A Comparison of Metamodeling Techniques via Numerical Experiments

Luis G. Crespo[*], Sean P. Kenny, Daniel P. Giesy,[†]

NASA Langley Research Center, Hampton, VA, 23681, USA

This paper presents a comparative analysis of a few metamodeling techniques using numerical experiments for the single input-single output case. These experiments enable comparing the models' predictions with the phenomenon they are aiming to describe as more data is made available. These techniques include (i) prediction intervals associated with a least squares parameter estimate, (ii) Bayesian credible intervals, (iii) Gaussian process models, and (iv) interval predictor models. Aspects being compared are computational complexity, accuracy (i.e., the degree to which the resulting prediction conforms to the actual Data Generating Mechanism), reliability (i.e., the probability that new observations will fall inside the predicted interval), sensitivity to outliers, extrapolation properties, ease of use, and asymptotic behavior. The numerical experiments describe typical application scenarios that challenge the underlying assumptions supporting most metamodeling techniques.

Abbreviations

CI	Credible Interval
CFI	Confidence Interval
DGM	Data Generating Mechanism
GP	Gaussian Process
IPM	Interval Predictor Model
LS	Least Squares
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimate
PI	Prediction Interval
PDF	Probability Density Function
RBF	Radial Basis Function

[*] Corresponding Author, Luis.G.Crespo@nasa.gov, AIAA Associate Fellow.

[†] Aerospace Technologists, Dynamic Systems and Control Branch.

I. Introduction

The problem statement of interest is as follows. A *Data Generating Mechanism* (DGM) is postulated to act on a vector of input variables, $x \in \mathbb{R}^{n_x}$, to produce an output, $y \in \mathbb{R}^{n_y}$. Let $X \subseteq \mathbb{R}^{n_x}$ and $Y \subseteq \mathbb{R}^{n_y}$ be the sets of input and output variables associated with the DGM. Denote by $\mathbb{P}$ the *unknown* probability distribution of the DGM. A general $\mathbb{P}$ yields a random process in which y is an arbitrary random variable for a fixed value of x. The particular case in which y is a deterministic function of x corresponds to a $\mathbb{P}$ that is concentrated over the function. Assume that N independent input-output pairs are obtained from a stationary DGM, and denote by $\mathsf{z} = \{z_i\}$, with $z_i = (x_i, y_i)$ for $i = 1, \ldots, N$, the corresponding data sequence. It is desired to build a metamodel of the DGM based on z which will predict the output corresponding to an unobserved realization of the input. The *only* information available to build such a model is the data. The presence of intrinsic variability, and uncertainty about the DGM's structure makes it unrealistic to build a model that will predict a single output for a fixed input. Instead, we want to predict an interval into which unobserved data is expected to fall. In this setting the two main problems of interest can be stated as follows. First, we want to find a metamodel that, when evaluated at a new value x_{N+1} of the input, returns an informative prediction of the unobserved output y_{N+1}. Second, we want to quantify the probability of y_{N+1} falling within the predicted interval. In this setting an informative prediction can be interpreted as a narrow output interval of high probability (given by $\mathbb{P}$). Note that the second objective implies that the prediction must conform to the DGM for *any* value of N without having *any* knowledge about its underlying structure.

A few metamodeling techniques will be used to address the above objectives. Metamodeling [7] refers to the process of creating a mathematical representation of a phenomenon of interest. These models can have a parametric or non-parametric structure. In the former case the analyst first prescribes a model's structure, which is an equation prescribing how the output depends on the inputs and parameters, and then determines the parameter values for which a measure of the discrepancy between observations and predictions is minimized. The first step entails prescribing a suitable model structure. A plethora of model forms can be assumed including polynomials, neural networks, radial basis functions, etc. The second step is commonly referred to as model calibration or regression. In the presence of model-form uncertainty, measurement noise, and numerical error, the prescription of a fixed parameter value often yields inaccurate predictions. Consequently, it is preferable to prescribe a set of parameter values or a joint random vector from which parameter points are chosen such that the collective prediction resulting from evaluating the model at each of such points accurately represents the ensemble of observations. Conversely, in a nonparametric regression the predictor does not take a predetermined form but is constructed according to information derived from the data. Nonparametric regression requires larger sample sizes than parametric regression because the model structure as well as the model hyper-parameters must be inferred. Gaussian Process models or Kriging, non-parametric multiplicative regression, and Kernel regression are commonly used techniques.

For simplicity sake, we will be limiting our study to the single input-single output case, so $n_x = n_y = 1$. This paper addresses the above objectives by using several metamodeling techniques to create surrogates of four DGMs. DGM1 is an unknown deterministic function (i.e., the probabilistic cloud $\mathbb{P}$ has zero output variance and an unknown mean), DGM2 is a random process in which the aleatory component of the measured output depends on the value of the input, DGM3 is a Gaussian random process with unknown mean and variance functions, and DGM4 is a random process requiring extrapolation beyond the range of observations. The metamodeling techniques studied are *Prediction Intervals* (PI), *Bayesian Credible Intervals* (CI), *Interval Predictor Models* (IPM) and *Gaussian Process* (GPs) models. The parametric regression techniques will use the linear structure

$$y = M(x, p) = p^\top \varphi(x; q), \tag{1}$$

for $\varphi(x; q) = \{\varphi_1(x; q^1), \ldots, \varphi_{n_p}(x; q^{n_p})\}$, where each φ_i is a non-negative-valued *Radial Basis Function* (RBF). This basis was chosen because of its large flexibility and localized domain of influence. The parameters prescribing each RBF, q^i, comprise q. Whereas the parameters in $p \in \mathbb{R}^{n_p}$ are to be calibrated, those in $q \in \mathbb{R}^{n_q}$ will be fixed in advance. The values of q will be set empirically based on the data. By making all parametric techniques share the same model structure M, we ensure that the same amount of modeling effort is devoted to all of them. Conversely, a GP is a nonparametric technique for which the prescription of M as in Equation (1), thus, of p are not required. A brief summary of the techniques being compared is presented next.

A. Prediction Intervals (PI)

Parameter estimation is commonly carried out by solving for the parameter realization that minimizes the sum of squared errors between predictions and observations [6]. This approach yields the *Least Squares* (LS) parameter estimate p_{LS}. The precision of this estimate, which prescribes how much it can deviate from its "true value" within an epistemic framework (i.e., the true value of p is fixed and unknown), is often evaluated using prediction intervals and confidence intervals. In linear regression statistics, a prediction interval defines a range of values within which the output is likely to fall given a specified value of the input. Linearly regressed data are by definition non-normally distributed. Normally distributed data are statistically independent of one another whereas regressed data are dependent on x. Because of this dependency, prediction intervals applied to linear regression statistics are considerably more involved to calculate than are prediction intervals for normally distributed data. The uncertainty represented by a prediction interval includes not only the uncertainty/variation associated with the population mean and the new observations, but also the uncertainty associated with the regression parameter p_{LS}. Because the uncertainties associated with the population mean and new observation are independent of the observations used to fit the model, the uncertainty estimates of these three sources are combined. This technique is often applied to computational models having the structure

$$y = M(x, p) + \eta, \tag{2}$$

where the prediction error or residual η is a random variable. This technique often requires (i) assuming a distribution for the prediction error (e.g., η is a zero mean Normal with a fixed variance), (ii) M and η assuming mathematically convenient forms and/or (iii) a nonlinear M being accurately described by a linear approximation. As such, the suitability of the predicted interval strongly depends on the validity of such assumptions. The implementation of the PI method used is based on the MATLAB Statistics toolbox, R2015b.

B. Bayesian Credible Intervals (CI)

A common approach to model calibration is Bayesian inference. This method's objective is to describe the model's parameters as a vector of possibly dependent random variables using Bayes' rule. The resulting vector, called the posterior, depends on an assumed prior random vector, and the likelihood function; which in turn depends on the data in z, and on the structure of the mathematical model M. The *Maximum Likelihood Estimate* (MLE), which is the value of p where the likelihood function is maximal, is searched for by using sampling or optimization. Whereas the application of Bayesian inference to model calibration does not require making any limiting assumptions on the manner in which y depends on p and η nor on the structure of the resulting posterior, it requires that the calibrated variables in p be epistemic. The variables to be calibrated can be a combination of physical epistemic uncertainties and hyper-parameters of aleatory variables[a]. Note that the consideration of aleatory uncertainties requires assuming a structure for them, so they can be parameterized in terms of epistemic variables that can be calibrated. This modeling exercise often requires significant effort and insight into the structure of the DGM.

The presence of aleatory and model-form uncertainty yields calibrated models that might fail to properly describe the prediction error. As more data is available, the posterior approaches a Dirac delta, making the predicted output y converge to a deterministic function of the input. As such, the offset between this function and the data points is not captured by the calibrated model. This deficiency can be mitigated by adding a discrepancy term to M [4]. This term, which might depend on epistemic and aleatory variables, is calibrated along with the parameters of M using Bayesian inference. In our application of this method we will use Equation (2) and make the standard deviation of η an additional calibration variable. This is a common approach that greatly facilitates the evaluation of the likelihood function, thereby substantially reducing computational effort. Moreover, this modification prevents the CI prediction from converging to a function (i.e., an interval with zero spread) as N increases. When the posterior is multimodal or when its support covers a wide range, predictions based on the MLE estimate might not be sufficiently informative. In such cases, CI should be used instead [4]. Credible intervals result from sampling the posterior using *Markov Chain Monte Carlo* (MCMC) sampling. Despite its high computational demands, the need for including a discrepancy term, and of the potentially high sensitivity of the posterior to the assumed prior, Bayesian Calibration is commonly regarded as a benchmark in model calibration. The implementation of the CI method used in this study uses elements of the MATLAB Statistics toolbox R2015b, and non-commercial code developed by the authors.

C. Interval Prediction Models (IPM)

IPMs [1, 3, 2] are metamodels that cast the spread in the data as interval uncertainty in the parameters p of the computational model $M(x, p)$ without the need for a discrepancy term such as η in (2). IPMs are calculated via optimization. When the model M depends linearly on the parameters, the corresponding optimization program is convex, thus, numerically efficient to solve. The developments in [2] focus on models depending linearly on the

[a] For instance, if q contains the physical parameters of the model M, where q_1 is epistemic and q_2 is aleatory having a Gaussian distribution with zero mean and standard deviation σ, the vector $[q_1, \sigma]^\top$ of variables to be calibrated contains two epistemic variables, one physical and one non-physical.

parameters and arbitrarily on the input. This setting enables the calculation of IPMs that admit a rigorous description for the range of the predicted output, and of the uncertainty in the model's parameters. Furthermore, IPMs enable a rigorous bounding of the probability that a future observation will fall within the predicted interval. This certificate of correctness does not require making any assumption on the underlying structure of the DGM. This is a substantial benefit of IPMs over all other metamodeling techniques.

D. Gaussian Process (GP) Models

GP models [5] are the only non-parametric approach studied in this article. A GP is a stochastic process in which the predicted output is a normal random variable having a mean and a covariance that depend on the input. Moreover, every finite collection of those random variables has a multivariate normal distribution. The distribution of a Gaussian process is the joint distribution of all those (infinitely many) random variables, and as such, it is a distribution over functions. The calculation of GPs requires prescribing functional forms for the mean and the covariance functions. The hyper-parameters of these functions are often calibrated using a Bayesian framework, for which priors must be prescribed. Techniques such as MLE and the *Maximum A Posteriori* (MAP) estimate, which is the mode of the posterior distribution, are commonly used techniques for setting the value of the hyper-parameters of the covariance function. In spite of the high computational cost of GPs, which restricts their usage to a few thousand data points, and the high sensitivity of the prediction to the assumed mean and covariance structures, GPs are widely used due to their ability to (i) characterize complex functional relationships, and (ii) to account for the variation in the predicted output caused by the distribution of the data; e.g., predictions away from the data exhibit a larger predicted variance. The GP method was implemented using the software package *GPML* version 3.5 developed by Rasmussen et al., and available through the url `http://www.gaussianprocess.org/#code` for a zero mean function and a squared exponential covariance function.

II. Figures of Merit

Each of the metamodeling techniques above will be used to compute an interval-valued function of the input. This function will be denoted as $I(x) = [\underline{y}(x), \bar{y}(x)]$, where $\underline{y}(x)$ and $\bar{y}(x)$ are the lower and the upper limit of the interval respectively. In the case of PI, CI, and GP we will use the confidence level $\alpha = 0.01$ (i.e., an output range bounded by the 0.5 and 99.5 quantiles, so we expect them to contain 99% of the probability), whereas for the IPMs we will neglect the worst $100\alpha\%$ of the data (See [1] for techniques to detect and outliers). This will make all predictions comparable[b]. The following figures of merit will be used for

[b] In this study IPMs will be used to approximate the tightest interval value function of the input for which

$$\mathbb{P}\left[\cup_{x\in X}\ x \times I(x)\right] = 1 - \alpha. \tag{3}$$

This will enable comparing the IPM predictions with those generated based on a $1 - \alpha$ confidence interval. To this end, we will first calculate an IPM based on the full set of N observations. We then identify the $100(1 - \alpha)$ percent of the observations falling in the inner most portion of the predicted interval. This can be efficiently done by removing the observations attaining the largest values of ρ (See [1]) such that the removed set is uniformly distributed over X. A second IPM based on the retained observations will then evaluating the metamodels performance:

1. The *expected spread* of the interval prediction corresponding to metamodel $\mathcal{M}$, $\Delta(\mathcal{M})$, is defined as

$$\Delta(\mathcal{M}) = \frac{1}{\text{Vol(X)}} \int_X \delta(x) dx, \tag{5}$$

where Vol$(\cdot)$ is the volume operator, and $\delta(x) = \bar{y}(x) - \underline{y}(x)$ is the *spread.* $\Delta(\mathcal{M})$ is the expected value of $\delta(x)$ corresponding to a uniform distribution over the input range. A smaller expected spread denotes a more informative global prediction whereas a small spread denotes a more informative local prediction.

