Modeling and Uncertainty

Industrial Data Science – A Review of Machine Learning Applications for Chemical and Process Industries

Max Mowbray, Mattia Vallerio, Carlos Perez-Galvan,
Dongda Zhang, Antonio Del Río Chanona,
Francisco Navarro-Brull*

February 17, 2022

The construction of a model for the prediction of a real process is naturally subject to sources of uncertainty that can change over time. To better explore the roots of such uncertainties, we will examine the following problem. Consider that in a production process, it is often desired to infer the end quality of a product during processing. For example, in [1], the authors discuss the merits of monitoring melt viscosity, temperature profile, and flow index as indicators of product quality in the context of polymer processing. The inference of these material properties may be used to inform process operation, optimization, and control [2], however, direct measurement often proves operationally impractical. As a result, soft sensors may be constructed to infer these qualities from other available process measurements (such as screw speed, die melt temperature, feed rates, and pressures) either via first principles, data-driven, or hybrid modeling approaches.

Regardless of the model constructed, predictions will always be subject to two broad forms of uncertainty - *aleatoric* and *epistemic*. The model construction process, therefore, is ultimately an analysis of that uncertainty. Aleatoric elements can be thought of as that irreducible aspect of model uncertainty, which reflects the underlying variance in the data generating process. Whereas epistemic uncertainty can be thought of as the reducible part of model uncertainty, which arises due to a lack of information. As we will see, the expression of both forms of uncertainties, via a predictive distribution, is important to providing credibility to a predictive model. When the predictive distribution is non-commital, we can conceptualize model uncertainty to be high. However, when it is more certain, model uncertainty is low. In the following, four specific forms of (parametric) model uncertainty are considered, the expression of which provides the predictive distribution desired. In line with this [3], these sources are detailed as follows:

1. Given that data expressive of process evolution are always corrupted by some noise, the fundamental measurement or observation of the underlying process variables (e.g. screw speed, die melt temperature, feed rates, and pressures) is subject to $\mathcal{Y}$ uncertainty.

2. The nature of mathematical model construction is inherently an approximation to the underlying physics. Typically, a number of model structures exist that could well approximate the real process, and this is denoted $\mathcal{M}$ uncertainty. An intuitive example of this in the domain of data-driven modeling arises in the definition of topology (number of hidden layers, activation functions, etc.) for an ANN, or the number of principal components for a PLS model. In the setting of first-principles modeling, such a structural question is often present in, for example, bioprocess prediction and optimization, where identification of the correct semi-empirical expression (i.e. Monod, Aiba, Droop, etc.) often poses a considerable challenge [4]. This situation is however independent of system complexity, as is demonstrated by the different models characteristic of terminal velocity in Stoke's, Newton's, and the transitional drag regime of single-particle settling. In this sense, it is worth noting the well-known position of George Box: 'Essentially, all models are wrong but some are useful', which has given rise to statistically founded model selection and validation practices.

3. Despite the proper selection of a model structure, there exists some uncertainty about the set of parameter values that well describe the underlying process in the domain of interest. This is otherwise known as parametric or θ uncertainty. Such uncertainty arises in identification of parameters for both data-driven and first-principles models.

4. If the inference is derived as the result of a computation, typically some approximation of the solution will be introduced via spatial or temporal discretization of the model. This is otherwise known as h uncertainty. h uncertainty is particularly prominent in computational fluid dynamics (CFD), where interrogation of simulation results is often driven by analysis of the effects of spatial discretization (otherwise known as mesh construction).

Despite the presence of these uncertainties, often we are able to construct models that will capture the underlying physics of the process in the domain of interest. In relation to our product end-quality prediction problem, [1] reports many examples of data-driven and first principle models that were able to successfully predict the desired metric (melt viscosity, temperature profile, and flow index). This is primarily due to well-established statistical practices, as encompassed by data reconciliation and validation approaches, model selection and validation tools, data assimilation practice, and the field of estimation theory. The use of numerical methods common to these fields ultimately enables handling and expression of $\mathcal{Y}, \mathcal{M}$ and θ uncertainties.

In the following, we briefly explore one of the core concepts to parameter estimation for demonstration of the concepts previously discussed. Consider the case of point estimation of model parameter values from an available dataset. The high-level framework for (a frequentist approach to) parameter estimation practice proceeds via the following ideas. The data (i.e. offline measurements of melt viscosity and corresponding values of screw speed, die melt temperature, feed rates, and pressures, etc.) available for model construction are independently and identically distributed (i.i.d.) according to some underlying probability density function (pdf). We would like to find a pdf, which is parameterized by our model (i.e. resides within a parametric family), that is most *similar*[1] to this underlying generative pdf. The assumption as to a) the form (e.g. Gaussian) of the pdfs concerned and b) that of i.i.d. data enable the formation of the likelihood function as equivalent to the joint distribution of the residuals between our data and models predictions. The likelihood function itself provides a ranking of merit over parameter values. Therefore, the parameter setting that maximizes the likelihood function is the most probable parameter setting for our model, given our data. Such a procedure, known as maximum likelihood estimation (MLE), goes some way to handling $\mathcal{Y}$ uncertainty (which is aleatoric), but other approaches are required to handle outliers, missing data [5], latent variable models, and for expression of θ uncertainty.

