




Article

Bayesian³ Active Learning for the Gaussian Process Emulator Using Information Theory

Sergey Oladyshkin ^{1,*}, Farid Mohammadi ^{2,†}, Ilja Kroeker ^{1,†} and Wolfgang Nowak ^{1,†}

¹ Department of Stochastic Simulation and Safety Research for Hydrosystems, Institute for Modelling Hydraulic and Environmental Systems/SC SimTech, University of Stuttgart, Pfaffenwaldring 5a, 70569 Stuttgart, Germany; ilja.kroeker@iws.uni-stuttgart.de (I.K.); wolfgang.nowak@iws.uni-stuttgart.de (W.N.)

² Department of Hydromechanics and Modelling of Hydrosystems, Institute for Modelling Hydraulic and Environmental Systems/SC SimTech, University of Stuttgart, Pfaffenwaldring 61, 70569 Stuttgart, Germany; farid.mohammadi@iws.uni-stuttgart.de

* Correspondence: sergey.oladyshkin@iws.uni-stuttgart.de; Tel.: +49-711-685-60116

† These authors contributed equally to this work.

Received: 15 July 2020; Accepted: 10 August 2020; Published: 13 August 2020



Abstract: Gaussian process emulators (GPE) are a machine learning approach that replicates computational demanding models using training runs of that model. Constructing such a surrogate is very challenging and, in the context of Bayesian inference, the training runs should be well invested. The current paper offers a fully Bayesian view on GPEs for Bayesian inference accompanied by Bayesian active learning (BAL). We introduce three BAL strategies that adaptively identify training sets for the GPE using information-theoretic arguments. The first strategy relies on Bayesian model evidence that indicates the GPE's quality of matching the measurement data, the second strategy is based on relative entropy that indicates the relative information gain for the GPE, and the third is founded on information entropy that indicates the missing information in the GPE. We illustrate the performance of our three strategies using analytical- and carbon-dioxide benchmarks. The paper shows evidence of convergence against a reference solution and demonstrates quantification of post-calibration uncertainty by comparing the introduced three strategies. We conclude that Bayesian model evidence-based and relative entropy-based strategies outperform the entropy-based strategy because the latter can be misleading during the BAL. The relative entropy-based strategy demonstrates superior performance to the Bayesian model evidence-based strategy.

Keywords: machine learning; active learning; Gaussian process emulator; Bayesian inference; Bayesian model evidence; relative entropy; Kullback–Leibler divergence; information entropy

1. Introduction

The greatest challenge of the scientific modeling workflow is to construct reliable and feasible models that can adequately describe underlying physical concepts and, at the same time, account for uncertainty [1]. Due to the computational complexity of the underlying physical concepts, numerical simulation models are often too expensive for applications tasks related to uncertainty quantification, risk assessment and stochastic model calibration. The great difficulty here is to establish a consistent and feasible framework that can provide appropriate conceptual descriptions and can simultaneously maintain a reliable time frame of simulations. The latter is the primary reason why a vast majority of ongoing research has been focusing on accelerating the forward model using surrogate models, such as response surfaces, emulators, meta-models, reduced-order models, etc. Due to the high computational costs of the original numerical simulation required for training runs of such surrogates, constructing

surrogate models is still challenging. Classical machine learning approaches, such as artificial neural networks, require huge numbers of model evaluations.

A reasonably fast approach to quantify forward uncertainty has been established by Wiener [2], projecting a full-complexity model onto orthogonal polynomial bases over the parameter space. The conventional non-intrusive version of the polynomial chaos expansion [3,4] or its generalization towards data-driven descriptions [5,6] gained popularity during the last few decades because it can offer an efficient reduction of computational costs in uncertainty quantification [7–9]. Advanced extensions towards sparse quadrature [10], sparse integration rules [11–13], or multi-element polynomial chaos approaches [14,15] were applied to complex and computationally demanding applications.

Alternately to global [16] or local polynomial representation [17], other kernels functions have been widely used in applied mathematics [18] and machine learning [19]. Similar to polynomial chaos expansions, Gaussian process emulators (GPE), also known as Kriging for spatial prediction in the Geosciences [20], offer a linear representation through nonlinear kernels using fundamentals of probability theory. For that reason, GPEs are also known as Wiener–Kolmogorov prediction, after Norbert Wiener [2] and Andrey Kolmogorov [21]. GPEs also offer representation via various kernels, and have gained popularity for such machine learning tasks as classification [22] and regression problems [23]. A recent paper [24] compares various surrogate-based approaches using a common benchmark model for forward uncertainty quantification in carbon carbon dioxide storage.

Surrogate representation of the original physical model can be very helpful to accelerate forward modeling and assess prior uncertainty. Most versions of the surrogate methods named above need training runs of the original model to construct the surrogate. However, once additional information is available in the form of measurement data, then a reliable and feasible framework for inverse modeling is indispensable to account for the uncertainty that remains after model calibration. Bayesian inference [21] offers a rigorous stochastic framework for inverse modelling and for assessing the remaining uncertainty in model parameters and predictions [25]. Direct implementation of Bayesian principles for the original physical model is usually not feasible using Monte Carlo (MC) simulations [26] or even Markov chain Monte Carlo (MCMC) approaches [27]. Any advanced technique, such as thermodynamic integration [28], parallel tempering [29], nested sampling [30,31], subset simulation [32,33], or Gaussian mixture importance sampling [34] is still not feasible for applications where the original model is very expensive.

A recent trend toward stochastic calibration based on surrogate models offers iterative improvement of surrogate representations [35–39] within a limited simulation time budget. For example, the link between Bayesian inference and information theory introduced in [40] can help localize adaptively the relevant spots in the parameter space for surrogate training, according to information-theoretic arguments. Such a procedure of active learning should be very informative regarding the available observation data. It has the potential to adaptively improve the surrogate model in those regions of the parameter space that are most important for Bayesian inference, while including relevant information in an iterative manner.

To improve this procedure, the current paper will make use of the information theory [41–44], which is strongly linked to classical probability theory [21], information entropy [42] and cross entropy [41,45]. The latter have been widely used to measure expected uncertainty and information [46,47]. Relative entropy, also known as Kullback–Leibler divergence [43], quantifies the difference between two probability distributions. All aforementioned entropies have been widely used for model selection [48–50], optimal design of experiments [51–54] and machine learning [55–57]. A review on entropy, information theory, information entropy and Bayesian inference can be found in [58].

The key idea of GPEs is based on the assumption, that the model in the yet unexplored regions of the parameter space can be considered as a Gaussian process. Thus, besides application in geoscience, active learning for GPEs is also widely applied in computer sciences [59–61]. A comprehensive introduction into Gaussian processes and GPEs is provided in [62]. The idea of GPE-based surrogates was introduced in the context of Bayesian calibration of computer models [63] and extended for

optimal selection of training points in [64,65] and extended with active learning concepts [13,38] by several authors. Since GPEs are described by their mean and covariance, the choice of a parametric covariance function and estimation of the related hyper-parameters is decisive for constructing the GPE. There are various works available in the literature that are focusing on estimation of the related hyper parameters. Usually, there are several covariance functions known and used in literature, but in particular the squared exponential covariance and the Matérn covariance functions play an important role in geophysical applications [66–68].

GPE surrogates have often been used to replicate computationally demanding models. Employing various learning function for selecting GPE training points helps to assure the accuracy in that procedure (see [69–71]). Conventionally, learning approaches only focus on GPE training for an underlying model without considering measurement data in the context of Bayesian updating of model parameters. Contrary to that, the study [72] makes use of data only (no numerical model involved) and constructs a GPE model that directly represents the underlying phenomena. It employs information entropy to perform an optimal design of experiments for sensor placements and, in doing so, it exploits the multivariate Gaussian distribution of the GPE model. The study [73] also looks at model-based optimal design of experiments for sensor/measurement selection. Specifically, they look at sample placement for contaminant source identification problems, where the contaminant source geometry is covered by model parameters that are to be inferred. Within their optimization, they used a GPE-based MCMC simulation with local GPE refinement as an auxiliary tool. This means that their active learning strategy is made for planning real-world data collection, not for planning model runs during GPE training. Recently, the study [74] constructed a framework for accelerating MCMCs in Bayesian inference of model parameters. They introduced local approximations of these models into the Metropolis–Hastings kernel of MCMC. In such a setting, the greatest challenge is to replicate behaviors of the original numerical model while accounting for the available observation data at acceptable computational costs. Going to that same direction, the work [75] uses entropy as a learning function in the context of Bayesian assimilation of available data, but it approximates the posterior distribution of model parameters as log normal and relies on a multivariate Gaussian approximation of entropy. Similarly, learning functions for GPE training in Bayesian parameter inference can focus on the posterior mean and variance of model parameters obtained via assimilation of available measurement data. Following this route, the paper [76] suggested minimizing the mean-squared (averaged over the parameter posterior distribution) predictive error of the GPE. Alternatively, the study [77] suggested minimizing the integrated posterior variance of the GPE, again involving an average over the parameter posterior distribution. However, the posterior distribution resulting from Bayesian assimilation of measurement data is typically not multivariate Gaussian and, hence, any corresponding assumptions should be avoided whenever possible. Therefore, the current paper avoids such unnecessary assumptions. Instead, it offers a fully Bayesian view that relies on integral quantities such as BME, RE and IE.

Moreover, the accuracy of GPE depends strongly on the selection of training points [13], and GPEs with different training points can be seen as different surrogate models. Therefore, the main challenge of GPE-based surrogates in Bayesian inference consists of selecting training runs to capture the relevant features of the underlying full-complexity physical model, while focusing its accuracy of the regions of a good fit with the available observation data.

The current paper introduces a novel GPE-based machine learning framework to replicate behaviors of a computational demanding physical model using training runs of that model. The suggested framework focuses the training runs optimally for the parameter inference from observation data in a fully Bayesian view. The introduced framework makes use of Bayesian theory on the three different levels: the first construction of GPE from available training runs and identification of hyper parameters usually employs classical Bayesian principles; the second, incorporation of the available observation data, could be rigorously acceded via Bayesian updating; the third, training runs should be well identified using an adaptive Bayesian active learning strategy for GPE construction

that captures the relevant features of the full-complexity model and honor the available observation data. Therefore, the overall framework can be seen as fully Bayesian (i.e., Bayesian³) active learning or, in short, denoted as BAL in the paper. The novelty over the previous studies and our main focus lies in the theirs level where we suggest three novel active learning strategies incorporating the information theory.

The rest of the paper is structured as follows: Section 2 explores the connection between Bayesian inference and information theory for Gaussian process emulators. This section also emphasizes that information entropy and relative entropy for Bayesian parameter inference and for Bayesian active learning can be computed avoiding any assumption or unnecessary multidimensional integration. Section 3 summarises fundamental properties of Gaussian process emulators and offers the three novel Bayesian active learning strategies based on GPEs and information theory for Bayesian parameter inference. Section 4 demonstrates application of the suggested GPE-based BAL strategies using an analytical example and a carbon dioxide benchmark problem. Additionally, Section 4 shows evidence of convergence against a brute-force reference solution for all proposed BAL strategies and demonstrates parameter inference for the proposed active learning strategies.

2. Bayesian Inference with Information Theory for a Gaussian Process Emulator

2.1. Construction and Training of Gaussian Process Emulators

We will consider a full-complexity model \mathcal{M} producing model response $\mathbf{M}(\omega, x, y, z, t)$ that depends on the some multi-dimensional parameter input ω at each physical point in space (x, y, z) and time t . The uncertain modelling parameters ω form a vector of random variables $\omega = \{\omega_1, \dots, \omega_n\}$ from the parameter space Ω , where n is the number of uncertain parameters.

A Gaussian process emulator $\mathbf{S}(\omega, x, y, z, t)$ (i.e., surrogate model) provides an approximation of the full-complexity model $\mathbf{M}(\omega, x, y, z, t)$ over the parameter space Ω and for each point of space (x, y, z) and time t :

$$\mathbf{M}(\omega, x, y, z, t) \approx \mathbf{S}(\omega, x, y, z, t) = \underbrace{\sum_{l=1}^m \beta_l(x, y, z, t) h_l(\omega)}_{\text{trend}} + \underbrace{u(\omega, x, y, z, t)}_{\text{zero-mean GPE}}. \quad (1)$$

Here, $h_l(\omega)$ for $l = 1, \dots, m$ denote trend basis functions over the parameter space Ω , h_1 is a constant and $\beta_l(x, y, z, t)$ are unknown coefficients of the expansion that only depend on space (x, y, z) and time t . The last term $u(\omega, x, y, z, t)$ in Equation (1) indicates the Gaussian process (GP). To introduce this term, we will assume u to be a GP with zero mean $\mathbb{E}(u) = 0$ and covariance $\text{Cov}(\omega, \omega')$ between $u(\omega, x, y, z, t)$ and $u(\omega', x, y, z, t)$ given by:

$$\text{Cov}(\omega, \omega') = \mathbb{E} [u(\omega, x, y, z, t) u(\omega', x, y, z, t)] = k(\omega, \omega'). \quad (2)$$

Therefore, the GP term $u(\omega, x, y, z, t)$ is assumed to be Gaussian distributed $u(\omega, x, y, z, t) \sim \mathcal{N}(0, k(\omega, \omega'))$ according to the covariance kernel function $k(\omega, \omega')$ for each point of space (x, y, z) and time t . It is worth mentioning that there are different choices of the covariance kernel function $k(\cdot, \cdot)$ available. The most common ones are the squared exponential kernel $k_{\text{SE}}(\cdot, \cdot)$ (the same as the Gaussian kernel) and the Matérn kernel $k_{\text{Matérn}, \nu}(\cdot, \cdot)$, which are defined as follows:

$$k_{\text{SE}}(\omega, \omega') := \sigma^2 \exp \left(-\frac{1}{2} \sum_{j=1}^n \frac{(\omega_j - \omega'_j)^2}{\lambda_j^2} \right), \quad (3)$$

$$k_{\text{Matérn}, \nu}(\omega, \omega') := \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \cdot \left(\frac{\sqrt{2\nu} \|\omega - \omega'\|_2}{\lambda} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu} \|\omega - \omega'\|_2}{\lambda} \right).$$

Here, $\lambda_j \equiv \lambda_j(x, y, z, t)$ and $\lambda \equiv \lambda(x, y, z, t)$ represent the length scale of auto-correlation along each component ω_j with $j = 1, \dots, n$, the variance parameter is denoted by $\sigma^2 \equiv \sigma^2(x, y, z, t)$ and the gamma function is denoted by Γ . K_ν is the modified Bessel function of the second kind with the smoothness parameter $\nu \equiv \nu(x, y, z, t)$. For a more comprehensive description of several kernel functions, we refer the reader to [62]. The parameters $\sigma^2, \lambda, \lambda_j$ and ν as well as the trend coefficients β_l define the mean and correlation function, and are called hyper parameters of \mathbf{S} .