2. The *unreliability* of metamodel $\mathcal{M}$, $\gamma(\mathcal{M})$, is defined as

$$\gamma(\mathcal{M}) = \text{Prob}_{\mathbb{P}} \left[\{x, y\} \notin [x, I(x)] \right]. \tag{6}$$

Hence, $\gamma(\mathcal{M})$ is the probability that a future observation will fall outside the predicted interval. We chose the unreliability over the reliability, which is the probability of falling within the predicted interval, so different orders of magnitude can be clearly plotted in a logarithmic scale. When the prediction matches the DGM exactly we obtain the ideal value $\gamma(\mathcal{M}) = \alpha$. Note, however, that attaining such a value does not necessarily imply that the predicted interval coincides with the 'true' interval. The true interval-valued function, denoted by

$$I_{\text{true}}(x) = [\underline{y}_{\text{true}}(x), \bar{y}_{\text{true}}(x)], \tag{7}$$

is the output range between the $\alpha/2$ and the $1-\alpha/2$ quantiles of the DGM[c]. To evaluate how well $I(x)$ approximates $I_{\text{true}}(x)$, we use the *error* metrics introduced next.

3. The *error* incurred by approximating the upper limit of $I_{\text{true}}(x)$ is

$$\bar{e}(\mathcal{M}) = \int_X (\bar{y}_{\text{true}}(x) - \bar{y}(x))^2 dx. \tag{8}$$

4. The *error* incurred by approximating the lower limit of $I_{\text{true}}(x)$ is

$$\underline{e}(\mathcal{M}) = \int_X (\underline{y}_{\text{true}}(x) - \underline{y}(x))^2 dx. \tag{9}$$

5. The CPU *time*, $t(\mathcal{M})$, is the time required to generate $\mathcal{M}$ and evaluate $I(x)$ at the set of input values x_e.

be calculated. An upper bound of the probability of falling outside the predicted interval, whose boundaries approximate the $\alpha/2$ and 1-$\alpha/2$ quantiles of the DGM, is given by:

$$\gamma(\mathcal{M}) = \epsilon\left(\beta, \lfloor \alpha N \rfloor, d, N\right), \tag{4}$$

where $\epsilon(\beta, k, d, N)$ is given in Campi [1], β is the confidence parameter, k is the number of outliers, the $\lfloor \cdot \rfloor$ operator is the greatest integer less than or equal to its argument, and d is the number of design variables used to calculate the IPM.

[c] The true interval-valued function will be estimated by exhaustive sampling of the DGM in all but the DGM 4 case, in which an analytical form is available.

6. *Containment* of the data by the predicted interval is evaluated by

$$\mu(\mathcal{M}) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{I}[y_i \in I(x_i)], \tag{10}$$

where $\mathcal{I}[\cdot]$ is the indicator function. The value $\mu(\mathcal{M}) = 1$ indicates the predicted interval contains all N data points.

7. *Conservatism*, which degrades the informative value of the predicted interval by making it overly wide, can be evaluated using

$$\nu(\mathcal{M}) = \sum_{i=1}^{N} \min s(x_i, y_i)\mathcal{I}[y_i \in I(x_i)] + \eta \max s(x_i, y_i)\mathcal{I}[y_i \notin I(x_i)], \tag{11}$$

where $s(x_i, y_i) = [|\overline{y}(x_i) - y_i|, |\, y_i - \underline{y}(x_i)|] \in \mathbb{R}^2$, and $\eta > 1$ is a parameter used for penalizing intervals not containing the data. Whereas an observed output falling within the predicted interval yields a summand ranging between zero and one half of the spread at the corresponding input, a predicted interval not containing the output yields a summand greater than the spread. As such, $\nu(\mathcal{M})$ assumes large values when the predicted interval is overly wide and it fails to contain all the data.

Note that $\gamma(\mathcal{M})$, $\bar{e}(\mathcal{M})$, and $\underline{e}(\mathcal{M})$ evaluate the degree to which a prediction based on N observations conforms to the DGM as $N \to \infty$. Naturally, these metrics cannot be calculated in most practical applications due to the inability to exhaustively evaluate the DGM. Even though $\mu(\mathcal{M})$ and $\nu(\mathcal{M})$ are metrics of practical interest, the numerical experiments that follow focus on $\Delta(\mathcal{M})$, $\gamma(\mathcal{M})$, $\bar{e}(\mathcal{M})$, $\underline{e}(\mathcal{M})$ and $t(\mathcal{M})$ only.

Notice that metamodels for which $I_{\text{true}}(x) \subset I(x)$ might yield values for $\gamma(\mathcal{M})$ close to the targeted value α even though $I(x)$ might well be a poor representation of $I_{\text{true}}(x)$. This anomaly however will yield large values for $\bar{e}(\mathcal{M})$, and $\underline{e}(\mathcal{M})$. For instance, if $I(x) = 10I_{\text{true}}(x)$, $\gamma(\mathcal{M})$ might be close enough to α to regard $\mathcal{M}$ acceptable. However, the large values of $\bar{e}(\mathcal{M})$, and $\underline{e}(\mathcal{M})$ will indicate and overly conservative prediction.

III. Numerical Experiments

Four numerical experiments were conducted each based on a different DGM. Each DGM targets a particular phenomenon of interest. DGM1 is an unknown deterministic function (i.e., the probabilistic cloud $\mathbb{P}$ has zero output variance and an unknown mean), DGM2 is a random process in which the aleatory component of the measured output depends on the value of the input, DGM3 is a Gaussian random process with unknown mean and variance functions, and DGM4 is a random process requiring extrapolation beyond the range of observations. The structure of the DGMs will remain unknown to the metamodeler who will only have the data points extracted from it to build a surrogate.

For each experiment we have several ensembles of observations, each corresponding to a different value of N. Denote by E_1, E_2, … E_{n_e} data ensembles corresponding to N_1, N_2, …, N_{n_e} observations respectively, where $N_i < N_{i+1}$ for $i = 1, \ldots, n_e - 1$. The values of N chosen are selected arbitrarily. The ensembles are constructed such that the data points in E_i are also in E_j for all $j > i$. The data in each ensemble will be used to generate the family of metamodels $\mathcal{M}_1$, $\mathcal{M}_2$, $\ldots$, $\mathcal{M}_{n_e}$ for each metamodeling technique. The figures of merit Δ, r, $\bar{e}$, $\underline{e}$ and t will be used for analysis and comparison. The dependency on N will be studied by comparing members of the same family, whereas the dependency on the metamodeling technique will be studied by comparing members of different families for the same data ensemble.

Recall that all parametric methods will assume the same model structure M. As such, we are devoting the *same amount of modeling effort* to all of them. In particular, M is assumed to be a linear combination of $n_p = 10$ Gaussian RBFs:

$$y = M(x, p, q) = \sum_{i=1}^{10} p_i e^{-\left(\frac{x-\mu_i}{\sigma_i}\right)^2}. \tag{12}$$

The centers μ_i and spread parameters σ_i constituting q^i will be fixed in advance, so only the parameters in p will be calibrated (with the exception of CIs for which an additional calibration variable is used to prescribe a discrepancy term, as explained in Section I.B.).

A. Data Generating Mechanism 1 (DGM1)

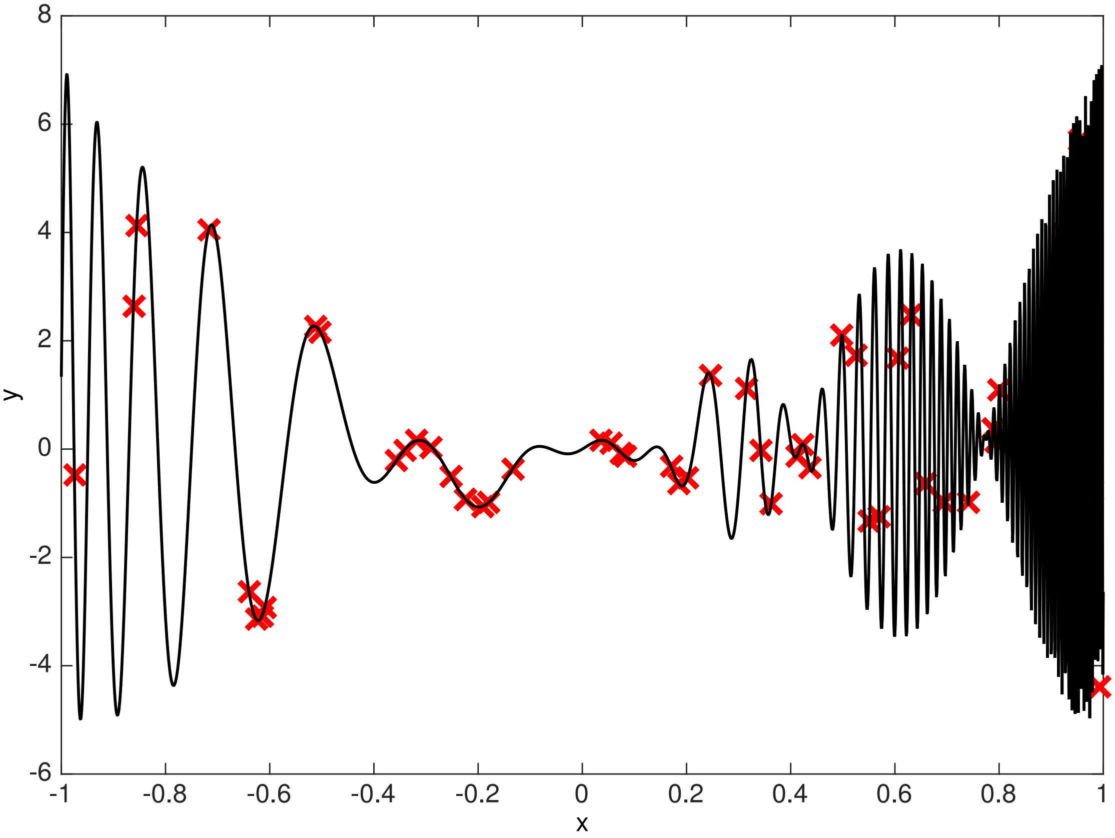

Figure 1. DGM1 and $N_1 = 50$ data points ($\times$).

Figure 1 shows DGM1, which is a deterministic function of the input, in the range $X = [-1, 1]$. Hence, all percentile curves are equal and $I_{\text{true}}(x)$ has zero width. A total of $n_e = 12$ data ensembles corresponding to the N values in $\{50, 100, \ldots, 600\}$ were obtained from DGM1. The points corresponding to first ensemble are shown as red crosses.

Figure 2 shows the predictions resulting from applying all metamodeling techniques for the $N_1 = 50$ case. The PI is shown in cyan, the CI in magenta, the IPM in blue, and the GP in green. The same color conventions apply throughout this paper. Note that the limits of the IPM contain $\lfloor 0.99 N_1 \rfloor$ observations by design whereas the other metamodels do not. Further notice that all metamodels capture most of the observations with a varying degrees of tightness. Figure 3 shows the predictions corresponding to $N_{12} = 600$. The comparison between Figures 2 and 3 illustrates the convergence properties of each metamodeling technique as the number of observations increases. Whereas the range of the IPM increased, that for all other methods decreased. The IPM provides the tightest interval value function of x containing 99 percent of the data. Note that the CI and GP prediction improved considerably as N increased becoming more tight and less oscillatory globally. Further notice that the GP matches well the oscillatory behavior of the DGM for $x < 0$ but it fails to do so for $x > 0$. The predictions corresponding to all parametric methods can be made tighter by choosing a better structure for M, say, through the selection of better q values, whereas improvements to the GP prediction require assuming more suitable mean and covariance functions, and

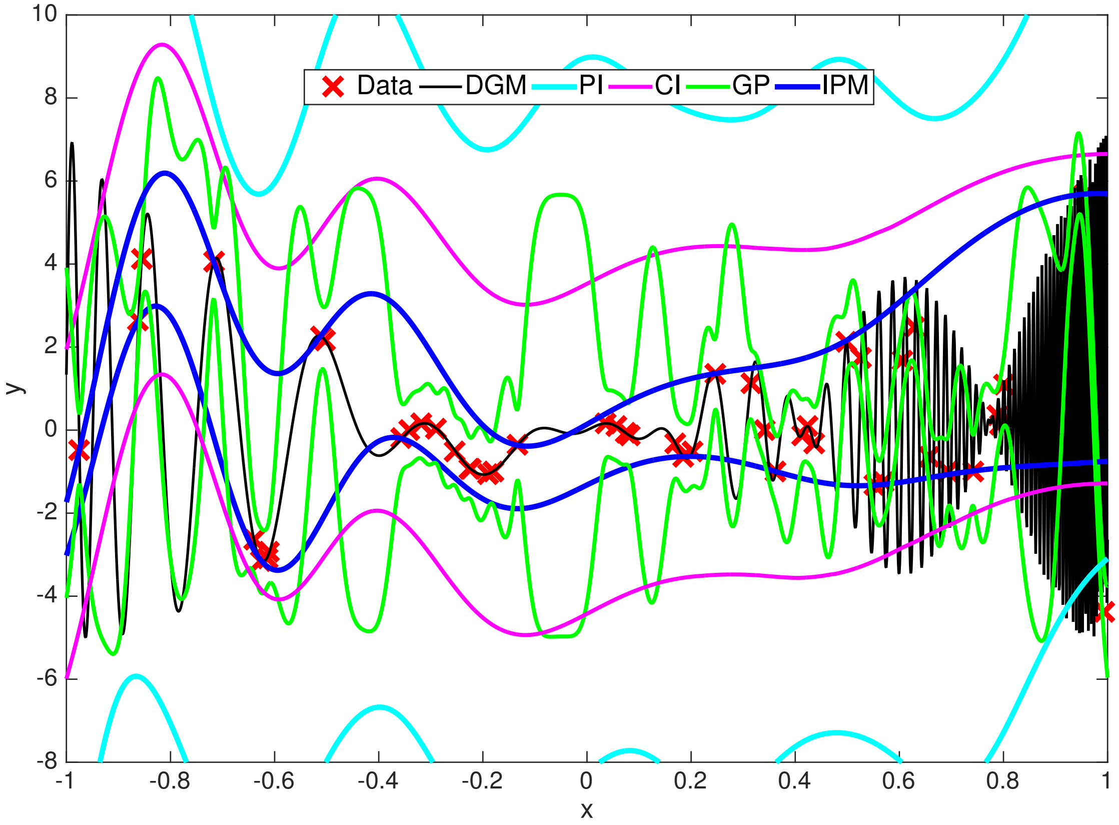

Figure 2. DGM1: predictions for $N_1 = 50$ observations.

corresponding priors. The latter task is far more cumbersome and non-intuitive than the former one.

