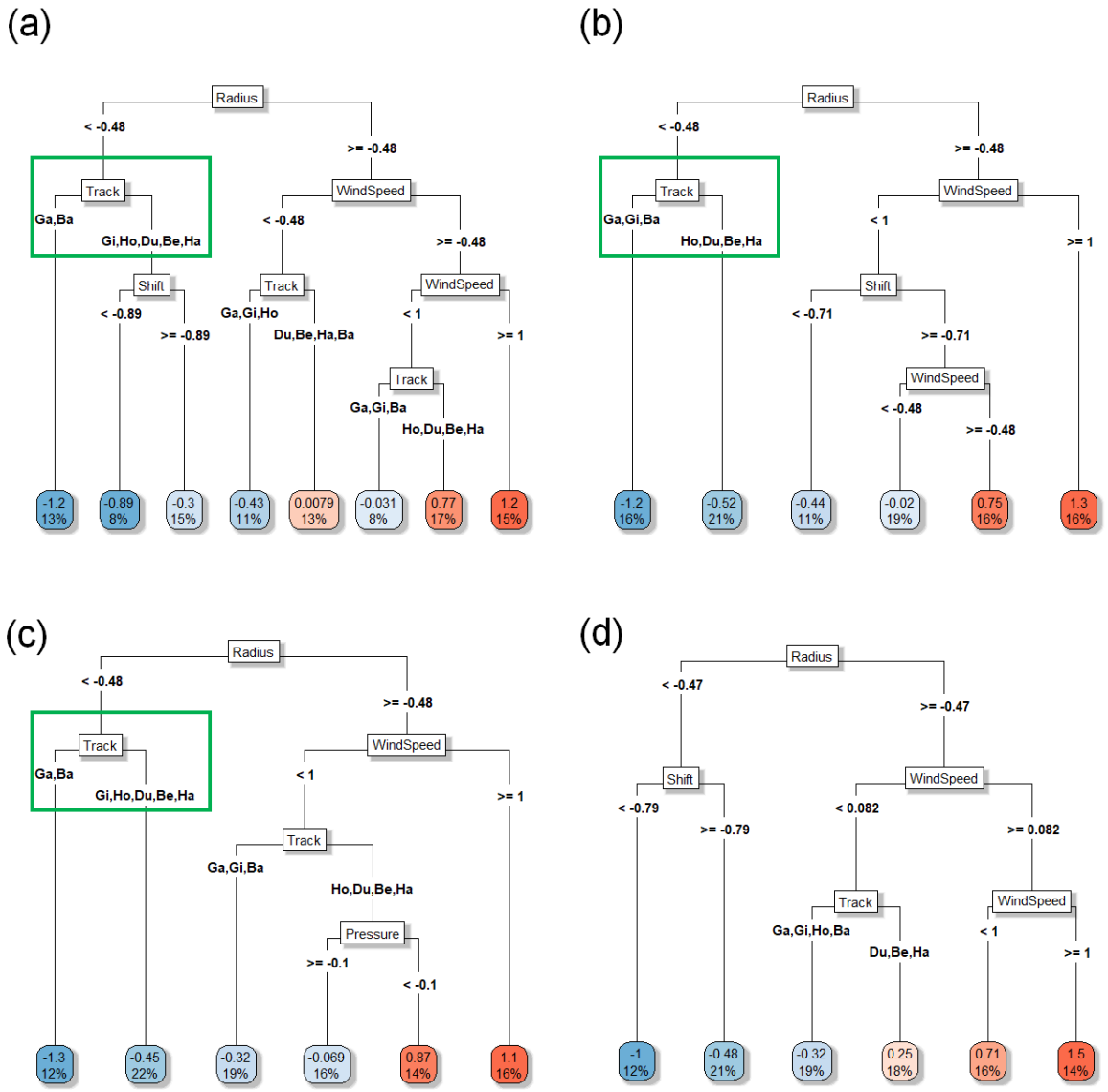

Figure 11. (a) Tree structures derived from the analysis of the cyclone real case; (b-d) Tree structures considering three iterations of the 5-fold cross validation procedure. The decision rule is provided on each respective branch. The leaf (bottom node) are coloured according to the mean of the variable of interest (scaled value of H_S); the number in percentage provides the number of samples falling in each leaf. The track name associated to the 'Track' node corresponds to the two first letters of the cyclone names provided in Fig. 4. The green rectangle outlines a particular grouping of tracks.