Constructing a Gaussian process emulator $\mathbf{S}(\omega, x, y, z, t)$ for the full-complexity physical model $\mathbf{M}(\omega, x, y, z, t)$ in Equation (1) is based on training runs of the full model. Let us denote the training input parameters (training points) by $\omega_T = \{\omega_{T1}, \dots, \omega_{TN_T}\}^T$ and the corresponding model responses by $\mathbf{M}_T = \{\mathbf{M}_{T1}, \dots, \mathbf{M}_{TN_T}\}^T$, where N_T is a value greater than zero representing the number of training points corresponding to the number of full-model evaluations. Thus, the data set $\{(\omega_{Ti}, \mathbf{M}_{Ti}), i = 1, \dots, N_T\}$ is the complete training set for the GPE. Here, and in the following, we drop the coordinates for space (x, y, z) and time t in our notation for the sake of readability.

First, we look at hyperparameter inference. According to the GP, we will assume that each instance of the model response \mathbf{M}_{Ti} for $i = 1, \dots, N_T$ can be modeled in a probabilistic sense as:

$$P(\mathbf{M}_{Ti} | u(\omega_{Ti}), \omega_{Ti}) \sim \mathcal{N}(\mathbf{M}_{Ti} | h(\omega_{Ti})^T \boldsymbol{\beta} + u(\omega), \sigma^2). \quad (4)$$

Introducing vector notation for $H = \{h(\omega_{T1}) \dots h(\omega_{TN_T})\}^T$, $U = \{u(\omega_{T1}, \dots, \omega_{TN_T})\}^T$, we can rewrite the GPE representation of a given parameter set ω in the following form:

$$P(\mathbf{S} | U) \sim \mathcal{N}(H\boldsymbol{\beta} + U, \sigma^2). \quad (5)$$

Furthermore, the joint distribution of the random vector U for a given parameter set ω is given by:

$$P(U | \omega) \sim \mathcal{N}(0, K(\omega, \omega')), \quad (6)$$

where the (co)variance matrix $K(\omega, \omega')$ for parameter sets ω, ω' is defined according to the covariance kernel functions as follows:

$$K(\omega, \omega') = \begin{pmatrix} k(\omega_1, \omega'_1) & \dots & k(\omega_1, \omega'_{N_T}) \\ \vdots & \ddots & \vdots \\ k(\omega_{N_T}, \omega'_1) & \dots & k(\omega_{N_T}, \omega'_{N_T}) \end{pmatrix}. \quad (7)$$

Equations (5)–(7) allow for estimating the GPEs hyper parameters using several GP-based methods and concepts that are available in the literature [13,62,65], such as the maximum likelihood method or more advanced Bayesian principles [62]. The mentioned Bayesian updating that identifies hyper parameters of the GPE representation is well-known in the literature [13,62,65] and will not be addressed in the current paper. Bayesian principles are again employed to train the Gaussian process emulator $\mathbf{S}(\omega, x, y, z, t)$ of the full-complexity physical model $\mathbf{M}(\omega, x, y, z, t)$ based on the available training points $\omega_T = \{\omega_{T1}, \dots, \omega_{TN_T}\}^T$ and the corresponding model responses $\mathbf{M}_T = \{\mathbf{M}_{T1}, \dots, \mathbf{M}_{TN_T}\}^T$ (see e.g., [78]). The training procedure provides the posterior multivariate Gaussian distribution $\mathcal{N}_\omega(\mu_S, \sigma_S)$ with a mean value μ_S and a standard deviation σ_S of $\mathbf{S}(\omega, x, y, z, t)$ for any given parameter set ω from the parameter space Ω . The accuracy of the GPE strongly depends on the number of training points and how they have been selected [13]. The question of how to select the training points properly is extremely relevant in general. When is even more challenging, but observation data should be incorporated into the full-complexity model via Bayesian inference of model parameters ω . Moreover, GPE with different training points can be seen as and indeed are different models. Therefore, the current paper will introduce a fully Bayesian view (Bayesian inference in Section 2.2 and Bayesian active learning in Section 3) on the construction of the GPE-based surrogate $\mathbf{S}(\omega, x, y, z, t)$ that must capture the main features of the full-complexity model $\mathbf{M}(\omega, x, y, z, t)$ and will be used to assist in Bayesian parameter inference.

2.2. Bayesian Updating on Observation Data Using GPE

Bayesian theory offers a statistically rigorous approach to deal with uncertainty during inference, providing probabilistic information on the remaining uncertainty in parameters and predictions while incorporating the available observation data \mathbf{D} (vector $N_D \times 1$ with N_D length of the observation data set) that is usually attributed at specific point in space (x, y, z) and time t . In the Bayesian framework, initial knowledge on modelling parameters ω is encoded in a prior probability distribution $p(\omega)$. After Bayesian parameter inference, one obtains a posterior probability distribution of the parameters $p(\omega|\mathbf{D})$, which is more informative than the prior distribution. Posterior probability distribution of the parameters $p(\mathbf{S}|\mathbf{D})$ could be obtained with the help of the full-complexity model \mathbf{M} (i.e., $p(\omega|\mathbf{D}, \mathbf{M})$) or with the help of the surrogate model \mathbf{S} (i.e. $p(\omega|\mathbf{D}, \mathbf{S})$) according to the approximation in Equation (1). Due to the high computational demand of the original full complexity model, we will employ the last option in the current paper, i.e., $p(\omega|\mathbf{D}) \equiv p(\omega|\mathbf{D}, \mathbf{S}) \approx p(\omega|\mathbf{D}, \mathbf{M})$ (see more details in [79]).

Formally, the posterior parameter distribution $p(\omega|\mathbf{D})$ of n uncertain parameters forming the vector of random variables $\omega = \{\omega_1, \dots, \omega_n\}$ is obtained by updating the prior parameter distribution $p(\omega)$ in the light of observed data \mathbf{D} according to Bayes' Theorem [21]:

$$p(\omega|\mathbf{D}) = \frac{p(\mathbf{D}|\mathbf{S})p(\omega)}{p(\mathbf{D})}, \quad (8)$$

where the term $p(\mathbf{D}|\omega)$ is the likelihood function that quantifies how well the surrogate model's predictions $\mathbf{S}(\omega, x, y, z, t)$ match the observed data \mathbf{D} (the full notation corresponding to $p(\mathbf{D}|\omega, \mathbf{S})$ will be avoided in the paper). The term $p(\mathbf{D})$ (i.e., $p(\mathbf{D}, \mathbf{S})$ in full notation) is called Bayesian model evidence (BME) and can be seen as a normalizing constant for the posterior distribution of the parameters ω .

In order to describe how well the GPE predictions $\mathbf{S}(\omega, x, y, z, t)$ in physical space $\{x, y, z, t\}$ match the observed data \mathbf{D} , we use the following likelihood function $p(\mathbf{D}|\omega)$ assuming independent and Gaussian distributed measurement errors:

$$p(\mathbf{D}|\omega) = (2\pi)^{-N_D/2} |\mathbf{R}|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} (\mathbf{D} - \mathbf{S}(\omega, x, y, z, t))^T \mathbf{R}^{-1} (\mathbf{D} - \mathbf{S}(\omega, x, y, z, t)) \right], \quad (9)$$

where \mathbf{R} ($N_D \times N_D$) is the diagonal (co)variance matrix of measurement error ϵ .

The performance of Bayesian updating on available observation data using GPE surrogate can be assessed by employing BME, relative entropy and information entropy [80] that are introduced in the upcoming Sections 2.3–2.5.

2.3. Bayesian Model Evidence

The BME value $p(\mathbf{D})$ in the denominator of Equation (8) indicates the quality of the model against the available data and can be obtained by integrating the Equation (8) over the parameter space Ω as:

$$\text{BME} \equiv p(\mathbf{D}) = \int_{\Omega} p(\mathbf{D}|\omega)p(\omega)d\omega, \quad (10)$$

or

$$\text{BME} = \mathbb{E}_{p(\omega)} (p(\mathbf{D}|\omega)). \quad (11)$$

Therefore, BME $p(\mathbf{D})$ can be directly estimated [81] from Equation (11) using Monte Carlo sampling techniques [82] on the GPE:

$$\mathbb{E}_{p(\omega)} (p(\mathbf{D}|\omega)) \approx \frac{1}{N} \sum_{i=1}^N (p(\mathbf{D}|\omega_i)), \quad (12)$$

where N is the size of Monte Carlo sample.

We remark that the GPE-based surrogate model $\mathbf{S}(\omega, x, y, z, t)$ contains approximation errors because the GPE is merely a surrogate model of the full-complexity model $\mathbf{M}(\omega, x, y, z, t)$. Therefore,

GPE-based BME value $p(\mathbf{D}, \mathbf{S})$ in Equation (11) is an approximation of full-complexity model BME value $p(\mathbf{D}, \mathbf{M})$, i.e., $\text{BME} = \mathbb{E}_{p(\omega)} (p(\mathbf{D}|\omega), \mathbf{S}) \approx p(\omega|\mathbf{D}, \mathbf{M})$. A correction factor for BME value can be incorporated similar to the one in [79].

2.4. Relative Entropy

Relative entropy (RE), also called Kullback–Leibler divergence, measures the difference between two probability distributions [43] in the Bayesian context. Relative entropy $D_{\text{KL}} [p(\omega|\mathbf{D}), p(\omega)]$ measures the so-called information geometry in moving from the prior $p(\omega)$ to the posterior $p(\omega|\mathbf{D})$, or the information lost when $p(\omega)$ is used to approximate $p(\omega|\mathbf{D})$:

$$D_{\text{KL}} [p(\omega|\mathbf{D}), p(\omega)] = \int_{\Omega} \ln \left[\frac{p(\omega|\mathbf{D})}{p(\omega)} \right] p(\omega|\mathbf{D}) d\omega, \quad (13)$$

Estimating the relative entropy in Equation (13) usually requires a multidimensional integration that is often infeasible for most applied problems. However, employing Equation (13) and definition (5) from the recent findings in the paper [40], we avoid this multidimensional integration:

$$D_{\text{KL}} [p(\omega|\mathbf{D}), p(\omega)] = -\ln \text{BME} + \int_{\Omega} \ln [p(\mathbf{D}|\omega)] p(\omega|\mathbf{D}) d\omega. \quad (14)$$

Therefore, relative entropy $D_{\text{KL}} [p(\omega|\mathbf{D}), p(\omega)]$ can be directly estimated from Equation (14) using Monte Carlo sampling techniques on the GPE:

$$D_{\text{KL}} [p(\omega|\mathbf{D}), p(\omega)] = -\ln \text{BME} + \mathbb{E}_{p(\omega|\mathbf{D})} (\ln [p(\mathbf{D}|\omega)]). \quad (15)$$

The expression for relative entropy in Equation (15) employs the prior-based estimation of BME values in Equation (11) and a posterior-based expectation of the likelihood $\mathbb{E}_{p(\omega|\mathbf{D})} (\ln [p(\mathbf{D}|\omega)])$ that could be obtained, e.g., using a rejection sampling technique or similar [26] as:

$$\mathbb{E}_{p(\omega|\mathbf{D})} (\ln [p(\mathbf{D}|\omega)]) \approx \frac{1}{N_p} \sum_{i=1}^{N_p} (\ln [p(\mathbf{D}|\omega_i)]), \quad (16)$$

where N_p is the size of posterior sample according to rejection sampling.

