Quantizing Rare Random Maps: Application to Flooding Visualization

Charlie Sire, Rodolphe Le Riche, Didier Rullière, Jérémy Rohmer, Lucie Pheulpin, and Yann Richet

1. Introduction

Political authorities typically support their decisions regarding flood risk management based on flood maps calculated with hydraulic simulations. Beven et al. (2015) provide a realistic application, and a review of the existing methods for graphical representations of flooding uncertainty is given by Seipel and Lim (2017).

The design and analysis of numerical experiments is a generic procedure to capture and explore the uncertainties related to such heavy numerical models (Santner, Williams, and Notz 2003). It consists in performing a series of runs of the hydrodynamic flow simulator where the forcing conditions (the river flow rate and the offshore wave characteristics), the breach characteristics (location and dimension of the breach) or the embankments characteristics (geometry, failures conditions, etc.) are one realization of a random law. Rohmer et al. (2018) provide an example of the setting up of the forcing scenarios in coastal simulations and specify a method to determine the probabilistic laws describing the natural phenomena.

Visualizing simulations for risk assessment is a real challenge because of their time-consuming aspect, the low probability of the flooding event, and the need to communicate the results to authorities who do not necessarily have a scientific background. To this end, a simple-but-efficient approach consists in providing a small size set of representative flood maps. Such prototype maps (Hastie, Tibshirani, and Friedman 2009, chap. 13) are chosen to best represent the probability law associated to the flooding event. Calculating the prototype maps is a quantization task (Gray and Neuhoff 1998), implemented here through the Lloyd's algorithm (Pagès 2014). Lloyd's algorithm searches for Voronoï cell centroids such that the mean distance (with respect to the uncertainty distribution) between an observation and its nearest centroid is minimized.

This procedure yields results that have both a probabilistic meaning and can be visually interpreted by a wide audience. For instance, in the flooding case tackled in this article, the four nonempty prototype maps and their probability masses expressed as frequencies are given in Figure 1. The empty prototype map occurs with 97% probability.

With this visualization, one can picture the most typical flooding scenarios, like Figure 1(c) where water invades a large area, or more intense and concentrated flooding like Figure 1(d).

Lloyd's algorithm is not a new quantization technique (Mac-Queen 1967), and has evolved since its introduction (Bock 2008). This article presents new extensions to quantization that address the following bottlenecks. First, the quantization is done here in the context of costly numerical simulations, which makes the computation of the centroids more difficult. Indeed, the flood maps are computed with hydraulic codes that can take several hours to simulate the water levels associated to each set of input parameters. Plus, flooding is a rare event. Classical Monte Carlo sampling is inefficient for extreme events with a very large return period (typically 1000–10,000 years). Finally, the quantization is performed in a space of pixelated maps (64 × 64)
which raises metamodeling and storage questions when a very large number (> $10^6$) of realizations has to be handled.

The article is structured as follows. Section 2 defines the quantization problem and the importance sampling estimator to compute the prototype maps. Section 3 introduces the metamodeling scheme built to overcome the duration of the hydraulic simulations. Section 4 presents the algorithm and the performance metrics that will be evaluated on the two applications described in Section 5 (analytical test case) and Section 6 (real coastal flooding). Finally, Section 7 discusses the main results and gives possible extensions to the method. Nomenclature is also provided at the end.

2. Finding Prototype Maps

2.1. Quantization

Let us consider a random vector $X \in \mathcal{X}$, with $\mathcal{X}$ a subset of $\mathbb{R}^d$, $d \in \mathbb{N}$. In our context, $X$ is a parametric description of the offshore hydrodynamic (like the tide and surge parameters) and the dyke breach conditions. We define a random function $Y$ from $\mathcal{X}$ the space of offshore conditions and breach parameters to $\mathcal{Y}$ the space of possible flood maps, with an inner product $\langle \cdot, \cdot \rangle_\mathcal{Y}$, and a metric $\| \cdot \|_\mathcal{Y}$ induced by the inner product:

$$Y : \mathcal{X} \to \mathcal{Y} \quad x \mapsto Y(x).$$

In other terms, with this general formulation, $Y$ associates a distribution of flood maps to every vector of input conditions $x$. In practice, $Y$ can be deterministic and thus associate a single flood map to every $x$ which we write $y(x)$.

The quantization problem is to find a set $\Gamma_\ell = \{\gamma_1, \gamma_2, \ldots, \gamma_\ell\} \in \mathcal{Y}^\ell$ of $\ell \in \mathbb{N}$ representatives of $Y(X)$ in a mathematically relevant way. The notation $Y(X)$ should be understood as the map $Y$ conditioned by the random vector $X$ as in regression, $Y(X) \equiv Y|X$. The prototype maps will be these representatives in our study. More precisely, we first define for $\Gamma_\ell \in \mathcal{Y}^\ell$ the closest candidate representative as the function $q_{\Gamma_\ell}$:

$$q_{\Gamma_\ell} : \mathcal{Y} \to \Gamma_\ell \quad y \mapsto q_{\Gamma_\ell}(y) = \arg\min_{y \in \Gamma_\ell} \| y - \gamma \|_\mathcal{Y}.$$

A measure of the quantization error is the average distance between a randomly chosen map and its closest representative (Pagès 2014),

$$e(\Gamma_\ell) = \left[ \mathbb{E} \left[ \| Y(X) - q_{\Gamma_\ell}(Y(X)) \|_\mathcal{Y}^2 \right] \right]^{\frac{1}{2}}.$$

To each centroid $\gamma_j$ in $\Gamma_\ell$, we associate a Voronoi cell $C_{\gamma_j}^{\Gamma_\ell}$ which is the subset of points in $\mathcal{Y}$ closest to $\gamma_j$,

$$C_{\gamma_j}^{\Gamma_\ell} = \{ y \in \mathcal{Y}, q_{\Gamma_\ell}(y) = \gamma_j \}.$$
The objective is to minimize this quantization error, that is, to find
\[
\Gamma_e^* = [\gamma_1^e, \ldots, \gamma_{\ell}^e] \in \arg\min_{\Gamma_e \in \mathcal{Y}^e} (e(\Gamma_e)) \\
\in \arg\min_{\Gamma_e \in \mathcal{Y}^e} \left[ \mathbb{E} \left[ \|Y(X) - q_{\Gamma_e}(Y(X))\|_2^2 \right] \right]^{1/2}
\in \arg\min_{\Gamma_e \in \mathcal{Y}^e} \left[ \mathbb{E} \left[ \min_{j \in \{1, \ldots, \ell\}} \|Y(X) - y_j\|_2^2 \right] \right]^{1/2}.
\]

Finding the optimal centroids is in general a NP hard problem (Mahajan, Nimbhorkar, and Varadarajan 2012). However, the solution has a strong attribute (Pagès and Yu 2013):

**Theorem 2.1 (Optimal quantization, Kieffer, Cuesta-Albertos).**

- \( \mathcal{Y} \) is of finite dimension \( m \)
- \( \forall y, y' \in \mathcal{Y}, \langle y, y' \rangle_{\mathcal{Y}} = \sum_{i=1}^{m} \lambda_i y_i y_i' \) with \( \forall i, \lambda_i > 0 \)
- \( \mathbb{E} \left[ \|Y(X)\|_{\mathcal{Y}}^2 \right] < +\infty \)

then \( \forall j \in \{1, \ldots, \ell\}, \mathbb{E} \left[ Y(X) \mid Y(X) \in C_j^e \right] = \gamma_j^* \).

This means that the representatives of an optimal quantification coincide with the Voronoi cell centroids. The proof is given in Appendix A. In our context, the distance induced by the scalar product \( \langle y, y' \rangle_{\mathcal{Y}} = \sum_{i=1}^{m} \lambda_i y_i y_i' \) between two maps is the mean squared difference between the water depth at each pixel of the maps, weighted by \( \lambda_i \) which can represent the importance of the zone (e.g., the housing density, the presence of factories, etc.).