However, we note, from the analysis of Fig. 11, that several differences between the structure constructed using the DT trained using the whole dataset (Fig. 11a) and the ones at each iteration of the cross-validation procedure (i.e. using three DT models setup with "perturbed" training database, Fig. 11b-d); in particular for the leftmost part related to the track variable (outlined by a green rectangle), the grouping of tracks are similar for Fig. 11a and Fig. 11c, but differs for Fig. 11b and is even absent for Fig.11d. This high sensitivity of the derived structure (to the changes in the training dataset) has already been identified in the literature (Breiman, 2001): in addition to bringing some confusion regarding the scenario dependencies, the drawback is also a poorer predictability: this is shown by the low Q^2 values for each test case in Table 2.

Table 2. Predictability measured by the Q^2 indicator for different statistical models. For the synthetic case with different points per level m, Q^2 is derived from the leave-one-out cross-validation procedure. For the application cases 1 and 2, Q^2 is derived from the 5fold cross-validation procedure (repeated 25 times): the median values are given together with and the 25th and 75th indicated in brackets.

Model	Regression Decision Tree	Regression Random Forest	Gaussian Process
Synthetic case (m=4)	-0.12 [-0.17, -0.08]	0.61 [0.58, 0.65]	0.96 [0.94, 0.97]
Synthetic case (m=5)	0.07 [-0.01, 0.19]	0.72 [0.68, 0.74]	0.98 [0.97, 0.98]
Synthetic case (m=6)	0.26 [0.18, 0.31]	0.77 [0.76, 0.80]	0.98 [0.97, 0.99]
Application case 1	0.51 [0.47, 0.55]	0.60 [0.58, 0.62]	0.92 [0.91, 0.93]
Application case 2	0.23 [0.15, 0.28]	0.44 [0.42, 0.46]	0.93 [0.92, 0.94]

On the other hand, RF achieves a higher predictive capability by adding a random character to the DT construction process at two levels: (1) each tree is constructed using a different bootstrap sample; (2) each node is split using the best among a subset of input parameters randomly

chosen at that node: this is confirmed by Table 2. Yet, the high predictability comes at the expense of losing some interpretability, i.e. the ability to represent the structure via the easily understandable tree representation (RF being an ensemble of randomized DT models) though some developments are available to extract some meaningful rules from RF (see e.g. Fokkema, 2020). We acknowledge that there is room for improving the comparison exercise; in particular, we have used commonly-used parametrisations of the tested tree-based methods: further works should include more advanced developments like the enhancement of RF for smooth non-linear relations (Friedberg et al., 2018), and adequate splitting rules for categorical inputs (see an extended discussion by Wright and König, 2019).

To conclude, GP models can be seen as a good compromise considering the results on the test cases, because:

(1) They clearly achieve the higher predictability with Q^2 value >90% given a moderate size of the training dataset (typically 100-200), whatever the considered case (Table 2), while achieving features of high interest (explainability, simplicity, stability) provided that the kernel function is appropriately selecting;
(2) The correlation matrices (Figs. 2,5,9) derived from the kernel function provide a concise and graphical way to get insights into the inter-dependencies among the scenarios (using the formal interpretation provided in Sect. 2.2).

These results highlight the added values of our GP-based approach and show that it can be considered a valuable tool to complement the toolbox of any geo-scientist in order to both explore and characterize model uncertainties related to scenario-like inputs.

Acknowledgements

This research was conducted within the frame of the Chair in Applied Mathematics OQUAIDO, gathering partners in technological research (BRGM, CEA, IFPEN, IRSN, Safran, Storengy) and academia (CNRS, Ecole Centrale de Lyon, Mines Saint-Etienne, University of Grenoble, University of Nice, University of Toulouse) around advanced methods for Computer Experiments.

Software availability

Software name: kergp

Developers: Yves Deville, David Ginsbourger, Olivier Roustant.