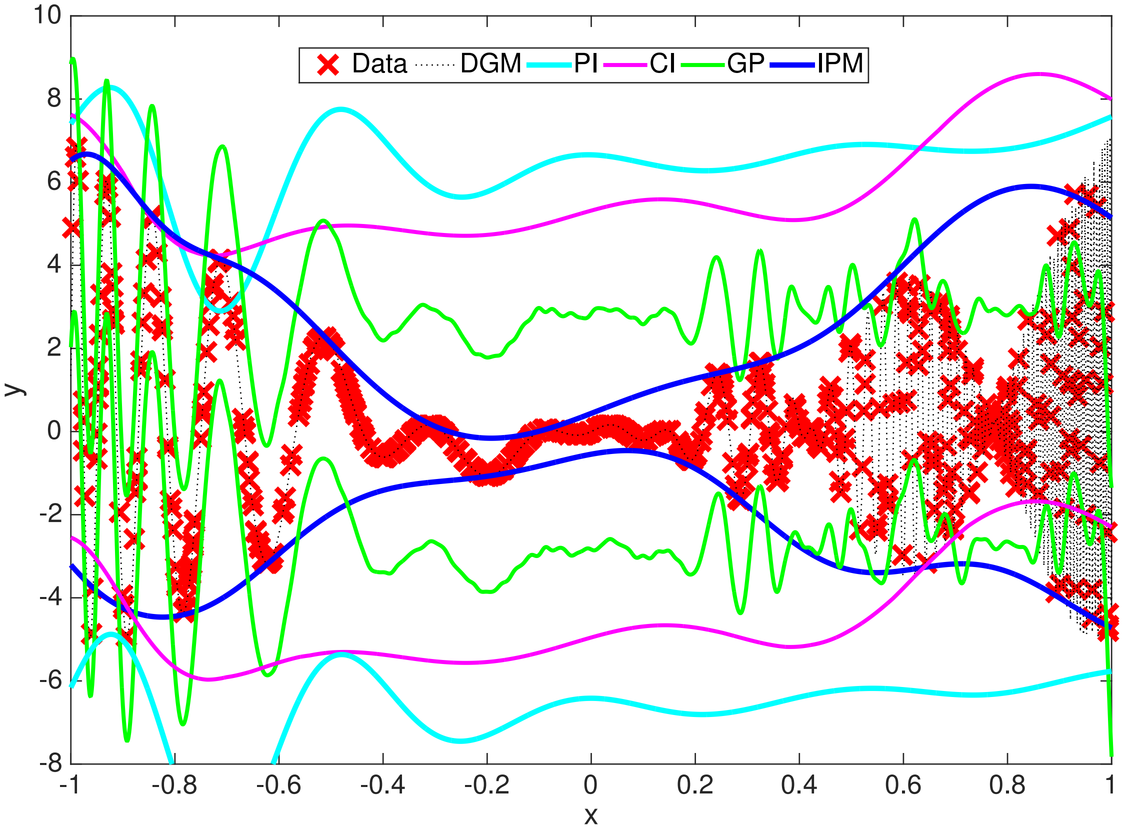

Figure 3. DGM1: predictions based on $N_{12} = 600$ observations.

A comparison of the metamodeling techniques above using the figures of merit in Section II is presented next. Recall that the expected spread Δ evaluates the tightness of the predicted interval globally. Figure 4 shows the dependency of Δ on N. The IPM attains the smallest expected spreads consistently, whereas CIs and GPs exhibit a wider range of variation. In the case of the CI, this behavior occurs even though the Markov Chain Monte Carlo (MCMC) sampling of the posterior appears to converge. The curve corresponding to the IPM is generally non-decreasing by virtue of including the data from one ensemble into the larger ensembles (it would be strictly non-decreasing for $\alpha = 0$). The predictions corresponding to PI and CI are highly conservative for most input values, i.e., they yield overly wide output ranges deviating considerably from the intrinsic aleatory variation of the DGM (See Figures 2 and 3). Conservatism can be evaluated by using the figures of merit μ and ν introduced earlier.

Figure 5 shows the error in the upper and lower limits of the predicted interval. Recall that these errors measure the discrepancy between a prediction based on N observations and the true quantiles of DGM. As such, we can only hope for the errors to converge to zero asymptotically. The IPMs and GPs yield the best approximations to $I_{\text{true}}(x)$. These two methods also exhibit the good convergence properties.

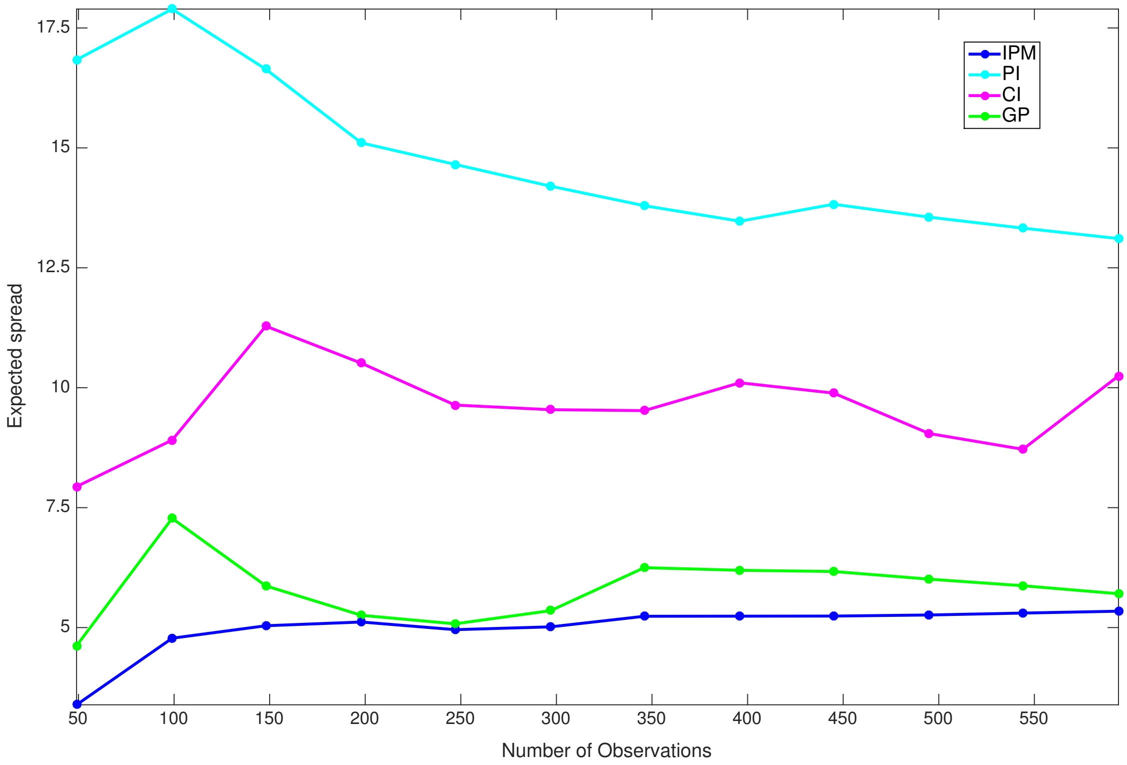

Figure 4. DGM1: expected spread Δ vs. number of observations.

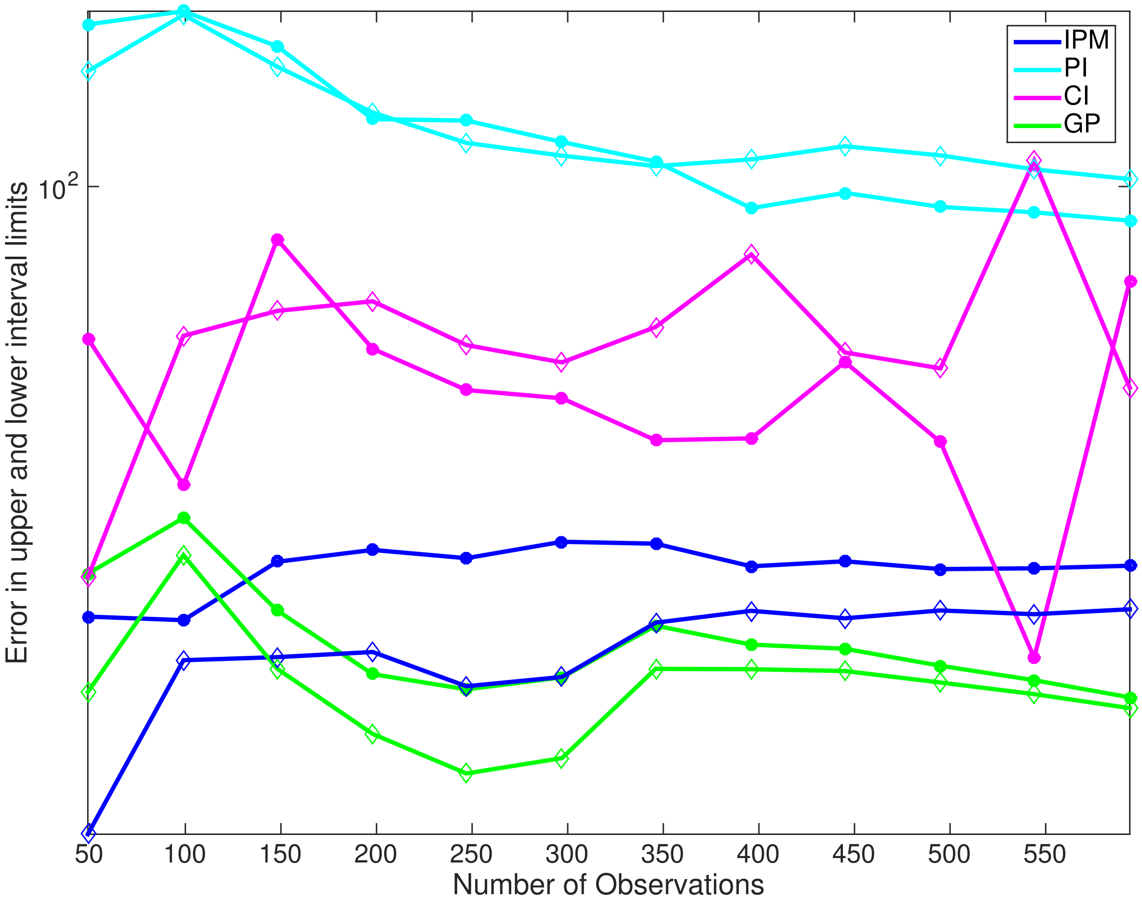

Figure 5. DGM1: Error in the upper limit $\bar{e}$ (circle), and lower limit $\underline{e}$ (diamond).

Figure 6 shows the CPU time required by each metamodeling technique. Bayesian methods, which are the most time consuming, are based on a random walk with a fixed number of steps. As such, the increment in the CPU time is solely caused by having to perform more evaluations of the likelihood function. The computational cost of GP grows the fastest. This growth, which is mainly a result of a matrix inversion that scales as $\mathrm{O}(N^3)$, restricts its usage for non-sparse formulations to relatively low values of N (i.e., a few thousands). The time requirements for the IPM, which entails solving a convex optimization problem, grows polynomially with N. As such it can handle problems with hundreds of thousand of data points. Lastly, PIs only require solving a linear least squares optimization problem and using statistical assumptions on the DGM to yield a prediction. As such, their computational demands are the lowest.

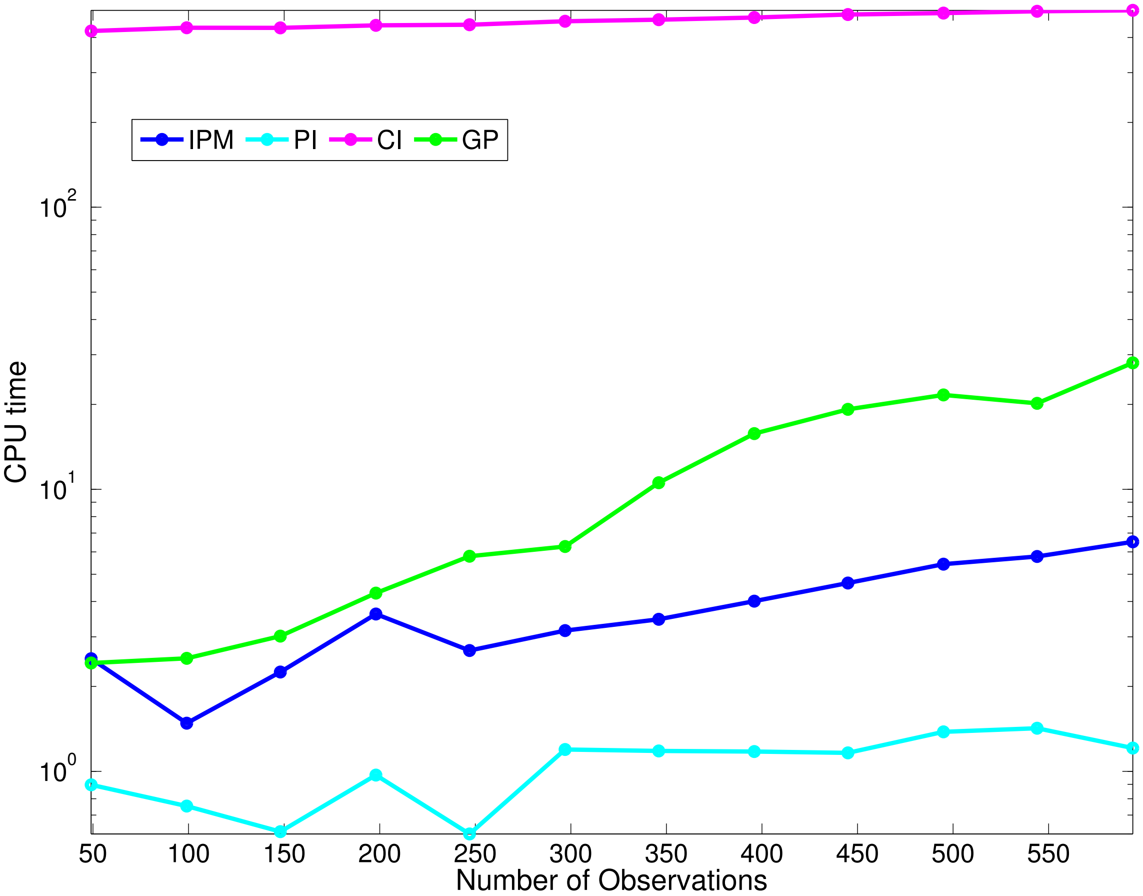

Figure 6. DGM1: Dependency of the CPU time t in seconds on number of observations.