The popular Lloyd’s algorithm, also called K-means, is based on Theorem 2.1. The stopping criterion of the algorithm can be a number of iterations or a lower bound on a distance between \( \Gamma_e^k \) and \( \Gamma_e^{k+1} \). The principle of the iterations is illustrated in Figure 2 in 2D: the dots are five possible starting points \{\gamma_1^0, \ldots, \gamma_5^0\}, their associated Voronoi cells are delimited by the black lines, the centroids of these cells are computed for a uniform distribution and are drawn as the triangles. Figure 2 shows only one iteration and does not represent the convergence of the algorithm.

**Algorithm 1 Lloyd’s algorithm**

\[
\Gamma_e^0 \leftarrow \{\gamma_1^0, \ldots, \gamma_\ell^0\}, k \leftarrow 0 \\
1: \textbf{while} \text{ stopping criterion not met} \textbf{do} \\
2:\quad \gamma_j^{(k+1)} \leftarrow \mathbb{E} \left[ Y(X) \mid Y(X) \in C_j^{(k+1)} \right], j \in \{1, \ldots, \ell\}. \\
3:\quad k \leftarrow k + 1 \\
4: \textbf{end while}
\]

\( \{\gamma_1^0, \ldots, \gamma_\ell^0\} \), the Voronoi cells associated to initial points and their centroids in 2D.

**2.2. Dealing with Low Probability Events**

As stated before, Lloyd’s algorithm requires to compute at each iteration and for each Voronoi cell \( C_j^e \), \( j \in \{1, \ldots, \ell\} \),
\[
\mathbb{E} \left[ Y(X) \mid Y(X) \in C_j^{(k)} \right].
\]

The most natural approach would be to sample \( Y(X) \) from its density and then to calculate the empirical mean of \( Y(X) \) given \( Y(X) \in C_j^{(k)} \). However, flooding being a rare event, we expect to have one Voronoi cell with almost only empty maps associated to a probability close to 1, and other Voronoi cells with more critical floodings associated to very low probability maps. Thus, the straightforward sampling is not relevant to locate the Voronoi cells of low probabilities because their member maps will almost surely be missed. Importance sampling is thus needed to perform a sampling less concentrated on a single cluster.

We will need to estimate centroids and probabilities, which reduces to calculating \( \mathbb{E} [h(Y(X))] \) with \( h: \mathcal{Y} \mapsto \mathcal{H} \) for some set.
To compute the probability that $Y$ is in $A$, $P(Y(X) \in A)$, $h$ is the indicator function $h : y \mapsto 1_{y \in A}$ with $A \subset \mathcal{Y}$ and $\mathcal{H} = \{0, 1\}$. Alternatively $h$ is $h : y \mapsto y 1_{y \in A}$ with $\mathcal{H} = \mathcal{Y}$ to calculate the mean of $Y$ when it belongs to $A$.

The importance sampling technique is based on the fact that the representation of $\mathbb{E}[h(Y(X))]$ as an expectation is not unique (L’Ecuyer, Mandjes, and Tuffin 2009). Indeed, if we consider $X$ a random variable with density function $g$ with a support such that $\text{supp}(f_X) \subset \text{supp}(g) \subset X$, then

$$
\mathbb{E}[h(Y(X))] = \mathbb{E}\left[h(Y(\hat{X})) \frac{f_X(\hat{X})}{g(\hat{X})}\right].
$$

The importance sampling estimator of the mean comes from this last representation:

$$
\hat{E}_n^{IS}(h) = \frac{1}{n} \sum_{k=1}^{n} h(Y(\hat{X}^k)) \frac{f_X(\hat{X}^k)}{g(\hat{X}^k)} \in \mathbb{R}^p,
$$

with $(\hat{X}^k)_{k=1}^n$ iid random variables of density function $g$. It is well known that such importance sampling estimator is unbiased and consistent (Tabandeh, Jia, and Gardoni 2022).

The covariance matrix of the estimator is

$$
\mathbb{V} \left( \hat{E}_n^{IS}(h) \right) = \frac{1}{n} \mathbb{V} \left( h(Y(\hat{X})) \frac{f_X(\hat{X})}{g(\hat{X})} \right).
$$

The idea of the importance sampling method is to choose a relevant density function $g$, that is, a density function that minimizes the terms of the diagonal of the above covariance matrix.

### 2.3. Importance Sampling Combined with Quantization

In Lloyd’s algorithm, we need to repeatedly compute the expectations of $Y$ within the candidate Voronoi cells. At the $p$th iteration, these expectations are

$$
\gamma_p^{\{j\}} = \mathbb{E}\left[ Y(X) \mid Y(X) \in C_j^{p-1} \right] = \mathbb{E}\left[ Y(X) 1_{Y(X) \in C_j^{p-1}} \right] / \mathbb{P}(Y(X) \in C_j^{p-1}).
$$

As for a given $\Gamma$, $\mathbb{P}(Y(X) \in C_j^{\Gamma}) = \mathbb{E}\left[ 1_{Y(X) \in C_j^{\Gamma}} \right]$ is not known, we rely on the importance sampling estimators for both the numerator and the denominator and we introduce

$$
\hat{P}_p(\Gamma, j, Y) = \frac{1}{n} \sum_{k=1}^{n} 1_{Y(\hat{X}^k) \in C_j^{\Gamma}} \frac{f_X(\hat{X}^k)}{g(\hat{X}^k)},
$$

$$
\hat{E}_p(\Gamma, j, Y) = \frac{1}{n} \sum_{k=1}^{n} \frac{f_X(\hat{X}^k)}{g(\hat{X}^k)} \frac{1_{Y(\hat{X}^k) \in C_j^{\Gamma}}}{\hat{P}_p(\Gamma, j, Y)},
$$

the estimators of $\mathbb{P}(Y(X) \in C_j^{\Gamma})$ and $\mathbb{E}\left[ Y(X) \mid Y(X) \in C_j^{\Gamma} \right]$, respectively. The choice of the distribution $g$ is delicate and will be further discussed in Section 6.

The overall approach was described so far in the very general context of a random field $Y$. In the following, these estimators will be computed with $y$ the deterministic output of the hydraulic simulators, and $\hat{y}$ its predictor.

### 3. Dealing with Time-Consuming Experiments

In order to have an estimator of $\mathbb{E}\left[ y(X) \mid y(X) \in C_j^{\Gamma} \right]$ of sufficiently small variance for all cells $j \in \{1, \ldots, \ell\}$, we need to compute $y(\hat{X}^k)$ for a sample $(\hat{X}^k)_{k=1}^n$ of large size $n$, with $y$ the deterministic output of the hydraulic simulator. Realistic physical simulations would be too costly for that purpose, and a metamodel is required to approximate $y(X)$. Gaussian processes (GP) (Williams and Rasmussen 2006) are used here to replace some of the simulations. GPs are readily applicable to low dimensional outputs (most often scalars) but not to $y(X)$, the flood map, which is in $\mathbb{R}^{4096}$ ($64 \times 64$). To overcome that, Perrin et al. (2021) implemented a method based on functional principal component analysis (FPCA). Their strategy, which we rely on, is to work with the projections of the maps in a space induced by a basis of $n_{pc} \ll 4096$ maps only. The predictions are expressed as coordinates of a map in this basis, through a Gaussian process approach. In the following, we call “metamodel” the whole procedure: the FPCA followed by a Gaussian process regression.

#### 3.1. Projection on a Basis of Maps

The objective here is to find a basis of $n_{pc}$ maps, with $n_{pc}$ small enough, and to compute the projection of the maps onto this basis, thus representing every map $y(x)$ as a linear combination of the $n_{pc}$ basis maps:

$$
y_{\text{proj}}(x) := t_1(x) y_1^{\text{pca}} + t_2(x) y_2^{\text{pca}} + \cdots + t_{n_{pc}}(x) y_{n_{pc}}^{\text{pca}}.
$$

For this purpose, a functional principal components analysis (FPCA) is carried out on a training database of maps $(y(x^i))_{i=1,...,n_{\text{train}}}$. The steps are the following, and further details are given in Appendix B.1.:

- Find a decomposition of these maps, seen as realizations of a spatial function, on a functional orthonormal basis $\Phi = (\Phi_1, \ldots, \Phi_K)^T : y(x^i) = \Phi^T \alpha(x^i)$
- Reduce the number of basis functions by keeping the $K$ most important ones, $\Phi$, in terms of their contribution to the energy of the coordinates in the database $(\hat{\alpha}(x^i))_{i=1,...,n_{\text{train}}}$.
- Perform a PCA on the selected coordinates to finally get the principal components $(t(x^i))_{i=1,...,n_{\text{train}}}$ with $t(x^i) = (t_1(x^i), \ldots, t_{n_{pc}}(x^i))^T$ and the $n_{pc} \times K$ projection matrix $\Omega$.

