Revealing the interlevel dependence structure of categorical inputs in numerical environmental simulations with kernel model selection

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

▶ **To cite this version:**

Jeremy Rohmer, O Roustant, Sophie Lecacheux, Jean-Charles Manceau. Revealing the interlevel dependence structure of categorical inputs in numerical environmental simulations with kernel model selection. Environmental Modelling and Software, 2022, 151, pp.105380. 10.1016/j.envsoft.2022.105380 . hal-03687171

HAL Id: hal-03687171
https://brgm.hal.science/hal-03687171

Submitted on 22 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S136481522200086X
Manuscript_d0016da8c505aa95fee736d3c39ba677

Revealing the interlevel dependence structure of categorical inputs in numerical environmental simulations with kernel model selection

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 are generally integrated in environmental long-running numerical simulators via a categorical variable. By focusing on Gaussian process (GP) models, we show how different categorical kernel models (exchangeable, ordinal, group, etc.) can bring valuable insights into the correlation of the simulator output values computed for different levels of the categorical variable, i.e., the interlevel dependence structure. Supported by two real case applications (cyclone-induced waves and reservoir modeling), we have proposed a cross-validation approach to select the most appropriate kernel by finding a trade-off between predictability, explainability, and stability of the covariance coefficients. This approach can be used effectively to support some physical assumptions regarding the categorical variable. 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 interlevel dependence structure.

© 2022 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

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

Abbreviations: CS, compound symmetry; DT, decision tree; E, expert-based; Gen, general; GP, Gaussian process; LR, low rank; O, ordinal; RF, random forest.

1 Introduction

High-resolution numerical simulators are key components of environmental science that help to obtain deeper insights into the behavior of natural systems. Some examples are Veeck et al. (2020) for hydrologic modeling; Zhao et al. (2013) for agricultural modeling; Vandromme et al. (2020) for landslide modeling; Abily et al. (2016) for urban flooding; and Idier et al. (2020) for marine flooding. To model the natural system, these simulators all have in common to involve a large spectrum of assumptions related to the choice in the structure/form of the model (e.g., 1D versus 2D modeling, 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 modeling assumptions, the simulation results can differ, hence resulting in model uncertainty termed *structural uncertainty* (e.g., Hill et al., 2013).

Some of these modeling assumptions can be modeled by means of continuous variables (such as geotechnical properties of a given soil formation or time series of rainfall conditions at a given location, etc.), some of them involve categorical variables, i.e., multilevel indicators that take up a finite number of discrete values; each discrete level of the categorical variable is associated with a different modeling assumption (e.g., level a is associated with modeling assumption a). Some real case applications are provided 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 and Helbert (2020), etc.

Quantifying the correlation of the simulator output values computed for two levels (i.e., two modeling assumptions) of the categorical variable (whatever the values of the other input variables) is of high interest to measure the impact of structural uncertainty, because it informs whether each level should be treated equivalently with respect to the numerically simulated variable of interest or if there is any dependence among the levels like a group structure. This type of structure, named the "interlevel dependence structure" in this study, can be useful to identify modeling assumptions that should be considered a priority in complementary simulation-based studies. For instance, a level showing pronounced impact on the variable of interest suggests where to focus the simulation-based exploration, and a group structure suggests simplifying the analysis by restricting the analysis to a single member of the group. Beyond structural uncertainty, such information may be of interest to support the analysis of deep uncertainty based on scenario discovery (Kwakkel and Jaxa-Rozen, 2016).

Our objective is to develop a statistical procedure to learn the interlevel dependence structure. Due to the high computation time cost of numerical environmental simulations, we aim to learn the dependence structure with only a few model runs (on the order of 100) by relying on the design and analysis of computer experiments (Santner et al., 2003). The key element of the proposed procedure is the use of Gaussian process models, denoted GP models (Williams and Rasmussen, 2006) with covariance functions (also known as kernels) adapted to handle mixed continuous/categorical inputs and combined by tensor products (e.g., Roustant et al., 2020; Qian et al., 2008; Zhang et al., 2020).

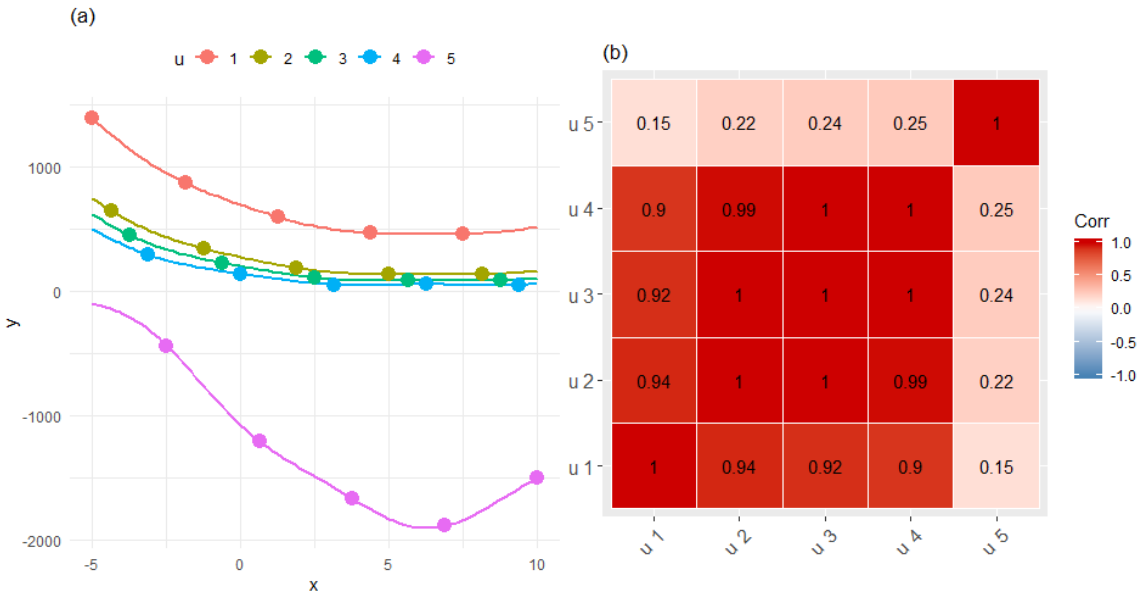

Figure 1 (a). Synthetic test function with continuous input variable x. Each color indicates a different level of the categorical variable u. Each dot corresponds to a model run. (b) Correlation matrix for u derived from the GP-based analysis using 25 model runs.

The kernels 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 categorical input). This "similarity" function can be encoded in different manners depending on the assumptions regarding the categorical variable (nominal/ordinal, interdependence between the levels, interactions between given levels, etc.); see, for instance, Roustant et al. (2020) and Lauvernet and Helbert (2020). Once fitted, the resulting correlation matrix provides the mapping of the dependence structure.

To illustrate the type of results that can be derived, Figure 1(a) depicts an unknown relationship between a continuous and a categorical input variable with 5 levels (i.e., five modeling assumptions). Figure 1(b) depicts the correlation matrix derived from the GP-based analysis given 25 model runs. 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 relationship in Figure 1(a) had been known, this result would have been straightforward, but here the structure is unknown and can be learned only with a limited number of numerical results (here with only 25 model runs). This result (that is further discussed in Sect. 4.2) depends on how the kernel model is defined, which raises the question of model selection that is addressed in the present study by relying on a multicriterion analysis.

The paper is organized as follows. Section 2 describes the different steps of the proposed procedure as well as the statistical methods. In this latter, we provide the formal definition of the interlevel dependence structure related to the GP correlation matrix (Sect. 2.2) by relying on the tensor product of kernels for continuous and categorical variables (Sect. 2.3). A multicriterion approach for selecting the categorical covariance kernel model is also detailed (Sect. 2.4). Section 3 describes the application cases, i.e., the synthetic case (described in Figure 1) and two real cases, namely, cyclone-induced wave numerical modeling (Rohmer et al., 2016) and reservoir modeling of CO_2 storage (Manceau and Rohmer, 2016). Both real cases are representative of two distinct situations. The cyclone case illustrates a situation where a physical intuition on an *a priori* influence of the categorical variable is available, whereas the reservoir case illustrates the opposite situation where the physical intuition is harder to give. The procedure is then applied to each of these cases in Section 4. In Section 5,

we further discuss the results by comparing the GP-based procedure to a popular alternative approach based on tree-based techniques. On this basis, practical recommendations are then defined in Section 6.

2 Methods

2.1 Description of the procedure

The proposed procedure is as follows:

- Step 1: We follow the approach of design and analysis of computer experiments (Santner et al., 2003). A series of runs of the expensive-to-evaluate environmental simulator were performed by considering a limited number of randomly selected input variable 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 modeled by means of different kernel (covariance) formulations (see further details in Sect. 2.2 and 2.3);
- Step 3: The question of selecting the most appropriate kernel is examined by analyzing different aspects, i.e., by considering different criteria as described in Sect. 2.4. The objective is to select the resulting GP model that can achieve a trade-off between the different criteria.
- Step 4: Since the practitioner is preferably interested in the dependence structure between the modeling assumptions, the correlation matrix derived from the covariance matrix is analyzed (see further explanation in Sect. 2.2). This result can be confronted with some *a priori* physically based interpretation of the categorical variable influence that the practitioner may have before analyzing 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 $n_{L_1},\ldots,n_{L_J}$ levels that represent the categorical 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) modeling (also named kriging, Williams and Rasmussen, 2006), the function f(.) is assumed to be a realization of a GP ($Y(\mathbf{w})$) with a constant mean μ 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 us denote $(\mathbf{w}^1, \ldots, \mathbf{w}^n)$ the training samples and $\boldsymbol{y} = (\mathrm{y}^1 = \mathrm{f}(\mathbf{w}^1), \ldots, \mathrm{y}^n = \mathrm{f}(\mathbf{w}^n))$ denote the corresponding results. The prediction at a new observation $\mathbf{w}^*$ is given by the kriging mean $\hat{Y}(\mathbf{w}^*)$ as follows:

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

where $\mathbf{C}$ is the covariance matrix between the points $Y(\mathbf{w}^1)$,…,$Y(\mathbf{w}^n)$ 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 $Y(\mathbf{w}^1)$,…, $Y(\mathbf{w}^n)$, and $\mathbf{I}$ is the identity vector of length n.

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

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

Using the GP mean and variance (Equations 2 and 3), the prediction interval PI_α at the given confidence level α (e.g., α=95%) can be computed as follows:

$$PI_\alpha(\mathbf{w}^*) = [\hat{Y}(\mathbf{w}^*) - \hat{\sigma}(\mathbf{w}^*) . \mathrm{q}_{N(0,1)}(\tfrac{1+\alpha}{2}), \hat{Y}(\mathbf{w}^*) + \hat{\sigma}(\mathbf{w}^*) . \mathrm{q}_{N(0,1)}(\tfrac{1+\alpha}{2})] \quad (4)$$

where $\mathrm{q}_{N(0,1)}$is the quantile of order $\frac{1+\alpha}{2}$ of the standard normal distribution.