2.5. Information Entropy

Information entropy (IE) is a measure of the expected missing information and can also be seen as uncertainty of a random variable ω . According to Shannon [42], the information entropy $H [p(\omega|\mathbf{D})]$ for a random variable ω with (posterior) parameter distribution $p(\omega|\mathbf{D})$ is defined as:

$$H [p(\omega|\mathbf{D})] = - \int_{\Omega} \ln [p(\omega|\mathbf{D})] p(\omega|\mathbf{D}) d\omega. \quad (17)$$

However, information entropy $H [p(\omega|\mathbf{D})]$ in Equation (17) can not be computed directly from a posterior sample because the posterior density values $p(\omega|\mathbf{D})$ are unknown. To overcome this situation, we will employ the definition of $D_{\text{KL}} [p(\omega|\mathbf{D}), p(\omega)]$ in Equation (13) to express the information entropy as:

$$H [p(\omega|\mathbf{D})] = - \int_{\Omega} \ln [p(\omega)] p(\omega|\mathbf{D}) d\omega - D_{\text{KL}} [p(\omega|\mathbf{D}), p(\omega)]. \quad (18)$$

Therefore, employing Equation (15), information entropy can be directly estimated according to Equation (A3) in the paper [40] using Monte Carlo sampling techniques on the GPE:

$$H [p(\omega|\mathbf{D})] = \ln \text{BME} - \mathbb{E}_{p(\omega|\mathbf{D})} (\ln [p(\omega)]) - \mathbb{E}_{p(\omega|\mathbf{D})} (\ln [p(\mathbf{D}|\omega)]). \quad (19)$$

Equation (19) does not contain any assumptions and avoids all multidimensional density estimations and integrals in Equation (17). It employs the prior-based estimation of BME values in Equation (10) and a posterior-based expectation of prior densities $\mathbb{E}_{p(\omega|\mathbf{D})}(\ln[p(\omega)])$ and likelihoods $\mathbb{E}_{p(\omega|\mathbf{D})}(\ln[p(\mathbf{D}|\omega)])$. The posterior-based expectation of prior log-densities could be obtained as well using rejecting sampling techniques [26]:

$$\mathbb{E}_{p(\omega|\mathbf{D})}(\ln[p(\omega)]) \approx \frac{1}{N_p} \sum_{i=1}^{N_p} (\ln[p(\omega_i)]). \quad (20)$$

3. Bayesian Active Learning for Gaussian Process Emulators in Parameter Inference

Section 2 described the standard construction of GPE $\mathbf{S}(\omega, x, y, z, t)$ based on available training runs of the physical model $\mathbf{M}_T = \{\mathbf{M}_{T1}, \dots, \mathbf{M}_{TN_T}\}^T$. As we want to infer the model parameters of the underlying physical model assisted by the constructed GPE surrogate, the training runs of the full model must ensure appropriate convergence. Specifically, such the GPE surrogate $\mathbf{S}(\omega, x, y, z, t)$ must captures the main global features of the full-complexity model $\mathbf{M}(\omega, x, y, z, t)$ and, at the same time, have local accuracy in the region of high posterior density $p(\omega|\mathbf{D})$ that will emanate during Bayesian inference. However, these regions for local accuracy are unknown a priori. Therefore, the current Section 3 focuses on iterative selection of training points. It employs the link between Bayesian inference and information theory [40] similar to Section 2 in order to perform Bayesian active learning (BAL). The later will iteratively select new training point as the regions with required local accuracy becomes progressively clear during the Bayesian updating of a Gaussian process emulator described in Section 2.

We will consider that the GPE surrogate $\mathbf{S}(\omega, x, y, z, t)$ in Equation (1) has been constructed based on at least one training point ($N_T \geq 1$) in the parameter space $\omega_T = \{\omega_{T1}, \dots, \omega_{TN_T}\}^T$ using the corresponding model responses $\mathbf{M}_T = \{\mathbf{M}_{T1}, \dots, \mathbf{M}_{TN_T}\}^T$. The goal of the current Section 3 is to identify the next training point ω_T^{BAL} that should be incorporated into the GPE surrogate $\mathbf{S}(\omega, x, y, z, t)$. To do so, we will introduce three Bayesian active learning strategies that are based on Bayesian model evidence (Section 3.2), relative entropy (Section 3.3), and information entropy (Section 3.4). These strategies avoid unnecessary approximations and assumptions (such as maximum likelihood estimation, multivariate Gaussian posterior, etc.). Once a new training point ω_T^{BAL} has been identified, the model should be evaluated in that new point and the GPE in Equation (1) should be updated with typical GPE-inherent methods. In that way, the presented GPE-based fully Bayesian approach could help to calibrate the physical model to the available measurement data at the reduced computational costs. However, the GPE representation could never be better than the underlying physical model.

3.1. Bayesian Inference of Gaussian Process Emulator Incorporating Observation Data

The GPE is a collection of random functions over the parameter space Ω , i.e., a random model response $\mathbf{S}(\omega, x, y, z, t)$ for each point of space (x, y, z) and time t . The Bayesian identification of hyper parameters during the training on the available model runs $\mathbf{M}_T(\omega, x, y, z, t)$ in Section 2.1 provides the multivariate Gaussian distribution $\mathcal{N}_\omega(\mu_S, \sigma_S)$ of the model response $\mathbf{S}(\omega, x, y, z, t)$ forming response space \mathcal{S} for any given parameter set ω from the parameter space Ω . Here, μ_S is a mean value and σ_S is a standard deviation of model response $\mathbf{S}(\omega, x, y, z, t)$ at each point of space (x, y, z) and time t . Therefore, we can explore the parameter space Ω using the exploration parameter set ω_E and we can assess how the obtained multivariate Gaussian distribution $\mathcal{N}_{\omega_E}(\mu_S, \sigma_S)$ meet the observation data \mathbf{D} . According to the Bayesian framework [21], we can obtain a posterior probability distribution $p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})$ of the model response for the given parameter set ω_E , incorporating the observed data \mathbf{D} :

$$p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}) = \frac{p_{\omega_E}^{BAL}(\mathbf{D}|\mathbf{S})\mathcal{N}_{\omega_E}(\mu_S, \sigma_S)}{p_{\omega_E}^{BAL}(\mathbf{D})}, \quad (21)$$

where the term $p_{\omega_E}^{BAL}(\mathbf{D}|\mathbf{S})$ is the likelihood function that quantifies how well the GPE predictions $\mathbf{S}(\omega_E, x, y, z, t)$ drawn from the multivariate Gaussian $\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})$ match the observed data \mathbf{D} and the term $p_{\omega_E}^{BAL}(\mathbf{D})$ is BME value of GPE for the given parameter set ω_E .

Assuming independent and Gaussian distributed measurement errors, the likelihood function $p_{\omega_E}^{BAL}(\mathbf{D}|\mathbf{S})$ can be written as:

$$p_{\omega_E}^{BAL}(\mathbf{D}|\mathbf{S}) = (2\pi)^{-N_D/2} |\mathbf{R}|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} (\mathbf{D} - \mathbf{S}(\omega_E, x, y, z, t))^T \mathbf{R}^{-1} (\mathbf{D} - \mathbf{S}(\omega_E, x, y, z, t)) \right], \quad (22)$$

where $\mathbf{S}(\omega_E, x, y, z, t) \sim \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})$.

3.2. Model Evidence-Based Bayesian Active Learning

As already mentioned in Section 2.2, BME is often used for model selection in order to identify the most suitable model among a set of competing models or to rank the competing models. During the active learning procedure, one has to identify the best “model” in the sense of the next trained version of GPE, i.e., the best position d of sampling point ω_E . This point ω_E can be chosen such that it provides the highest BME of the next trained GPE. Therefore, BME value $p_{\omega_E}^{BAL}(\mathbf{D})$ for each point ω_E in the prior parameter space providing models responses $\mathbf{S}(\omega, x, y, z, t)$ that forms a response space Y can be obtained using the following equation:

$$\text{BME}^{BAL} \equiv p_{\omega_E}^{BAL}(\mathbf{D}) = \int_{\mathbf{S}} p_{\omega_E}^{BAL}(\mathbf{D}|\mathbf{S}) \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) d\mathbf{S}. \quad (23)$$

Equation (23) shows that BME^{BAL} is equal to the expected value $\mathbb{E}_{\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})}$ of the likelihood $p_{\omega_E}^{BAL}(\mathbf{D}|\mathbf{S})$ over the prior $\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})$ that GPE provides after training:

$$\text{BME}^{BAL} = \mathbb{E}_{\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})} \left(p_{\omega_E}^{BAL}(\mathbf{D}|\mathbf{S}) \right). \quad (24)$$

The value BME^{BAL} can be directly estimated from Equation (24) using Monte Carlo sampling techniques [82] on the GPE:

$$\mathbb{E}_{\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})} \left(p_{\omega_E}^{BAL}(\mathbf{D}|\mathbf{S}) \right) \approx \frac{1}{N^{BAL}} \sum_{i=1}^{N^{BAL}} \left(p_{\omega_{E_i}}^{BAL}(\mathbf{D}|\mathbf{S}) \right), \quad (25)$$

where N^{BAL} is the size of Monte Carlo sample for Bayesian active learning.

Therefore, by formal maximization of the model evidence BME^{BAL} , one can find the next training point ω_T^{BAL} from the parameter space Ω :

$$\omega_T^{BAL} = \arg \max_{\omega_E \in \Omega} \text{BME}^{BAL}(\omega_E). \quad (26)$$

3.3. Relative Entropy-Based Bayesian Active Learning

Relative entropy is usually employed for Bayesian experimental design [51] to maximize the expected (marginalized) utility [53]. In the current paper, we will introduce the relative entropy $D_{\text{KL}}^{BAL} [p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})]$ to assess the information geometry in moving the GPE from the multivariate Gaussian prior $\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})$ to its posterior $p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})$ during the active learning procedure. Formally, the relative entropy $D_{\text{KL}}^{BAL} [p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})]$ can be defined for each sampling point ω_E from the parameter space Ω as following:

$$D_{\text{KL}}^{BAL} [p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})] = \int_{\mathbf{S}} \ln \left[\frac{p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})}{\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})} \right] p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}) d\mathbf{S}. \quad (27)$$

Similar to Section 2.4, we can avoid multidimensional integration in Equation (27) using Equation (13) and definition (5) from the paper [40]:

$$D_{KL}^{BAL} \left[p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] = -\ln \text{BME}^{BAL} + \mathbb{E}_{p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})} \left(\ln \left[p_{\omega_E}^{BAL}(\mathbf{D}|\mathbf{S}) \right] \right). \quad (28)$$

The posterior-based expectation of the log-likelihood could be obtained using a rejection sampling technique [26] on the GPE:

$$\mathbb{E}_{p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})} \left(\ln \left[p_{\omega_E}^{BAL}(\mathbf{D}|\mathbf{S}) \right] \right) \approx \frac{1}{N_p^{BAL}} \sum_{i=1}^{N_p^{BAL}} \left(\ln \left[p_{\omega_{E_i}^{BAL}}(\mathbf{D}|\mathbf{S}) \right] \right), \quad (29)$$

where N_p^{BAL} is the size of the posterior sample according to rejection sampling for Bayesian active learning.

Therefore, during the active learning procedure, we will identify the sampling point ω_T^{BAL} from the parameter space Ω that corresponds to maximum relative entropy $D_{KL}^{BAL} \left[p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right]$:

$$\omega_T^{BAL} = \arg \max_{\omega_E \in \Omega} D_{KL}^{BAL} \left[p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right]. \quad (30)$$

It is evident that the optimization problem for RE value in Equation (30) relies not only on BME^{BAL} values from Equation (24). It also relies on the cross entropy represented by the term $\mathbb{E}_{p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})} \left(\ln \left[p_{\omega_E}^{BAL}(\mathbf{D}|\mathbf{S}) \right] \right)$ that reflects how likelihood informative for the posterior (see details in [40]). The last term could be obtained using a rejection sampling technique using the GPE evaluations.

3.4. Information Entropy-Based Bayesian Active Learning Criterion

Minimizing the expected information loss in terms of information entropy [42] has been suggested to identify the best fitting model [83] and is often used in machine learning. Again seeing the GPE with different training points as different models, we will introduce the information entropy $H^{BAL} \left[p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}) \right]$ to assess information loss for each parameter set ω_E :

$$H^{BAL} \left[p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}) \right] = - \int_{\mathcal{S}} \ln \left[p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}) \right] p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}) d\mathbf{S}. \quad (31)$$

Similar to Section 2.5, using Equation (A3) from the paper [40], the information entropy in Equation (31) can be written as follows:

$$H^{BAL} \left[p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}) \right] = \ln \text{BME}^{BAL} - \mathbb{E}_{p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})} \left(\ln \left[\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \right) - \mathbb{E}_{p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})} \left(\ln \left[p_{\omega_E}^{BAL}(\mathbf{D}|\mathbf{S}) \right] \right), \quad (32)$$

where the posterior-based expectation $\mathbb{E}_{p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})} \left(\ln \left[\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \right)$ could be obtained as well using a rejection sampling technique [26] on the GPE:

$$\mathbb{E}_{p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})} \left(\ln \left[\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \right) \approx \frac{1}{N_p^{BAL}} \sum_{i=1}^{N_p^{BAL}} \left(\ln \left[\mathcal{N}_{\omega_{E_i}}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \right), \quad (33)$$

All terms in Equation (32) could be obtained directly avoiding any multidimensional integration using prior-based or posterior-bases sampling from the GPE, such as rejecting sampling techniques. Therefore, to perform active learning, we will rely on the parameter set ω_T^{BAL} that corresponds to the minimum of information entropy $H^{BAL} \left[p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}) \right]$:

$$\omega_T^{BAL} = \arg \min_{\omega_E \in \Omega} H^{BAL} \left[p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D}) \right]. \quad (34)$$

Equation (34) that minimizes the IE value in Equation (32). It makes use of BME^{BAL} in Equation (24) and cross entropy $\mathbb{E}_{p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})} \left(\ln \left[p_{\omega_E}^{BAL}(\mathbf{D}|\mathbf{S}) \right] \right)$ similar to Equation (28). Moreover, Equation (32) shows

that the information entropy $H^{BAL} [p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})]$ relies on cross entropy represented and a posterior-based expectation of multivariate Gaussian prior $\mathbb{E}_{p_{\omega_E}^{BAL}(\mathbf{s}|\mathbf{D})} (\ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{s}}, \sigma_{\mathbf{s}})])$.

4. Application of GPE-Based Bayesian Active Learning

In the previous section, we have introduced three strategies for Bayesian³ active learning during the GPE-assisted Bayesian updating of model parameters as described in Section 2. The current section will make use of an analytical example in Section 4.1 and a carbon dioxide benchmark problem in Section 4.2 to illustrate the suggested active learning strategies from Section 3.