The obtained FPCA-basis maps are $\Omega \Phi$. The two basis maps $y_1^{\text{pca}}$ and $y_2^{\text{pca}}$ of the coastal flooding example are shown in Figure 3. Negative values can appear on these FPCA maps, but they are selected here is induced by the usual inner product $\langle y, y' \rangle^y = \sum_{i=1}^{n} y_i y_i'$.

One can see that even in the worst case, the map $Y$ and its projection on the FPCA basis $y_{\text{proj}}$ are quite close. This
3.2. Gaussian Process Regression

We now aim at predicting the flood map for a new input \( x^* \) by predicting its coordinates in the FPCA basis \( (t_1(x^*), \ldots, t_{n_{pc}}(x^*)) \). For simplicity these coordinates are assumed independent, but the study of their dependency may be relevant in some cases (Gu and Shen 2020). Therefore, \( n_{pc} \) one-dimensional Gaussian processes define the metamodel for the flood maps. Let \( t: \mathcal{X} \subset \mathbb{R}^d \rightarrow \mathbb{R} \) be a deterministic function. \( t \) is approximated as a realization of a Gaussian process \( Z(x) \) of prior mean \( \mu(x) \) and prior covariance kernel \( k(x, x') \). The kernel contains the spatial dependencies between \( y(x) \) and \( y(x') \). In our study we use the stationary Matérn 5/2 kernel (Palar et al. 2023).

We consider a design of experiments \( \mathcal{X}_{train} = (x^1, \ldots, x^{n_{train}}) \) and the observations \( Z_{train} = (z_i)_{i \in [I_{train}]} = (t(x'))_{i \in [I_{train}] \text{ with } I_{train} = \{1, \ldots, n_{train}\}} \). \( t \) can be approximated using the distribution of \( Z(x) \mid Z(\mathcal{X}_{train}) = Z_{train} \), which is a Gaussian process with the following mean (Williams and Rasmussen 2006):

\[
\hat{t}(x) = \mu(x) + k(x, \mathcal{X}_{train})k(\mathcal{X}_{train}, \mathcal{X}_{train})^{-1}(Z_{train} - \mu(\mathcal{X}_{train}))
\]

where \( k(\mathcal{X}_{train}, \mathcal{X}_{train}) = [k(x_i, x_j)]_{i,j \in [I_{train}]} \) and \( k(x, \mathcal{X}_{train}) = [k(x, x_i)]_{i \in [I_{train}]} \). This method is applied to all the FPCA components \( i \in \{1, \ldots, n_{pc}\} \), for which the GP mean is a predictor \( \hat{t}_i(x^*) \) of \( t_i(x^*) \). Then, \( \hat{y}(x^*) \), a predictor of \( y(x^*) \) is obtained by inverse FPCA as explained in Appendix B.2. The metamodel errors are investigated in the following sections.

4. Algorithm and Performance Metrics

Sections 2 and 3 described how to estimate the expectations in Lloyd’s algorithm and how to predict a large number of maps that cannot be computed with the time-consuming hydraulic simulators. This section details the overall scheme for calculating the prototype maps and their probability masses, and then explains how to evaluate its robustness.

4.1. Prototype Maps Algorithm

The Prototype Maps Algorithm incorporates, inside Lloyd’s algorithm, the metamodelling of spatial output and the importance sampling scheme.

The expression for \( \hat{P}_\text{maps}(\Gamma^*, j, \hat{y}) \) and \( \hat{E}_\text{maps}(\Gamma^*, j, \hat{y}) \) stems from applying (4) to the Gaussian Process prediction of the maps. The calculation of \( \hat{P}_\text{maps}(\Gamma^*, j, \hat{y}) \) at step 8 could be extracted from the last iteration since it is the denominator of \( \hat{E}_\text{maps}(\Gamma^*, j, \hat{y}) \). However, these quantities are the estimators of the probabilities that a flood map is part of each Voronoi cell, for which we want to be as precise as possible, we thus recompute \( \hat{P}_\text{maps}(\Gamma^*, j, \hat{y}) \) at the end with \( n > n_{\text{maps}} \).

The overall complexity of the Prototype Maps Algorithm is \( O((\bar{n} + n_{\text{maps}})n_{\text{train}}^3n_{\text{pc}} + \bar{n}^2(n_{\text{maps}}n_{\text{fit}} + \bar{n})) \) with \( n_{\text{fit}} \) the number of iterations. The calculation of the complexity is detailed in Sire et al. (2023a). The algorithm requires the storage of
Algorithm 2 Prototype Maps Algorithm

Require: \((y(x^j))_{j=1,...,n_{\text{train}}}, f_X, g, \text{minDistance}, \ell\)
1: Sample \((\hat{x}_i^j)_{i \in [1,...,n_{\text{maps}}]}\) iid of density function \(g\)
2: Compute \((\hat{y}(\hat{x}^j))_{1 \leq k \leq n_{\text{maps}}}\) from \((y(x^j))_{j=1,...,n_{\text{train}}}(\text{GP \\& \\ FPCA})\)
3: Compute \((f_k(\hat{x}^j))_{1 \leq k \leq n_{\text{maps}}}\)
4: Initialize \(\Gamma_\ell^0 = \{y_0, \ldots, y_\ell, \ldots, y_\ell\} \in \gamma^\ell\)
5: while \(\|\Gamma_{\ell+1} - \Gamma_\ell\|^2 \geq \text{minDistance}\) do
   \(y_j = \hat{E}_{n_{\text{maps}}}(\Gamma_\ell^j, j, \hat{y}), j = 1, \ldots, \ell\)
   \(k = k + 1\)
6: end while
7: \(\Gamma_\ell^\ast = \Gamma_k^j\)
8: Compute \(\hat{P}_n(\Gamma_\ell^\ast, j, \hat{y}), j = 1, \ldots, \ell\)

Output: \(\Gamma_\ell^\ast\) and \(\hat{P}_n(\Gamma_\ell^\ast, j, \hat{y})\)

\((\hat{y}(\hat{X}^j))_{1 \leq k \leq n_{\text{maps}}}\) of size \(n_{\text{maps}}^2\) (with \(s = 64\) here), which is a limiting factor.

If the map simulations are not expensive, the algorithm can of course be performed directly with the true maps \((y(x^j)) = 1, \ldots, n_{\text{maps}}\), instead of the GP predictions.

\(\text{minDistance} = 10^{-10}\) is chosen here in the applications.

\[\epsilon_{\ell}^{\text{MM}} = \frac{\hat{e}(\Gamma_\ell^\ast) - \hat{e}(\Gamma_\ell^j)}{\hat{e}(\Gamma_\ell^j)}, \]

where \(\hat{e}(\Gamma_\ell)\) is empirical quantization error. We take \(n_e = 10^6\) here. This error can only be computed in the analytical case (see Section 5.2) as \(\Gamma_\ell^\ast\) can not be computed in the flooding case whose computational cost is too high.