Accounting for a mixture of input variable types - continuous or categorical (ordinal or nominal) - is made via the covariance function $k(\mathbf{w}, \mathbf{w}')$. Here, it is assumed to be the tensor product of the covariance function for the continuous inputs $k_{\mathrm{cont}}(\mathbf{x}, \mathbf{x}')$ and the ones for the categorical inputs $k_{\mathrm{cat}}(\mathbf{u}, \mathbf{u}')$ as $k(\mathbf{w}, \mathbf{w}') = k_{\mathrm{cont}}(\mathbf{x}, \mathbf{x}') \prod_{i=1}^{J} k_{\mathrm{cat}}^i(u_i, u'_i)$. Other combination approaches are possible (as discussed by Roustant et al., 2020).

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 twice differentiable in the mean square Matérn 5/2 model (Williams and Rasmussen, 2006). 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 categorical input, as described in Sect. 2.3.

In practice, 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 $\mathbf{x}$ and $\boldsymbol{u}_{-i}$ =

$(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots u_J)$ are fixed. In practice, we prefer using the scaled form of the covariance, i.e., the correlation to ease the interpretation of the dependencies of the simulator output on interactions between the levels of the categorical variable (i.e., of the modeling assumptions). With the same notations, the correlation kernel (derived from k_{cat}^i) is interpreted as the correlation of the 1D section $u_i \to Y(\boldsymbol{x}, u_i, \boldsymbol{u}_{-i})$, regardless of $\boldsymbol{x}$ and $\boldsymbol{u}_{-i}$. Thus, the inspection of k_{cat}^i reveals the correlation of the simulator output explained by the i$^{\text{th}}$ categorical input, the others being fixed. Note that such a correlation does not depend on the remaining inputs, which is a result of constructing $k(.)$ by tensor product.

2.3 Covariance kernel models for categorical inputs

Table 1 summarizes the different options for defining a categorical covariance function, their interpretation and their practical implementation. The interested reader can refer to Roustant et al. (2020) and references therein for a more formal presentation, in particular with an analysis of the positive definiteness of covariance matrices. Note that we restrict the presentation to the case of a single input with n_L levels, which is hereafter denoted by u. The generalization to multiple categorical inputs can be done by following the tensor product formulation described in Sect. 2.2.

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

Type of covariance matrix	Symbol	Assumption	Implementation	Equation
Compound Symmetry	CS	No difference in influence across levels	A unique correlation coefficient	5
General	Gen	All interlevel dependencies are considered	Each level has its own variance coefficient and a between-level structure exists	6
Group	E	Levels are gathered by	Block structure	7

		groups defined with expert information		
Low rank approximation	LR	Level dependencies are explained by a few key latent continuous variables	The covariance matrix is related to a matrix of lower rank q with typical value of 2 or 3	8
Ordinal	O	The levels can be ordered, i.e., they are seen as discretized values of a latent ordinal variable.	Combination of a continuous kernel with a nondecreasing warping function modeled with splines.	9

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

$$k_{\text{cat}}^{\text{CS}}(u,u') = \begin{cases} \sigma^2 & \text{if } u = u' \\ \rho.\sigma^2 & \text{if } u \neq u' \end{cases} \quad (5)$$

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

When the practitioner assumes that the variable of interest will act differently depending on the considered level 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:

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

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 levels perform similarly and that they can be grouped. Assume that the n_L levels of u are partitioned into G groups, and denote $g(u)$ as the group number

corresponding to a given level u. Then, the covariance function can be written as (Roustant et al., 2020):

$$k_{\text{cat}}^{\text{E}}(u,u') = k_{\text{cat}}^{\text{Gen}}(g(u),g(u')) = \begin{cases} c_{g(u),g(u')} \text{ if } \mathrm{g}(u) \neq g(u') \\ v_{g(u)} \text{ if } g(u) = g(u') \end{cases} \quad (7)$$

where for all $\mathrm{i},\mathrm{j} \in \{1,\ldots,\mathrm{G}\}$, the terms $\frac{c_{i,i}}{v_i}$ are within-group correlations, 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 explain the level dependencies by a few key latent continuous variables using a low rank approximation (Roustant et al. 2020) so that the covariance matrix $\boldsymbol{k}_{\text{cat}}^{\text{LR}}$ can be defined as:

$$\boldsymbol{k}_{\text{cat}}^{\text{LR}} = \mathbf{Q}.\mathbf{Q}^{\text{T}} \quad (8)$$

where the matrix $\mathbf{Q}$ is of size n_L×q where q is low. The higher q is, the more likely the danger of overfitting. In practice, typical values of q are 2 or 3; a value of 1 is not recommended (see Zhang et al., 2020).

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:

$$k_{\text{cat}}^{\text{O}}(u,u') = \tilde{k}_{\text{cont}}(F(u),F(\mathrm{u}')) \quad (9)$$

where $\tilde{k}_{\text{cont}}$ is a one-dimensional continuous kernel (such as the Matérn 5/2 model), and $F(.)$ is a one-dimensional nondecreasing function (also called warping) $F:\{1,\ldots,n_L\} \to \mathbb{R}$. The warping function can be modeled via a parametric model (e.g., the cdf of a flexible probability distribution such as Normal or Beta) or a nonparametric model (e.g., a piecewise-linear or a quadratic spline) as detailed by Roustant et al. (2020).

2.4 Kernel model selection

As mentioned above, different kernel modeling choices can be made to represent the categorical variable, raising the question of selecting the most appropriate kernel covariance

model. Depending on the modeling 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 for prediction. Therefore, we propose to examine different viewpoints on the problem of kernel selection by analyzing different criteria. This multicriterion approach shares similarities with the data science framework proposed by Yu and Kumbier (2020), who advocate analyzing 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. This concept 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, which can be assessed by using independent test samples, bootstrap or 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:

$$Q^2 = 1 - \frac{\sum_{i \in \boldsymbol{T}} (y_i - \hat{y}_i)^2}{\sum_{i \in \boldsymbol{T}} (y_i - \bar{y})^2} \tag{10}$$

where $\hat{y}_i$ is the ith 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. For the sake of comparability between the different criteria, we consider 1-Q^2.

The second criterion is related to the coverage (denoted *CA*) of the prediction intervals PI_α (Eq. 4) at the given confidence level α calculated on the test set $\boldsymbol{T}$, defined by:

$$CA = \frac{1}{|\boldsymbol{T}|} \sum_{i \in \boldsymbol{T}} \mathbf{1}_{(y_i \in PI_\alpha(\boldsymbol{w}^i))} \tag{11}$$

where $\mathbf{1}_A$ is the indicator function. *CA* evaluates whether the model output y_i is within the bounds of the prediction interval. The GP-derived PI_α is "optimal" when *CA* is close to the theoretical value of α. For the sake of comparability between the different criteria, we consider the "error in coverage" defined as 1-*CA*.

Explainability and simplicity. The former concept relates to whether the considered GP model is capable of representing the data, for instance, by analyzing 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 a simpler model. Here, simplicity refers to the number of GP model parameters, but alternative definitions exist (see, for instance, Rougier and Priebe (2020), who introduce the concept of flexibility). By assuming that the true GP model exists and that it is among the set of candidate GP models, we propose relying on the Bayesian information criterion BIC (Schwarz, 1978) described as follows:

$$BIC = 2\log(l) + k.\log(n) \tag{12}$$

where k is the number of parameters and n is the number of observations.

Stability. We explore to what extent the kernel correlation matrix (derived from the covariance matrix) is stable to the perturbations in the training dataset. We evaluate an error measure between the correlation matrices $\widehat{R_0}$, estimated using the GP model fitted with the whole training dataset and $\widehat{R_j}$, estimated at the jth iteration of the n_{cv}-fold cross validation procedure. For instance, when n_{cv}=5, the training dataset corresponds to the whole training dataset, from which 20% of the observations have been randomly removed. The error measure is defined by:

$$err = \frac{1}{n_{cv}} \sum_{k=1}^{n_{cv}} \left\| \widehat{R_0} - \widehat{R_k} \right\|_F^2 \tag{13}$$

where $\|.\|_F$ is the Frobenius matrix norm. By restricting to the nondiagonal elements in the upper triangular part of $\hat{R}$, Eq. 12 is calculated as $\frac{1}{n_{cv}} \sum_{k=1}^{n_{cv}} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} (\hat{r}_{i,0} - \hat{r}_{i,k})^2 \right)$, where $\hat{r}_{i,0}$ is the ith coefficient of $\widehat{R_0}$ (read in column order for instance), $\hat{r}_{i,k}$ is the ith coefficient of $\widehat{R_k}$, and n_p is the number of terms of $\hat{R}$. Note that alternative formulations could also be proposed to better account for the correlation matrix structure, such as the diagonally weighted matrix norm proposed by Cressie and Hardouin (2019).

3 Description of the application cases

In this section, we describe the application cases, namely, a synthetic test function (Sect. 3.1) and two real cases in the domain of cyclone-induced wave modeling (Sect. 3.2) and reservoir engineering (Sect. 3.3).

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:

$$y = \begin{cases} h(x,-20), & if\ u = 1 \\ h(x,-10), & if\ u = 2 \\ h(x,-7.5), & if\ u = 3 \\ h(x,-5.0), & if\ u = 4 \\ -5.h(x,20), & if\ u = 5 \end{cases} \quad (14)$$

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). We consider the different categorical kernels defined in Table 1 with the following assumptions:

- Expert-based groups. Two experts have given their opinions: the first expert (denoted *E*) indicates a 'realistic' grouping of levels (i.e., consistent with the true function), and the second one (denoted *EW*) indicates an unrealistic grouping (i.e., (u_1,u_3), and (u_2,u_4,u_5));
- *Low rank* approximation. A matrix with rank of q=2 is tested (denoted *LR2*);
- *Ordinal* variable. The levels are ordered by following the level index. For the expert-based grouping, we assume that another expert does not know the correct ordering and assumes an unrealistic order (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 x-points per u-level with m=4, 5 and 6, i.e., with different training dataset sizes of 20, 25 and 30.

3.2 Real case application 1: cyclone-induced waves

The first real case is based on Rohmer et al. (2016) and deals with the modeling of waves induced by cyclones at Sainte Suzanne city located in the northeast of Reunion Island (Fig. 2b). The aim is to analyze the evolution of the significant wave height H_s (maximum value over time) as a function of the cyclone characteristics. These parameters are modeled by means of five scalar continuous input parameters, namely, the maximum wind speed, the

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

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, which 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 tracks are randomly shifted (via the continuous input variable modeling the relative position of the track) from their original track so that they cross the center of Reunion Island (see Fig. 2b): a categorical variable is defined here, with each level corresponding to a given track. A series of 100 computer experiments were performed by randomly sampling the inputs using a Latin hypercube sampling approach combined with a *maximin* criterion (Johnson et al., 1990). The design of the experiments is presented in Appendix A.