Finally, Figure 7 shows the unreliability defined in (6). Recall that the theoretical value for the unreliability is equal to the confidence level $\alpha = 0.01$. Whereas the value of γ for all metamodels is calculated empirically by sampling the DGM and evaluating the probability of being outside the predicted interval, IPMs enable the calculation of an analytical upper bound to γ. This upper bound enables making an assessment on the reliability of the prediction solely based on the data available (i.e., we don't need to (i) use additional data for validation, (ii) assume a distribution for the data, (iii) make asymptotic claims requiring $N \rightarrow \infty$). As such, this is a *distribution-free* result (i.e., we do not need to know/assume anything about the probability distribution $\mathbb{P}$ governing the DGM) which is applicable to all data ensembles regardless of the value of N. This upper bound is shown as a blue dashed line. Note that both the upper bound and the empirical estimate of the IPMs unreliability decrease monotonically with N, and that such a bound is a close approximation of the actual unreliability. The empirical calculation of γ is often infeasible in practice due to the high cost of obtaining a sufficiently large sample set of the DGM. As such, the ability to calculate a tight reliability upper bound without having to make any assumptions on $\mathbb{P}$ constitute a significant benefit of IPMs over all other methods.

As before, the unreliability of the CI varies irregularly whereas the conservative character of the PI prediction makes it to be much closer to the theoretical value than the other methods. The small value of γ is the result of over-bounding the true 99% quantile curves. As such, this is not an indication of goodness but instead of a large amount of conservatism in the prediction, as shown in Figure 3.

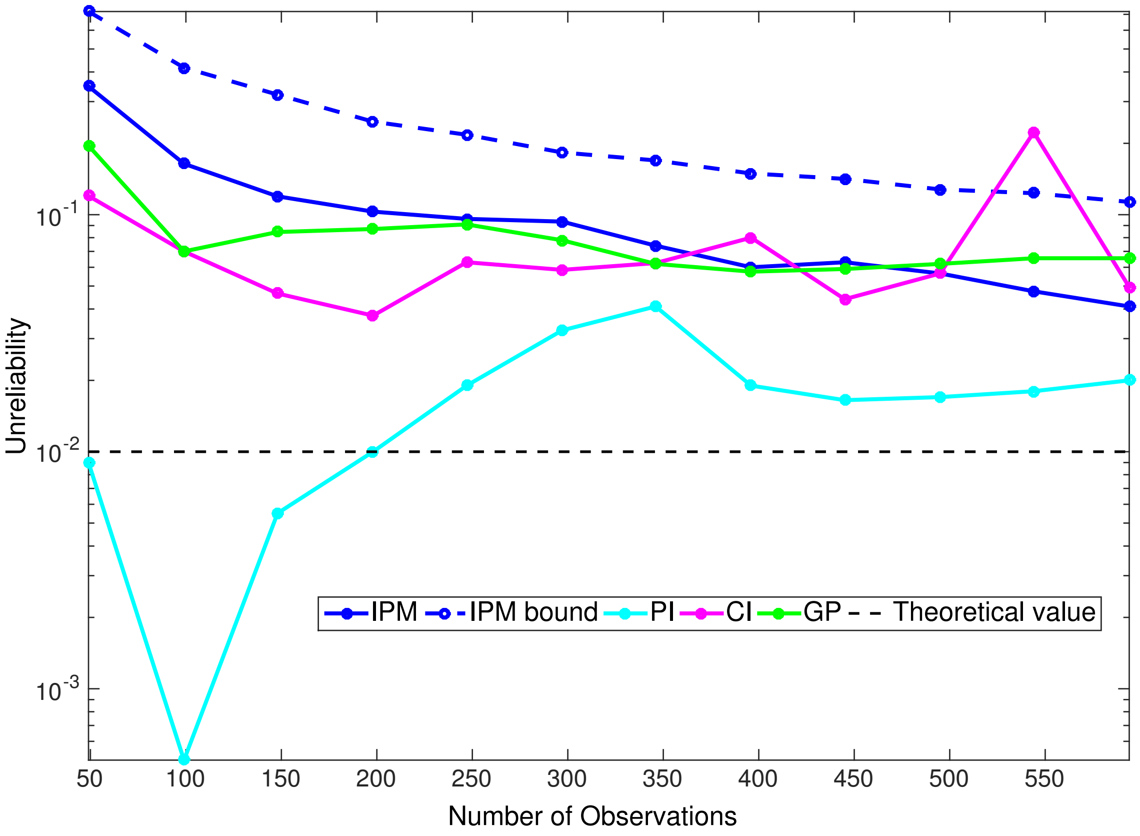

Figure 7. DGM1: Dependency of the unreliability $1 - r$ on number of observations.

B. Data Generating Mechanism 2 (DGM2)

Figure 8 shows two-percentiles curves of DGM2 in the input range $X = [-1, 1]$. This figure is obtained by performing a large Monte Carlo simulation of the DGM. Note that the aleatory component of the output y depends on the value of the input x. The output y converges to a deterministic function as x decreases, whereas it becomes a strongly skewed random process with input dependent moments as x increases. As such, metamodeling techniques assuming that the output depends additively on aleatory variables having a input-independent structure will incur significant error (e.g., the structure of the discrepancy term used by CIs and the underlying structure of GP models). DGM2 is the result of a process in which the output y is a deterministic function of two variables: the input x shown and an input z absent from the "experiment" used to collect the data. As such, the perceived spread in the measured outputs is not caused by measurement noise or aleatory variability in y but instead it is the result of model-form uncertainty.

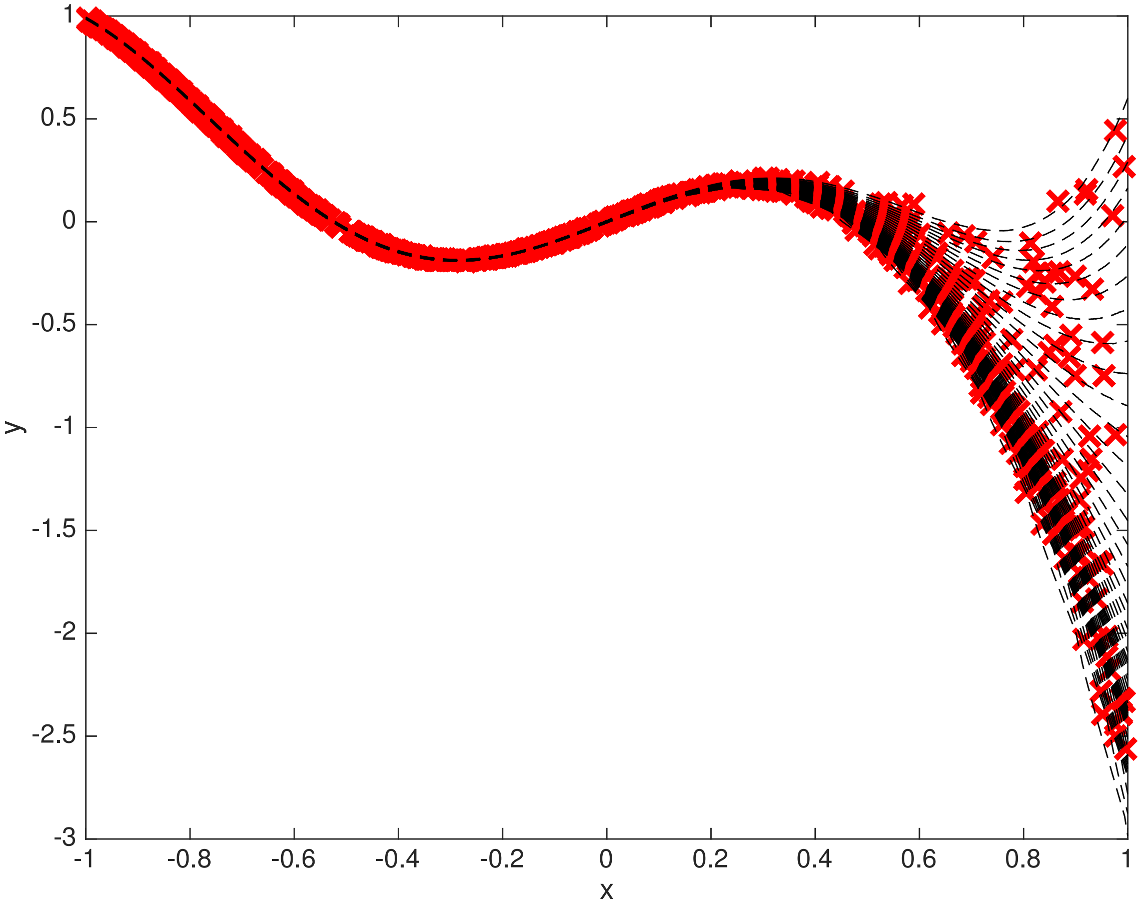

Figure 8. DGM2: Two percentile lines and $N_1 = 1000$ data points ($\times$).

A total of $n_e = 12$ data ensembles corresponding to values of N in $\{1000, 2000, \ldots, 12000\}$ were obtained from DGM2. Figure 8 also shows the data corresponding to the first ensemble. Figures 9 and 10 show the interval predictions corresponding to $N_1 = 1000$ and $N_{12} = 12000$.

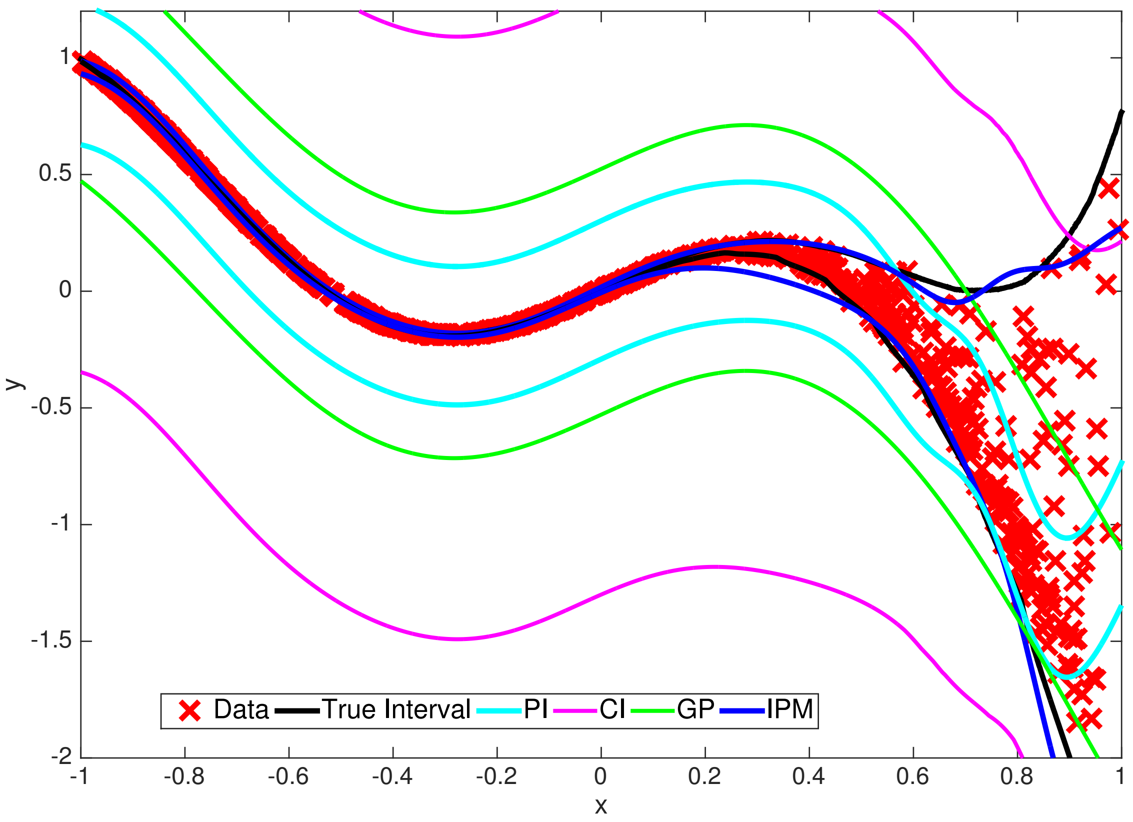

Figure 9. DGM2: predictions for $N_1 = 1000$ observations.

An empirical approximation to the limits of $I_{\mathrm{true}}(x)$, calculated based on a large Monte Carlo sample of DGM2, is shown as a thick black line. Note that the IPM provides the best

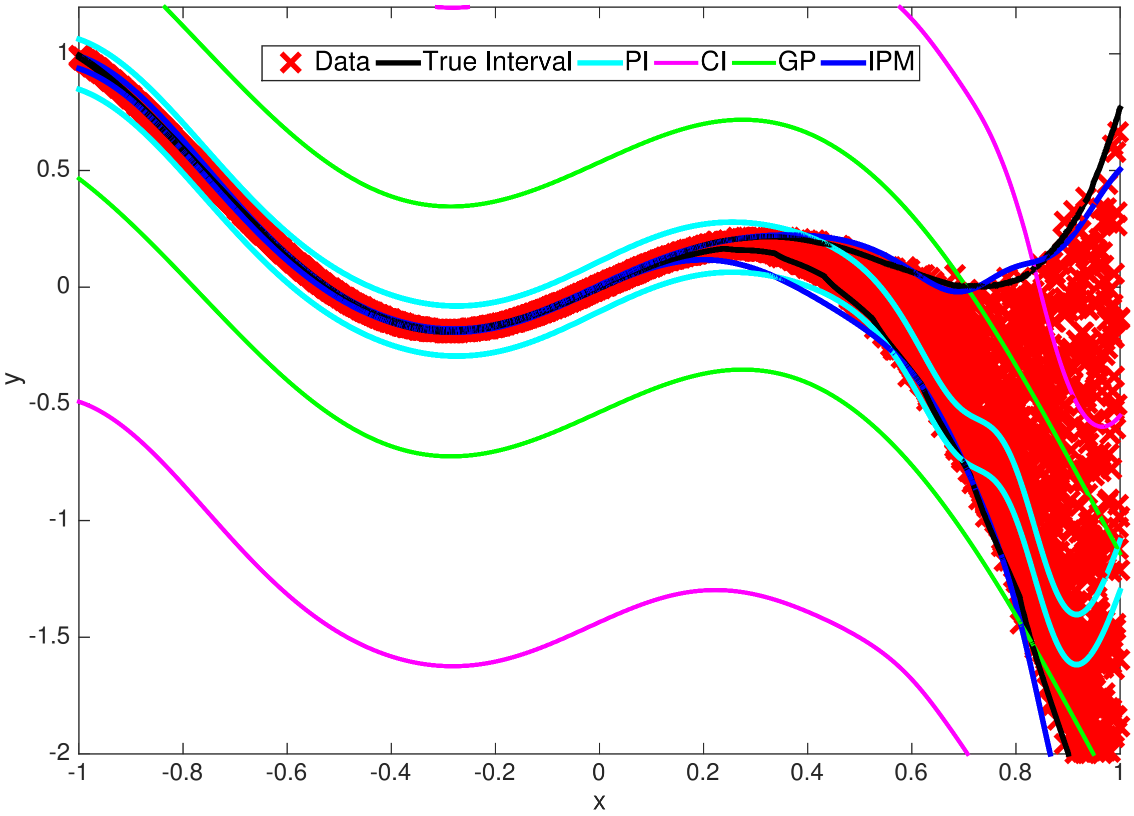

Figure 10. DGM2: predictions for $N_{12} = 12000$ observations.

description of the true interval throughout X. All other methods fail to discriminate between these two regions. Even though the range of the PI decreases as N increases, the converged interval grossly misrepresent the DGM. This is apparent near $x = 0.85$, where the midpoint of the predicted interval diverges from the high-probability region. The predictions resulting from CI and GP are poor representations of the DGM. Both methods describe the overall trend of the data acceptably well (i.e., they contain a region where most of the probability is), but not their input-dependent spread. Whereas CI yields the most conservative prediction, the GP method rapidly becomes insensitive to additional data. Further notice that the GP kind of traces the region in Y where most of the probability is (the skewness of y is a positive increasing function of x).