Contributors: Nicolas Durrande

Maintainer: Olivier Roustant roustant@insa-toulouse.fr

System requirements: Windows, Linux, Mac

Program language: R

Availability: https://cran.r-project.org/web/packages/kergp/index.html

License: GPL-3.0

Documentation: https://cran.r-project.org/web/packages/kergp/kergp.pdf

References

Abily, M., Bertrand, N., Delestre, O., Gourbesville, P., Duluc, C. M., 2016. Spatial Global Sensitivity Analysis of High Resolution classified topographic data use in 2D urban flood modelling. Environmental Modelling & Software 77, 183-195.

Au, T. C., 2018. Random forests, decision trees, and categorical predictors: the" absent levels" problem. The Journal of Machine Learning Research, 19(1), 1737-1766.

Breiman, L., Friedman, J., Olshen, R., Stone, C. 1984. Classification and Regression Trees, Chapman & Hall, New York.

Breiman, L. 2001. Random forests. Machine learning 45(1), 5–32.

Burnham, K. P., Anderson, D. R. 2002. Model Selection and Inference A Practical Information Theoretic Approach, 2nd ed. Springer, New York.

Deville, Y., Ginsbourger, D., Roustant, O., 2018. kergp: Gaussian process laboratory. https://CRAN.R-project.org/package=kergp. Contributors: N. Durrande. R package version 0.4.0 (last access 24 November 2020).

Fokkema, M., 2020. Fitting prediction rule ensembles with R package pre. Journal of Statistical Software 92(12), 1-30.

Friedberg, R., Tibshirani, J., Athey, S., Wager, S., 2018. Local linear forests. arXiv preprint arXiv:1807.11408.

Ginsbourger, D., Helbert, C., Carraro, L., 2008. Discrete Mixtures of Kernels for Kriging-based optimization. Quality and Reliability Engineering International 24(6), 681-691.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer-Verlag, New York.

Hill, L., J., Sparks, R., S., J., Rougier, J., C., 2013. Risk assessment and uncertainty in natural hazards, in: Rougier, J. C., Sparks, R. S. J., Hill, L. J., (Eds.), Risk and uncertainty assessment for natural hazards. Cambridge University Press, New York, pp 1–18.

Höge, M., Wöhling, T., Nowak, W., 2018. A primer for model selection: The decisive role of model complexity. Water Resour. Res. 54, 1688–1715.

Idier, D., Rohmer, J., Pedreros, R., Le Roy, S., Lambert, J., Louisor, J., et al., 2020. Coastal flood: a composite method for past events characterisation providing insights in past, present

and future hazards—joining historical, statistical and modelling approaches. Natural Hazards 101(2), 465-501.

Johnson, M. E., Moore, L. M., Ylvisaker, D., 1990. Minimax and maximin distance designs. Journal of statistical planning and inference 26(2), 131-148.

Lauvernet, C., Helbert, C., 2020. Metamodeling methods that incorporate qualitative variables for improved design of vegetative filter strips. Reliability Engineering & System Safety 204, 107083.

Leandro, J., Chen, A. S., Djordjević, S., Savić, D. A., 2009. Comparison of 1D/1D and 1D/2D coupled (sewer/surface) hydraulic models for urban flood simulation. Journal of hydraulic engineering 135(6), 495-504.

Le Cozannet, G., Rohmer, J., Cazenave, A., Idier, D., van De Wal, R., De Winter, R., et al., 2015. Evaluating uncertainties of future marine flooding occurrence as sea-level rises. Environmental Modelling & Software 73, 44-56.

Liu, S., Shao, Y., Kunoth, A., Simmer, C., 2017. Impact of surface-heterogeneity on atmosphere and land-surface interactions. Environmental Modelling & Software 88, 35-47.

Mishra, B. K., Rafiei Emam, A., Masago, Y., Kumar, P., Regmi, R. K., Fukushi, K., 2018. Assessment of future flood inundations under climate and land use change scenarios in the Ciliwung River Basin, Jakarta. Journal of Flood Risk Management 11, S1105-S1115.

Manceau, J. C., Rohmer, J., 2016. Post-injection trapping of mobile CO 2 in deep aquifers: Assessing the importance of model and parameter uncertainties. Computational Geosciences 20(6), 1251-1267.