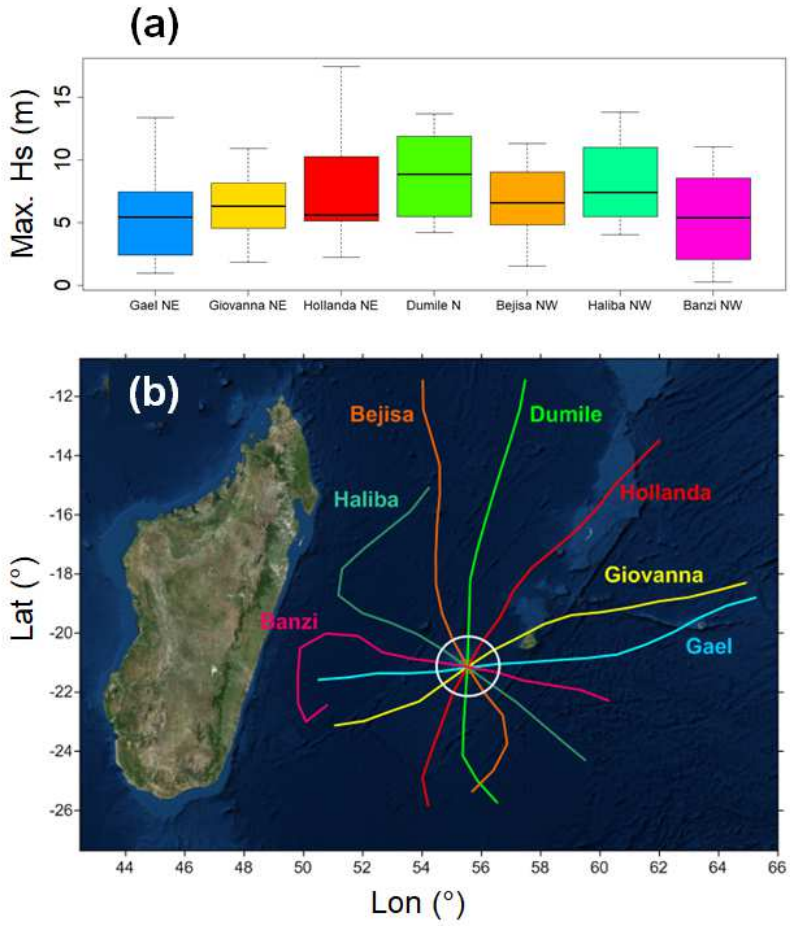

Figure 2. (a) Boxplot of the maximum significant wave height H_s (m) considering each cyclone track ordered according to the angle of approach from 0° to 180° (from the east – leftmost part to the west – rightmost part); (b) Cyclone tracks used for modeling the waves at Saint Suzanne city (Reunion Island).

Accounting for the spatial variability of the track (as illustrated in Fig. 2b) requires integrating each cyclone's spatial position along the track as an input variable of the GP models, which is hardly feasible in practice. 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° is north, confirmed by the sensitivity analysis of Rohmer et al. (2016). This means that the track effect can be summarized by a single scalar continuous input, i.e., in relation to the angle of approach. The analysis of the boxplots in Fig. 2a 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). However, 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 quasilinear shape of the others (Fig. 2b). To support the evidence of the influence of the angle of approach, a more rigorous analysis is needed here: the validity of this hypothesis is further investigated using a GP model with different categorical kernels as described in Table 1 with the following assumptions:

- *Expert-based* groups. A first assumption relies on the selection of two groups: one composed of the 4 cyclones (Gael, Giovanna, Hollanda, and Dumile) coming from the northeastern (NE) quadrant and another group composed of 3 tracks (Bejisa, Haliba, and Banzi) coming from the northwestern (NW) quadrant (Fig. 2). A second assumption based on three groups is also tested by differentiating the track whose angle is almost at 90° (North), namely, the Dumile cyclone (Fig. 2). 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. Ranks of q=2 and q=3 are tested (kernels denoted *LR2* and *LR3*, respectively);
- *Ordinal* variable. The levels are ordered following the angle of approach.

3.3 Real case application 2: CO_2 geological storage

The second real case application corresponds to the modeling 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 over 30 years in the Lower Triassic sandstone formation at a depth of approximately 1,000 m was numerically simulated. The evolution of the quantity of mobile CO_2 for a time period of 150 years after the injection stopped 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 assumptions for permeability anisotropy ("minor", "medium", and "large"), regional hydraulic gradient ("absence" and "activated"), capillary effect ("absence" and "activated"), and the choice in the physical law used to model the relative permeability as a function of CO_2 saturation (ten choices), as depicted in Fig. 3. 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 was performed by randomly sampling the inputs using a Latin hypercube sampling approach combined with a *maximin* criterion. A base 10 logarithm transformation was applied to the quantity of mobile CO_2 due to the large asymmetry of its distribution. The design of the experiments is presented in Appendix A.

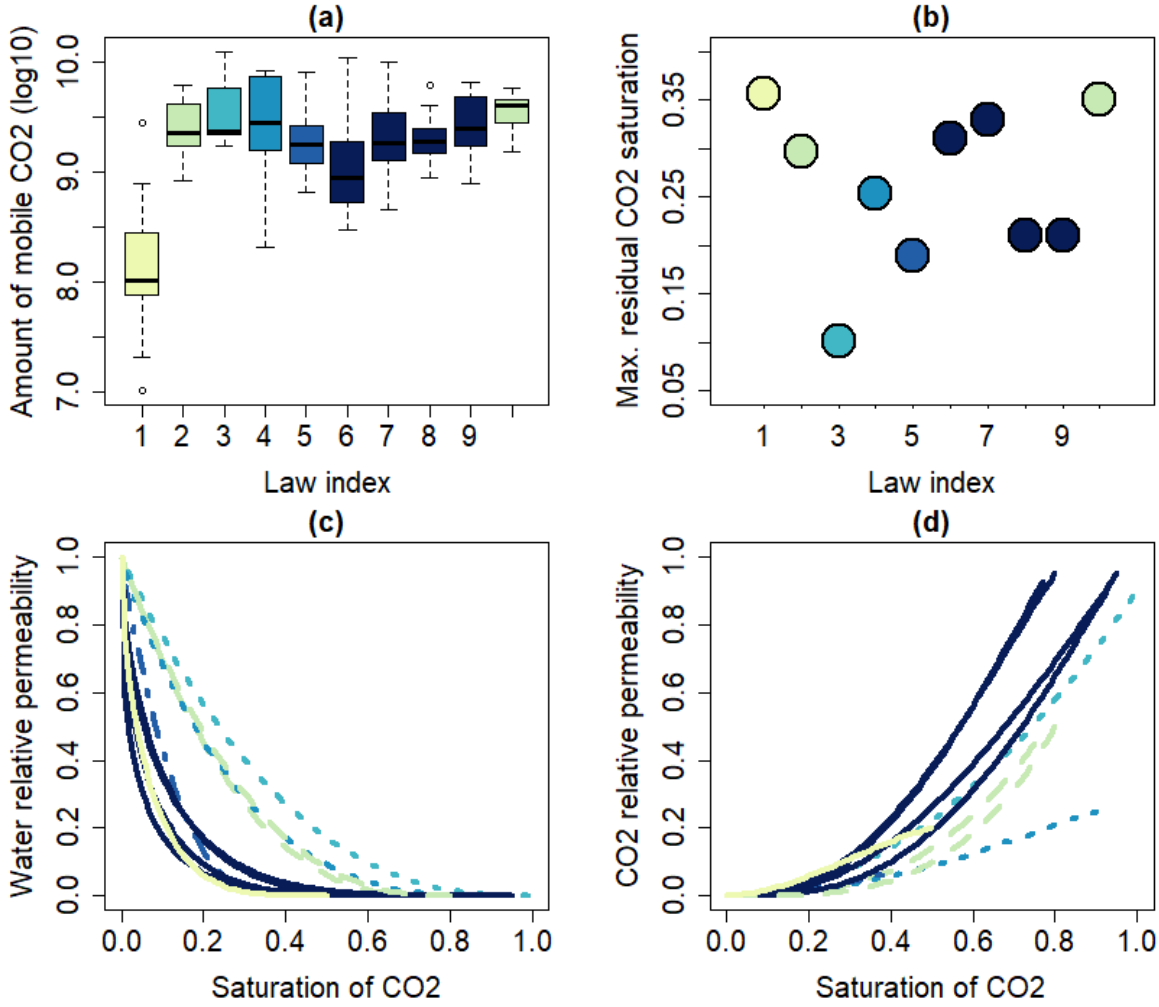

Figure 3. (a) Boxplot of the amount of mobile CO_2 considering each physical law index (1-10). Colors indicate the expert-based grouping of the laws. (b) Maximum residual saturation of CO_2 for each law. Relative permeability law used in the reservoir test case: (c) water; (d) CO_2. See Sect. 3.3 for a description of the physical processes related to these laws.

Unlike the case described in Sect. 3.2, it is harder to give a physical intuition on an *a priori* influence of the categorical variable related to the relative permeability laws. This lack of intuition is related to the richness of the information associated with 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 (called relative permeability) is represented as a function of the gas saturation in the porous medium (see examples in Fig. 3c). This capacity is associated with a potential hysteretic effect resulting in a nonunique dependence during the flow over time (see further details in Juanes et al., 2006); (2) the

capacity of the alternative phase (water) that is represented by another function of the saturation (Fig. 3d); and (3) the considered phase (gaseous or aqueous) progressively becomes isolated when its saturation decreases in a porous medium, leading to saturation that cannot be reduced: these specific situations are called residual saturations for the gaseous phase (Fig. 3b) and irreducible saturations for the aqueous phase.

To help formulate a physically based assumption about the interdependencies, the boxplots in Fig. 3a allow us to identify some specific law behaviors, 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 obtain a clearer picture, we test the validity of these observations via the proposed GP-based approach by considering the different categorical kernels described in Table 1 with the following assumptions:

- *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. 3b-d: dissimilarities in both the imbibition, drainage curve shape and 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 denoted *E* and *Ecs,* respectively);
- *Low rank* approximation. A rank of q=2 and of q=6 (i.e., of the same number of expert-based groups) are tested (kernels denoted *LR2* and *LR6*);
- *Ordinal* variable. The levels are ordered with respect to the value of the maximum CO_2 residual saturation (Fig. 3b).

The three other categorical variables are modeled as follows. The regional hydraulic gradient and the capillary effect are both assigned a compound symmetric kernel. An ordinal kernel with spline-based warping is defined for the permeability anisotropy because there is a "natural" ordering of the levels.