A comparison based on the expected spread, not shown, and the errors, shown Figure 11, indicates that IPMs yields the best approximation. Note that all methods but the CI settle fairly rapidly, and values corresponding to the first member of the ensemble are not much different from those corresponding to the last one. Finally, Figure 12 shows the unreliability as a function of N. As before, IPM provides unreliability estimates that are the closest to the theoretical prediction and the corresponding upper bound is tight. Note that the GP prediction corresponding to the first ensemble is practically equal to that for the last

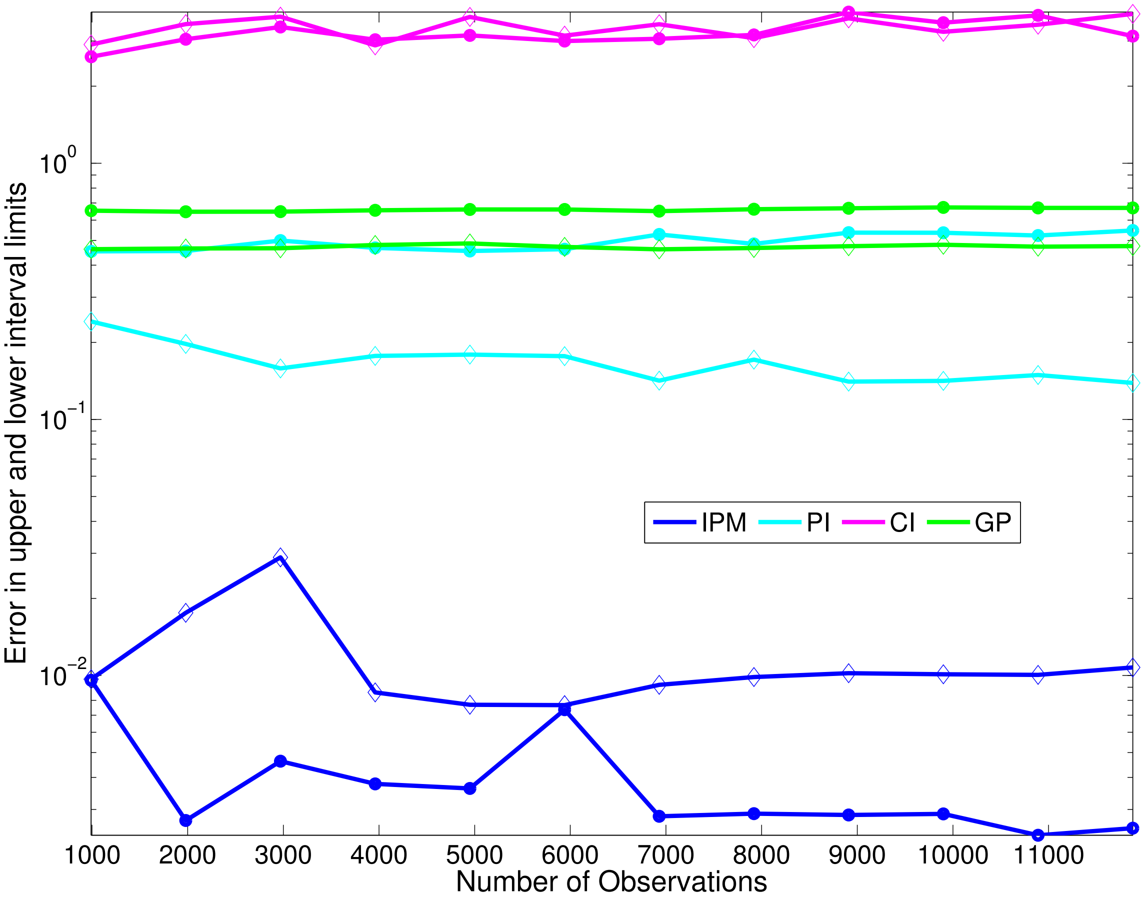

Figure 11. DGM2: Error in the upper limit $\bar{e}$ (circle), and lower limit $\underline{e}$ (diamond).

ensemble. As such, the substantial increment in computational effort required to calculate and evaluate the GP model (data not shown) is not justified by the corresponding changes in the prediction.

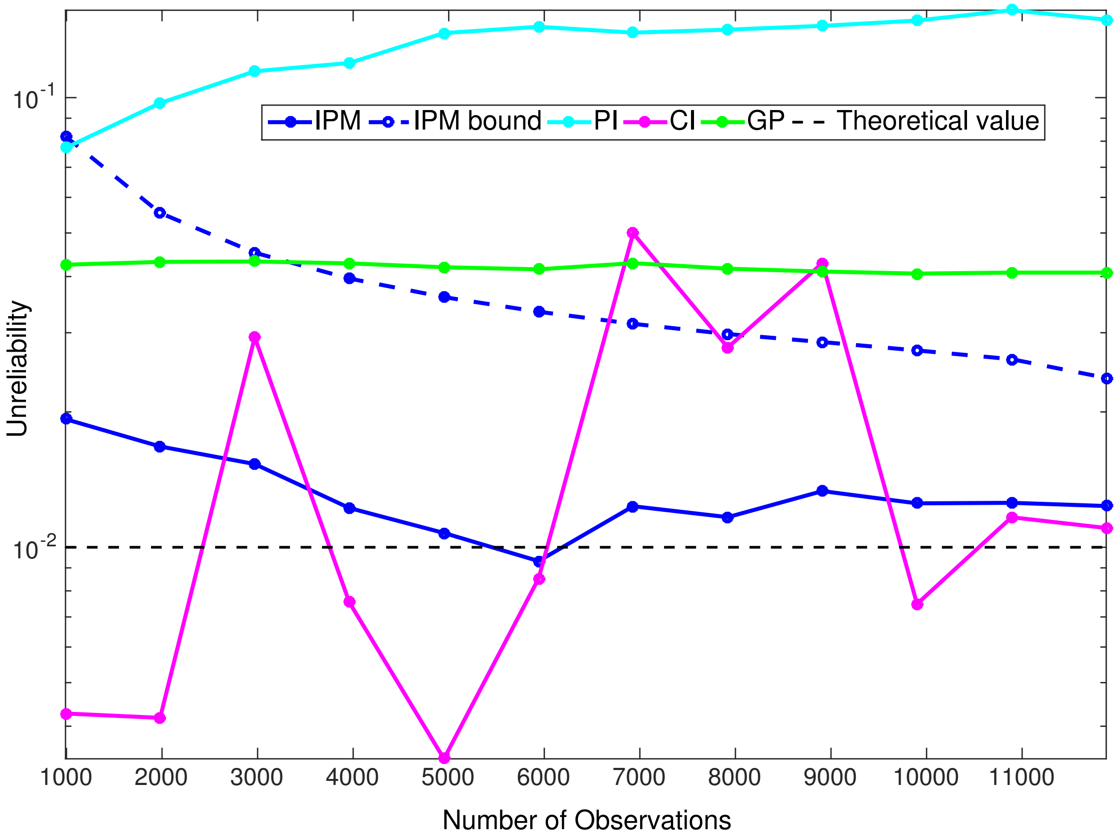

Figure 12. DGM2: Dependency of the unreliability $1 - r$ on number of observations.

C. Data Generating Mechanism 3 (DGM3)

Figure 13 shows two-percentile curves of DGM3 in the input range $X = [-1, 1]$. This is a Gaussian random process admitting analytical expressions for the mean and variance functions. Note that both of these functions vary with x in a nonlinear fashion. A total of $n_e = 11$ data ensembles corresponding to values of N in the set $\{25, 50, 75, 100, 250, 500, 750, 1000, 5000, 10000, 25000\}$ were extracted from DGM3. Figures 14 and 15 show the observations, the true interval value function, and the predictions corresponding to $N_6 = 500$ and $N_9 = 5000$ respectively. Note that the range of the predicted interval for all methods, with the notable exception of IPMs, is fairly independent on the value of x. The CI and IPM predictions change as more data points are available, whereas the PI and GP predictions are practically insensitive. As before, the IPM prediction is the closest to the true interval-valued function.

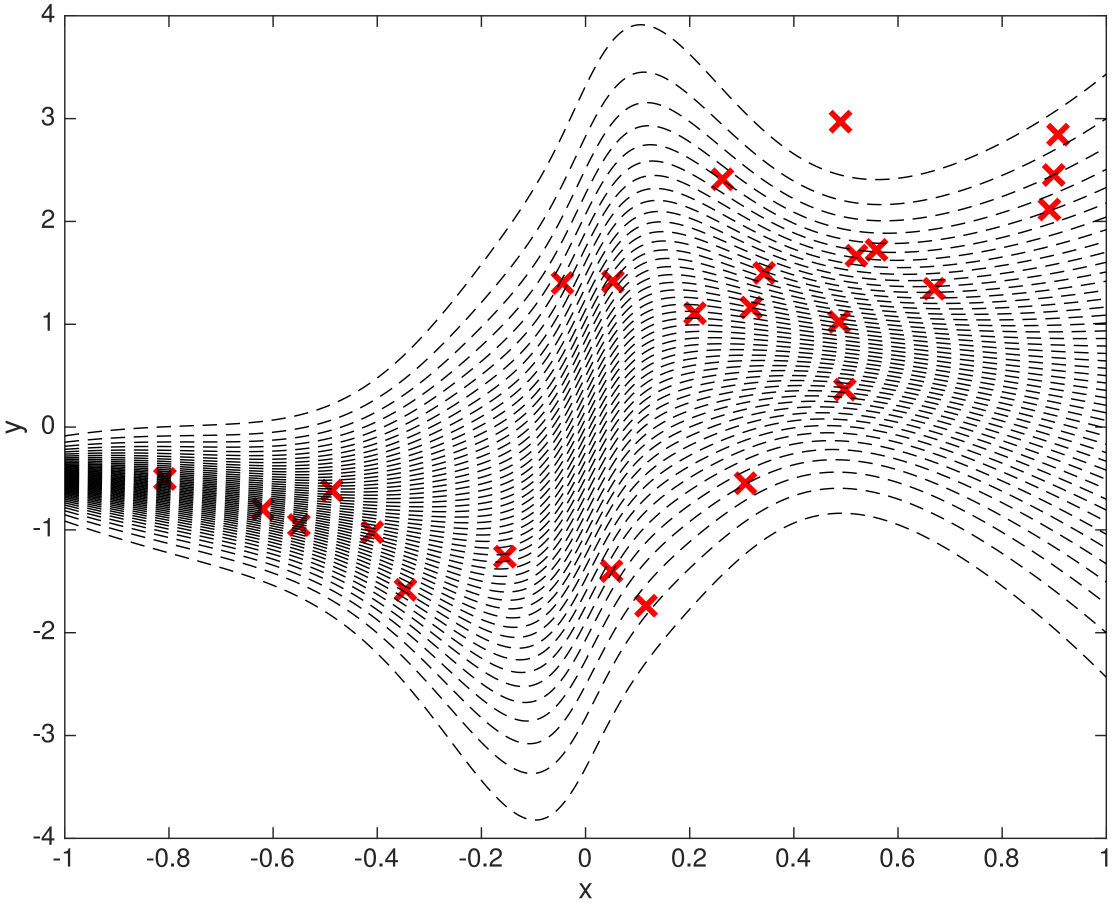

Figure 13. DGM3: Two percentile lines and $N_1 = 25$ data points ($\times$).

Figure 16 shows the error in the upper and lower limits of the predicted intervals as a function of N. The errors attained by the IPM predictions are the smallest of all methods by a sizable margin. At $N_{11} = 25000$ the IPM errors are at least one order of magnitude smaller than the rest. Notice that the GP fails to describe the DGM well (e.g., the predicted variance

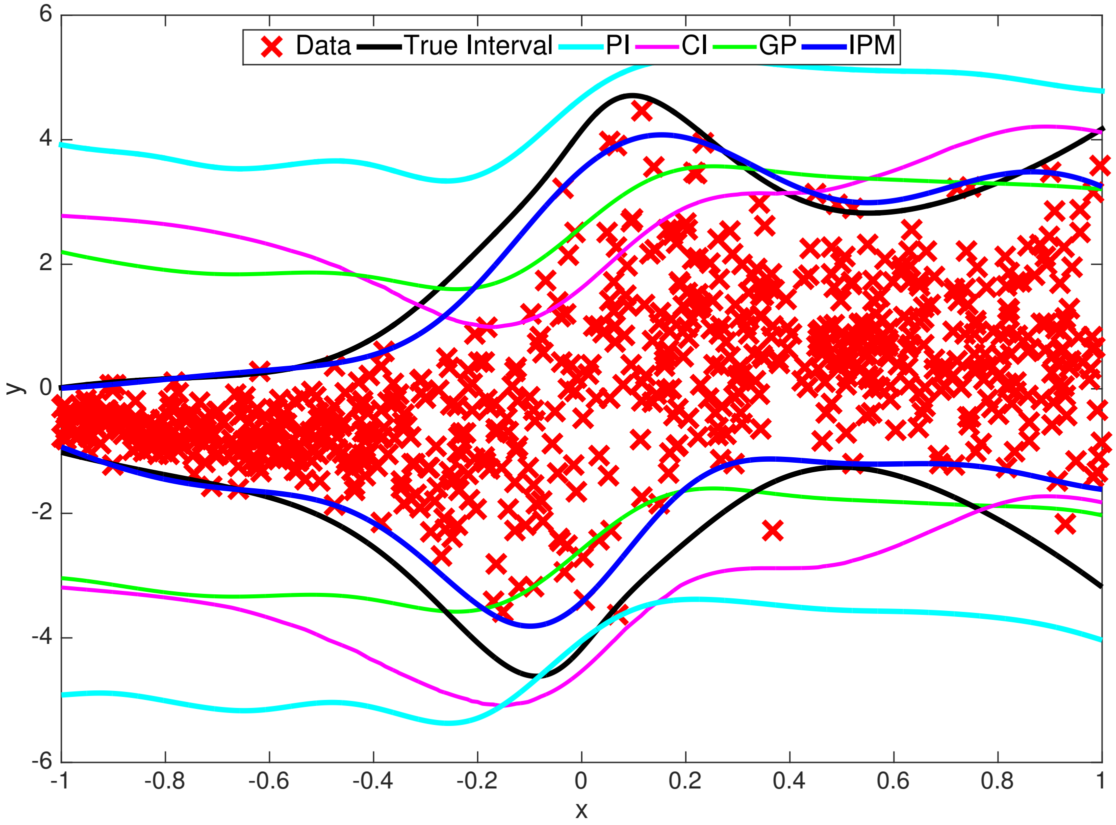

Figure 14. DGM3: predictions for $N_6 = 500$ observations.