In the present work, we use the Matlab *fitrgp* function [78] to obtain the values of GPE parameters and hyper parameters introduced in Section 2.1 via a traditional Bayesian training on the available model runs. The proposed Bayesian active learning strategies in Section 3 for Bayesian inference in Section 2 have been implemented as an extension of the existing Matlab *fitrgp* function. The fully Bayesian³ active learning extension of *fitrgp* function is available online for the reader through Matlab file exchange [84]. For the sake of consistency, in the current publication, we have used the *fitrgp* function together with the squared exponential kernel $k_{SE}(\cdot, \cdot)$ as defined in Equation (3) for all examples. However, various kernel functions could be easily selected within Matlab *fitrgp* function using various training options. Therefore, the reader is invited to test the suggested Bayesian active learning strategies for own needs exploring the full range of Matlab *fitrgp* functionality.

4.1. Bayesian Active Learning for an Analytical Test Case

4.1.1. Scenario Set up

We will consider a test case scenario in the form of a nonlinear analytical function $\mathbf{M}(\omega, t)$ of ten ($n = 10$) uncertain parameters $\omega = \{\omega_1, \dots, \omega_n\}$ from the paper [40]:

$$\mathbf{M}(\omega, t) = (\omega_1^2 + \omega_2 - 1)^2 + \omega_1^2 + 0.1\omega_1 \exp(\omega_2) - 2\omega_1 \sqrt{0.5t} + 1 + \sum_{i=3}^n \frac{\omega_i^3}{i}. \quad (35)$$

The uncertain parameters ω in Equation (35) are considered to be independent and uniformly distributed with $\omega_i \sim \mathcal{U}(-5, 5)$ for $i = \overline{1, 10}$. The prior assumptions on the parameters will be updated using synthetic observation data $\mathbf{D} = \mathbf{M}(\omega, t_k)$ with $t_k = (k - 1)/9$ and $k = \overline{1, 10}$ that correspond to the parameter set $\omega_i = 0 \forall i$. The standard deviation of the measurement error is considered to be $\sigma_{\mathbf{D}} = 2$.

4.1.2. Likelihood Reconstruction during Bayesian Active Learning

We will construct the Gaussian process emulator $\mathbf{S}(\omega, t)$ in Equation (1) for the test case problem in Equation (35) to approximate the full model $\mathbf{M}(\omega, t)$ in the parameter space ω for each point of time t . We will start the Bayesian active learning with one training point only ($N_T = 1$). The starting training point corresponds to the mean value of the uncertain parameters ω , i.e., $\omega_T = \mathbb{E}_{p(\omega)}(\omega)$. We will perform the Bayesian updating in Equation (8) using Monte Carlo sampling [26] on the constructed GPE surrogate $\mathbf{S}(\omega, t)$ with sample size $N = 10^5$ (alternative approaches can be used similarly). In order to identify the next training points for the GPE iteration, we employ the three Bayesian active learning strategies introduced in Section 3: the model evidence-based strategy in Equation (26), the relative entropy-based strategy in Equation (30) and the information entropy-based strategy in Equation (34).

Let us illustrate how the GPE-based likelihood function updates during the Bayesian active learning procedure. Additionally, we will assess the corresponding computational costs in terms of number of full model runs. For illustrative purposes, we will reduce the 10D problem (35) to a 2D problem with only two parameters, i.e., $\omega_i = 0$ for $i = \overline{3, 10}$. Figures 1–3 show how the GPE's likelihood function cover the 2D parameter space during active learning based on a BME-based

strategy, RE-based strategy and IE-based strategy, respectively. Moreover, Figures 1–3 show Monte Carlo reference solutions that have been obtained directly using Monte Carlo sampling on the original model, introduced in Section 4.1.1.

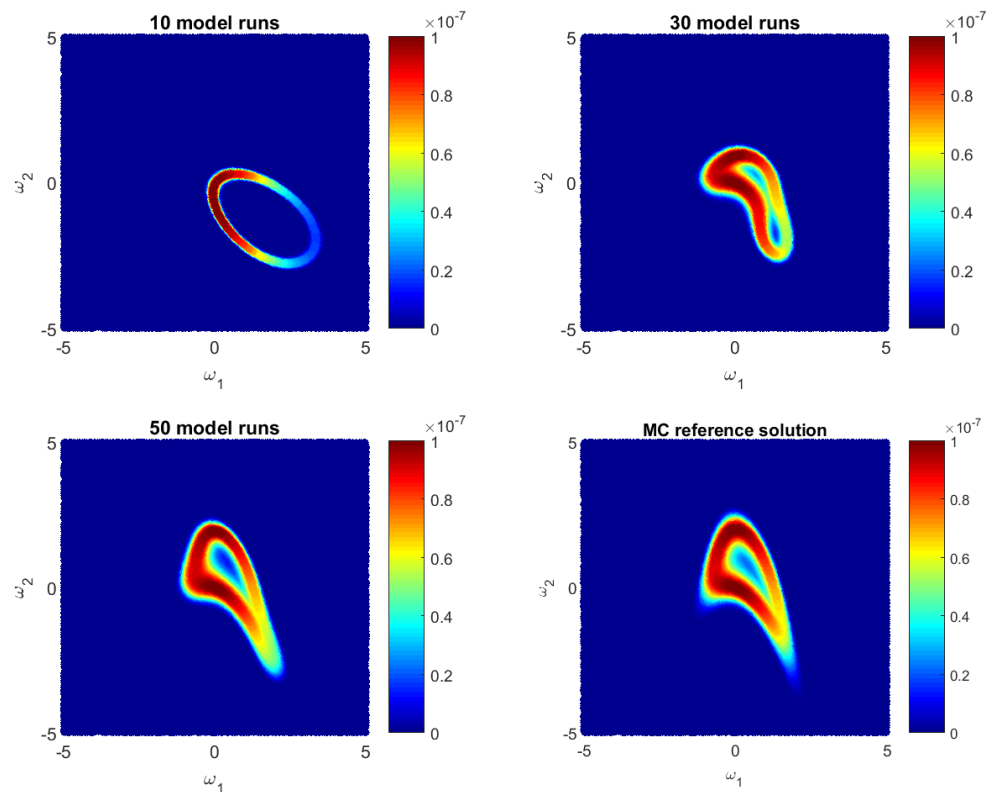


Figure 1. Likelihood values during Bayesian **BME-based** active learning as approximate by the Gaussian process emulator and by a reference Monte Carlo solution for a 2D reduction of the 10D problem.

The RE-based Bayesian active learning captures the non-Gaussian aspects of the analysed problem in a remarkably effective manner in comparison to the BME-based and the IE-based approaches. The RE-based active learning provides a likelihood estimation that is practically identical to the MC reference solution after 25 model runs. The BME-based strategy captures the main features only in the beginning and requires a longer learning procedure to reflect details of the reference solution. The reason is that the RE relies on both BME and cross entropy that indicates how informative the likelihood for the posterior are (see Equation (28)). Contrary to the BME-based and the RE-based strategies, the IE-based active learning manages to cover only partially the likelihood function in the 2D parameter space for the given computational budget. Additionally, it shows a stagnation during the learning procedure, where 50-model-run training shows no significant improvement in comparison to 30-model-run training (see Figure 3). The difference between the RE and the IE-based active learning strategies consists in the second term in Equation (32). That term denotes the cross entropy and reflects how informative the trained multivariate Gaussian distribution $\mathcal{N}_{\omega_E}(\mu_S, \sigma_S)$ of GPE is, for the posterior $p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})$. Formally, it can be seen as a posterior-based expectation of multivariate Gaussian distribution $\mathbb{E}_{p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})}(\ln[\mathcal{N}_{\omega_E}(\mu_S, \sigma_S)])$ and it can overcome the RE value in Equation (32). Apparently, once the trained distribution $\mathcal{N}_{\omega_E}(\mu_S, \sigma_S)$ is extremely informative, then the IE-based active learning suggests to add new training point where $\mathcal{N}_{\omega_E}(\mu_S, \sigma_S)$ is already very similar to the posterior $p_{\omega_E}^{BAL}(\mathbf{S}|\mathbf{D})$. Therefore, the last property could lead to a stagnation of the information entropy-based active learning, similar to Figure 3.

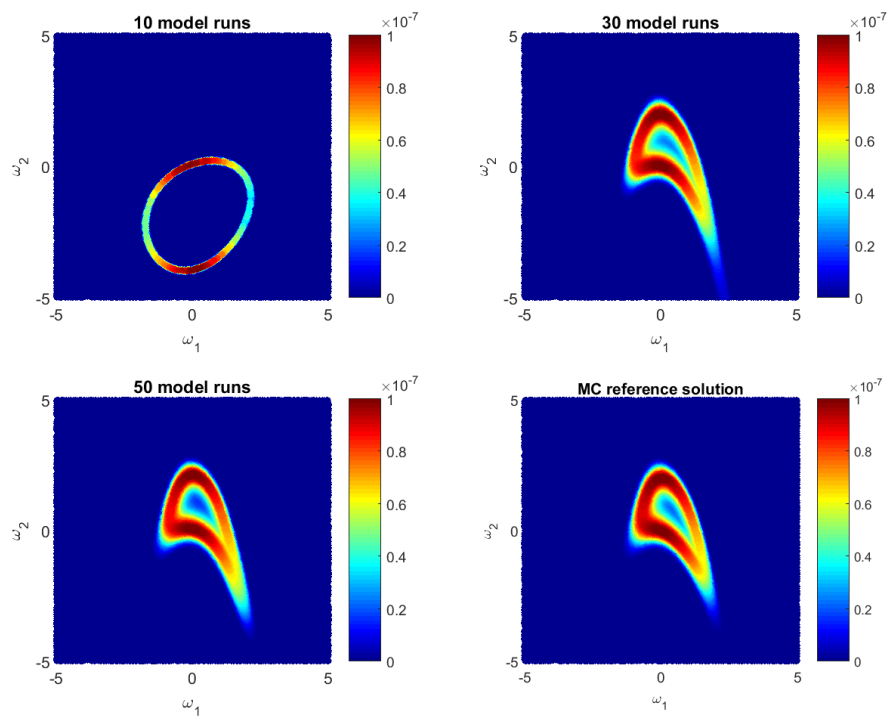


Figure 2. Likelihood values during Bayesian **Relative Entropy-based** active learning as approximate by the Gaussian process emulator and by a reference Monte Carlo solution for a 2D reduction of the 10D problem.

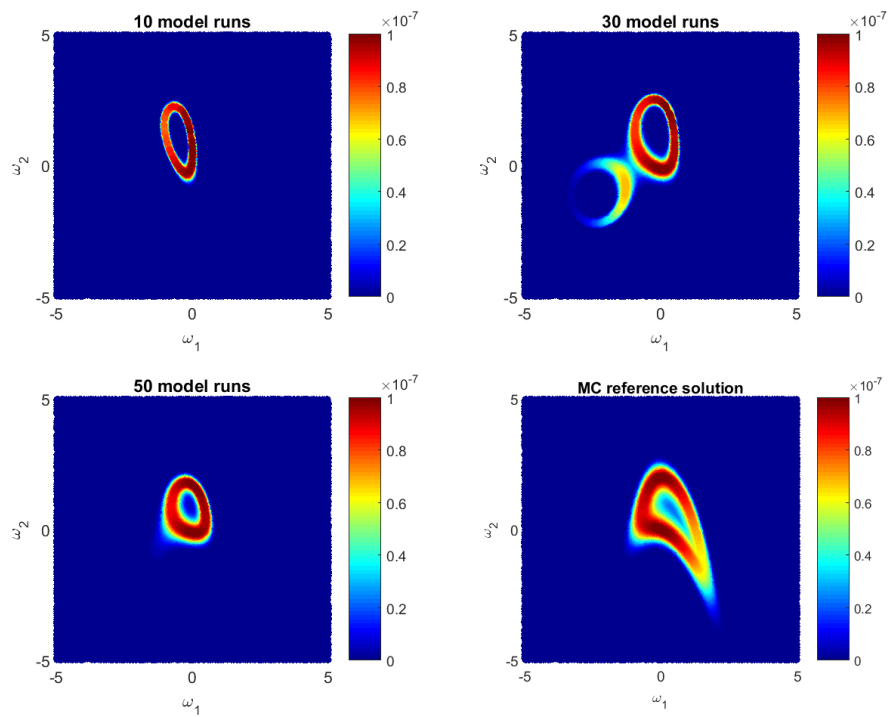


Figure 3. Likelihood values during Bayesian **Entropy-based** active learning as approximate by the Gaussian process emulator and by a reference Monte Carlo solution for a 2D reduction of the 10D problem.

4.1.3. Assessment of Information Arguments during Bayesian Active Learning

To assess the overall performance of the introduced Bayesian active learning strategies in Section 3, we compute Bayesian model evidence $p(\mathbf{D})$, relative entropy $D_{\text{KL}} [p(\omega|\mathbf{D}), p(\omega)]$ and information entropy $H [p(\omega|\mathbf{D})]$. To do so, we will employ the resulting GPE surrogates in Equation (11), (15) and (19), correspondingly. In that way, we avoid any unnecessary multidimensional integration or density estimation via Monte Carlo integration. Figure 4 illustrates how the Bayesian model evidence, the information entropy and the relative entropy adjust their value during Bayesian active learning for the discussed 2D reduction of the original 10D problem. Figure 4 shows the results of the BME-based, the IE-based and the RE-based active learning using red, green and blue lines, respectively. All three approaches reach their plateaus after approximately 20–30 active learning steps that corresponds to 20–30 runs of the original model.