The following performance metrics quantify the errors caused by the metamodel (MM) on the quantization and on the cluster probability:

- The excess in quantization error due to the metamodel,
  \[\epsilon_{\ell}^{\text{MM}} = \frac{\hat{e}(\Gamma_\ell^\ast) - \hat{e}(\Gamma_\ell^j)}{\hat{e}(\Gamma_\ell^j)}, \]

- The relative probability error,
  \[\epsilon_{\ell}^{\text{PM}}(n, \Gamma_\ell^j) = \frac{\hat{P}_n(\Gamma_\ell^\ast, j, \hat{y}) - \hat{P}_n(\Gamma_\ell^j, j, \hat{y})}{\hat{P}_n(\Gamma_\ell^j, j, \hat{y})}. \]

computed for all \(r = 1, \ldots, n_{\text{maps}}\) and \(j = 1, \ldots, \ell\). Notice that the prototypes are identical in both probabilities, and only the effect of the metamodel on the cluster probability is assessed.

**Importance sampling performance metrics.** The effect of the importance sampling (IS) on the quantization is quantified, for a given Voronoi cell \(C_j^{\Gamma_{\ell}^j}\), through two metrics:

- The IS coefficient of variation of the membership probability, \(\frac{\hat{e}_{\ell}^{\text{IS}}(\hat{n}(\Gamma_{\ell}^j)), j)}{\hat{e}(\hat{n}(\Gamma_{\ell}^j), j)}\), which is an estimator of \(\frac{\text{Var}(\hat{P}_n(\Gamma_{\ell}^j, j, y))}{\hat{P}_n(\Gamma_{\ell}^j, j, y)}\). It measures an error due to the importance sampling in the estimation of the probability of the Voronoi cell number \(j\).
- The IS centroid standard deviation, \(\epsilon_{\ell}^{\text{IS}}(\hat{n}(\Gamma_{\ell}^j), j)\), which is the 90%-quantile of the diagonal terms of the covariance matrix \(\text{Cov}(\hat{P}_n(\Gamma_{\ell}^j, j, y))\). It characterizes the error coming from importance sampling in the estimation of the centroid of the Voronoi cell number \(j\). Among other things, it will show potential variations due to the denominator being close to zero in the centroid formula (3).

These quantities are calculated for \(r = 1, \ldots, n_{\text{maps}}\), and \(j = 1, \ldots, \ell\) and their distributions serve as the metrics for the importance sampling performance. Details about the computation of these errors are given in Sire et al. (2023b).

**5. Application to an Analytical Case**

**5.1. Campbell Test Case**

The performance of the method is first evaluated on an analytical test case, called the Campbell2D function (Marrel et al. 2011). Contrarily to realistic flooding simulators, a very high number of calls to the Campbell function are possible, which makes it an appropriate case for testing the Prototype Maps Algorithm. This function has eight inputs \((d = 8)\) and a spatial map as output, that is, its output is a function which depends on two inputs \((z = (z_1, z_2))\) corresponding to spatial coordinates,

\[h: [-1,5]^8 \rightarrow L^2([-90,90]^2)\]
\[x = (x_1, \ldots, x_8) \mapsto h(x)\],

where \(z = (z_1, z_2) \in [-90,90]^2\), \(x_j \in [-1,5]\) for \(j = 1, \ldots, 8\) and

\[h(x)(z_1, z_2) = x_1 \exp \left[-\frac{(0.8z_1 + 0.2z_2 - 10x_2)^2}{60x_1^2}\right] + (x_2 + x_4) \exp \left[-\frac{(0.5z_1 + 0.5z_2)x_1}{500}\right] + x_5(x_3 - 2) \exp \left[-\frac{(0.4z_1 + 0.6z_2 - 20x_6)^2}{40x_3^2}\right] + (x_6 + x_8) \exp \left[-\frac{(0.3z_1 + 0.7z_2)x_7}{250}\right].\]

Examples of Campbell maps are shown in Figure 5.

As in the forthcoming flooding case, the output space is \(Y = M_{64,64}(\mathbb{R})\). \(f_X\), the density function of \(X\), is the same as the density function of the input variables of the flooding
application. In order to have the same input domain as in the forthcoming flooding example, we impose \( x_8 = -1 \) and we introduce an affine transformation, denoted \( a_i \) from \( \text{supp}(f_X) \) to \([-1, 5] \), so that the input space is \( \mathcal{X} = \text{supp}(f_X) \subset \mathbb{R}^7 \). Details are given in Appendix C. In the end, the true map function is

\[
y: \mathcal{X} \rightarrow \mathcal{Y}
\]

\[
x \mapsto (h_{i_1(x_1), \ldots, a_i(x)_7}(-1)(z))_{z \in \mathcal{Z}},
\]

with \( \mathcal{Z} = \{(-90 + \ell \frac{180}{64}, -90 + \frac{j}{\ell} \cdot 180) \}_{j = 1, \ldots, 64}^{\ell = 1, \ldots, 5} \).

The sampling density, \( g \), is the same as in the flooding case and is described in Section 6.1. The number of prototypes \( \ell = 5 \) is chosen a priori here but its selection is discussed in Section 7.

5.2. Quantizing the Campbell Function

Campbell metamodel errors. As explained is Section 4.2, we gauge the performance of the empirical quantization with the two following metrics.

First, the excess in quantization error due the metamodel, \( \epsilon_{\mu}^M(5) \), is equal to 3.6% when evaluated from \( n_{\text{maps}} = 10^6 \) maps.

Regarding the error on the probability estimators, we evaluate \( \epsilon_{\mu}^M(10^5, \Gamma^{\ell}_r, j) \) for all the cells \( j = 1, \ldots, \ell \) of all the tested quantization \( \Gamma^{\ell}_r \), \( r = 1, \ldots, n_{\Gamma^{\ell}} \). Figure 6(a) illustrates the distribution of these errors. The results are really satisfying: the median and the 75th percentile are below 1%. The tests on the Campbell function thus show that the metamodel is sufficiently precise in terms of both the prototype maps and the computed probability mass of the Voronoi cells.

Importance sampling errors. As stated in Section 4.2, the errors related to the importance sampling scheme are measured through the IS coefficient of variation of the membership probability, \( \epsilon_{\text{IS}}^M(\tilde{n}, \Gamma^{\ell}_r, j) \), and the IS centroid standard deviation, \( \epsilon_{\text{IS}}^M(n_{\text{maps}}, \Gamma^{\ell}_r, j) \), for all choices of prototypes \( r = 1, \ldots, n_{\Gamma^{\ell}} \) and all clusters \( j = 1, \ldots, \ell \).

The distributions of these metrics are presented in Figure 6(b) and (c) for \( \tilde{n} = 10^7 \) and \( n_{\text{maps}} = 10^6 \). \( \epsilon_{\text{IS}}^M(\tilde{n}, \Gamma^{\ell}_r, j) \) is around 1%, which is of the same order of magnitude as the error from the metamodel, and which is sufficiently accurate. \( \epsilon_{\text{IS}}^M(n_{\text{maps}}, \Gamma^{\ell}_r, j) \) is around 0.04, which is also very precise. These results validate the precision of the importance sampling estimators.

Campbell prototype maps. The outputs of the adapted Lloyd’s algorithm are the \( \ell = 5 \) representative maps \( \Gamma^{\ell}_r = \{\gamma_1^r, \ldots, \gamma_5^r\} \) drawn in Figure 7 which were obtained by using the true maps. The probability mass of each Voronoi cell is given below each map with two estimates, one calculated with the true maps and the other with the metamodel maps. The closeness of the results confirms the precision of the metamodel-based approach. The same broad backslash-like pattern runs across the 5 maps, which differ by their \( \ell \) levels. It is explained by the density function \( f_X \) illustrated in Appendix C, that concentrates \( x_2, x_4 \), and \( x_5 \) in specific zones that boost the occurrence of this pattern.

6. Application to Flooding

6.1. Description of the Flooding Test Case

Parameterization. The main application of this article is about the Boucholeurs district located on the French Atlantic Coast (Figure 8(a)) near the city of La Rochelle. This site was hit by the Xynthia storm in February 2010, which caused the inundation of several areas and severe human and economic damage (Bertin et al. 2012).