4 Results

4.1 Implementation procedure

In this section, we apply the GP-based procedure described in Sect. 2 to the cases described in Sect. 3. To ensure identifiability of the model defined with $k(.,.)$, we treat k_{cont} as a covariance kernel and each k_{cat}^{j} as a correlation kernel (j=1...,J). For all GP models, we consider a Matérn 5/2 covariance matrix for the continuous variables. For the ordinal kernel $k_{\text{cat}}^{\text{O}}$, a Matérn 5/2 continuous kernel is combined with a quadratic nondecreasing spline warping function F (see Eq. 9). We consider GP models with constant trends and fit them using the R package *kergp* (Deville et al., 2018) by applying a prescaling 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 optimizer by linear approximations named *COBYLA* developed by Powell (1994) with 250 randomly selected initial starts. The predictability and stability were assessed via a 5-fold cross-validation procedure repeated 25 times.

4.2 Application to the synthetic case

Considering the synthetic test case, Fig. 4 depicts the GP-derived correlation matrices for each of the kernel assumptions described in Sect. 3.1 using the training dataset of intermediate size (m=5). The application of the expert-based and ordinal kernel assumptions (Fig. 4c, e) shows some consistent structures among the levels, namely, the highly correlated group for u_1 to u_4 and the particular behavior of u_5. The magnitude of the interdependencies between the identified group and u_5 differs slightly: 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 interlevel dependencies is richer for the general (Fig. 4a) and LR2-based GP models (Fig. 4b). For these correlation matrices, the interpretation is also less straightforward than for E or O, for which some group structures are more easily identified. The visual inspection of both matrices suggests, however, some interesting features, i.e., the increasing correlation between $u_{1\text{-}4}$ (for Gen) and the particular behavior of u_5, as well as a possible grouping of $u_{2\text{-}4}$ (for LR2). Capturing the correlation structure can be hard for these cases given the size of the training dataset (of 25); this is further discussed below when analyzing the stability criterion. In this case, the exchangeable assumption (i.e., compound symmetry *CS*) leads to a low-to-moderate intercorrelation coefficient of ~40% (not shown).

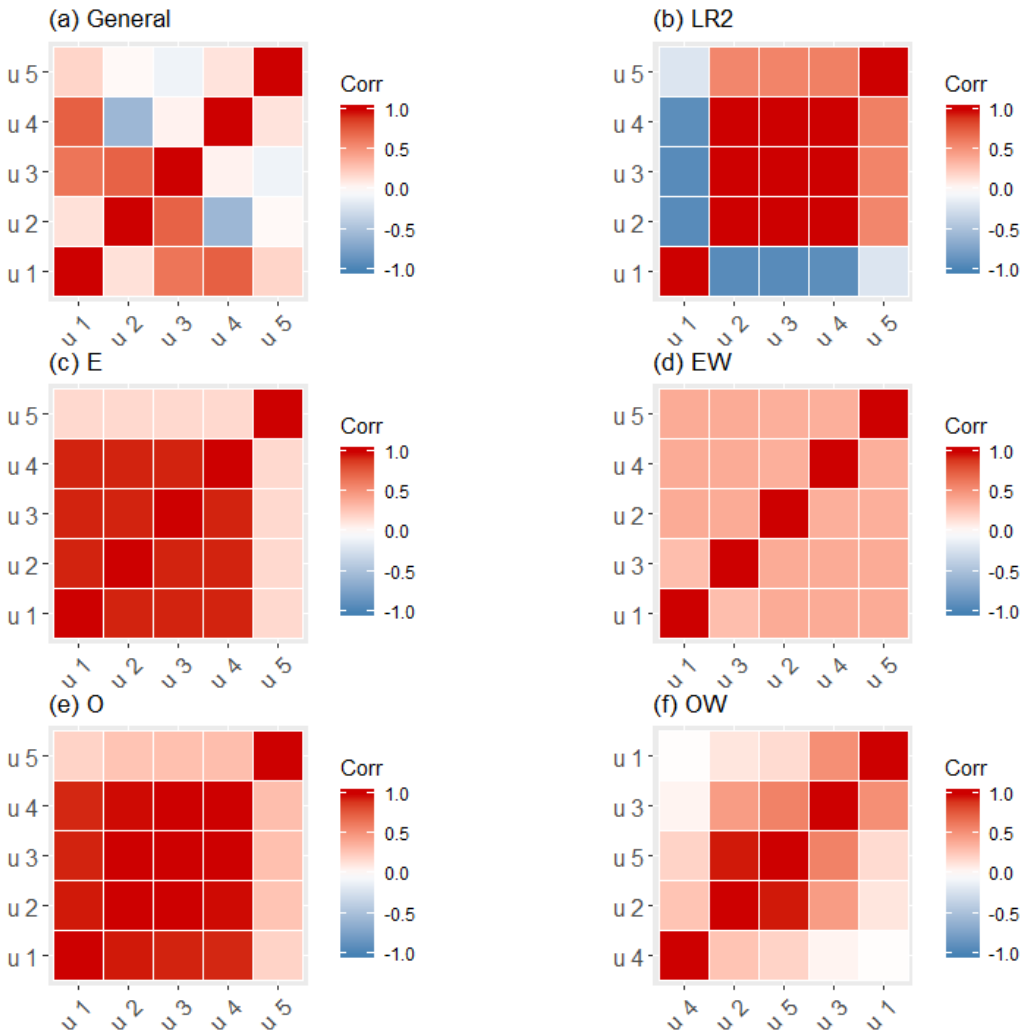

Figure 4. Correlation matrix for the synthetic test case (with m=5) considering different assumptions (see Sect. 3.1) regarding the kernel associated with the categorical input variable denoted u, namely: LR2, low rank kernel with rank = 2; E, expert-based kernel; EW, expert-based kernel with unrealistic grouping, O, ordinal kernel; OW, ordinal kernel with unrealistic order (see Sect. 2.3 for a formal description). For the expert-based assumptions (c and d), the ordering of matrix coefficients follows expert-based clustering, i.e., (u_1-u_4), and (u_5). Note that the compound symmetry kernel leads to a correlation matrix with a unique coefficient of ~40% (not shown).

The four criteria for kernel model selection are examined in Fig. 5. For the sake of comparability between the different assumptions for m, we preferably plot the difference of BIC with the minimum value over the whole experiment (named ΔBIC). Several observations can be made:

- Explainability measured by BIC (Fig. 5a) appears efficient here to exclude the unrealistic assumptions *OW* and *EW*: they present very large BIC differences regardless of *m*. For instance, Burnham and 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;
- The Q^2 criterion 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. 5b): selecting one of them is thus hardly achievable. Furthermore, the model associated with the unrealistic expert-based grouping *EW* has a satisfactory predictive capability at low *m* values (although the width of the confidence interval is larger than the others);
- For larger *m* value - here m=6 - (i.e., with the largest size of the training dataset), the Q^2 criterion allows us to identify model *O* as the most appropriate model with respect to the predictive capability ($\log_{10}(1\text{-}Q^2)$ is the lowest in Fig. 5b);
- The other facet of predictability related to the coverage of the prediction intervals suggests discarding 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 regardless of *m* (Fig. 5c);
- The stability criterion (Fig. 5d) tends to suffer from the same sensitivity as Q^2 to the size of the training dataset but has a higher discriminative power: at low *m* values (m=4), both kernel models *CS* and *O* are selected as very stable (because of low $\log_{10}(err)$ values in Fig. 5d). However, a high stability is also reached for *EW*. For a large *m* value, the ordinal assumption is then clearly selected as the assumption leading to the most stable correlation matrix.

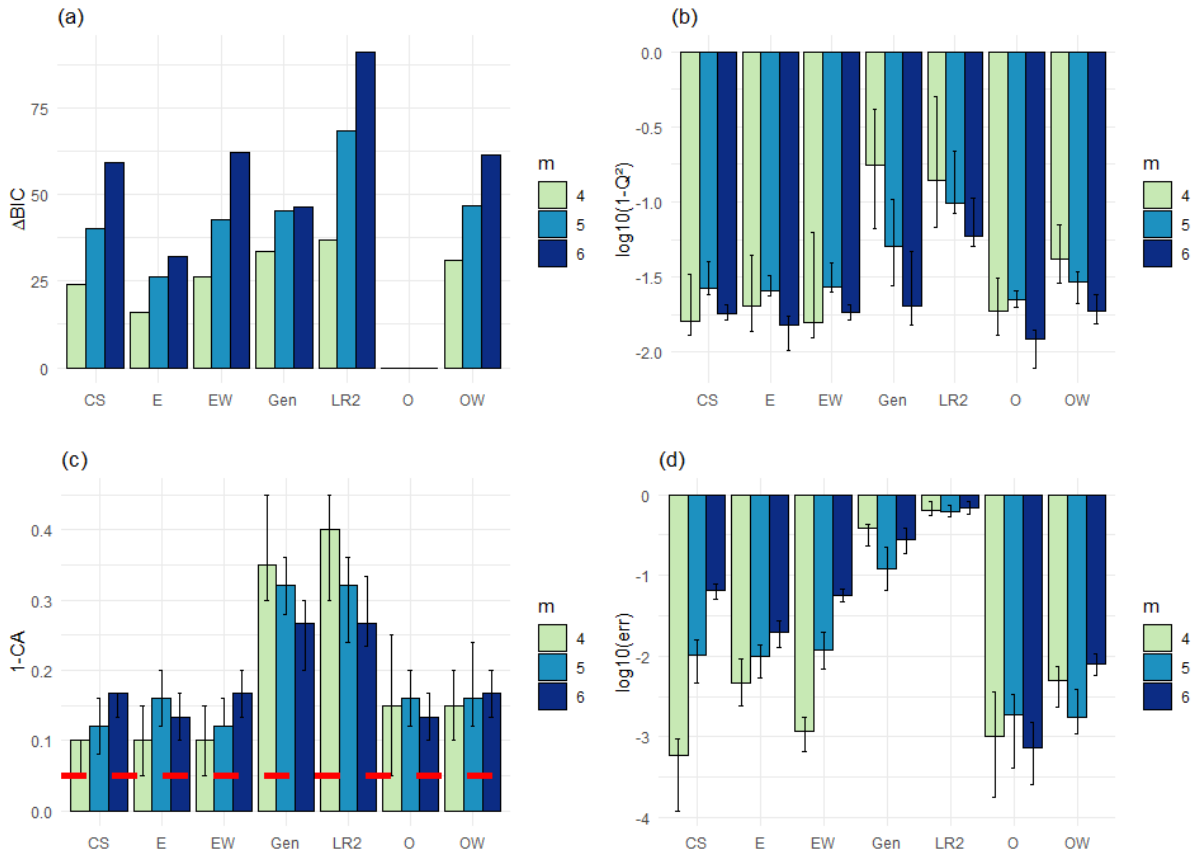

Figure 5. Selection criterion for the synthetic test case considering different numbers *m* of *x*-points per *u*-level: (a) BIC difference with respect to the minimum value; (b) predictability measured by 1-Q^2 ($\log_{10}$ scale); (c) coverage error measured by 1-*CA*. The horizontal dashed line corresponds to 5%, i.e., the threshold consistent with the level of the 95% prediction interval; (d) stability error *err* ($\log_{10}$ scale), considering different kernel models, namely: CS, compound symmetry kernel; E, expert-based kernel; EW, expert-based kernel with unrealistic grouping, Gen, general kernel; LR2, low rank kernel with rank = 2; O, ordinal kernel; OW, ordinal kernel with unrealistic order (see Sect. 2.3 for a formal description). The three latter criteria are derived from the 5-fold cross validation repeated 25 times: the height of the bar plot 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 us to successfully fulfill three of the criteria (explainability, predictability, and stability), especially at sufficiently high m values (m≥5). For coverage, the ordinal assumption is not ranked first, but *CA* appears of a reasonable order of magnitude (median value of ~85%), i.e., with moderate

deviation from the level of the 95% predictive interval. For the small size of the training dataset (m=4), unambiguously rejecting selecting *EW* and *CS* becomes difficult if the explainability criterion (BIC criterion) is not considered, which can partly be explained by the analysis of the *EW*-based correlation matrix (Fig. 4d), which reveals a (quasi)homogeneous structure with correlation coefficients ranging from 44 to 56%, i.e., of the same order of magnitude as the *CS* assumption (of ~40%), hence indicating that both GP models should perform similarly, suggesting 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).