A proper conclusion, however, could be drawn once the obtained BME, $D_{\text{KL}} [p(\omega|\mathbf{D}), p(\omega)]$ and $H [p(\omega|\mathbf{D})]$ are compared against their reference solutions. Therefore, we compute all the reference values denoted here as BME^{Ref} , $D_{\text{KL}}^{\text{Ref}} [p(\omega|\mathbf{D}), p(\omega)]$ and $H^{\text{Ref}} [p(\omega|\mathbf{D})]$ employing the Equations (11), (15) and (19) avoiding any assumptions or density estimations. To do so, we evaluate the original model \mathbf{M} instead of the surrogate \mathbf{S} in the Equations (11), (15) and (19) for the available Monte Carlo 10^5 samples in parameter space. Figure 5 illustrates the convergence of the Bayesian model evidence, the information entropy and the relative entropy estimates using GPE to the reference Monte Carlo solution during the Bayesian active learning. The RE-based active learning (blue line) convergences faster to the reference values than BME-based (red line) and RE-based (green line) active learning for all three indicators (BME, $D_{\text{KL}} [p(\omega|\mathbf{D}), p(\omega)]$ and $H [p(\omega|\mathbf{D})]$). Figure 5 aligns well with the results and discussion presented in Section 4.1.2.

Now, we will consider the full 10D setup of the problem (35) from Section 4.1.1. Similar to our discussion above, we will start the Bayesian active learning procedure with one training point only ($N_{\text{T}} = 1$) corresponding to the mean value of uncertain parameters ω and we will employ again all three introduced strategies. Assessing the performance of the active learning procedures will also compute the BME values $p(\mathbf{D})$, the RE value $D_{\text{KL}} [p(\omega|\mathbf{D}), p(\omega)]$ and the IE value $H [p(\omega|\mathbf{D})]$ based on the GPE surrogate. We will compare them against the reference values obtained from the plain MC technique on the original model with sample size of 10^5 . The results presented in Figure 6 confirm the anticipations from above and demonstrate a superior performance of RE-based active learning (blue line) in comparison to BME-based (red line) and IE-based active learning (green line). From the computational point of view, Figure 6 shows that the RE-based strategy already reaches an acceptable precision after approximately 200 model runs. This precision for the BME-based and the IE-based strategy can be reached, however, only after 500 model runs. It is worth mentioning that the current 10D setup is extremely challenging for GPE surrogates because of parameter dimensionality and its strong nonlinearity. From the current section, one can conclude that the relative entropy-based Bayesian active learning demonstrates a highly acceptable performance and seems to be the most suitable one for practical applications.

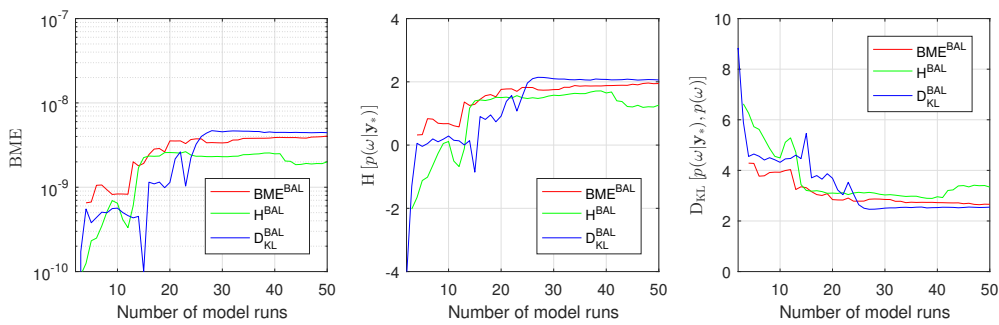


Figure 4. Bayesian model evidence, Information entropy and Relative entropy estimates during Bayesian active learning for Gaussian process emulator for a 2D reduction of the 10D problem: BME-based active learning (red line), IE-based active learning (green line) and RE-based active learning (blue line).

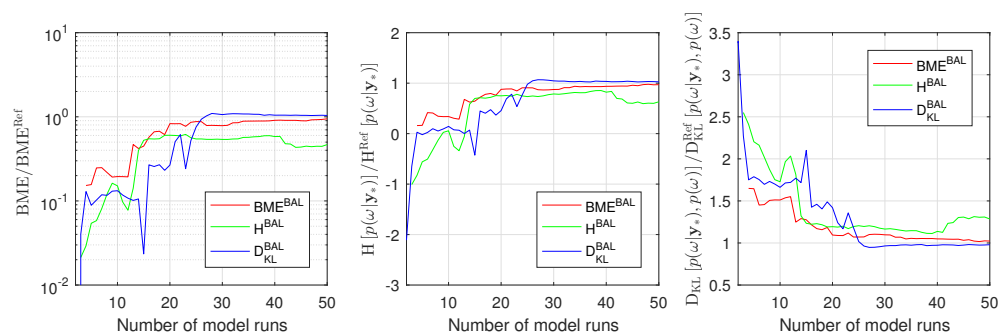


Figure 5. Convergence of Bayesian model evidence, Information entropy and Relative entropy estimates during Bayesian active learning for Gaussian process emulator to the reference Monte Carlo solution for a 2D reduction of the 10D problem: BME-based active learning (red line), IE-based active learning (green line) and RE-based active learning (blue line).

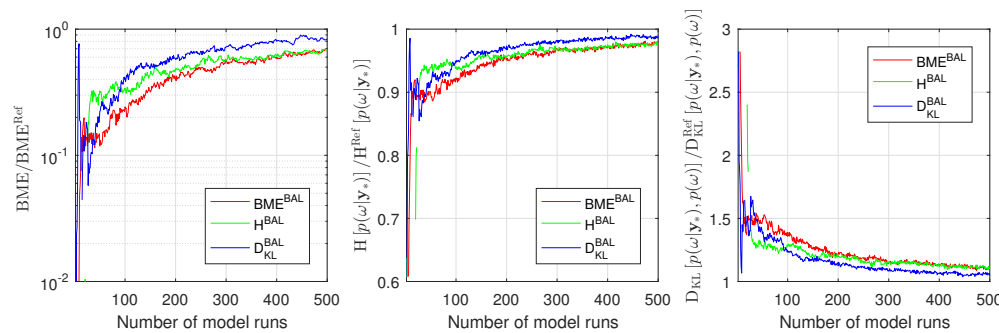


Figure 6. Convergence of Bayesian model evidence, Information entropy and Relative entropy estimates during active learning for Gaussian process emulator to the reference Monte Carlo solution for the 10D problem: BME-based active learning (red line), IE-based active learning (green line) and RE-based active learning (blue line).

4.2. Bayesian Active Learning for Carbon Dioxide Benchmark Problem

4.2.1. CO₂ Benchmark Set up

We will consider a multi-phase flow problem in porous media, where carbon dioxide (CO₂) is injected into a deep aquifer and then spreads in a geological formation. This yields a pressure build-up and a plume evolution. The CO₂ injection into the subsurface could be a possible practice to mitigate the CO₂ emission into the atmosphere. In this study, we use the deterministic model, provided by Köppel et al. [24], which is a reduced version of the model in a benchmark problem defined in the paper [85]. This reduction consists of a radial flow in the vicinity of the injection well, and made

primarily due to the high computational demand of the original CO₂ model. It is assumed that the fluid properties such as the density and the viscosity are constant, and all processes are isothermal. The CO₂ and the brine build two separate and immiscible phases, and the mutual dissolution is neglected. Additionally, the formation is isotropically rigid and chemically inert, and capillary pressure is negligible. Overall, the considered CO₂ benchmark problem is strongly nonlinear because the CO₂ saturation spreads as a strongly nonlinear front that could be challenging to capture via surrogates. For detailed information on the governing equations, the modeling assumption and the approaches, the reader is referred to the original publication [24].

Similar to [24], we consider the combined effects of three sources of uncertainty. We take into account the uncertainty of boundary conditions due to the injection rate, the uncertainty of parameters in the constitutive relations, introduced via uncertainty in the relative permeability definitions, and the uncertainty of material properties, i.e., the porosity of the geological layer. These three sources of uncertainty were introduced for the analysis in [24] using injection rate (IR), power theta (PT) and reservoir porosity (RP), i.e., $\omega = \{\text{IR, PT, RP}\}$.

We consider the CO₂ saturation to be the quantity of interest at a monitoring distance of 15 m from the injection well, measured each 10 days over a period of 100 days. We construct a scenario, in which the synthetic observed saturation values have been generated from the deterministic CO₂ benchmark model itself, with the uncertain parameters to be set as $\omega_{\text{Truth}} = \{1.0e - 04, 0.2, 0.3\}$. Additionally, we will assume that a measurement error of 0.02 ($\sigma_{\mathbf{D}} = 0.02$) exists for each synthetic observation data. Using the synthetic measurement data, we construct the reference solution conducting a Bayesian updating of the original CO₂ benchmark model. Namely, reference values of Bayesian model evidence (BME^{Ref}), the information entropy ($H^{\text{Ref}} [p(\omega|\mathbf{D})]$) and the relative entropy $D_{\text{KL}}^{\text{Ref}} [p(\omega|\mathbf{D}), p(\omega)]$ have been obtained based on 10⁴ Monte Carlo simulations using Equations (11), (15) and (19), correspondingly. Additionally, the posterior distribution of modeling parameters has been obtained via the same 10⁴ Monte Carlo simulations. One model run of the analyzed CO₂ benchmark problem required approximately 3–7 min on a standard computer, depending strongly on the values of modeling parameters. In what follows, we present the results and analyze the performance of the three Bayesian active learning methods, introduced earlier in this paper, applied to the aforementioned CO₂ benchmark set-up.

4.2.2. Assessment of Information Arguments during Bayesian Active for CO₂ Benchmarks

We start the Bayesian active learning process for the CO₂ benchmark model with one training point only ($N_T = 1$) using $\omega_T = \mathbb{E}_{p(\omega)}(\omega)$. Similar to the previous applications in Section 4.1, we perform the Bayesian active learning procedure by using the BME-based, RE-based and IE-based strategies. Analogously, the performance of the active learning process is analyzed by comparing the BME values $p(\mathbf{D})$, relative entropies $D_{\text{KL}} [p(\omega|\mathbf{D}), p(\omega)]$ and information entropies $H [p(\omega|\mathbf{D})]$ based on the GPE surrogates against their corresponding MC reference values.

Figure 7 illustrates the convergence of the BME value, the IE value and the RE value obtained during GPE-based Bayesian active learning against the reference Monte Carlo values. The results, presented in this figure, demonstrate that the RE-based (blue line) and the BME-based (red line) active learning shows again a superior performance compared to IE-based active learning (green line). Here, the RE-based strategy catches the reference BME values slightly better than the BME-based approach, and both approaches perform similarly well for other quantities of interest. The IE-based active learning demonstrates very similar behaviors to Section 4.1 and confirms the findings that have been reported for the 10D and its 2D reduction problems.

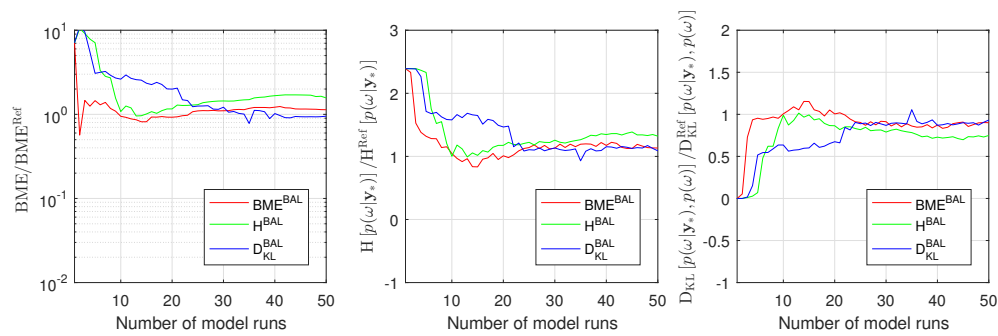


Figure 7. Convergence of Bayesian model evidence, Information entropy and Relative entropy estimates during active learning for Gaussian process emulator to the reference Monte Carlo solution for the CO₂ Benchmark problem: BME-based active learning (red line), IE-based active learning (green line) and RE-based active learning (blue line).

4.2.3. Posterior Distribution of Modeling Parameters for CO₂ Benchmarks

We will consider how good the introduced Bayesian active learning strategies can capture the posterior distribution of modeling parameters. To do so, we will illustrate the posterior distributions and correlations of modeling parameters for the CO₂ Benchmark problem obtained after 50 active learning iterations. Figure 8 presents the results obtained using the BME-based active learning (Figure 8a), the IE-based active learning (Figure 8b) and the RE-based active learning (Figure 8c), and it compares them with the reference Monte Carlo solution (Figure 8d). The BME-based strategy in Figure 8a and the RE-based strategy in Figure 8c capture very well the posterior distributions of all analyzed parameters and their correlations in comparison to the MC reference solution. The information entropy-based strategy in Figure 8b captures acceptably the distributions and the correlations of the injection rate (IR) and the reservoir porosity (RP) parameters. However, it could not properly capture the distribution of the PT parameter controlling the relative permeability distribution. Figure 8b illustrates a very strong overestimation of high-value probabilities of this parameter. Very similar posterior distributions for all strategies have already been observed after 25 iterations of active learning, which corresponds to the convergence shown in Figure 7.