The main coastal flooding processes correspond to overflow and are simulated with the numerical code MARS (Lazure and Dumas 2008), where adaptations were made by the BRGM to take into account specificities of local water circulation flows around connections (nozzles, spillways, etc. and breaching phenomena). We focus on the interplay between tide and storm surge on the spatial distribution of the maximum water depth after flooding in the urban area. As shown in the Figure 8(b), the tide is simplified and assumed to be represented by a sinusoidal signal, parameterized by the high-tide level \( T \). The surge signal is modeled as a triangular function defined by four parameters: the peak amplitude \( S \), the phase difference between the surge peak and the high tide \( t_0 \), the time duration of the rising part \( t_- \) and the time duration of the falling part \( t_+ \). In total, five variables describe the offshore conditions, \( X = (T, S, t_0, t_-, t_+) \).
Each of these input variables is assumed to be independent and described by an empirically defined probability distribution based on the analysis performed by Rohmer et al. (2018) of tide gauge measurements from Ile d’Aix ¹ (the tide gauge closest to the study site) and La Rochelle-La Pallice. A constant surge level offset (chosen at 65 cm that is, compatible with future sea level rise in this region ²) is added to its empirical distribution to ensure that a sufficient number of flooding events are induced. Appendix C shows the obtained probability distributions.

Regarding the variables representing a possible breach, two characteristics are taken into account. First, the location of the breaches. Ten different locations are selected. Six of them (B1–B6, Figure 8(a)) are related to natural dunes and are selected based on observations made during the Xynthia event (Muller et al. 2016). Four other locations (B7–B10) are related to artificial dykes and are added because they are near vulnerable zones.

The other breach-related variable is the topographic level after failure, parameterized as the fraction \( p \) (erosion rate) of the initial crest level left. The distribution of the breach characteristics (location and erosion rate) depends on the offshore conditions. If the maximum of the sum of the signals (tide and surge) over time is

---

¹ data available on data.shom.fr
² see for example, https://sealevel.nasa.gov/data_tools/17
higher than 70% of the embankment's height, the probability of breaching is 0.5. Vice versa, the probability of breaching is assumed to be $10^{-4}$. Finally, we consider that if there is a breach, then the location and the erosion rate follow a uniform distribution $U(1,10)$ and $U(0,1)$, respectively. This breach model is a compromise between simplicity and realism.

Dataset generation. Our study relies on 1300 simulations of flood maps generated as follows:

- The offshore conditions are sampled as the beginning of the Sobol Sequence in five dimensions (Cheng and Druzdzel 2013).
- 500 of the 1300 simulations are run without breach, and the breach parameters of the 800 others are sampled uniformly according to $U(1,10)$ for a single breach location and $U(0,1)$ for the associated erosion rate.

These 1300 maps form the training database, and each map belongs to the space $\mathcal{Y} = \mathcal{M}_{64,64}(\mathbb{R})$. The inner product $\langle \cdot, \cdot \rangle_\mathcal{Y}$ of Theorem 2.1 is such that all $\lambda_i$'s are set to 1 so that $\forall (y, y') \in \mathcal{Y}^2$, $\langle y, y' \rangle_\mathcal{Y} = \sum_{i=1}^{64} \sum_{j=1}^{64} y_i y'_i$.

**IS sampling density.** The biased density function is $g = g_1 \times \cdots \times g_7$ with $g_i = \frac{1_{I_i}}{\delta_0}$, $i = 1, \ldots, 5$, where $I_i$ is the support of the marginal density function of the $i$th input variable (see, Appendix C), $g_6 = \frac{1}{12} \sum_{k=1}^{64} \delta_k$ and $g_7 = \frac{5}{12} \delta_0 + \left( \frac{3}{12} \right) I_{[0,1]}$. The idea is to choose a single density for all the estimators related to all the Voronoi cells.

The optimal biased density for which the IS estimators have zero variance is theoretically known (L’Ecuyer, Mandjes, and Tuffin 2009) to be a normalized version of $\mathbb{1}_{y(\tilde{x}) \in C_i} f_\tilde{X}(\tilde{x})$ or $y(\tilde{x}) \mathbb{1}_{y(\tilde{x}) \in C_i} f_\tilde{X}(\tilde{x})$. It changes with each cell $j$, affecting a non-zero probability mass only to the cell. It is not practical as it requires the knowledge of the result of the estimation (probability or mean map) for the normalization. The choice of a uniform distribution for the variables related to the offshore conditions is an approximation to these optimal densities that doesn’t depend on the cell. The uniform density is relevant for quantization as it provides maps covering the whole domain irrespective of their occurrence probability.

**Dealing with the concentration of empty maps.** One of the main difference between the Campbell function and the flooding case is a Dirac mass in the output distribution. Indeed, the probability that the map is free of water (0 at every pixel of the matrix) is close to 1. As Gaussian processes are not adapted to predict empty maps, their direct use in the Prototype Maps Algorithm would overestimate the flooding probability. As a counter-measure, a random forest classification step is added (Ayyadevara 2018), creating a hurdle model (McDowell 2003):

- A first step predicts if the water level is below a threshold $t$ (it is a parameter to tune) or not with the classification method.
- If the water level is forecasted to be above $t$, then use the spatial metamodel described in Section 3. Otherwise, the predicted map is empty.

A final post-processing step is to convert all negative water depth values into null values.

Another difference with the Campbell case is the existence of two behaviors with or without breach. Two metamodels are built independently for each of these two situations.

**Initializing Lloyd.** Another implementation aspect is the choice of the initial prototype maps at the beginning of Lloyd’s algorithm (Step 4, Algorithm 2). Our a priori is that the future Voronoi cells will be associated with diverse flooding levels. The Prototype Maps Algorithm is first run with only the 1300 realistic maps instead of the $n_{\text{maps}} = 10^6$ metamodeled maps. But this degraded version of the algorithm, which provides the initial centroids, also needs to be initialized, which is done with the maps associated to the $\ell$ equally spaced quantiles of the training water volumes.
6.2. Quantizing Floodings

Flooding metamodel error. As presented in Section 4.2, to evaluate the impact of the metamodel on the precision of the probability estimators, we compute the relative probability error, $\epsilon_p^{\text{MM}}(1300, \Gamma'_r, j)$. Note that, unlike the analytical Campbell function, the excess in quantization error, $\epsilon_q$, cannot be calculated since the true centroids are not known.

The calculation of the relative probability error is composed of the following steps:

1. The flood predictions, $(\hat{y}(\tilde{X}^0))_{1 \leq j \leq 1300}$, are computed from the $\{y(\tilde{X}^0))_{1 \leq j \leq 1300}$ hydraulic simulations within a 10-fold cross-validation procedure. Each of the fold involves a classification and a metamodeling from 1170 maps followed by a prediction of 130 maps.
2. The probabilities $P_{1300}(\Gamma_r, j, y)$ and $\hat{P}_{1300}(\Gamma_r, j, \hat{y})$, $j \in \{1, \ldots, 5\}$, are computed for the perturbed centroids (generated as explained in Section 4.2) $\Gamma_r \in (\Gamma'_r)_{r \in \{1, \ldots, 5\}}$.
3. The relative probability error $\epsilon_p^{\text{MM}}(1300, \Gamma'_r, j)$ is computed following (6) for $\Gamma_r \in (\Gamma'_r)_{r \in \{1, \ldots, 5\}}$ and $j = 1, \ldots, 5$.

The distribution of $\epsilon_p^{\text{MM}}(1300, \Gamma'_r, j)$, the relative errors, is shown in Figure 9(a) for $n_{\Gamma'_r} = 100$. The median of the error is around 2.5%. It reveals that the FPCA and the Gaussian process together are effective in predicting the Voronoi cell in which a flood map is located, that is, in predicting the type of flooding of a new input vector. This performance is obtained with only two FPCA basis maps that explain 98% of the variance, see Figure 3, which highlights the low intrinsic dimension of the phenomenon.