4.3 Application to the real case application 1

Considering the cyclone test case, Fig. 6 depicts the GP-derived correlation matrices for each of the kernel assumptions described in Sect. 3.2. Fig. 6a reflects the hypothesis of a single group without difference in effect on maximum significant wave height between the tracks. The derived correlation appears to be high (~80%). The application of alternative kernel assumptions shows some consistent structures among the cyclone tracks. Two groups of cyclones appear to be highly correlated (coefficient >75%), namely, those coming from NE and those coming from NW, as shown in Fig. 6b (general formulation) and Fig. 6 e,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. 6c, d) - but these assumptions lead to a richer interdependency structure; (2) to a lesser extent, the expert-based assumption *E3cs* (Fig. 6 h). However, we note that there is ambiguity regarding the Dumile cyclone (see in particular, Fig. 6g), which is highly correlated with either NE cyclones (*E2*, *E2cs*) or NW cyclones (*LR3*, *E3cs*). The different assumptions all suggest a moderate correlation (on the order of 40-60%) between groups of cyclones coming from NE and NW. These observations are consistent with the ordinal assumption (Fig. 6h), which indicates a decreasing correlation from Gael (the track with the lowest angle of approach) to the Banzi cyclone (the track with the largest angle of approach) and a central role of Dumile, with a decreasing correlation either with the NE cyclone or with the NW cyclone.

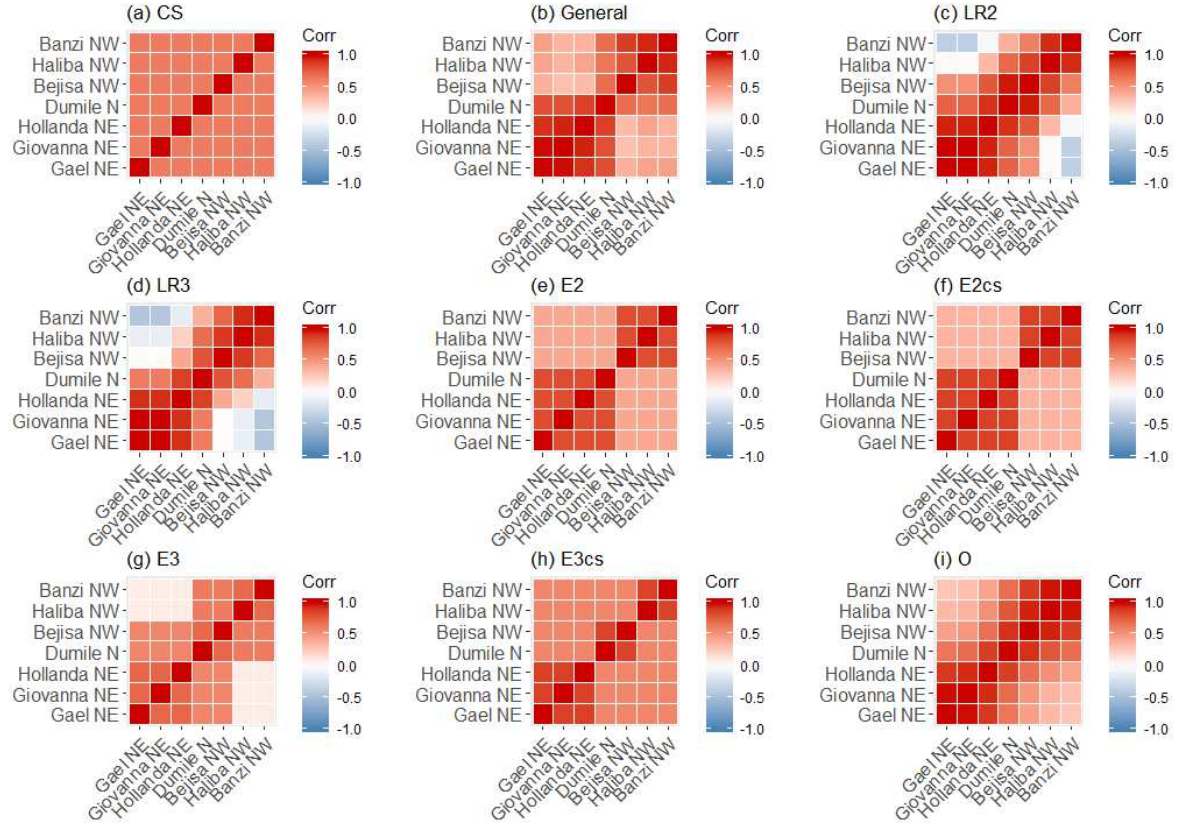

Figure 6. Correlation matrix for the cyclonic test case considering different assumptions (see Sect. 3.2) regarding the kernel associated with the categorical input variable, namely: CS, compound symmetry kernel; LRj, low rank kernel with rank = j; Ej, expert-based kernel with j groups; Ejcs, expert-based kernel with j groups and CS assumption; O, ordinal kernel (see Sect. 2.3 for a formal description).

The four criteria of kernel model selection are examined in Fig. 7. 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 are large; here, they are larger than 20. Regarding predictability, the ordinal assumption also leads to the best model regarding this criterion. However, the predictability of the expert-based model candidates (*E2*, *E2cs*, *E3*, *E3cs*) remains of moderate-to-high degree (with a median value of Q^2 of approximately 90%). Regarding *CA*, the use of *E3cs* allows us to reach the level of the 95% prediction interval, but alternative assumptions (*E2*, *E2*cs, *E3*cs, *CS* and *O*) lead to satisfactory coverage of the prediction intervals. Finally, Fig. 6d indicates the poor stability of the *Gen* and *LR3* kernels, i.e., the high sensitivity of the estimates of the correlation coefficients, which may be related to the high number of coefficients to estimate

(21 and 11, respectively), also showing the satisfactory stability of two alternative assumptions, i.e., the expert-based ones and the ordinal one.

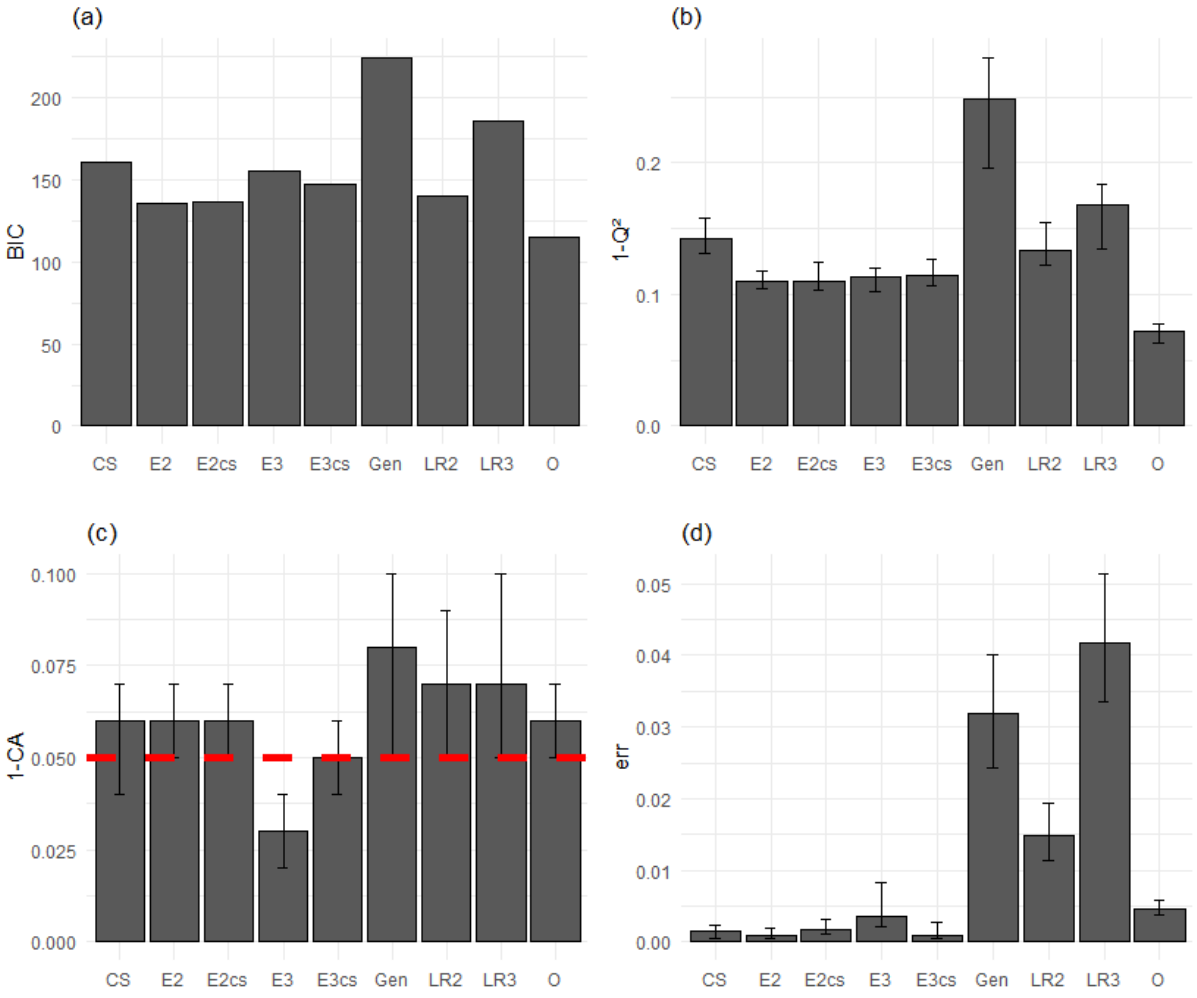

Figure 7. Selection criterion for application case 1: (a) BIC; (b) predictability measured by 1-Q^2; (c) coverage error measured by 1-*CA*. The horizontal dashed line corresponds to 5%, i.e., the threshold consistent with the level of the 95% prediction interval; (d) stability error *err*, considering different kernel models, namely: CS, compound symmetry kernel; Ej, expert-based kernel with j groups; Ejcs, expert-based kernel with j groups and CS assumption; Gen, general kernel; LRj, low rank kernel with rank = j; O, ordinal kernel (see Sect. 2.3 for a formal description). Criteria (b-d) are derived from a 5-fold cross validation repeated 25 times: the height of the bar plot is the median value, and the lower and upper bounds are defined using the 25th and 75th percentiles.