We have used 50 interactions for the GPE-based Bayesian active learning for the demonstrative purposes only. Apparently, such a high number of active learning iterations is unnecessary for practical applications. In the Bayesian context, a stabilization of the posterior distributions indicates convergence of the surrogate representation to the original model in the region of high posterior density. That property could be useful, especially once the reference solution cannot be constructed due to computation reasons, see [35,36]. The convergence of posterior distributions in Figure 8 aligns well with convergence of the information-theoretic indicators shown in Figure 7.

Overall, Section 4.2 indicates that the RE-based Bayesian active learning demonstrates a slightly superior performance over the BME-based strategy, and both strategies are superior to the IE-based approach. Obviously it is very easy to judge the performance of BAL once the reference solution is available. However, as in many practical cases, the reference solution is not available due to a very high computational demand and it is relevant to draw the conclusion without the reference solution. Apparently, all mentioned indicators such as BME value $p(\mathbf{D})$, RE value $D_{\text{KL}}[p(\omega|\mathbf{D}), p(\omega)]$, and IE value $H[p(\omega|\mathbf{D})]$ reflect relevant information for a Bayesian inference. Therefore, the active learning procedure can be stopped once all such information-theoretic indicators stagnate and reach a plateau because the best possible surrogate representation of the original model have been (almost) reached.

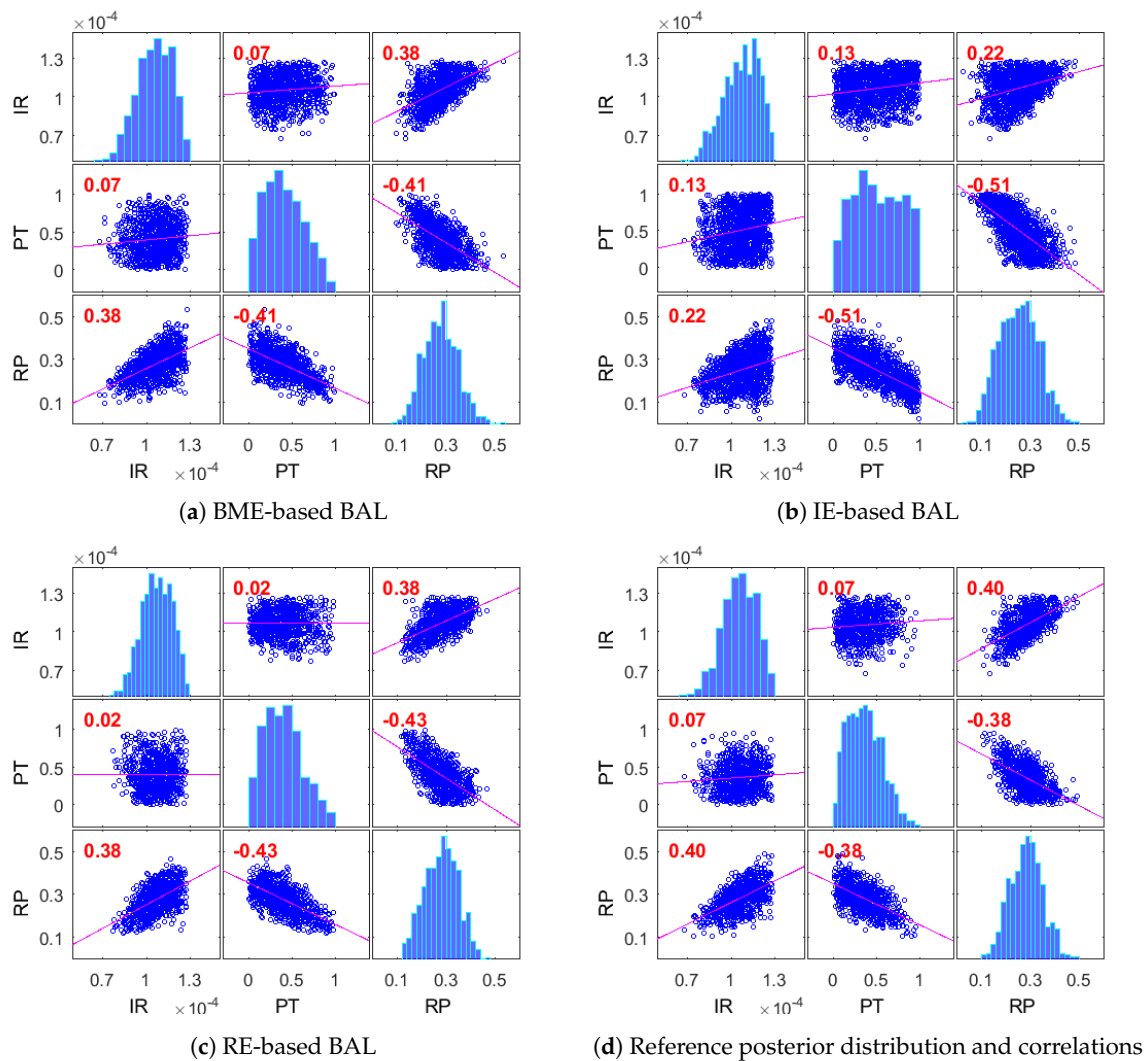


Figure 8. Posterior distributions and correlations of modeling parameters for the CO₂ Benchmark problem after 100 active learning iterations: BME-based active learning (a), IE-based active learning (b), RE-based active learning (c) and reference Monte Carlo solution (d).

4.3. Discussion

We use the link between Bayesian inference and information theory introduced in [40]. This means that we compute BME, RE and IE values while avoiding the criticized multi-Gaussian assumption. Instead, we perform prior-based sampling, i.e., Monte Carlo accompanied by rejecting sampling. Doing so, we consider that the main computations are related to running the original model and we assume feasibility of Monte Carlo sampling on the GPE surrogate.

Alternatively, any posterior-based sampling algorithm (e.g., MCMC) could be used during the Bayesian active learning procedure. However, when any posterior-based sampling algorithm is used, then the values for BME, RE and IE become quite rough approximations because relatively strong assumptions have to be taken. To provide cheap alternatives, various estimates of BME, RE and IE based on known criteria from information theory are now listed in Appendix A for the sake of completeness.

However, the estimates in Appendix A have to be used with care. According to [40], the harmonic mean estimate and the maximum-likelihood estimate for BME provide very unreliable results. Therefore, only rough guesses of the true BME value can be obtained from the maximum a posteriori estimate, Chib's estimate [86], Bayesian information criterion [87] and the Akaike information criterion (with [83] and without second-order bias correction [88]) due to their strong assumptions.

While estimates for BME, RE and IE based on the Kashyap information criterion [89] demonstrated unsatisfactory performance as well, we re-scaled them to a proper scale in [40]. However, the re-scaled Kashyap information criterion still includes unnecessary simplifications of the involved cross entropies. Among all these simplified estimates, the multivariate Gaussian posterior estimate [40] avoids the most unreasonable simplifications for posterior-based sampling and includes the least assumptions for estimating BME, RE and IE. The Gelfand and Dey approach [90] includes assumptions similar to the multivariate Gaussian, but provides slightly inferior results in the cases tested by us. Thus, for posterior-based sampling during Bayesian active learning, we suggested in our 2019 paper [40] to use the multivariate Gaussian estimate that includes least assumptions (Considering multivariate Gaussian distribution for active learning approaches focusing on GPE training on an underlying model only without considering measurement data is fully appropriate. It can be seen as Bayesian² active learning and approximation signs in Equations (A25), (A26) and (A27) turn to equality containing no assumptions by definition of GPE.).

Finalizing the discussion, we would like to remark that straightforward applications of the suggested Bayesian active learning strategies (or even already existing approaches) to GPE representations could be computationally very demanding once the problem dimensionality increases. This is mainly caused by the structure of GPE surrogates that is represented via localized kernels. These localized kernels require a lot of training for high-dimensional cases. Increasing the amount of measurement data will help to localize the relevant spots better. However, for high-dimensional problems, the structure of the surrogate should be constructed adaptively and sparse representations will be very beneficial. Alternatively, a preliminary sensitivity analysis (see e.g., [91,92]) could be conducted to partially overcome the problem of dimensionality.

5. Summary and Conclusions

The current paper deals with Gaussian process emulator that replicates a computational demanding physical model and honors the available observation data establishing fully Bayesian³ active learning framework. We elaborate the connection between Bayesian inference and information theory and offer a fully Bayesian view on a Gaussian process emulator through a Bayesian inference accompanied by a Bayesian active learning.

The paper employs the fundamental properties of Gaussian process emulator and introduces, in Section 3, three Bayesian active learning strategies. These strategies adaptively identify training sets, for which the full-complexity model must be evaluated. The first Bayesian active learning strategy, relying on Bayesian model evidence, indicates the quality of representation against the available measurements data. The second Bayesian active learning strategy, based on the relative entropy, seeks a relative information gain. The third Bayesian active learning strategy, based on information entropy, considers the expected missing information. The introduced strategies improve the Gaussian process emulator-based surrogate representation of a full-complexity physical model in the region of high posterior density. We employ the information-theoretic arguments to incorporate adaptively the measurements data. We emphasize in the paper that the information entropy and the relative entropy can be computed avoiding any assumption or unnecessary multidimensional integration.

We illustrate the performance of the suggested Bayesian active learning strategies using an analytical example and a carbon dioxide benchmark. Section 4 shows how the suggested approaches capture the likelihood values during an active learning procedure. We also show a visual comparison with the reference Monte Carlo solution for a 2D reduction of the 10D problem. We demonstrate rigorous evidence of convergence against the reference Monte Carlo values for the Bayesian model evidence, the information entropy and the relative entropy obtained via the three Bayesian active learning strategies. Additionally, Section 4 shows the evidence of convergence for the carbon dioxide benchmark problem against the reference solution for all proposed Bayesian active learning strategies. We also illustrate how the suggested Bayesian active learning strategies manage to quantify the post-calibration uncertainty in comparison to available Monte Carlo reference solutions.

Overall, we conclude that the introduced Bayesian active learning strategies for Gaussian process emulators could be very helpful for applied tasks where underlying full-complexity models are computationally very expensive. Moreover, the employed information-theoretic indicators can be used as stop criteria for Bayesian active learning once a reference solution is not available due to a very high computational demand. Our analysis indicates that the Bayesian model evidence-based and the relative entropy-based strategy demonstrate more reliable results in comparison to the information entropy-based strategy, which could be misleading. Additionally, the relative entropy-based strategy demonstrates a superior performance relative to the Bayesian model evidence-based strategy and seems to provide very sensitive arguments for the active learning.

Author Contributions: All authors have substantially contributed to this work, All authors have been involved in writing and revising the paper, and they have all read and approved the submitted version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This paper was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project Number 327154368—SFB 1313. The authors would like to thank the German Research Foundation for Financial support of the project within the Cluster of Excellence “Data-Integrated Simulation Science” (EXC 2075) at the University of Stuttgart and for supporting this work by funding SFB 1313, Project Number 327154368.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Appendix A. List of Approximative Active Learning Strategies

The current section briefly summarizes the alternative learning strategies that are based on certain assumptions. The approximates for BME, RE and IE values have been adopted from the paper [40] to Bayesian active learning introduced in Section 3 and we refer the reader to the original work for the complete details. All estimates mentioned here are less powerful than the strategies introduced in Section 3 due to their definitions. The most promising approximates are based as well on the multivariate Gaussian assumptions of the posterior $p_{\omega_E}^{BAL}(\omega|\mathbf{D})$ in Equation (21).

As it has been mentioned already in Section 1, learning functions for GPE training in Bayesian parameter inference can focus on the posterior mean and variance of model parameters obtained via assimilation of available measurement data. Doing so, we have adopted the idea of minimizing the integrated posterior variance (IVAR) of the GPE [77] to perform an active learning in the context of Bayesian assimilation of available data. Figure A1 illustrates convergence of active learning based on the IVAR strategy for the 10D setup discussed in Section 4.1.3. The IVAR-based active learning strategy (teal line) shows promising results while relying on low-order moments only. However, all three strategies introduced in the current paper demonstrate faster convergence, capturing the relevant spots at low computational budget and, as it has been already pointed out in Section 4, the RE-based active learning strategy demonstrates superior results. The reason is that the posterior distribution $p_{\omega_E}^{BAL}(\omega|\mathbf{D})$ resulting from Bayesian assimilation of measurement data is typically not multivariate Gaussian and corresponding assumptions slow down the active learning procedure.

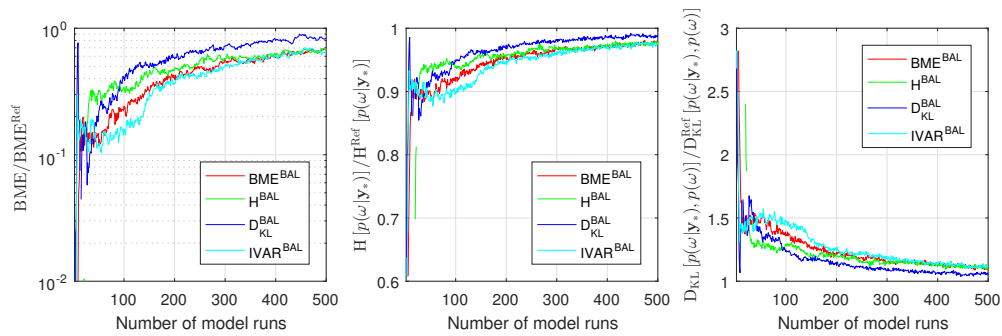


Figure A1. Convergence of Bayesian model evidence, Information entropy and Relative entropy estimates during active learning for Gaussian process emulator to the reference Monte Carlo solution for the 10D problem: BME-based active learning (red line), IE-based active learning (green line), RE-based active learning (blue line) and IVAR-based active learning (teal line).