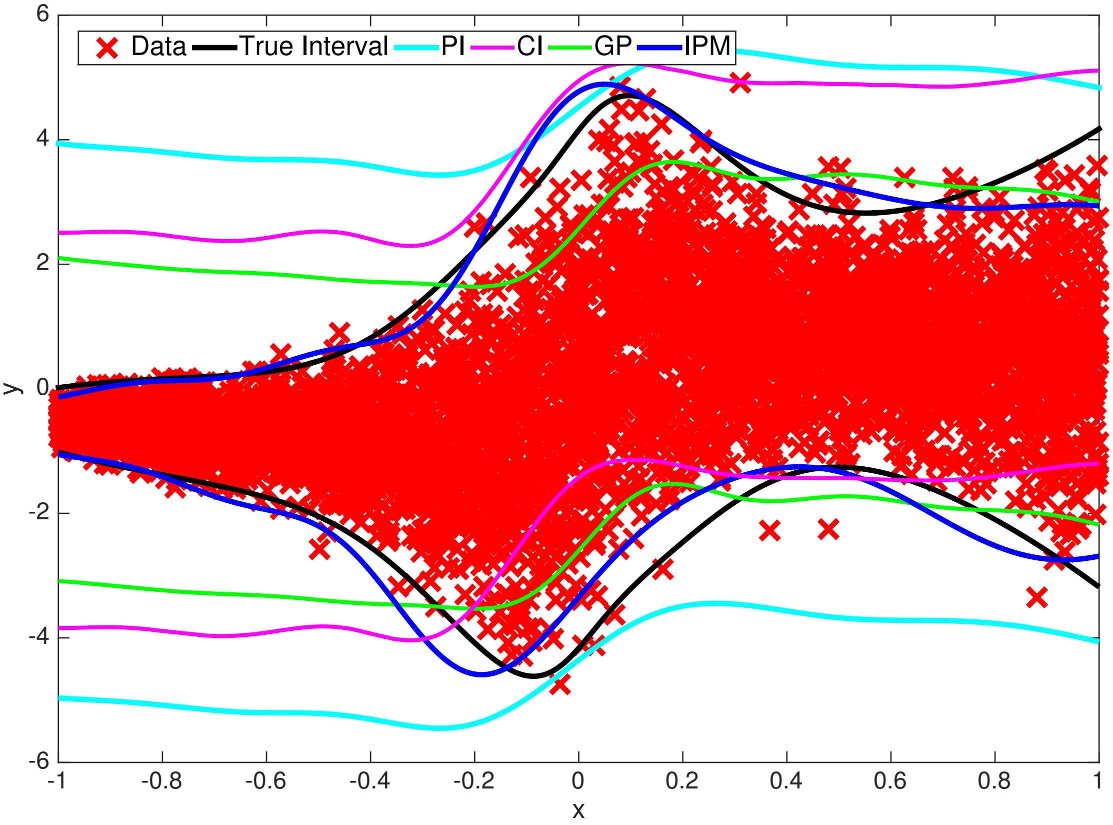

Figure 15. DGM3: predictions for $N_9 = 5000$ observations.

becomes excessively large as x decreases) even though both the DGM and the metamodel are Gaussian processes. The error points missing in the GP curve are the result of the high computational demands of the method, which prevented calculating the GP model.

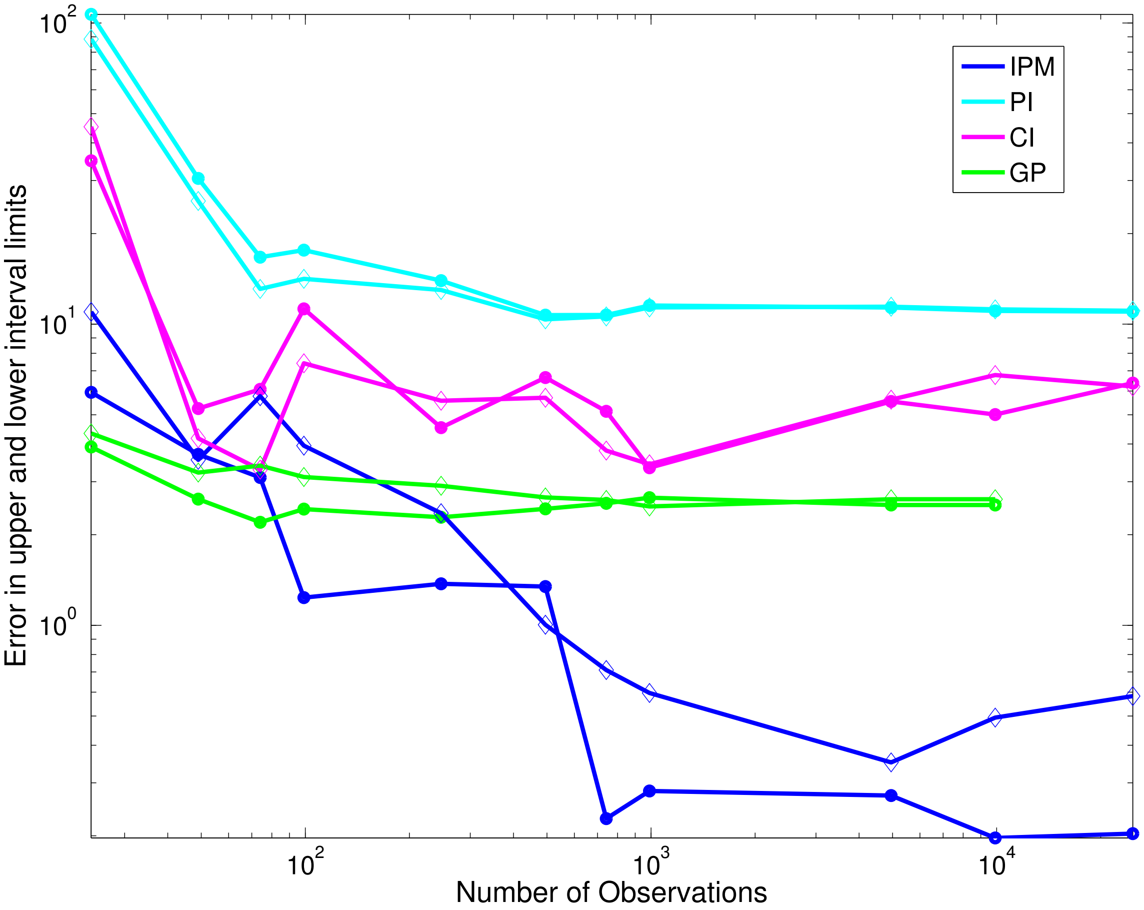

Figure 16. DGM3: upper limit error $\bar{e}$ (circle), and lower limit error $\underline{e}$ (diamond).

Figure 17 shows the CPU time of generating and evaluating a metamodel as a function of N. The figure illustrates the rapid growth in the computational cost of GPs, which prevents its usage for data ensembles having more than a few thousand data points, thus, larger input dimensions. As explained before, this cost prevented the calculation of GPs for $N = 10000$, and $N = 25000$ whose CPU time points are missing from the figure. The CPU time for the IPM includes an empirical procedure for identifying the data points falling outside the tightest output domain containing 99% of the N observations, and the solution to 2 convex optimization programs. Such a program, solved using the interior point algorithm LIPSOL (the Linear Interior Point Solver used by the MATLAB linear program solving function `linprog`), is solvable in polynomial time; i.e., iteration complexity $O(\sqrt{n}L)$ and computational complexity $O(n^3L)$. The first program solves for an IPM based on the original N observations whereas the other one solves for an IPM based on the observations falling in the 0.05 and 99.5 quantiles. This figure illustrates that the computational requirements of the IPMs are about two-orders of magnitude smaller than those of GPs for $N > 10^3$.

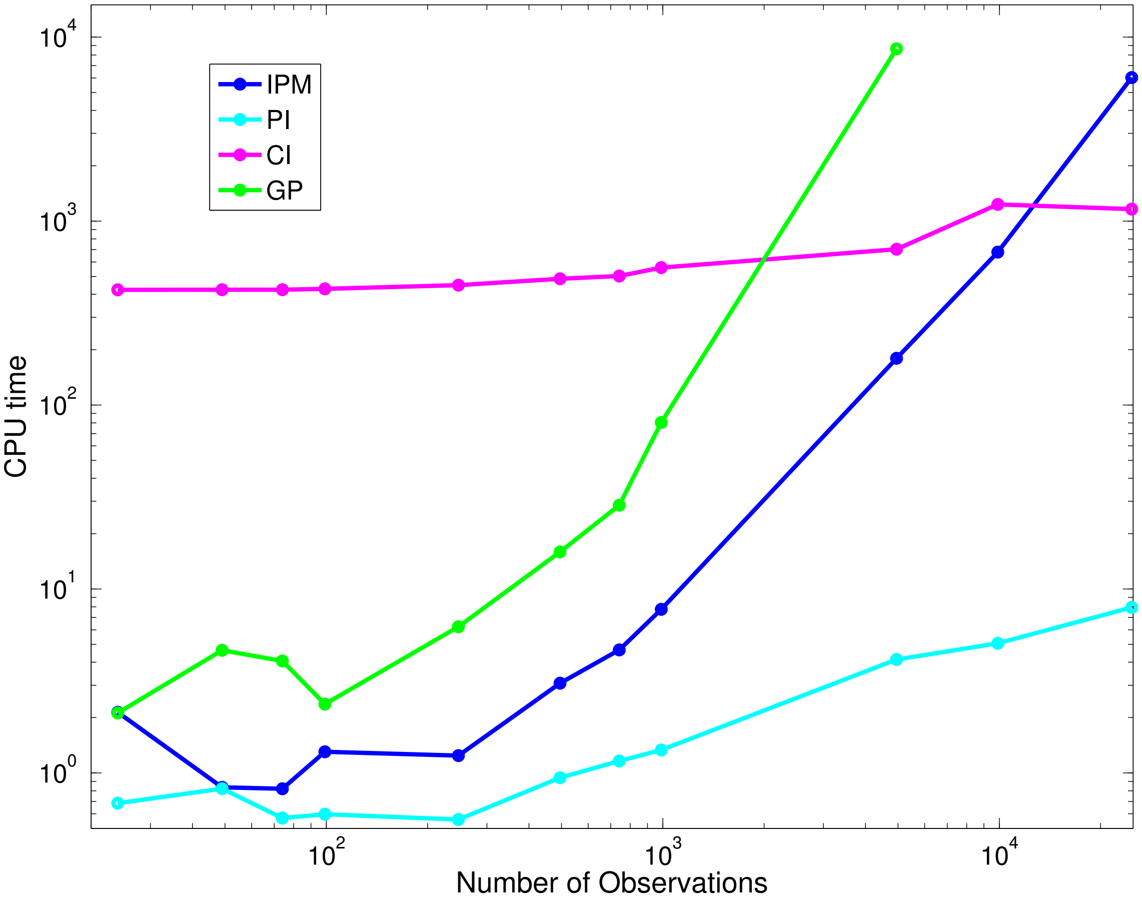

Figure 17. DGM3: Dependency of the CPU time t on the number of observations N.

Figure 18 shows the unreliability as a function of N. As before, the conservatism in the PI yields an overly small unreliability, the CI has an erratic behavior, and both the IPM's unreliability and its upper bound approach the theoretical value. As before, the upper bound on the unreliability closely approximates the actual reliability, and the approximation error converges to zero as N increases.

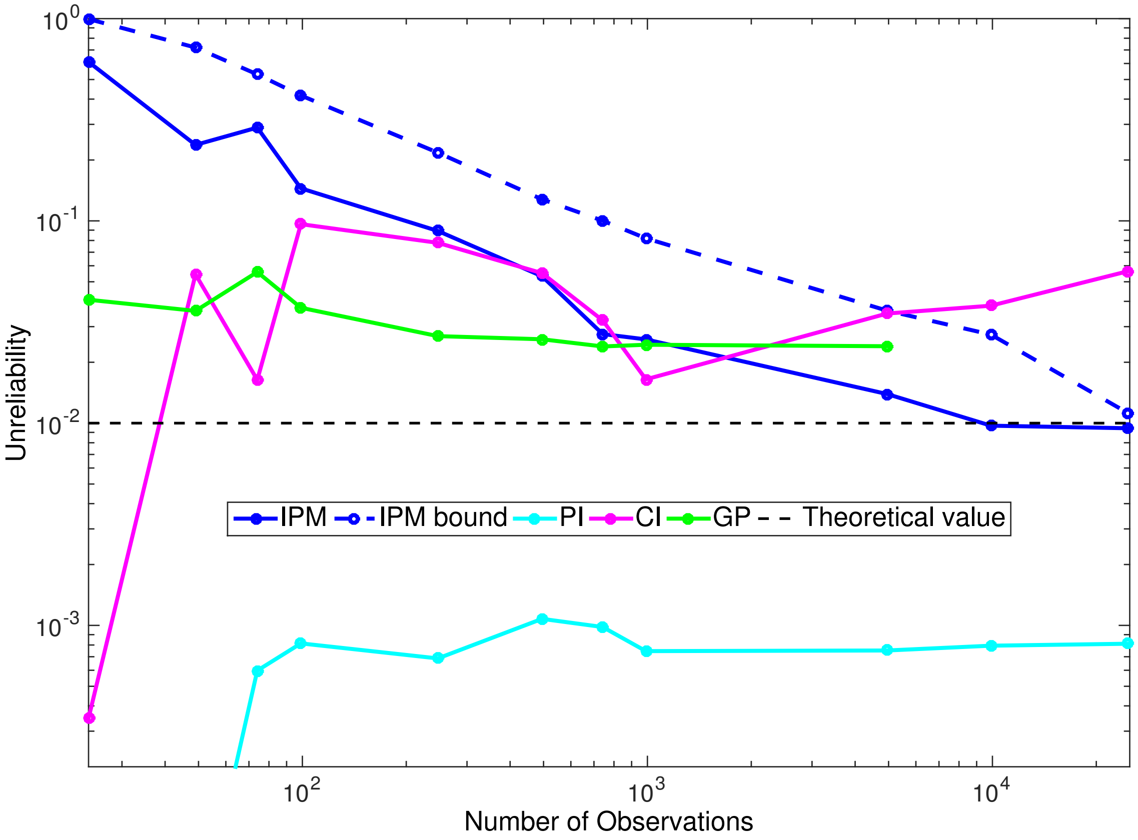

Figure 18. DGM3: Dependency of the unreliability $1 - r$ on the number of observations N.