Flooding importance sampling error. We complete this study by an analysis of the importance sampling error as explained in Sire et al. 2023b. A bootstrap approach provides, for each of the $j = 1, \ldots, \ell$ centroids of each $\Gamma'_r$, the IS coefficient of variation of the membership probability, $\epsilon_p^{\text{IS}}(\tilde{n}, \Gamma'_r, j)$, and the IS centroid standard deviation, $\epsilon_p^{\text{IS}}(n_{\text{maps}}, \Gamma'_r, j)$. The distributions of these metrics are given in Figure 9(c) and (d) for $n = 10^7$ and $n_{\text{maps}} = 10^6$. $\epsilon_p^{\text{IS}}(\tilde{n}, \Gamma'_r, j)$ is lower than 1%, showing that the error on the probability is mainly driven by the error of the metamodel. $\epsilon_p^{\text{IS}}(n_{\text{maps}}, \Gamma'_r, j)$ is less than 0.02 meters for $n_{\text{maps}} = 10^6$ which is quite satisfying in a flooding context.

Flooding prototype maps. The application of the Prototype Maps Algorithm to Les Boucholeurs with $n_{\text{maps}} = 10^6$ gives the 5 representative maps of Figure 10. Each of them represents a type of flooding. The probability masses are computed with $\tilde{n} = 10^7$ to increase precision, which was not possible in the Lloyd's algorithm because of the computational cost of the multiple iterations. The most serious the flooding, the lowest the probability, except for the Voronoi cell number 3 (Figure 10(c)) which is a bit less likely than Voronoi cell number 4 (Figure 10(d)), even if it is associated to more moderate floodings. Although this seems counter-intuitive, it is possible that a specific type of small floodings is less likely than another type of more severe flooding. Among the Voronoi cells associated to the strongest floodings, the prototype number 4 (Figure 10(d)) is a map with shallow water covering a large area, while the flooded area is smaller for cell number 5 (Figure 10(e)) but with a greater water depth near the coastline. These five Voronoi cells can be visualized on the two first FPCA axes that summarize 98% of the inertia of the 1300 flooding maps. Figure D.1 illustrates this (and Appendix D discusses these results more thoroughly), showing that the fourth and fifth cells associated to the most extreme floodings contains very different types of maps.

7. Summary and Perspectives

This work proposes a new visualization method based on quantization adapted to rare events and for which a limited database is available, as it occurs with computationally costly simulations. Without loss of generality, the method, called Prototype Maps Algorithm, is developed around an application to flood maps. The visualization takes the form of a set of prototype maps and their probability masses that best represent the continuous probability distribution on the space of maps. Importance sampling and a metamodel scheme based on Functional PCA combined with Gaussian process regression are used to address the challenges of the scarcity of the (flooding) events and the small dataset size. The obtained prototype maps along with their probabilities are an intuitive representation of the flooding risk while remaining meaningful mathematically. Indeed, the prototype maps are a discrete approximation of the continuous random flooding event. The distance between the discrete and the continuous distribution is the quantization error, which

![Figure 9](image-url) - Boxplots of the distributions of the various errors in the flooding case. The distributions correspond to all $r = 1, \ldots, 100$ and $j = 1, \ldots, 5$. (a) $\epsilon_p^{\text{MM}}(1300, \Gamma'_r, j)$, the relative probability error. (b) $\epsilon_p^{\text{IS}}(10^7, \Gamma'_r, j)$, the IS coefficient of variation of membership probability. (c) $\epsilon_p^{\text{IS}}(10^6, \Gamma'_r, j)$, the IS centroid standard deviation.
Figure 10. Centroids calculated by the Prototype Maps Algorithm with $n_{\text{maps}} = 10^6$ predicted maps, and their probability mass estimated with $n = 10^7$. The frequency, that is, the inverse of the probability is provided for an easier interpretation. The water depths (in meters) are drawn with colors ranging from light to dark blue. Blue contour lines represent a water depth of 15 cm.

Tuning of the algorithm. First, a more sophisticated biased density function of importance sampling could be sought. We have integrated importance sampling inside Lloyd's algorithm to deal with the low flooding probability when estimating the centroid and the probabilities of each cluster. The choice of a uniform density $g$ proposed here for the offshore conditions is a compromise between simplicity and the exploration of the whole domain and was empirically seen to keep the variance of the estimators small. Yet a real optimization of this density has not been performed. While it is a difficult problem with functional unknowns, the search for a better biased density, defined independently for each cell, is a relevant continuation to this work.
Second, the number of prototype maps, $\ell$, is chosen a priori to remain usable in practice while capturing sufficiently diverse floodings. Strategies to identify the number of Voronoi cells could be implemented based, for instance, on the evolution of the quantization error with the number of representatives.

**Complementary approaches.** The method described here characterizes representative floodings. It would now be valuable to know more about the forcing and breaching conditions leading to these scenarios, that is, to study how the inputs are related to each prototype.

An inherent feature of the obtained quantization is that none of the prototype maps is the no-flood map. The centroid of the first Voronoi cell has a negligible water depth (3 × 10^{-5} meters on average), but it is not really empty. However, fortunately, it is known that a flood map is empty is close to 1. The formulation of quantization could then be adapted so as to force the no-flood map (or more generally a Dirac of a continuous probability law) to be one of the representatives.

In this article, the metamodel is the kriging mean. The uncertainties of the Gaussian process were not considered. Another future development could account for this metamodel uncertainty in the computation of the conditional expectations and the probabilities.

**Appendix A. Proof of Theorem 2.1**

In this Appendix, we generalize the proof of optimal quantization of Kieffer and Cuesta-Albertos (Pagès and Yu 2013) to random fields $Y$. Let us consider a probability space $(\Omega, F_0, P)$, a random variable $Y: \Omega_0 \to \mathcal{Y}$ and the optimal quantization of $\ell$ prototype maps $\Gamma^*_\ell = \{Y_1^*, \ldots, Y_n^*\} \in \arg\min \{\alpha(\Gamma_\ell)\}$.

We introduce the inner product of the space $L^2_\gamma(F_0)$: $\langle W, Y \rangle = \mathbb{E} \left[ \langle W, Y \rangle \right]$. Let us show that $\forall j \in \{1, \ldots, \ell\}$, $\mathbb{E} \left[ \langle Y \mid Y \in \mathcal{C}_j^* \rangle \right] = Y^*_j$.

We introduce the inner product of the space $L^2_\gamma(F_0)$: $\langle W, Y \rangle = \mathbb{E} \left[ \langle W, Y \rangle \right]$. We set $F = \mathcal{F}(\Omega, F_0, \Omega_0)$ and the $\sigma$-algebra induced by $q_{\Gamma^*_\ell}(Y)$.

**Proposition A.1.** Let $W \in L^2_\gamma(F_0)$, $\pi_{L^2_\gamma(F)}(W) = \mathbb{E} \left[ \langle W, \mathcal{F} \rangle \right]$, $\pi_{L^2_\gamma(F)}(W)$ the orthogonal projection of $W$ on $L^2_\gamma(F)$ (Durrett 2010, p. 229)

**Proof.** Let $W \in L^2_\gamma(F_0), \forall Z \in L^2_\gamma(F)$,

\[
\mathbb{E} \left[ \langle Z, W \rangle \right] = \mathbb{E} \left[ \mathbb{E} \left[ \langle Z, W \rangle \mid \mathcal{F} \right] \right] = \mathbb{E} \left[ \sum_{i=1}^{n} \lambda_i z_i w_i \mid \mathcal{F} \right] = \mathbb{E} \left[ \sum_{i=1}^{n} \lambda_i \mathbb{E} \left[ z_i w_i \mid \mathcal{F} \right] \right] = \mathbb{E} \left[ \sum_{i=1}^{n} \lambda_i z_i \mathbb{E} \left[ w_i \mid \mathcal{F} \right] \right] \text{ because } \forall i, z_i \in \mathcal{F} = \mathbb{E} \left[ \langle Z, \mathbb{E} \left[ W \mid \mathcal{F} \right] \rangle \right].
\]

We have $q_{\Gamma^*_\ell}(Y) - \mathbb{E} \left[ Y \mid q_{\Gamma^*_\ell}(Y) \right] \in \mathcal{F} = \sigma(q_{\Gamma^*_\ell}(Y))$, so $Y - \mathbb{E} \left[ Y \mid q_{\Gamma^*_\ell}(Y) \right]$ is orthogonal to $q_{\Gamma^*_\ell}(Y) - \mathbb{E} \left[ Y \mid q_{\Gamma^*_\ell}(Y) \right]$ for the inner product of the space $L^2_\gamma(F_0)$ according to Proposition A.1. Then, the Pythagorean theorem gives

\[
\mathbb{E} \left[ \|Y - q_{\Gamma^*_\ell}(Y)\|_2^2 \right] = \mathbb{E} \left[ \|Y - \mathbb{E} \left[ Y \mid q_{\Gamma^*_\ell}(Y) \right]\|_2^2 \right] + \mathbb{E} \left[ \|q_{\Gamma^*_\ell}(Y) - \mathbb{E} \left[ Y \mid q_{\Gamma^*_\ell}(Y) \right]\|_2^2 \right].
\]

We set $\tilde{\Gamma}_\ell = \{\tilde{\gamma}_1, \ldots, \tilde{\gamma}_\ell\}$ such that $\forall j \in \{1, \ldots, \ell\}$, $\tilde{\gamma}_j = \mathbb{E} \left[ Y \mid Y \in \mathcal{C}_j^* \right]$.