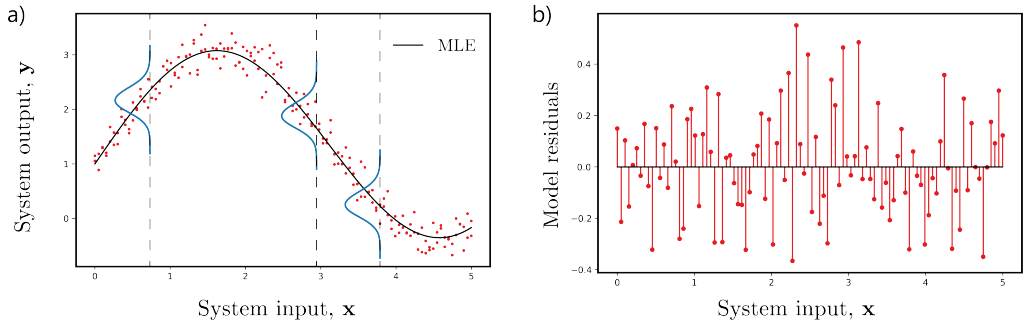

Figure 1: Maximum likelihood estimation of parameters: (a) Assumption of a Gaussian noise model (b) Minimisation of projected distance from data to estimator.

At a high level, there are two approaches to expressing θ uncertainty. The first of those is provided by a Frequentist approach, which typically estimates a confidence region, by considering the model parameters jointly [6]. The confidence region can be conceptualized as the set of physical model parameters that could describe our data with a given *confidence*. Parameters may then be sampled from the confidence region and the corresponding θ uncertainty expressed via simulation of the model under those parameter settings to provide a predictive distribution. It is generally accepted that the greater the cardinality of the set, the larger the epistemic uncertainty associated with the model.

Alternatively, parameter estimation may proceed via a Bayesian (or distributional) approach. The primary idea of the Bayesian approach is encompassed by Bayes' Rule, which (in the setting of parameter estimation) infers a posterior probability density over parameters given our data

[1] In this context, similarity can be quantified via the Kullback-Liebler divergence or a valid distance metric, e.g. the Wasserstein distance

(and model). At a very high level, the posterior density is obtained by conditioning a prior belief (pdf) over parameters on the data available for model construction. The use of a prior probability distribution enables the utilization of information known *a priori* within the parameter estimation procedure. As in the frequentist procedure, the non-committedness of the posterior distribution expresses the epistemic uncertainty of the model, and a predictive distribution may be generated by sampling parameters and performing Monte Carlo simulation. Typically, it is not possible to compute the posterior analytically, although estimates can be gained via various sampling strategies [7, 8]. Despite the estimation of a probability distribution over parameters, point estimates for parameters can be found via the *maximum a posteriori* (MAP) estimate, which is the most likely parameter setting under the posterior. A nice connection between Bayesian and Frequentist approaches is underpinned by the equivalence of the MAP estimate to L1 and L2 regularised MLE estimates, under the assumption of Laplacean and Gaussian priors, respectively [9]. For more information on parameter estimation, we direct the interested reader to [10, 6].

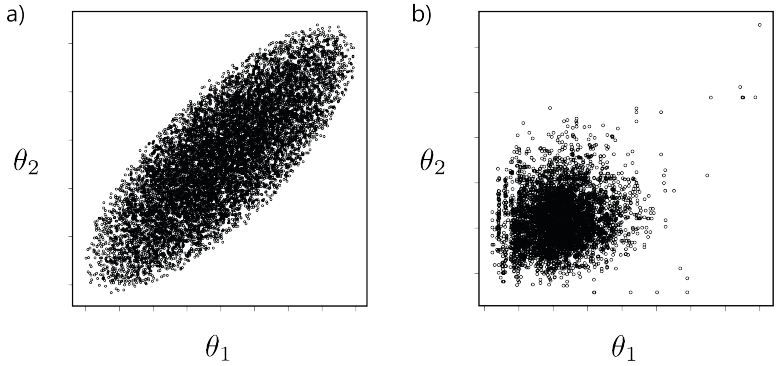

Figure 2: Quantification of uncertainty in parameter estimates a) Frequentist or set-based approach, b) Bayesian or distributional approach. The plots demonstrate correlations between two parameters within a model (plot a) and b) do not share the same range).

It is also worth explicitly stating what has been implicit in the discussion so far. The parameter estimation procedure for a model is specific to the domain one has data in. If one was to estimate the parameters for a given model structure in two distinct regions of the domain, it is highly likely that the estimation procedure will return different parameters. Consider our problem of predicting end-product quality from process data with a model of a fixed structure. If we change our process screw speed to another operational domain, it is likely that the model predictive of our process will require re-estimation in order to predict product end-quality accurately. This provides a warning for use of predictive models in regions of the input domain that the model did not observe in 'training'. Some first principles models can (to an extent) handle such extrapolation demands due to the possession of a model structure that is derived from physical understanding. However, this poses a significant challenge for the use of data-driven models. As a result, it is important to know when one is extrapolating. This is discussed further in the subsequent section.