Munoz Zuniga, M., & Sinoquet, D., 2020. Global optimization for mixed categorical-continuous variables based on Gaussian process models with a randomized categorical space exploration step. INFOR: Information Systems and Operational Research 1-32.

Pelamatti, J., Brevault, L., Balesdent, M., Talbi, E. G., Guerin, Y., 2019. Efficient global optimization of constrained mixed variable problems. Journal of Global Optimization 73(3), 583-613.

Pinheiro, J., Bates, D., 2006. Mixed-effects models in S and S-PLUS. Springer Science & Business Media.

Powell, M. J. D., 1994. A direct search optimization method that models the objective and constraint functions by linear interpolation, in: Gomez, S., Hennart, J.-P., (Eds.), Advances in Optimization and Numerical Analysis, Springer, Dordrecht, pp 51–67.

Qian, P. Z. G., Wu, H., Wu, C. J., 2008. Gaussian process models for computer experiments with qualitative and quantitative factors. Technometrics 50(3), 383-396.

Williams, C. K., Rasmussen, C. E., 2006. Gaussian processes for machine learning. MIT press, Cambridge, MA.

Rohmer, J., Douglas, J., Bertil, D., Monfort, D., Sedan, O., 2014. Weighing the importance of model uncertainty against parameter uncertainty in earthquake loss assessments. Soil Dynamics and Earthquake Engineering 58, 1-9.

Rohmer, J., Lecacheux, S., Pedreros, R., Quetelard, H., Bonnardot, F., Idier, D., 2016. Dynamic parameter sensitivity in numerical modelling of cyclone-induced waves: a multi-look approach using advanced meta-modelling techniques. Natural Hazards 84(3), 1765-1792.

Roustant, O., Padonou, E., Deville, Y., Clément, A., Perrin, G., Giorla, J., Wynn, H., 2020. Group kernels for Gaussian process metamodels with categorical inputs. SIAM/ASA Journal on Uncertainty Quantification 8(2), 775-806.

Santner, T. J., Williams, B. J., Notz, W. I., & Williams, B. J., 2003. The design and analysis of computer experiments (Vol. 1). Springer, New York.

Schwarz, G., 1978. Estimating the Dimension of a Model, Ann. Stat. 6, 461–464.

Silva Ursulino, B., Maria Gico Lima Montenegro, S., Paiva Coutinho, A., Hugo Rabelo Coelho, V., Cezar dos Santos Araújo, D., Cláudia Villar Gusmão, A., et al., 2019. Modelling soil water dynamics from soil hydraulic parameters estimated by an alternative method in a tropical experimental basin. Water 11(5), 1007.

Storlie, C. B., Reich, B. J., Helton, J. C., Swiler, L. P., Sallaberry, C. J., 2013. Analysis of computationally demanding models with continuous and categorical inputs. Reliab. Eng. Syst. Saf. 113, 30–41.

Vandromme, R., Thiery, Y., Bernardie, S., Sedan, O. 2020. ALICE (Assessment of Landslides Induced by Climatic Events): A single tool to integrate shallow and deep landslides for susceptibility and hazard assessment. Geomorphology 367, 107307.

Veeck, S., da Costa, F. F., Lima, D. L. C., da Paz, A. R., Piccilli, D. G. A., 2020. Scale dynamics of the HIDROPIXEL high-resolution DEM-based distributed hydrologic modeling approach. Environmental Modelling & Software 104695.

Wright, M. N., König, I. R., 2019. Splitting on categorical predictors in random forests. PeerJ 7, e6339.

Yu, B., Kumbier, K. 2020. Veridical data science. Proceedings of the National Academy of Sciences 117(8), 3920-3929.

Zhang, Y., Tao, S., Chen, W., Apley, D. W., 2020. A latent variable approach to Gaussian process modeling with qualitative and quantitative factors Technometrics 62(3), 291-302.

Zhang, J., Taflanidis, A. A., 2019. Bayesian model averaging for Kriging regression structure selection. Probabilistic Engineering Mechanics 56, 58-70.

Zhao, G., Bryan, B. A., King, D., Luo, Z., Wang, E., Bende-Michl, U., et al., 2013. Large-scale, high-resolution agricultural systems modeling using a hybrid approach combining grid computing and parallel processing. Environmental Modelling & Software 41, 231-238.