**Appendix B. Projection on a Basis of Maps**

**B.1. Functional Principal Components Analysis**

Every flood map $y(x)$ that we want to model can be considered as a matrix indicating the water depth at each pixel, that is, as a function $y_{x_i}: \mathbb{R}^2 \to \mathbb{R}$.

We work with a training database of $n_{\text{train}}$ maps $(y_{x_i}(z))_{i=1, \ldots, n_{\text{train}}}$ and the objective here is to compute the projection of the flood maps in a FPCA basis of $n_{\text{PC}}$ maps. We first consider a functional orthonormal basis $\Phi(z) = (\Phi_1(z), \ldots, \Phi_{n_{\text{PC}}}(z))^\top$, D4 Daubechies wavelets (Daubechies 1988), which are widely used in image processing, are chosen here, as Perrin et al. (2021) shows their accuracy for a similar problem. The procedure consists in the following steps:

1. Find the coordinates $(\alpha(x), i) = 1, \ldots, n_{\text{train}}$ in the $\Phi$ basis: $y_{x_i}(z) = \sum_{k=1}^{K} \alpha_i(x)\Phi_k(z) = \Phi(z)^\top \alpha(x)$

2. Compute $\lambda_k = \mathbb{E} \left[ \alpha_i(x)^2 \right] / \sum_{i=1}^{n_{\text{train}}} \alpha_i(x)^2$, $k = 1, \ldots, K$, order them by decreasing values and select the $K \ll K$ most important vectors of the basis (the $K$ vectors associated with the highest $\lambda_k$) such that $\sum_{k=1}^{K} \lambda_k \geq p$, with $p$ a given proportion of energy. Then, we have $y_{x_i}(z) = \sum_{k=1}^{K} \alpha_i(x)\Phi_k(z) = \Phi(z)^\top \tilde{\alpha}(x)$

3. Apply PCA in $\mathbb{R}^K$ on the dataset of coefficients evaluated at the design points $(\tilde{\alpha}(x))_{i=1, \ldots, n_{\text{train}}}$. Denote $t_1(x^0), \ldots, t_{n_{\text{PC}}}(x^0), i = 1, \ldots, n_{\text{PC}}, x^0$ a set of $n_{\text{PC}}$ training points.

\[
\tilde{t}_i(x) = \sum_{k=1}^{K} \alpha_i(x)\Phi_k(x) = \Phi(x)^\top \tilde{\alpha}(x).
\]
1, ..., ntrain, the coordinates of the first npc principal components and ω1, ..., ωnpc the associated eigenvectors in \( \mathbb{R}^K \). We have
\[
\tilde{\alpha}(x^i) \simeq \sum_{j=1}^{npc} t_j(x^i)w_j = \Omega^T t(x^i),
\]
with \( \Omega \) thenpc \( \times \tilde{K} \) projection matrix.

Finally we can write \( \tilde{y}_x(z) \simeq (\Omega\tilde{\Phi}(z))^T t(x^i) \), with \( \Omega\tilde{\Phi} \) a basis ofnpc maps. \( \tilde{K} \) (or, equivalently, \( p \)) andnpc are tuned by cross-validation.

### B.2. Inverse FPCA

As described in Section 3, Gaussian processes are used to predict \( \hat{t}(x^i) \) for a new input \( x^* \). The following steps provide the map \( \hat{y}(x^*) \) from the coordinates \( \hat{t}(x^*) \):

1. Compute the coefficients of the vector \( \alpha(x^*) \)
   - for \( k = \tilde{K} + 1, \ldots, K \), \( \hat{\alpha}_k(x^*) = \frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} \alpha_k(x^i) \)
   - for \( k = 1, \ldots, \tilde{K} \), \( \hat{\alpha}_k(x^*) = \sum_{j=1}^{npc} t_j(x^*)w_j \)
2. Compute \( \hat{y}_x(z) \) with \( \hat{y}_x(z) = \sum_{k=1}^{K} \hat{\alpha}_k(x^*)\Phi_k(z) \)

The computation of \( \hat{y}_x(z) \) at the targeted family of pixels provides the predicted map \( \hat{y}(x^*) \).

### Appendix C. Probability Distributions

The density functions of the offshore variables are obtained through empirical observations as described in Section 6.1. They are shown in Table C.1, along with their support. The location of the breach is an integer in \([1, \ldots, 10]\), and the erosion rate is in \([0, 1]\).

The affine transformation \( a \) from \( \text{supp}(f_X) \) to \([-1, 5]^7 \), introduced in the Campbell case, is then
\[
a: \text{supp}(f_X) \rightarrow [-1, 5]^7 \quad x \mapsto \begin{bmatrix}
-1 + 6\frac{x-0.52}{3.59-0.52} \\
-1 + 6\frac{x-0.65}{2.5-0.65} \\
-1 + 6\frac{x-8.05}{8.05+8.05} \\
-1 + 6\frac{x+12}{12+12} \\
-1 + 6\frac{x-12}{12-12} \\
-1 + 6\frac{x-\frac{13}{9}}{13-\frac{13}{9}}
\end{bmatrix}.
\]

### Appendix D. Voronoi Cells on the FPCA Axes

In the flooding case, the two first FPCA axes summarize 98% of the inertia of the training 1300 maps. It is relevant to visualize the quantization
Table C.1. (Continued)

| $t_+$ Duration falling part | [0,12.2] | Empirical |
|-----------------------------|---------|-----------|

Figure D.1. (a) Distribution of $10^5$ points sampled with density $g$ in the different Voronoi cells in the two first FPCA axes of maps without breach. (b) 2D histogram where for each bin, the quantity $\log\left(\sum_{k=1}^{n_{bin}} \frac{f(x_k)}{g(x_k)}\right)$ is displayed with the color scale. The prototype maps are represented by the red squares. The limits of cells 2, 3, and 4 are drawn with black lines. Grey pixels are associated to a null weight $\left(\sum_{k=1}^{n_{bin}} \frac{f(x_k)}{g(x_k)}\right)$. Among other things, cells through their two first FPCA coefficients. Among other things, this allows to have a better understanding of their size. Figure D.1(a) shows the distribution of $10^5$ maps sampled with the uniform density $g$ and the prototype maps in the FPCA basis calculated for the maps.
without breach. As expected, cell 1 is almost a Dirac at the empty map whose coefficients are (−27, 8). Moreover, the largest cells are the ones associated to the most extreme floodings. For instance, cell 5 contains very diversified maps in terms of water depth magnitude and spatial distribution. It is an expected property of the optimal Voronoi cells: the cells with high probabilities are small and vice versa. However, this characteristic may not be true in some specific contexts, as shown in Appendix E. Figure D.1(b) shows the probability density of the maps expressed in the 2 first FPCA coefficients. It is a histogram where each bin has a color corresponding to the weight \( \log \left( \sum_{k=1}^{n_{\text{bin}}} \frac{f(X_k)}{g(Y)} \right) \) associated to the IS density ratio of the maps, \( n_{\text{bin}} \) being the number of maps falling within each bin. It is essential to understand the position of the cell 5, which is located near the left limit. This is due to the low compared to the other cells. It also explains the position of the centroid of the cell 5, which is located near the left limit. This is due to the low weight terms at the right of the cell.