D. Data Generating Mechanism 4 (DGM4)

Figure 19 shows the two-percentile curves of DGM4 in the input range $X = [-1, 1.5]$ as well as a few data points. Note the data gaps in $[-0.25 < x < 0.15]$ and $x > 1$. These gaps will be present in all data ensembles, which correspond to the 10 values of N in the set $\{100, 200, \ldots, 1000\}$. The objective of this example is to evaluate the predictions during extrapolation. Without making any further assumptions on the DGM, there is no basis to expect any prediction to properly represent the DGM outside the range of measured inputs $\hat{X} = [-1, -0.25] \cup [0.15, 1]$. The quality of a prediction within $\hat{X}$ can be objectively assessed by comparing it with the data. Any notion of goodness outside this domain, however, is always subjective. Whereas conjectures about the behavior of the DGM outside $\hat{X}$ are often used to justify extrapolation, say, because of expert opinion, continuity in the moments, or the understanding of the underlying physics driving the DGM, the resulting prediction will likely not conform to $\mathbb{P}$. The behavior of DGM4 outside $\hat{X}$ can be anything and expecting the resulting predictions to conform to such an unknown (to the metamodeler in this paper, and to everyone in practical applications) is unreasonable. Note that the existence of data gaps, say, because our inability to take measurements outside $\hat{X}$, is equivalent to having a DGM based on the conditional probabilistic cloud $\{\mathbb{P} | x \in \hat{X}\}$. Inferences and theoretical predictions such as the predicted intervals resulting from all metamodeling techniques, including the upper bound on a IPM's reliability, must conform to this conditional DGM to be valid.

Figures 20 and 21 show the predictions corresponding to a set with $N_1 = 100$ and $N_{10} = 1000$ observations respectively. The best predictor within $\hat{X}$ is the IPM, followed by the GPs and the PIs. The CI prediction is too conservative to have informative value. As a result from the RBF structure assumed, the spread of the IPM approaches zero as the separation between the input x and the data set $\hat{X}$ increases. This is apparent at $x = 1.5$ where the spread of $I(x)$ is zero. This would not have been the case for other bases. Conversely, the PI prediction has constant spread throughout the input range. Further notice that whereas the spread of the GP prediction increases rapidly during extrapolation, the spread within $\hat{X}$ is fairly constant. Such an spread is driven by the region of the DGM exhibiting the largest aleatory variation, which in this case occurs somewhere in $x \in [0.15, 1]$. Hence, the predicted variance in $x \in [-1, -0.25]$ is overly conservative. This anomaly can be created by a single outlier in the data set. As such, having a single observation deviating considerably from the rest of the observations will degrade the tightness of the GP prediction globally by an extent proportional to the magnitude of such a deviation. This problem can be prevented by removing outliers from the data set upfront, as it was done with IPMs. The spread of the GP prediction outside the domain of the data increases rapidly. Even though this behavior is somewhat intuitive, there is no rational basis supporting the belief this practice is indeed correct (e.g., the physics driving the DGM might imply that the output must converge to zero as the input increases, so the diverging spread predicted by a GP during extrapolation violates the underlying physics of the phenomenon of interest). Numerical experiments showed that alternative priors for the hyper-parameters of the mean and covariance functions (results not shown) affect the prediction outside $\hat{X}$ without noticeably altering the prediction within $\hat{X}$.

Extensions to the IPM framework that enable manipulating the prediction during extrapolation are being developed. These extensions have been kept separate from the data-based

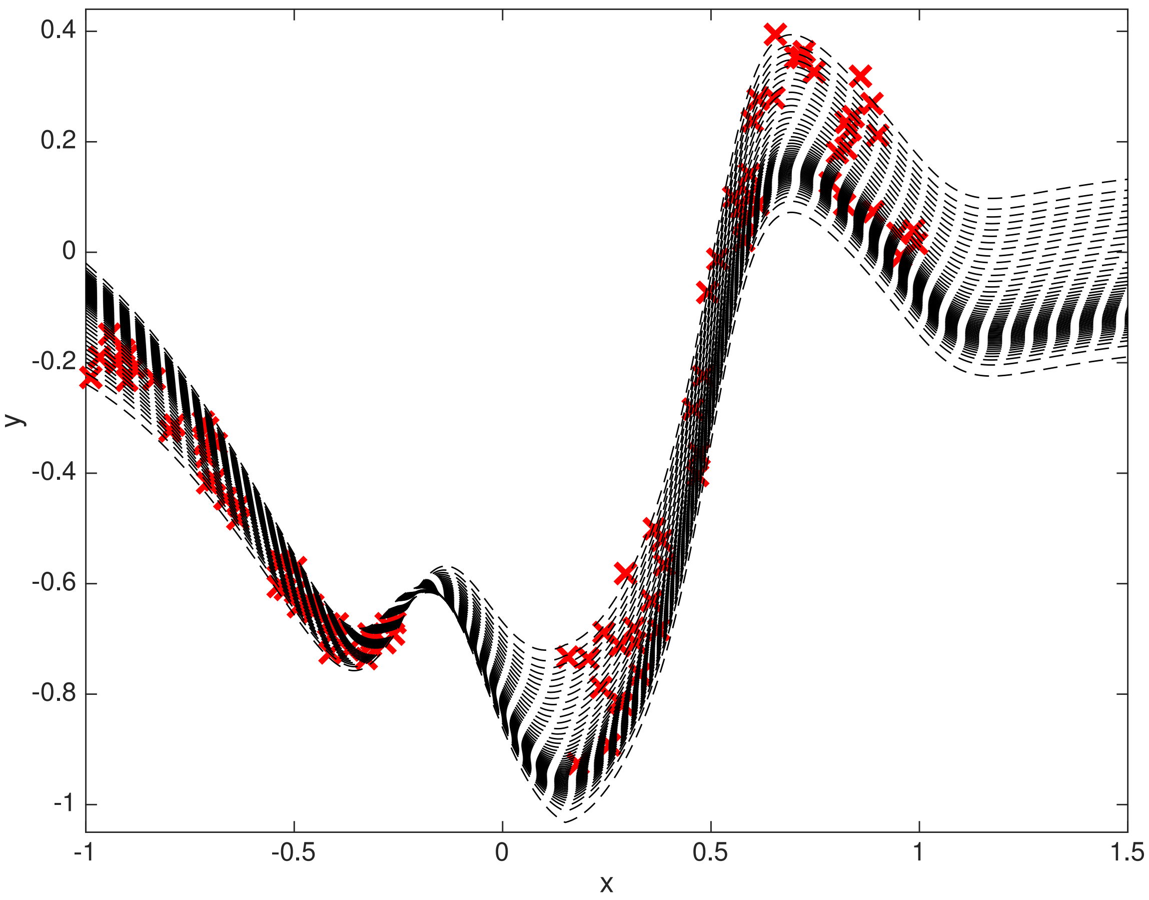

Figure 19. DGM4: Two percentile lines and $N_1 = 100$ data points ($\times$).

IPMs shown thus far in order to separate objective, data-based information from subjective, belief-based information. These extensions, omitted in this article, will be presented elsewhere.

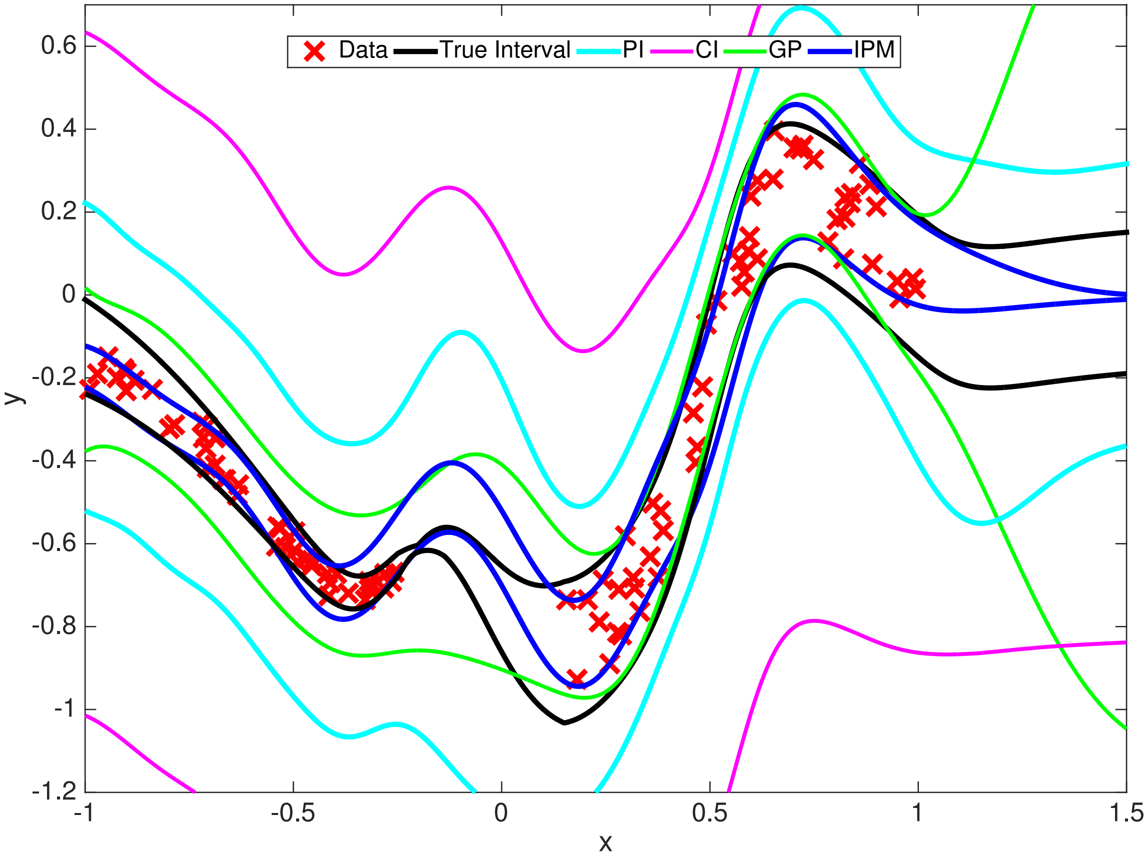

Figure 20. DGM4: predictions for $N_1 = 100$ observations.

Figure 22 shows the error in the limits of the predicted interval as a function of N. As before, IPMs attain the smallest error in all cases. The results in this figure are somewhat misleading since they account for the behavior of the prediction outside $\hat{X}$, a domain where there is no basis to expect any method to perform better than any other. The trends in the errors observed previously apply to DGM4 as well.

Finally, Figure 23 shows the unreliability of the predictions. The line discontinuities in this figure are the result of not being able to plot $\gamma = 0$ in the log scale. Note that the upper bound of the reliability is below the empirical reliability. This anomaly is the result of using $\{\mathbb{P}|x \in X\}$ rather than $\{\mathbb{P}|x \in \hat{X}\}$ in the analysis, i.e., calculating the empirical reliability using input values outside the data range $\hat{X}$. As expected, this anomaly disappears when the empirical reliability is calculated based on the data set $\hat{X}$ used to build the IPM.

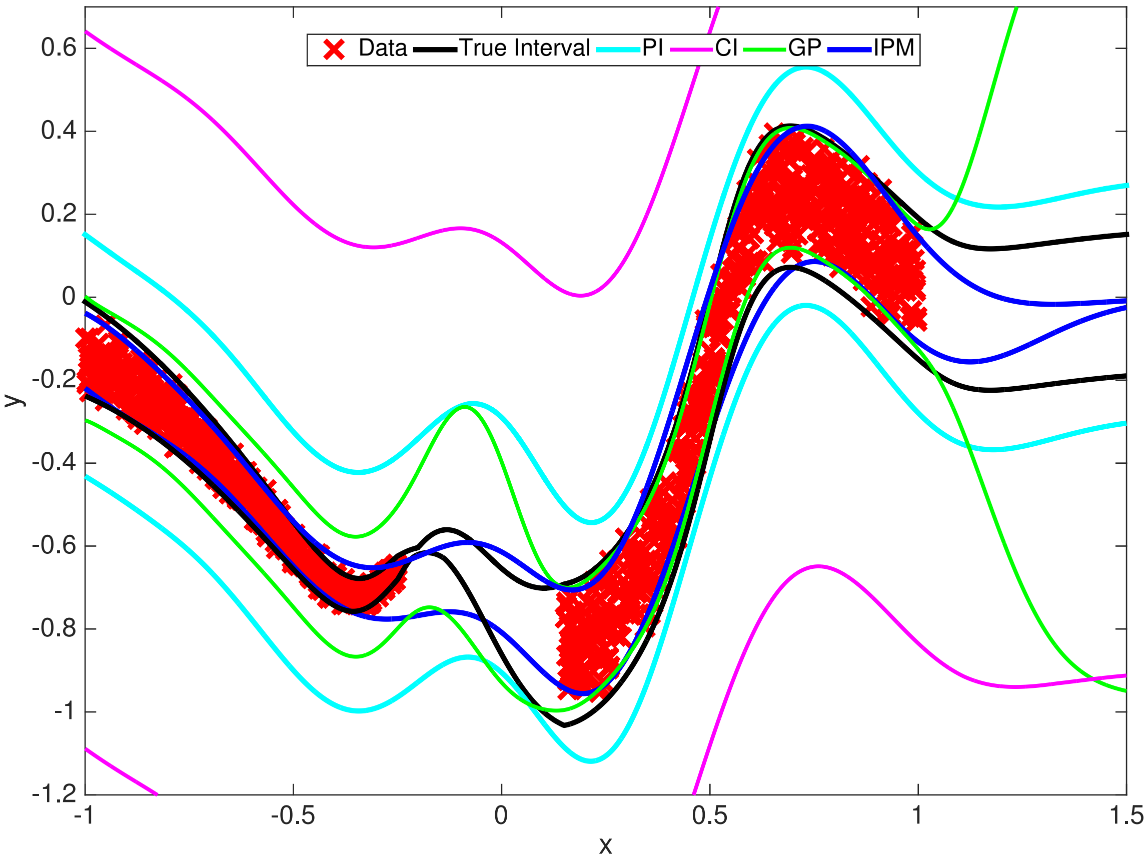

Figure 21. DGM4: predictions for $N_{10} = 1000$ observations.