Appendix A.1. Maximum a Posteriori Estimates

The BME, RE and IE values could be approximated as follows using the maximum a posteriori (MAP) estimates [40]:

$$\ln \text{BME}_{\text{MAP}}^{\text{BAL}} \approx \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{D}|\mathbf{S}) \right] \right) + \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})] \right) - \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right], \quad (\text{A1})$$

$$\text{D}_{\text{KL-MAP}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \approx -\mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})] \right) + \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right], \quad (\text{A2})$$

$$\text{H}_{\text{MAP}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}) \right] \approx -\ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right], \quad (\text{A3})$$

where \mathbf{S}_{MAP} is a maximum a posteriori response of the surrogate \mathbf{S} .

Appendix A.1.1. Chib’s Estimates

Following the idea of Chib [86], a single point estimate could be used in the following way [40]:

$$\ln \text{BME}_{\text{CHIB}}^{\text{BAL}} \approx \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{D}|\mathbf{S}_{\text{MAP}}) \right] + \ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})(\mathbf{S}_{\text{MAP}})] - \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right]. \quad (\text{A4})$$

$$\text{D}_{\text{KL-CHIB}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \approx -\ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})(\mathbf{S}_{\text{MAP}})] + \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right], \quad (\text{A5})$$

$$\text{H}_{\text{CHIB}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}) \right] \approx -\ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right]. \quad (\text{A6})$$

Appendix A.1.2. Estimates via Akaike Information Criterion

The MAP approximation could be extended while employing the Akaike information criterion (AIC) [83] as follows [40]:

$$\ln \text{BME}_{\text{AIC}}^{\text{BAL}} \approx \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{D}|\mathbf{S}) \right] \right) + \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})] \right) - \frac{1}{n} \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right] + 1, \quad (\text{A7})$$

$$\text{D}_{\text{KL-AIC}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \approx -\mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})] \right) + \frac{1}{n} \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right] - 1, \quad (\text{A8})$$

$$\text{H}_{\text{AIC}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}) \right] \approx -\frac{1}{n} \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right] + 1. \quad (\text{A9})$$

Appendix A.1.3. Estimates via Second-Order bias Correction for Akaike Information Criterion

Second-order bias correction [88] for a limited sample size s (length of vector \mathbf{D}) extends the Akaike information criterion to the following from [40]:

$$\ln \text{BME}_{\text{AICc}}^{\text{BAL}} \approx \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{D}|\mathbf{S}) \right] \right) + \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})] \right) - \frac{1}{n} \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right] + \frac{s}{s-n-1}. \tag{A10}$$

$$\text{D}_{\text{KL-AICc}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \approx - \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})] \right) + \frac{1}{n} \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right] - \frac{s}{s-n-1}, \tag{A11}$$

$$\text{H}_{\text{AICc}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}) \right] \approx - \frac{1}{n} \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right] + \frac{s}{s-n-1}. \tag{A12}$$

Appendix A.1.4. Estimates via Bayesian Information Criterion

Employing the Bayesian information criterion (also known as Schwarz information criterion) introduced by Schwarz [87] leads to the following estimates [40]:

$$\ln \text{BME}_{\text{BIC}}^{\text{BAL}} \approx \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MLE}}|\mathbf{D}) \right] + \frac{n}{2} \ln s, \tag{A13}$$

$$\text{D}_{\text{KL-BIC}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \approx \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{D}|\mathbf{S}_{\text{MLE}}) \right] - \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MLE}}|\mathbf{D}) \right] - \frac{n}{2} \ln s, \tag{A14}$$

$$\text{H}_{\text{BIC}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}) \right] \approx \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MLE}}|\mathbf{D}) \right] - \ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})(\mathbf{S}_{\text{MLE}})] - \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{D}|\mathbf{S}_{\text{MLE}}) \right] + \frac{n}{2} \ln s, \tag{A15}$$

where \mathbf{S}_{MLE} is a maximum a posteriori response of the surrogate \mathbf{S} .

Appendix A.1.5. Estimates via Kashyap Information Criterion

Approximations based on the Kashyap Information Criterion [89] offer the following estimates [40]:

$$\ln \text{BME}_{\text{KIC}}^{\text{BAL}} \approx \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right] + \ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})(\mathbf{S}_{\text{MAP}})] + \frac{1}{2} \ln [(2\pi)^n |\mathbf{C}|]. \tag{A16}$$

$$\text{D}_{\text{KL-KIC}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \approx - \ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})(\mathbf{S}_{\text{MAP}})] - \frac{1}{2} \ln [(2\pi)^n |\mathbf{C}|], \tag{A17}$$

$$\text{H}_{\text{KIC}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}) \right] \approx \frac{1}{2} \ln [(2\pi)^n |\mathbf{C}|]. \tag{A18}$$

Appendix A.1.6. Estimates via Re-Scaled Kashyap Information Criterion

The re-scaled correction of Kashyap Information Criterion according to paper [40] leads to the following:

$$\ln \text{BME}_{\text{KICr}}^{\text{BAL}} \approx \ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}_{\text{MAP}}|\mathbf{D}) \right] + \ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})(\mathbf{S}_{\text{MAP}})] + \frac{1}{2} \ln [(2\pi e)^n |\mathbf{C}|]. \tag{A19}$$

$$\text{D}_{\text{KL-KICr}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \approx - \ln [\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})(\mathbf{S}_{\text{MAP}})] - \frac{1}{2} \ln [(2\pi e)^n |\mathbf{C}|], \tag{A20}$$

$$\text{H}_{\text{KICr}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}) \right] \approx \frac{1}{2} \ln [(2\pi e)^n |\mathbf{C}|], \tag{A21}$$

where \mathbf{C} is the posterior (co)variance matrix.

Appendix A.1.7. Estimates via Gelfand and Dey Sampling

Following the idea of Gelfand and Dey [90], an importance sampling with density $\tau(\mathbf{S})$ leads to the following approximates [40]:

$$\ln \text{BME}_{\text{GD}}^{\text{BAL}} \approx \ln \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})}^{-1} \left(\frac{\tau(\mathbf{S})}{p_{\omega_E}^{\text{BAL}}(\mathbf{D}|\mathbf{S}) \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})} \right), \quad (\text{A22})$$

$$\begin{aligned} \text{D}_{\text{KLGD}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] &\approx \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{D}|\mathbf{S}) \right] \right) - \\ &- \ln \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})}^{-1} \left(\frac{\tau(\mathbf{S})}{p_{\omega_E}^{\text{BAL}}(\mathbf{D}|\mathbf{S}) \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})} \right), \end{aligned} \quad (\text{A23})$$

$$\begin{aligned} \text{H}_{\text{GD}}^{\text{BAL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}) \right] &\approx - \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{D}|\mathbf{S}) \right] \right) - \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln \left[\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \right) + \\ &+ \ln \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})}^{-1} \left(\frac{\tau(\mathbf{S})}{p_{\omega_E}^{\text{BAL}}(\mathbf{D}|\mathbf{S}) \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}})} \right), \end{aligned} \quad (\text{A24})$$

where the importance sampling density $\tau(\mathbf{S})$ is often assumed to be multivariate Gaussian or t -distributed.

Appendix A.1.8. Multivariate Gaussian Estimates

Assumption on multivariate Gaussian (MG) distribution [93,94] of the posterior distribution $p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})$ leads to the following approximates [40]:

$$\ln \text{BME}_{\text{MG}}^{\text{BAL}} \approx \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln \left[p_{\omega_E}^{\text{BAL}}(\mathbf{D}|\mathbf{S}) \right] \right) + \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D})} \left(\ln \left[\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \right) + \frac{1}{2} \ln \left[(2\pi e)^n |\mathbf{C}| \right], \quad (\text{A25})$$

$$\text{D}_{\text{KL}} \left[p_{\omega_E}^{\text{BAL}}(\mathbf{S}|\mathbf{D}), \mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] = - \mathbb{E}_{p_{\omega_E}^{\text{BAL}}(\omega|\mathbf{D})} \left(\ln \left[\mathcal{N}_{\omega_E}(\mu_{\mathbf{S}}, \sigma_{\mathbf{S}}) \right] \right) - \frac{1}{2} \ln \left[(2\pi e)^n |\mathbf{C}| \right], \quad (\text{A26})$$

$$\text{H} \left[p_{\omega_E}^{\text{BAL}}(\omega|\mathbf{D}) \right] \approx \frac{1}{2} \ln \left[(2\pi e)^n |\mathbf{C}| \right]. \quad (\text{A27})$$

References

1. Wirtz, D.; Nowak, W. The rocky road to extended simulation frameworks covering uncertainty, inversion, optimization and control. *Environ. Model. Softw.* **2017**, *93*, 180–192. [\[CrossRef\]](#)
2. Wiener, N. The homogeneous chaos. *Am. J. Math.* **1938**, *60*, 897–936. [\[CrossRef\]](#)
3. Ghanem, R.G.; Spanos, P.D. *Stochastic Finite Elements: A Spectral Approach*; Springer: New York, NY, USA, 1991.
4. Lin, G.; Tartakovsky, A. An efficient, high-order probabilistic collocation method on sparse grids for three-dimensional flow and solute transport in randomly heterogeneous porous media. *Adv. Water Res.* **2009**, *32*, 712–722. [\[CrossRef\]](#)
5. Oladyshkin, S.; Nowak, W. Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion. *Reliab. Eng. Syst. Saf.* **2012**, *106*, 179–190. [\[CrossRef\]](#)
6. Oladyshkin, S.; Nowak, W. Incomplete statistical information limits the utility of high-order polynomial chaos expansions. *Reliab. Eng. Syst. Saf.* **2018**, *169*, 137–148. [\[CrossRef\]](#)
7. Foo, J.; Karniadakis, G. Multi-element probabilistic collocation method in high dimensions. *J. Comput. Phys.* **2010**, *229*, 1536–1557. [\[CrossRef\]](#)
8. Zhang, Y.; Liu, Y.; Pau, G.; Oladyshkin, S.; Finsterle, S. Evaluation of multiple reduced-order models to enhance confidence in global sensitivity analyses. *Int. J. Greenh. Gas Control* **2016**, *49*, 217–226. [\[CrossRef\]](#)
9. Oladyshkin, S.; Class, H.; Helmig, R.; Nowak, W. An integrative approach to robust design and probabilistic risk assessment for CO₂ storage in geological formations. *Comput. Geosci.* **2011**, *15*, 565–577. [\[CrossRef\]](#)
10. Keese, A.; Matthies, H.G. Sparse quadrature as an alternative to Monte Carlo for stochastic finite element techniques. *Proc. Appl. Math. Mech.* **2003**, *3*, 493–494. [\[CrossRef\]](#)
11. Blatman, G.; Sudret, B. Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach. *C. R. Mécanique* **2008**, *336*, 518–523. [\[CrossRef\]](#)
12. Ahlfeld, R.; Belkouchi, B.; Montomoli, F. SAMBA: Sparse approximation of moment-based arbitrary polynomial chaos. *J. Comput. Phys.* **2016**, *320*, 1–16. [\[CrossRef\]](#)