On this basis, we can conclude that the ordinal assumption allows us to reach a satisfactory trade-off between the four criteria, which appears to be consistent with the aforementioned physical intuition; the "latent" continuous variable here is related to the angle of approach. Figure 8a 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. 9): this shows a strong link between NE cyclones Gael-Giovanna (and to a lesser extent, for Hollanda as well), and a quasilinear increasing influence to Banzi. This warping is the basis for computing the correlation matrix (Fig. 6i). To ease the interpretation, let us focus on a single row of Fig. 6i, i.e., the pairwise correlation between the Gael track and the others. Fig. 8b provides a clear indication of a highly correlated group of cyclones coming from NE (within the red-colored envelope), with a correlation coefficient exceeding 80% and a decreasing correlation with those coming from NW. The analysis of the correlations derived from the cross-validation iterations (for each repetition) - gray lines in Fig. 8b - 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. 6e).

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

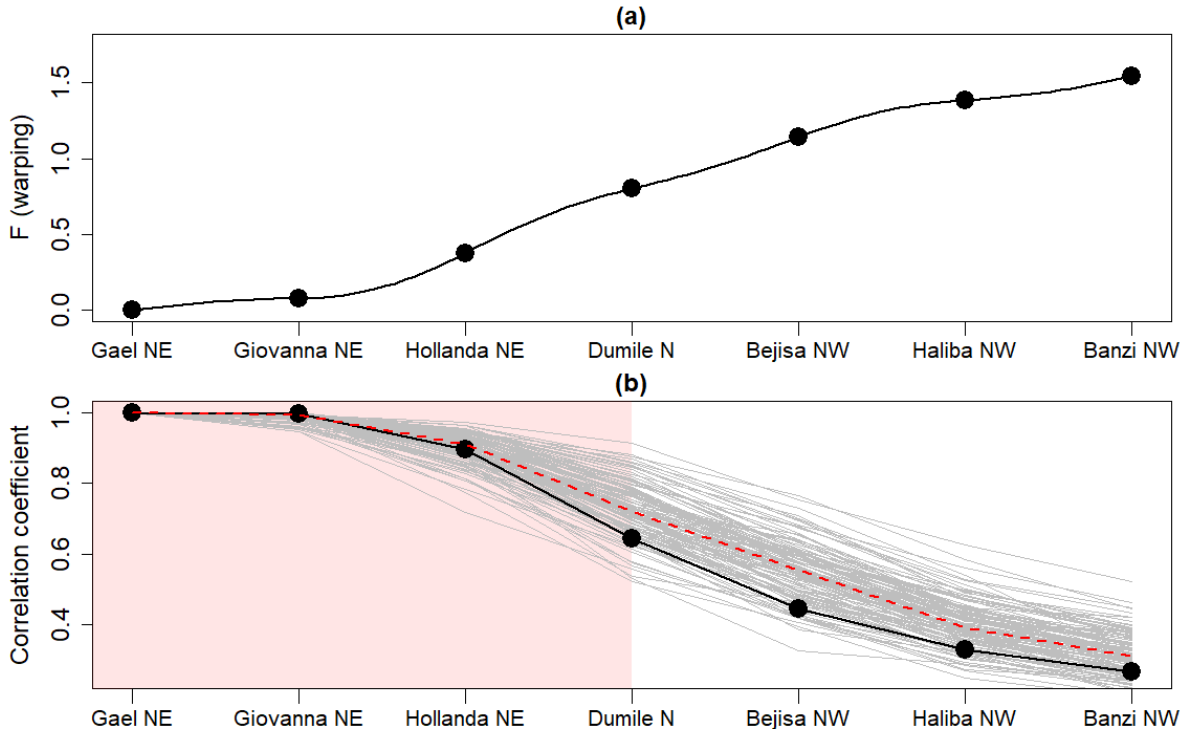

Figure 8. (a) Spline-based (unnormalized) warping function F (see Eq. 8) used in the ordinal assumption for the cyclone case. (b) Pairwise correlation between the Gael track and the other tracks. The red envelope indicates cyclones coming from the northeast (denoted NE). The gray-colored curves are the correlations derived at each iteration of the 5-fold cross-validation procedure (repeated 25 times). The red dashed line indicates the median value.

4.4 Application to the real case application 2

Considering the CO_2 geological storage test case, Fig. 9 depicts the GP-derived correlation matrices for each of the kernel assumptions described in Set. 3.3. Some consistent structures can be noticed regarding law 1, which appears to be anticorrelated with the others, as shown in Fig. 9a (general assumption) and in Fig. 9 b,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%, 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 decrease. Although disagreeing on the correlation magnitude, the *LR2*, *Gen*, and *E* models all suggest a moderate correlation among all laws except for law 1. The assumption *LR6* also suggests the particular behavior of law 5, which goes in the same direction as the expert-based clustering of considering it as 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 laws 6-9 (as suggested by the experts, see also the laws outlined in dark blue in Fig. 3), 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 laws 2 and 4 as well).

To summarize, the inspection of the correlation matrices is more difficult here than for the cyclonic application case (Sect. 3.2), where all assumptions more or less agree regarding the information supplied by the modelers. Nevertheless, this inspection highlights the specificities of law 1 and law 5 (LR6 assumption), which both strongly differ from the other laws: this is suggested by the irreducible water saturation (yellow and medium blue curve in Fig. 3c,d): law 1 is even more different with an irreducible water saturation associated with a large maximum gas residual saturation, which seems to explain why the model behaves so specifically when this law is accounted for.

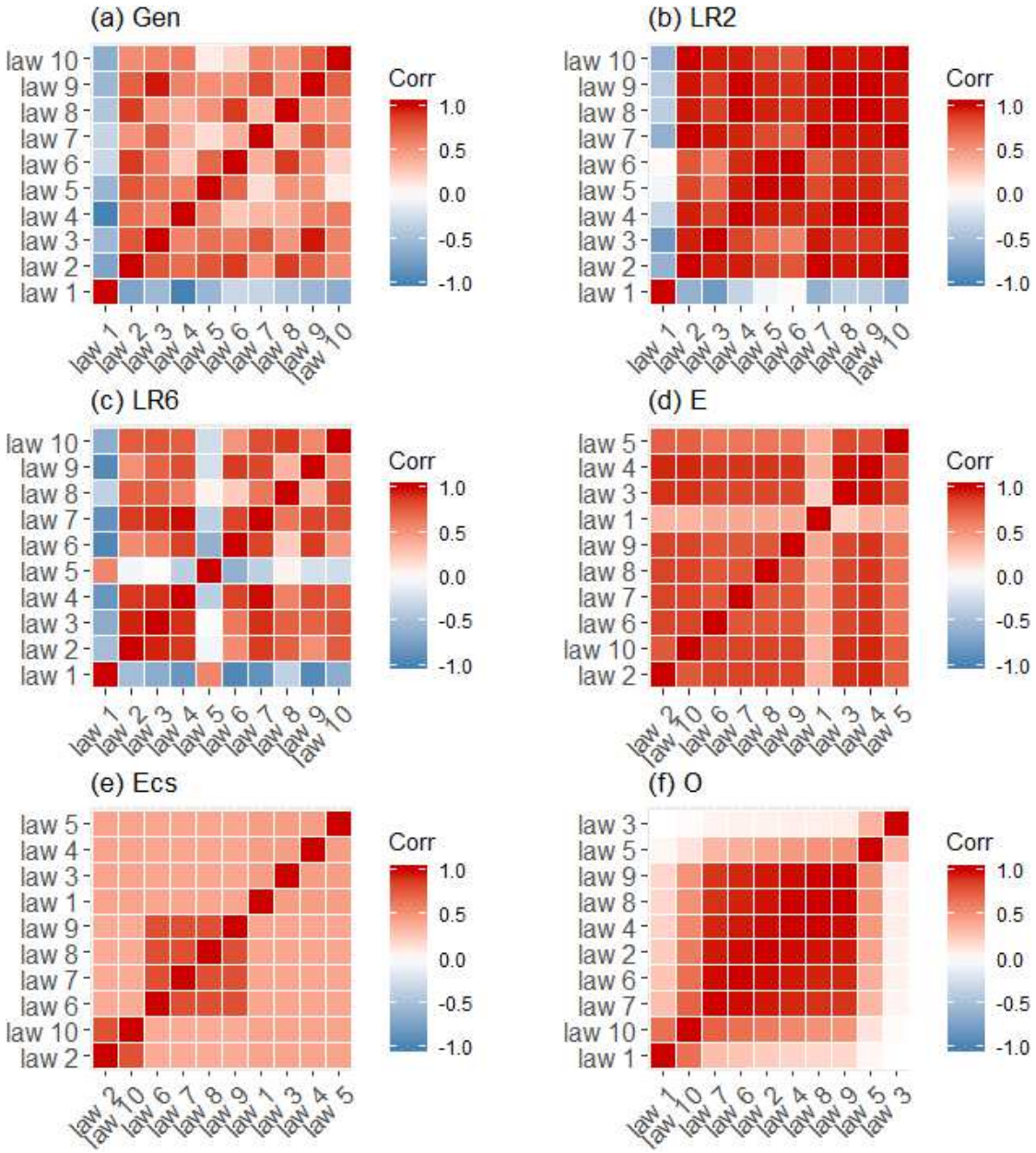

Figure 9. Correlation matrix for the reservoir test case considering different assumptions (see Sect. 3.3) regarding the kernel associated with the categorical input variable, namely: Gen, general kernel; LRj, low rank kernel with rank = j; E, expert-based kernel; Ecs, expert-based kernel with compound symmetry assumption; O, ordinal kernel (see Sect. 2.3 for a formal description). For the expert-based assumptions (d and e), the ordering of matrix coefficients follows expert-based clustering, i.e., (law 2; law 10); (law 6-9), (law 1); (law 3); (law 4); (law 5). For ordinal assumption (f), the 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*, where the expert-based grouping of laws, in particular laws 6-9 (in dark blue 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 stability (due to the lowest number of *CS* correlation coefficients, namely, 1). Note that due to the high number of covariance parameters (relative to the 100 runs), *Gen* and *LR6* score poorly here.