### Appendix E. Relation between Cells Size and Probability

As the Lloyd’s algorithm minimizes the quantization error \( e(\Gamma_\ell) = \left[ \mathbb{E} \left[ \left\| Y - d_{\Gamma_\ell}(Y) \right\|^2 \right] \right]^2 \), we can expect that the largest Voronoi cells are associated to a small probability mass. However, it may not be the case with specific random variables \( Y \).

For instance, let consider a random variable \( Y \in [−20, 20] \), with a density function \( d = \frac{1}{10} [−20, −10] + \frac{9}{10} [0, 20] \).

If we perform the Lloyd’s algorithm with two centroids \( \gamma_1 \) and \( \gamma_2 \), it gives \( \gamma_1 = −15 \) and \( \gamma_2 = 10 \) and two Voronoi cells: \( C_1 = [−20, −2.5] \) with probability mass \( p_1 = \frac{1}{10} \) and \( C_2 = [−2.5, 20] \) with probability mass \( p_2 = \frac{9}{10} \).

### Supplementary Materials

**Supplementary material description:** Note describing the git repository (readme.md).

**Codes related to the Campbell2D case:** Git repository containing R notebooks to reproduce all the experiments related to the Campbell2D function that are described in the article. ([https://github.com/charliesire/quantization_Campbell2D.git](https://github.com/charliesire/quantization_Campbell2D.git))

**R package FunQuant:** R package containing functions to perform the Prototype Maps Algorithm and compute the associated performance metrics. ([https://github.com/charliesire/FunQuant.git](https://github.com/charliesire/FunQuant.git))

**Importance sampling performance metrics:** Note providing details about the computation of the importance sampling performance metrics. (in_perf_metrics.pdf).

**Complexity of the Prototype Maps Algorithm:** Note explaining the computation of the complexity of the Prototype Maps Algorithm. (complexity_pma.pdf).

### Acknowledgments

This research was conducted with the support of IRSN and BRGM, through the consortium in Applied Mathematics CIROQUO ([https://doi.org/10.5281/zenodo.6581217](https://doi.org/10.5281/zenodo.6581217)), gathering partners in technological and academia in the development of advanced methods for Computer Experiments.

### Disclosure Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.
References

Ayyadevara, V. K. (2018). *Random Forest*, pp. 105–116, Berkeley, CA: Apress. [1564]

Bertin, X., Bruneau, N., Breilh, J.-F., Fortunato, A. B., and Karpytchev, M. (2012), “Importance of Wave Age and Resonance in Storm Surges: The Case of the Bay of Biscay,” *Ocean Model*, 42, 16–30. [1562]

Beven, K., Lamb, R., Leedal, D., and Hunter, N. (2015), “Communicating Uncertainty in Flood Inundation Mapping: A Case Study,” *International Journal of River Basin Management*, 13, 285–295. [1556]

Bock, H.-H. (2008), “Origins and Extensions of the k-means Algorithm in Cluster Analysis,” *Journal Électronique d’Histoire des Probabilités et de la Statistique [electronic only]*, 4, Article 14. [1556]

Cheng, I., and Druzdzel, M. J. (2013), “Computational Investigation of Low-Discrepancy Sequences in Simulation Algorithms for Bayesian Networks,” *CoRR* abs/1301.3841, pp. 72–91. [1564]

Daubechies, I. (1988), “Orthonormal Bases of Compactly Supported Wavelets,” *Communications on Pure and Applied Mathematics*, 41, 909–996. [1567]

Durrant, R. (2010), *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. [1567]

Gray, R., and Neuhoff, D. (1998), “Quantization,” *IEEE Transactions on Information Theory*, 44, 2325–2383. [1556]

Gu, M., and Shen, W. (2020), “Generalized Probabilistic Principal Component Analysis of Correlated Data,” *Journal of Machine Learning Research*, 21, 1–41. [1560]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd ed.), New York: Springer. [1556]

Lazure, P., and Dumas, F. (2008), “An External-Internal Mode Coupling for a 3D Hydrodynamical Model for Applications at Regional Scale (MARS),” *Advances in Water Resources*, 31, 233–250. [1562]

L’Ecuyer, P., Mandjes, M., and Tuffin, B. (2009), *Rare Event Simulation using Monte Carlo Methods*, pp. 17–38. Chichester: Wiley. [1559,1564]

MacQueen, J. B. (1967), “Some Methods for Classification and Analysis of Multivariate Observations,” in *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), eds. L. M. L. Cam and J. Neyman, pp. 281–297, Berkeley: University of California Press. [1556]

Mahajan, M., Nimbhorkar, P., and Varadarajan, K. (2012), “The Planar k-means Problem is NP-hard,” *Theoretical Computer Science*, 442, 13–21. Special Issue on the Workshop on Algorithms and Computation (WALCOM 2009). [1558]

Marrel, A., Iooss, B., Jullien, M., Laurent, B., and Volkova, E. (2011), “Global Sensitivity Analysis for Models with Spatially Dependent Outputs,” *Environmetrics*, 22, 383–397. [1561]

McDowell, A. (2003), “From the Help Desk: Hurdle Models,” *The Stata Journal*, 3, 178–184. [1564]

Muller, H., Van Rooijen, A., Idier, D., Pedreros, R., and Rohmer, J. (2016), “Assessing Storm Impact on a French Coastal Dune System Using Morphodynamic Modeling,” *Journal of Coastal Research*, 33, 254–272. [1563]

Pagès, G., and Yu, J. (2013), “Pointwise Convergence of the Lloyd Algorithm in Higher Dimension,” working paper or preprint. [1558,1567]

Pagès, G. (2014), “Introduction to Optimal Vector Quantization and its Applications for Numerics,” Technical report, LPMA. 54 pp. [1560,1557,1567]

Palar, P. S., Parussini, L., Bregant, L., Shimoyama, K., and Zuhal, L. R. (2023), “On Kernel Functions for bi-fidelity Gaussian Process Regressions,” *Structural and Multidisciplinary Optimization*, 66, 37. [1560]

Perrin, T., Roustant, O., Rohmer, J., Alata, O., Naulin, J., Idier, D., Pedreros, R., Moncoulon, D., and Tinard, P. (2021), “Functional Principal Component Analysis for Global Sensitivity Analysis of Model with Spatial Output,” *Reliability Engineering and System Safety*, 211, 107222. [1559,1567]

Rohmer, J., Idier, D., Paris, F., Pedreros, F., and Louisier, J. (2018), “Casting light on forcing and breaching scenarios that lead to Marine Inundation: Combining Numerical Simulations with a Random-Forest Classification Approach,” *Environmental Modelling & Software*, 11, 64–80. [1556,1563,1564]

Santner, T., Williams, B., and Notz, W. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer. [1556]

Seipel, S., and Lim, N. J. (2017), "Color Map Design for Visualization in Flood Risk Assessment," *International Journal of Geographical Information Science*, 31, 2286–2309. [1556]

Sire, C., Le Riche, R., Rullière, D., Rohmer, J., Phelpin, L., and Richet, Y. (2023a), "Complexity of the Prototype Maps Algorithm." Available at https://github.com/charliesire/quantization_Campbell2D.git. [1560]

Sire, C., Le Riche, R., Rullière, D., Rohmer, J., Phelpin, L., and Richet, Y. (2023b), "Importance Sampling Performance Metrics." Available at https://github.com/charliesire/quantization_Campbell2D.git. [1561,1565]

Tabandeh, A., Jia, G., and Gardoni, P. (2022), "A Review and Assessment of Importance Sampling Methods for Reliability Analysis," *Structural Safety*, 97, 102216. [1559]

Williams, C. K., and Rasmussen, C. E. (2006), *Gaussian Processes for Machine Learning* (Vol. 2), Cambridge, MA: MIT Press. [1559,1560]