13. Sinsbeck, M.; Nowak, W. Sequential Design of Computer Experiments for the Solution of Bayesian Inverse Problems. *SIAM/ASA J. Uncertain. Quantif.* **2017**, *5*, 640–664. [[CrossRef](#)]
14. Alkhateeb, O.; Ida, N. Data-Driven Multi-Element Arbitrary Polynomial Chaos for Uncertainty Quantification in Sensors. *IEEE Trans. Magn.* **2017**, *54*, 1–4. [[CrossRef](#)]
15. Kröker, I.; Nowak, W.; Rohde, C. A stochastically and spatially adaptive parallel scheme for uncertain and nonlinear two-phase flow problems. *Comput. Geosci.* **2015**, *19*, 269–284. [[CrossRef](#)]
16. Oladyshkin, S.; Class, H.; Helmig, R.; Nowak, W. A concept for data-driven uncertainty quantification and its application to carbon dioxide storage in geological formations. *Adv. Water Res.* **2011**, *34*, 1508–1518. [[CrossRef](#)]
17. Köppel, M.; Kröker, I.; Rohde, C. Intrusive uncertainty quantification for hyperbolic-elliptic systems governing two-phase flow in heterogeneous porous media. *Comput. Geosci.* **2017**, *21*, 807–832. [[CrossRef](#)]
18. Wendland, H. *Scattered Data Approximation*; Cambridge University Press: Cambridge, UK, 2005; Volume 17.
19. Schölkopf, B.; Smola, A. *Learning with Kernels*; The MIT Press: Cambridge, MA, USA, 2002.
20. Cressie, N.A. Spatial prediction and kriging. *Statistics for Spatial Data*, Cressie NAC, ed.; John Wiley & Sons: New York, NY, USA, 1993; pp. 105–209.
21. Kolmogorov, A.N.; Bharucha-Reid, A.T. *Foundations of the Theory of Probability: Second English Edition*; Courier Dover Publications: Mineola, NY, USA, 2018.
22. Xiao, S.; Oladyshkin, S.; Nowak, W. Reliability analysis with stratified importance sampling based on adaptive Kriging. *Reliab. Eng. Syst. Saf.* **2020**, *197*, 106852. [[CrossRef](#)]
23. Williams, C.K.; Rasmussen, C.E. Gaussian processes for regression. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1996; pp. 514–520.
24. Köppel, M.; Franzelin, F.; Kröker, I.; Oladyshkin, S.; Santin, G.; Wittwar, D.; Barth, A.; Haasdonk, B.; Nowak, W.; Pflüger, D.; et al. Comparison of data-driven uncertainty quantification methods for a carbon dioxide storage benchmark scenario. *Comput. Geosci.* **2019**. [[CrossRef](#)]
25. Lia, O.; Omre, H.; Tjelmeland, H.; Holden, L.; Egeland, T. Uncertainties in reservoir production forecasts. *AAPG Bull.* **1997**, *81*, 775–802.
26. Smith, A.F.; Gelfand, A.E. Bayesian statistics without tears: A sampling–resampling perspective. *Am. Stat.* **1992**, *46*, 84–88.
27. Gilks, W.; Richardson, S.; Spiegelhalter, D. *Markov Chain Monte Carlo in Practice*; Chapman & Hall: London, UK, 1996.
28. Liu, P.; Elshall, A.S.; Ye, M.; Beerli, P.; Zeng, X.; Lu, D.; Tao, Y. Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods. *Water Resour. Res.* **2016**, *52*, 734–758. [[CrossRef](#)]
29. Xiao, S.; Reuschen, S.; Köse, G.; Oladyshkin, S.; Nowak, W. Estimation of small failure probabilities based on thermodynamic integration and parallel tempering. *Mech. Syst. Signal Process.* **2019**, *133*, 106248. [[CrossRef](#)]
30. Skilling, J.; others. Nested sampling for general Bayesian computation. *Bayesian Anal.* **2006**, *1*, 833–859. [[CrossRef](#)]
31. Elsheikh, A.; Oladyshkin, S.; Nowak, W.; Christie, M. Estimating the probability of co2 leakage using rare event simulation. In Proceedings of the ECMOR XIV-14th European Conference on the Mathematics of Oil Recovery, Catania, Italy, 8–11 September 2014.
32. Au, S.K.; Beck, J.L. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Eng. Mech.* **2001**, *16*, 263–277. [[CrossRef](#)]
33. Zuev, K.M.; Beck, J.L.; Au, S.K.; Katafygiotis, L.S. Bayesian post-processor and other enhancements of Subset Simulation for estimating failure probabilities in high dimensions. *Comput. Struct.* **2012**, *92*, 283–296. [[CrossRef](#)]
34. Volpi, E.; Schoups, G.; Firmani, G.; Vrugt, J.A. Sworn testimony of the model evidence: Gaussian mixture importance (GAME) sampling. *Water Resour. Res.* **2017**, *53*, 6133–6158. [[CrossRef](#)]
35. Oladyshkin, S.; Class, H.; Nowak, W. Bayesian updating via Bootstrap filtering combined with data-driven polynomial chaos expansions: Methodology and application to history matching for carbon dioxide storage in geological formations. *Comput. Geosci.* **2013**, *17*, 671–687. [[CrossRef](#)]
36. Oladyshkin, S.; Schroeder, P.; Class, H.; Nowak, W. Chaos expansion based Bootstrap filter to calibrate. CO₂ injection models. *Energy Procedia* **2013**, *40*, 398–407. [[CrossRef](#)]

37. Li, J.; Marzouk, Y.M. Adaptive construction of surrogates for the Bayesian solution of inverse problems. *SIAM J. Sci. Comput.* **2014**, *36*, A1163–A1186. [[CrossRef](#)]
38. Sinsbeck, M.; Cooke, E.; Nowak, W. Sequential Design of Computer Experiments for the Computation of Bayesian Model Evidence. Submitted.
39. Beckers, F.; Heredia, A.; Noack, M.; Nowak, W.; Wieprecht, S.; Oladyshkin, S. Bayesian Calibration and Validation of a Large-Scale and Time-Demanding Sediment Transport Model. *Water Resour. Res.* **2020**, *56*, e2019WR026966. [[CrossRef](#)]
40. Oladyshkin, S.; Nowak, W. The Connection between Bayesian Inference and Information Theory for Model Selection, Information Gain and Experimental Design. *Entropy* **2019**, *21*, 1081. [[CrossRef](#)]
41. Wiener, N. *Cybernetics*; John Wiley & Sons Inc.: New York, NY, USA, 1948.
42. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
43. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
44. Good, I. Some terminology and notation in information theory. *Proc. IEE-Part C Monogr.* **1956**, *103*, 200–204. [[CrossRef](#)]
45. Shannon, C.E.; Weaver, W. The mathematical theory of communication. *Ill. Press. Urbana I* **1949**, *11*, 117.
46. Murari, A.; Peluso, E.; Cianfrani, F.; Gaudio, P.; Lungaroni, M. On the use of entropy to improve model selection criteria. *Entropy* **2019**, *21*, 394. [[CrossRef](#)]
47. Gresele, L.; Marsili, M. On maximum entropy and inference. *Entropy* **2017**, *19*, 642. [[CrossRef](#)]
48. Cavanaugh, J.E. A large-sample model selection criterion based on Kullback’s symmetric divergence. *Stat. Probab. Lett.* **1999**, *42*, 333–343. [[CrossRef](#)]
49. Vecer, J. Dynamic Scoring: Probabilistic Model Selection Based on Utility Maximization. *Entropy* **2019**, *21*, 36. [[CrossRef](#)]
50. Cliff, O.; Prokopenko, M.; Fitch, R. Minimising the Kullback–Leibler divergence for model selection in distributed nonlinear systems. *Entropy* **2018**, *20*, 51. [[CrossRef](#)]
51. Chaloner, K.; Verdinelli, I. Bayesian experimental design: A review. *Stat. Sci.* **1995**, *10*, 273–304. [[CrossRef](#)]
52. Lindley, D.V. On a measure of the information provided by an experiment. *Ann. Math. Stat.* **1956**, *27*, 986–1005. [[CrossRef](#)]
53. Fischer, R. Bayesian experimental design—studies for fusion diagnostics. *Am. Inst. Phys.* **2004**, *735*, 76–83.
54. Nowak, W.; Guthke, A. Entropy-based experimental design for optimal model discrimination in the geosciences. *Entropy* **2016**, *18*, 409. [[CrossRef](#)]
55. Richard, M.D.; Lippmann, R.P. Neural network classifiers estimate Bayesian posterior probabilities. *Neural Comput.* **1991**, *3*, 461–483. [[CrossRef](#)]
56. Rubinstein, R.Y.; Kroese, D.P. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*; Springer Science & Business Media: Berlin, Germany, 2013.
57. Granzio, D.; Ru, B.; Zohren, S.; Dong, X.; Osborne, M.; Roberts, S. MEME: An accurate maximum entropy method for efficient approximations in large-scale machine learning. *Entropy* **2019**, *21*, 551. [[CrossRef](#)]
58. Mohammad-Djafari, A. Entropy, information theory, information geometry and Bayesian inference in data, signal and image processing and inverse problems. *Entropy* **2015**, *17*, 3989–4027. [[CrossRef](#)]
59. Laws, F.; Schätze, H. Stopping criteria for active learning of named entity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*; Association for Computational Linguistics: Strassburg, PA, USA, 2008; pp. 465–472.
60. Fu, L.; Grishman, R. An efficient active learning framework for new relation types. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan, 14–18 October 2013; pp. 692–698.
61. Schreiter, J.; Nguyen-Tuong, D.; Eberts, M.; Bischoff, B.; Markert, H.; Toussaint, M. Safe Exploration for Active Learning with Gaussian Processes. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2015)*, Porto, Portugal, 7–11 September 2015.
62. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; The MIT Press: Cambridge, MA, USA, 2006.
63. Kennedy, M.C.; O’Hagan, A. Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2001**, *63*, 425–464. [[CrossRef](#)]
64. O’Hagan, A. Bayesian analysis of computer code outputs: A tutorial. *Reliab. Eng. Syst. Saf.* **2006**, *91*, 1290–1300. [[CrossRef](#)]

65. Busby, D. Hierarchical adaptive experimental design for Gaussian process emulators. *Reliab. Eng. Syst. Saf.* **2009**, *94*, 1183–1193. [[CrossRef](#)]
66. Handcock, M.S.; Stein, M.L. A Bayesian Analysis of Kriging. *Technometrics* **1993**, *35*, 403–410. [[CrossRef](#)]
67. Diggle, P.J.; Ribeiro, P.J.; Christensen, O.F. An Introduction to Model-Based Geostatistics. In *Spatial Statistics and Computational Methods*; Møller, J., Ed.; Springer: New York, NY, USA, 2003; pp. 43–86. [2](#). [[CrossRef](#)]
68. Minasny, B.; McBratney, A.B. The Matérn function as a general model for soil variograms. *Geoderma* **2005**, *128*, 192–207. [[CrossRef](#)]
69. Echard, B.; Gayton, N.; Lemaire, M. AK-MCS: An active learning reliability method combining Kriging and Monte Carlo simulation. *Struct. Saf.* **2011**, *33*, 145–154. [[CrossRef](#)]
70. Sundar, V.; Shields, M.D. Reliability analysis using adaptive kriging surrogates with multimodel inference. *ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng.* **2019**, *5*, 04019004. [[CrossRef](#)]
71. Sun, Z.; Wang, J.; Li, R.; Tong, C. LIF: A new Kriging based learning function and its application to structural reliability analysis. *Reliab. Eng. Syst. Saf.* **2017**, *157*, 152–165. [[CrossRef](#)]
72. Krause, A.; Singh, A.; Guestrin, C. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *J. Mach. Learn. Res.* **2008**, *9*, 235–284.
73. Zhang, J.; Li, W.; Zeng, L.; Wu, L. An adaptive Gaussian process-based method for efficient Bayesian experimental design in groundwater contaminant source identification problems. *Water Resour. Res.* **2016**, *52*, 5971–5984. [[CrossRef](#)]
74. Conrad, P.R.; Marzouk, Y.M.; Pillai, N.S.; Smith, A. Accelerating asymptotically exact MCMC for computationally intensive models via local approximations. *J. Am. Stat. Assoc.* **2016**, *111*, 1591–1607. [[CrossRef](#)]
75. Wang, H.; Li, J. Adaptive Gaussian process approximation for Bayesian inference with expensive likelihood functions. *Neural Comput.* **2018**, *30*, 3072–3094. [[CrossRef](#)]
76. Gramacy, R.B.; Apley, D.W. Local Gaussian process approximation for large computer experiments. *J. Comput. Graph. Stat.* **2015**, *24*, 561–578. [[CrossRef](#)]
77. Gorodetsky, A.; Marzouk, Y. Mercer kernels and integrated variance experimental design: Connections between Gaussian process regression and polynomial approximation. *SIAM/ASA J. Uncertain. Quantif.* **2016**, *4*, 796–828. [[CrossRef](#)]
78. MATLAB. Version 9.7.0.1216025 (R2019b). 2019. Available online: <https://www.mathworks.com/help/stats/fitrgp.html> (accessed on 10 July 2020).
79. Mohammadi, F.; Kopmann, R.; Guthke, A.; Oladyshkin, S.; Nowak, W. Bayesian selection of hydro-morphodynamic models under computational time constraints. *Adv. Water Resour.* **2018**, *117*, 53–64. [[CrossRef](#)]
80. Soofi, E.S. Information theory and Bayesian statistics. In *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellng*; John Wiley & Sons: New York, NY, USA, 1996; pp. 179–189.
81. Kass, R.E.; Raftery, A.E. Bayes Factors. *J. Am. Stat. Assoc.* **1995**, *90*, 773–795. [[CrossRef](#)]
82. Hammersley, J.M. Monte Carlo Methods for solving multivariable problems. *Ann. N. Y. Acad. Sci.* **1960**, *86*, 844–874. [[CrossRef](#)]
83. Akaike, H. A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*; Springer: Berlin, Germany, 1974; pp. 215–222.
84. Oladyshkin, S. BAL-GPE Matlab Toolbox: Bayesian Active Learning for GPE, MATLAB Central File Exchange. 2020. Available online: <https://www.mathworks.com/matlabcentral/fileexchange/74794-bal-gpe-matlab-toolbox-bayesian-active-learning-for-gpe> (accessed on 12 August 2020).
85. Class, H.; Ebigbo, A.; Helmig, R.; Dahle, H.K.; Nordbotten, J.M.; Celia, M.A.; Audigane, P.; Darcis, M.; Ennis-King, J.; Fan, Y.; et al. A benchmark study on problems related to CO₂ storage in geologic formations. *Comput. Geosci.* **2009**, *13*, 409. [[CrossRef](#)]
86. Chib, S. Marginal likelihood from the Gibbs output. *J. Am. Stat. Assoc.* **1995**, *90*, 1313–1321. [[CrossRef](#)]
87. Schwarz, G.; others. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [[CrossRef](#)]
88. Sugiura, N. Further analysts of the data by Akaike’s information criterion and the finite corrections: Further analysts of the data by Akaike’s. *Commun. Stat.-Theory Methods* **1978**, *7*, 13–26. [[CrossRef](#)]
89. Kashyap, R.L. Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Trans. Pattern Anal. Mach. Intell.* **1982**, *PAMI-4*, 99–104. [[CrossRef](#)]

90. Gelfand, A.E.; Dey, D.K. Bayesian model choice: Asymptotics and exact calculations. *J. R. Stat. Soc. Ser. B (Methodol.)* **1994**, *56*, 501–514. [[CrossRef](#)]
91. Oladyshkin, S.; De Barros, F.; Nowak, W. Global sensitivity analysis: A flexible and efficient framework with an example from stochastic hydrogeology. *Adv. Water Resour.* **2012**, *37*, 10–22. [[CrossRef](#)]
92. Xiao, S.; Oladyshkin, S.; Nowak, W. Forward-reverse switch between density-based and regional sensitivity analysis. *Appl. Math. Model.* **2020**, *84*, 377–392. [[CrossRef](#)]
93. Goldman, S. *Information Theory*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1953.
94. McEliece, R.; Mac Eliece, R.J. *The Theory of Information and Coding*; Cambridge University Press: Cambridge, UK, 2002.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).