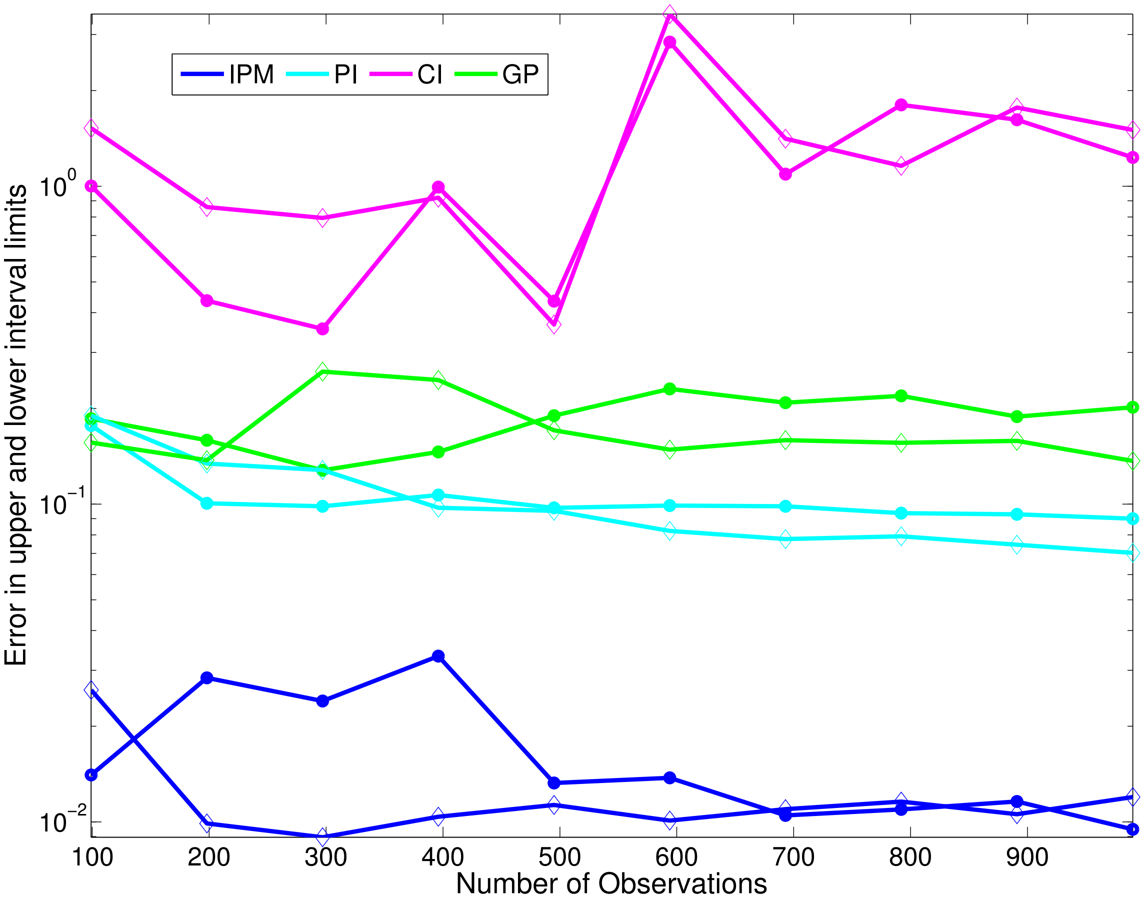

Figure 22. DGM4: Error in the upper and lower interval limits, $\bar{e}$ (circle), and $\underline{e}$ (diamond).

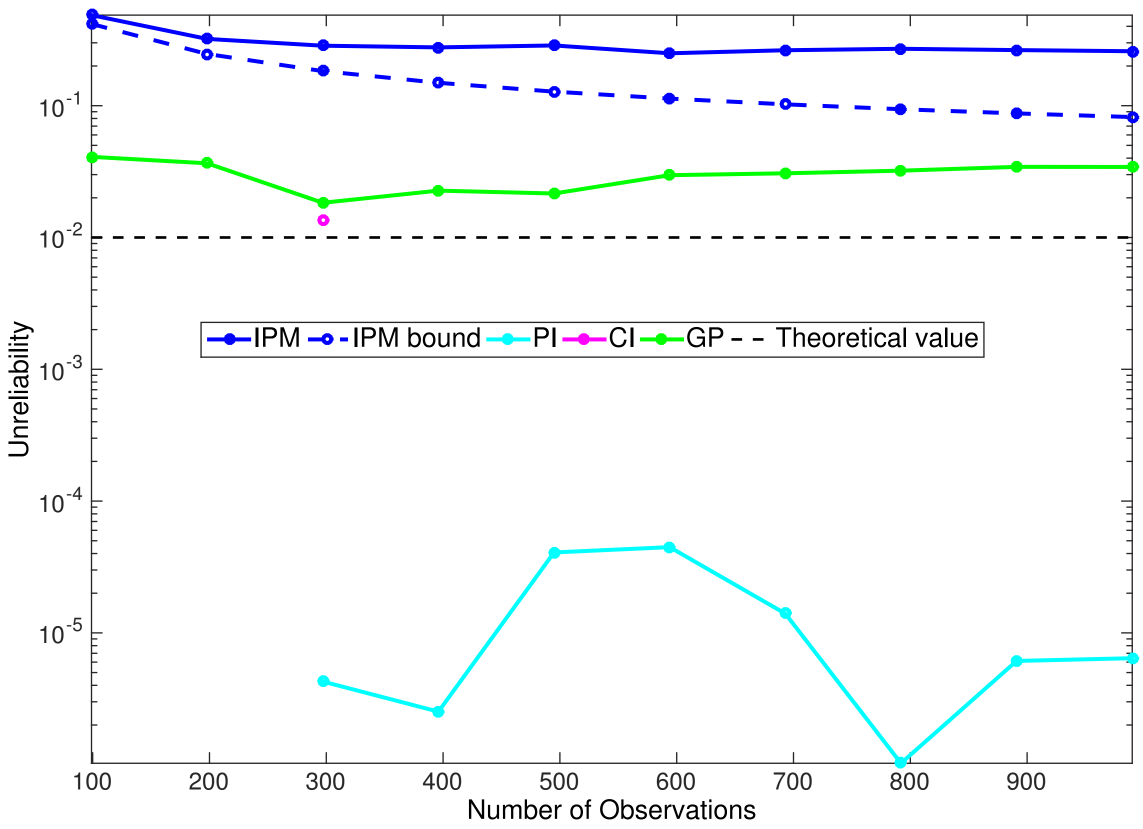

Figure 23. DGM4: Dependency of the unreliability on number of observations.

IV. Discussion

A few general remarks regarding the strengths and limitations of the metamodeling techniques compared are in order. A fundamental premise in this analysis was to use comparable levels of effort in the development and tuning of all metamodels. To this end we assumed that all parametric methods have the same computational model M, thus, they all share the same calibrating variables. The only deviation from this rule was the assumption of η being a Normal zero-mean random variable with a fixed standard deviation, which is an additional calibrating variable for the Bayesian CI (see details above). Even though better model structures could have been used instead, we purposely chose not to do so. Embarking into the process of using a model structure targeting the particular features of a DGM will not only violate the premise above, but might also significantly increase the computational demands of the method. For instance, if the structure of the likelihood function is no longer Gaussian, its evaluation will require creating a sampling-based approximation to a probability density function of unknown structure. This practice greatly increases the computational demands of calculating the MLE, and of sampling the posterior via MCMC. In this regard, the inclusion of GPs as an alternative metamodeling technique is unfair because its increased flexibility gives it a substantial advantage over the parametric metamodeling techniques.

The numerical experiments above consistently indicate that IPMs provide the tightest interval prediction of the data at a reasonable computational cost. This method does not have a single parameter to tune, and there is no need to iterate or evaluate the convergence of the solution. All IPM results shown in this article were obtained in the very first run. These features as well as the ability to rigorously bound the reliability of the prediction make the IPM technique preferable. Furthermore, the IPMs matched very well all four DGMs without having to adjust the structure M to the particular experiment.

The assumed Gaussian RBF structure dictates the extrapolating properties of the prediction for all parametric methods, whereas the assumed covariance function and the corresponding prior dictate that for the GPs. There is no basis to expect any method or model to describe well the DGM during extrapolation. This observation applies to all metamodeling techniques, parametric or not. Overall, PIs require the least computational effort but the conservatism in the resulting prediction made them practically useless. Bayesian CI exhibited high sensitivity to the support set of the assumed prior and were expensive to calculate. This sensitivity yielded considerable variation in the predicted interval.

GPs provided acceptable predictions overall. The GPs' performances were closely dependent on the number of observations N and on the input values where the GPs are evaluated. In regard to the dependency of N, and when x is sufficiently close to the range of observed inputs, there were three distinct operational ranges: a range of small values of N where the predicted interval has an input-dependent spread ($N < 10^3$), a range of moderate values of N where the predicted interval has an input-independent spread sized according to the strongest aleatory part of the DGM ($10^3 < N < 5 \times 10^3$), and a range of moderately large values of N where the calculation of the GP is numerically intractable ($N > 5 \times 10^3$). Hence, operating within the second range of N values, for which the GP model has saturated and it is insensitive to additional data, yields conservative predictions at all inputs where the aleatory spread of the DGM is not the largest. Hence, the presence of a few outliers at some input values has the potential to degrade the tightness of the prediction globally (this will also be the case for IPMs). The features characterizing these 3 operational ranges should be used to determine the number of input dimensions n_x for which a GP can render an acceptable model. In regard to the dependency of the GP prediction on the value of the input, we noted that the spread of the resulting interval grows rapidly with the distance between the testing input and the closest measured input. This makes the predicted interval non-informative during extrapolation. Furthermore, the tuning of the GP model by the selection of the mean and the covariance functions and the priors of the corresponding hyper-parameters was quite cumbersome and non-intuitive. Even though the results above are based on the squared exponential covariance function, several other functions were tried did not render better results. The referenced GP software used in this work allowed for extremely general definitions of these priors, leaving the user with a myriad of possible choices. Unfortunately, the apparent lack of causality and intuition with such choices obscured and complicated the usage of the method. The degraded performance of the GPs observed might also be the result of having to prescribe global values for the hyper-parameters of the mean and covariance functions. As Figure 3 illustrates, a value for the length scale parameter that suits a subdomain of the input range ($x < 0$) might be unacceptable for other subdomains ($x > 0$). Multi-model techniques that might be used to mend for this deficiency were not considered in this analysis.

There are several conceptual differences between standard metamodeling techniques and IPMs. One of them is that while most of these techniques aim at interpolating the data, IPMs aim at describing their spread. For instance, while each of the functions comprising a GP model for a zero-noise level passes through all the data points, there is at least one member of the family of infinitely many functions comprising an IPM passing through each data point. The differences are not only conceptual, but also practical. The calculation of an IPM requires solving an optimization problem, which for the linear program formulation used here can be done very efficiently for a large number of design variables and constraints; i.e., on the order of 10^5 using standard optimization algorithms. This enables considering problems with many more data points N, thus input dimensions n_x, than alternative approaches such as standard, non-sparse implementations of GPs which are restricted to a few thousand points.

A final remark regarding the usage of metamodels is in order. Each metamodeling technique should stipulate the application domain $\mathcal{X} \subset \mathbb{R}^{n_x}$ associated with the resulting computational model. In the context of this paper such a domain is the set of inputs where the prediction is an adequate representation of the DGM. As the examples above show, the application domain $\mathcal{X}$ might not even contain the entire data domain. The data, expert opinion, and engineering judgment must be used to prescribe and communicate $\mathcal{X}$ to the user of the model. This can be attained by making the predicted output range artificially wide outside $\mathcal{X}$. A reduction in the informative value of the prediction will discourage using the model outside its domain of application (this is a natural appeal of GP models, for which the predicted variance grows rapidly with the separation from the data). This practice, which requires modifying the data-based metamodel so the prediction during extrapolation is also dictated by belief- and physics-based arguments, will make the reliability bound of IPMs invalid.

References

[1] L. G. Crespo, S. P. Kenny, and D. P. Giesy. Interval predictor models with a formal characterization of uncertainty and reliability. In *53 IEEE Conference on Decision and Control*, pages 1–26, Los Angeles, CA, USA, December 2014.

[2] L. G. Crespo, S. P. Kenny, and D. P. Giesy. Interval predictor models with a linear parameter dependency. *ASME Journal of verification, validation and uncertainty quantification*, pages 1–20, 2016. Accepted.

[3] L. G. Crespo, S. P. Kenny, D. P. Giesy, R. B. Norman, and Steve R. Blattnig. Interval predictor models with a radial basis structure and their application to space radiation shielding. In *AIAA Scitech 2016*, pages 1–20, San Diego, CA, USA, January 2016.

[4] M.C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society B*, 63(3):425–464, 2001.

[5] Carl. E. Rasmussen and Christopher K. Williams. *Gaussian Processes for Machine Learning.* MIT Press, 2006.

[6] G. A. Seber and C. J. Wild. *Nonlinear Regression.* JohnWiley & Sons, Hoboken, New Jersey, USA, 2003.

[7] T. Simpson, J. Peplinski, P. Koch, and J. Allen. Metamodels for computer-based engineering design: survey and recommendations. *Engineering with Computers*, 17(1):129–150, 2001.