From this analysis, we can conclude that, given the 100 simulation results, there is only mild evidence supporting the assumption of the structure associated with the permeability laws that was intuited from the analysis of the boxplots in Fig. 3a and of the curve similarities (Fig. 3b-d). Similar to the synthetic case, additional simulation results should be performed to unambiguously discriminate the most appropriate kernel assumption.

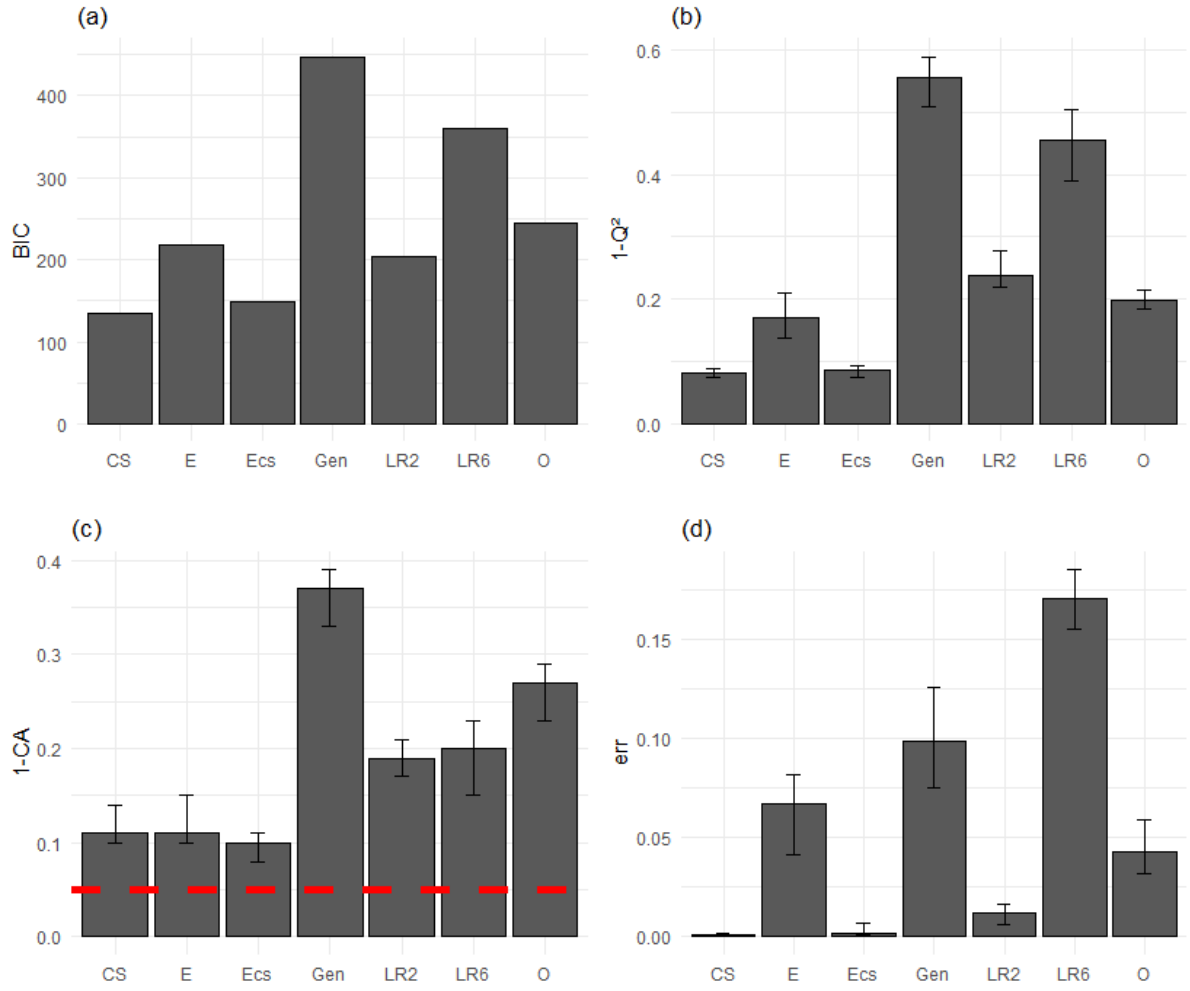

Figure 10. Selection criterion for application case 2: (a) BIC; (b) predictability measured by 1-Q^2; (c) coverage error measured by 1-*CA*. The horizontal dashed line corresponds to 5%, i.e., the threshold consistent with the level of the 95% prediction interval; (d) stability error *err*, considering different kernel models, namely: CS, compound symmetry kernel; E, expert-based kernel; Ecs, expert-based kernel with CS assumption; Gen, general kernel; LRj, low rank kernel with rank = j; O, ordinal kernel (see Sect. 2.3 for a formal description). Criteria (b-d) are derived from a 5-fold cross validation repeated 25 times: the height of the bar plot is the median value, and the lower and upper bounds are defined using the 25th and 75th percentiles.

5 Comparison to tree-based methods

In practice, GP models are not the first statistical modeling option that comes to mind when addressing the problem of categorical variables. A popular approach relies on tree-based

methods such as 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 interlevel dependence (as well as the interactions with the other input variables) 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.

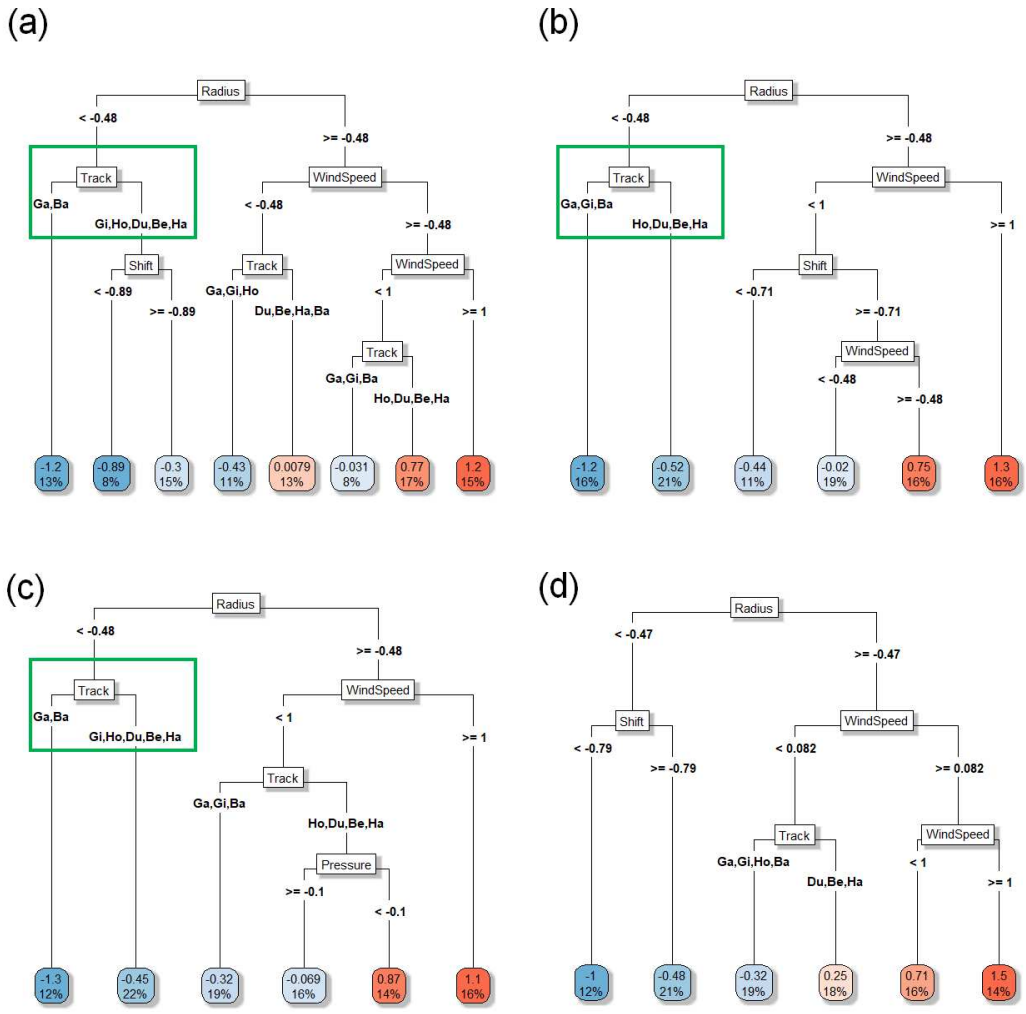

Figure 11. (a) Tree structures derived from the analysis of the cyclone real case; (b-d) Tree structures considering three iterations of the 5-fold cross validation procedure. The decision rule is provided on each respective branch. The leaf (bottom node) is colored according to the

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 with the 'Track' node corresponds to the first two letters of the cyclone names provided in Fig. 2. The green rectangle outlines a particular grouping of tracks.

However, from the analysis of Fig. 11, we note 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 model setups with a "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 is 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 dependencies between levels, 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 numbers m of x-points per u-level, Q^2 is derived from the leave-one-out cross-validation procedure. For application cases 1 and 2, Q^2 is derived from the 5-fold cross-validation procedure (repeated 25 times): the median values are given together, and the 25th and 75th are 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]

However, 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 *mtry* input parameters randomly chosen at that node, as confirmed by Table 2. However, 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), although some developments are available to extract some meaningful rules from RF (see, e.g., Fokkema, 2020).

This comparison exercise could be improved regarding different aspects; in particular, we have used commonly used parametrizations of the tested tree-based methods. First, tuning RF hyperparameters, namely, *mtry* and the minimum node size (which are, respectively set up to the root square of the total number of input variables and to 5), is expected to improve the results (Probst et al., 2019). Second, we used the default splitting rule based on squared residuals minimization by Breiman (2001): this splitting rule is known to favor the selection of variables with many possible splits (continuous variables or categorical variables with many levels) over variables with few splits, as shown by Strobl et al. (2007) for the estimates of variable importance measures. Improvements can either rely on some processing of the categorical inputs (see an extended discussion by Wright and König, 2019) or on alternative splitting rules, e.g., based on the randomized algorithm named "Extra-Trees" by Geurts et al. (2006) as tested by Jaxa-Rozen and Kwakkel (2018). Finally, it should be emphasized that one reason for the higher performance of GP models is that they have the ability to exploit the smoothness with respect to the continuous inputs. Further work may include recent developments on RF for smooth nonlinear relations (Friedberg et al., 2020).

6 Concluding remarks, Recommendations and Further work

Model uncertainties (related to the structure/form of the model or to the unambiguous choice of "appropriate" physical laws) are generally analyzed in computer-based studies by defining a categorical variable, i.e., a multilevel indicator. To obtain deeper insights into how the simulator output values computed for two levels of the categorical variable correlate, the

interlevel dependence structure is learned from a series of computer experiments (100-200) by means of a GP-based correlation matrix.

Table 3 summarizes the key aspects of the GP approach for three test cases together with a comparison to tree-based methods, showing that the GP-based approach can be seen as a satisfactory compromise because: (1) it clearly achieves higher predictability with a Q^2 value >90% given a moderate size of the training dataset (typically 100-200), whatever the considered case (Table 2), while reaching high performance with respect to different criteria (explainability, simplicity, stability); and (2) the correlation matrices provide a concise and graphical way of interlevel dependence structure (using the interpretation provided in Sect. 2.2). This matrix (estimated for three cases in Figs. 4,6,9) is useful to design complementary simulation-based studies by focusing on some levels that present some distinct behaviors (such as level u_5 for the synthetic case or the end members of the cyclone tracks) or by confirming that all levels should be considered (such as in the reservoir case). The GP-revealed interlevel dependence structure should also provide key elements to support the scenario discovery process, in particular to nuance the uniformity and independence assumptions of the scenarios (Quinn et al., 2020). Bringing the combined GP scenario discovery to an operational level necessitates further investigations.

Table 3. Synthesis and practical recommendations on the interlevel dependence structure detected by three regression methods

Criterion	Predictive capability	Interlevel dependence structure
Gaussian Process Regression (GP)	Higher predictability given a moderate size of the training dataset (typically 100-200), with Q^2 value >90% for considered test cases.	Graphical presentation through a correlation matrix (see examples in Figures 4,6,9) with practical interpretation described in Sect. 2.2 (by mixing the kernels by tensor product). A careful selection of the kernel model is necessary. This selection can be based on the multicriterion approach of Sect.

		2.4.
Regression Decision Tree (DT)	Lower predictability for all test cases.	The tree-like presentation provides the dependence structure as well as interaction with the other inputs. High sensitivity to the training dataset (see, e.g., Fig. 11).
Regression Random Forest (RF)	Higher predictability than DT but remains lower than GP by a factor ~1.3-2.1 for the considered test cases.	Extraction of a representative tree is possible through the application of adequate extraction rules techniques.

The cornerstone of the GP approach is, however, the careful selection of the kernel model that can take different forms (exchangeable, ordinal, group, etc.) depending on the physically based assumptions for the categorical variable. Table 1 provides the different options that the environmental modeler can implement. A multicriterion selection approach is proposed to question the different kernel options within a transparent procedure. 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. In the absence of a physical-based assumption, an optimization search procedure can be envisioned, for instance, by searching for all subsets of groups or for the rank dimension value of the LR kernel that minimizes all four proposed criteria (within a multifold cross validation approach).

One limitation of the proposed GP-based approach is that it does not ensure that a unique kernel model is selected: multiple assumptions may eventually turn out 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 expert-based kernel models perform similarly with respect to the four selection criteria. Although 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, which 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. However, this question deserves 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 fulfill 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 and Taflanidis (2019) for uncertainty quantification and by Ginsbourger et al. (2008) for optimization problems.

Acknowledgments

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. We are grateful to Prof. J. Rougier as well as to two anonymous reviewers for their comments that led to improvements of the article.

Software and data 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

Reproducible code: a Jupyter Notebook is available on Zenodo at (Rohmer, 2022). It showcases the implementation of Gaussian process models using mixed continuous/categorical inputs variables with an application on the real case in the domain of marine flooding (application case 1 described in Sect. 3.2, and results discussed in Sect. 4.3). The design of numerical experiments and the simulation results for application case 1 are provided; to re-use these numerical simulation results, please contact Jeremy Rohmer, BRGM (j.rohmer@brgm.fr). The notebook provides the R scripts to perform the fitting and the performance assessment, which can easily be adapted to any other application case.

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).

Cressie, N., Hardouin, C., 2019. A diagonally weighted matrix norm between two covariance matrices. Spatial statistics 29, 316-328.

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., 2020. Local linear forests. Journal of Computational and Graphical Statistics, 1-15.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach Learn 63, 3-42.

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.

Juanes, R., Spiteri, E. J., Orr Jr, F. M., & Blunt, M. J., 2006. Impact of relative permeability hysteresis on geological CO2 storage. Water resources research, 42(12).

Kwakkel, J. H., Jaxa-Rozen, M., 2016. Improving scenario discovery for handling heterogeneous uncertainties and multinomial classified outcomes. Environmental Modelling & Software 79, 311-321.

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.

Probst, P., Wright, M. N., Boulesteix, A. L., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301.

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.

Quinn, J. D., Hadjimichael, A., Reed, P. M., & Steinscheinder, S., 2020. Can exploratory modeling of water scarcity vulnerabilities and robustness be scenario neutral? Earth's Future 8(11). doi:10.1029/2020EF001650

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.

Rohmer, J., (2022). Notebook for revealing the interlevel dependence structure of categorical inputs with kernel model selection - application to cyclone-induced wave modelling (Version 1). Zenodo. https://doi.org/10.5281/zenodo.6090468 (accessed 15 February 2022).

Rougier, J., Priebe, C. E., 2020. The Exact Form of the “Ockham Factor” in Model Selection, The American Statistician, DOI: 10.1080/00031305.2020.1764865

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.

Strobl, C., Boulesteix, A. L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, 8(1), 1-21.

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.

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

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.

Appendix A. Design of experiments for the real cases

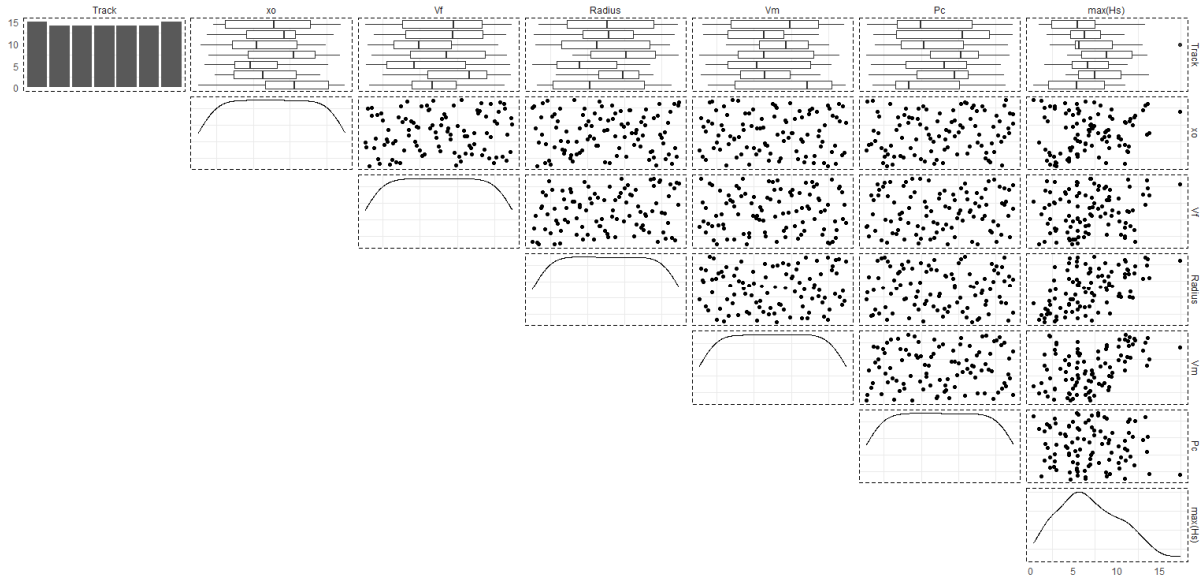

Figure A1. Design of the experiment for the real cyclone case described in Sect. 3.2. *Track* indicates the categorical variable related to the selection of the cyclone track (depicted in Fig. 2). The continuous variables (normalized between 0 and 1) are the landfall position x_o, the forward speed V_f, the radius of maximum winds *Radius*, the shift of the maximum wind speed V_m and the shift around the central pressure P_C. The first row provides the boxplots of the continuous variable given each level of *Track*. The last column provides the scatter plot of the variable of interest, i.e., the maximum significant wave height max (H_s) versus the continuous variables.

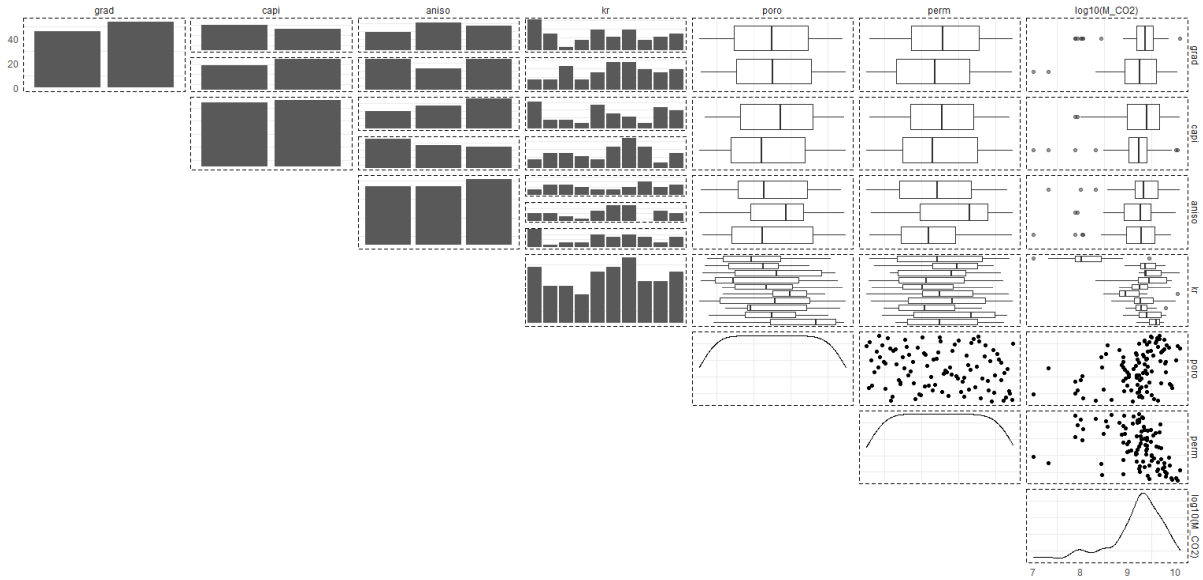

Figure A2. Design of the experiment for the reservoir real case described in Sect. 3.3. The categorical variables are the regional hydraulic gradient *grad*, the capillary effect *capi*, the permeability anisotropy and the physical laws for the relative permeability k_r. The last column provides the boxplot of the variable of interest, i.e., the log10 of the quantity of mobile CO_2 (denoted M_CO_2) given each level of the categorical variables and the scatter plot for the continuous variables (normalized between 0 and 1).