In view of the discussion of both $\mathcal{Y}$ and θ uncertainties, we now turn our attention to $\mathcal{M}$ and h uncertainties. $\mathcal{M}$ uncertainty is typically handled via model selection and validation processes. This can be conducted in one of two ways. The first proceeds via bias-variance analysis in the form of K-fold model validation. K-fold validation evaluates the model predictive performance by using k different splits of the available data for parameter estimation and validation. More information can be found in [11]. The alternative approach is provided by the Akaike (AIC) and Bayesian (BIC) information criteria. Both metrics evaluate model complexity and predictive performance with the best model corresponding to the minimizer of the criterion. The derivation of the two metrics is based on some fundamental assumptions. These assumptions may be used to select the criteria most suitable for the modeling problem at hand [12]. This leaves us with the presence of h uncertainty. Ultimately, minimization of h uncertainties is guided by the selection of appropriate numerical methods and discretization practice, with selection choice validated by interrogation of the solution.

So far, we have considered the expression of uncertainties specific to our model. In the following, we shall discuss the use of models for the propagation of uncertainties in the process variables and the use that can be found in such analysis with respect to product end-quality. Typically, process

variables (flow rates, pressures, etc.) are likely to observe some form of variation. This may arise from the presence of unquantified disturbance, sub-optimal control, variability in an upstream process, etc. This can be captured computationally by assuming process variables (screw speed etc.) are random variables distributed according to a distribution of choice, and subsequently performing Monte Carlo simulations to provide a hypothesis about the resultant effects of their variation on end-product quality. Appropriate analysis can help determine the variables with the strongest correlation to end-quality variation, which may ultimately guide process operation. This is shown in Fig. **ref**.

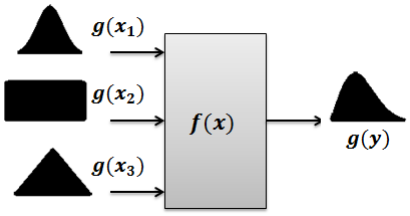

Figure 3: Propagation of input-output uncertainties

0.0.1 Uncertainty aware data-driven modeling

Despite the discussion in the previous section, the expression of uncertainty (via a predictive distribution) in the paradigm of parametric, data-driven modeling is no easy task. This is primarily because the conventional approaches for estimation of θ uncertainty become intractable as the dimensionality of the parameter space becomes large. This causes issues for the identification of θ uncertainties and has ultimately led to the development of innovative training approaches and foundation for the use of a variety of models.

With a focus on the paradigm of parametric (deep) learning, the first example of this is the use of an ensemble of models to provide a bootstrap approximation of the model uncertainty. This has been demonstrated in ANN [13], hybrid models [14], and in random forest [13]. Another approach to training ANNs that considers θ uncertainty is provided by the Bayesian learning paradigm. Bayesian neural networks (BNN) share the same topology as conventional neural networks, but instead of having point estimates for parameters, they instead have a (posterior) distribution over parameters. Given the dimensionality of the parameter space, the learning of such θ uncertainty is generally facilitated via approximate methods such as variational inference [15] and the use of the evidence lower bound (ELBO) [16, 17]. Similarly, Bayesian extensions to other models such as e.g. support vector machines (SVMs) [18] exist, which facilitate the expression of associated aleatoric and epistemic uncertainties.

One eloquent approach for expressing † (aleatoric) uncertainty leverages the likelihood function [13, 17]. Here, the authors proceed under the assumption of a heteroscedastic Gaussian noise model of the residuals. The use of a heteroscedastic model means the variance of the Gaussian distribution descriptive of the model residuals is conditional to the model input, (i.e. the description of the error in melt viscosity prediction changes with the values of screw speed, die melt temperature, feed rates and pressures in our end-quality prediction problem). The authors leverage this to identify a predictive model that expresses both a nominal and uncertainty prediction in closed form. Although such an approach does not account for θ uncertainty (epistemic), the uncertainty prediction is beneficial and the use of a heteroscedastic noise model provides greater flexibility over conventional homoscedastic approaches (where the variance of the noise distribution is considered constant and finite over the input space).

The ultimate contribution of all of these approaches is a probabilistic model, for which a predictive distribution (i.e. a distribution over predictions of melt viscosity given values of screw speed, die melt temperature, feed rates, and pressures) can be generated. This is either constructed in closed form or approximated via Monte Carlo. For example, in the case of BNNs, the predictive distribution is generated via Monte Carlo whereas in the heteroscedastic ANN model of [13] it is expressed in closed form. Clearly, Monte Carlo is less computationally convenient than an explicit uncertainty prediction. However, in BNN, this is a compromise for arguably greater expressivity of model uncertainties as well as robustness to overfitting as naturally inherited via the Bayesian framework [19].

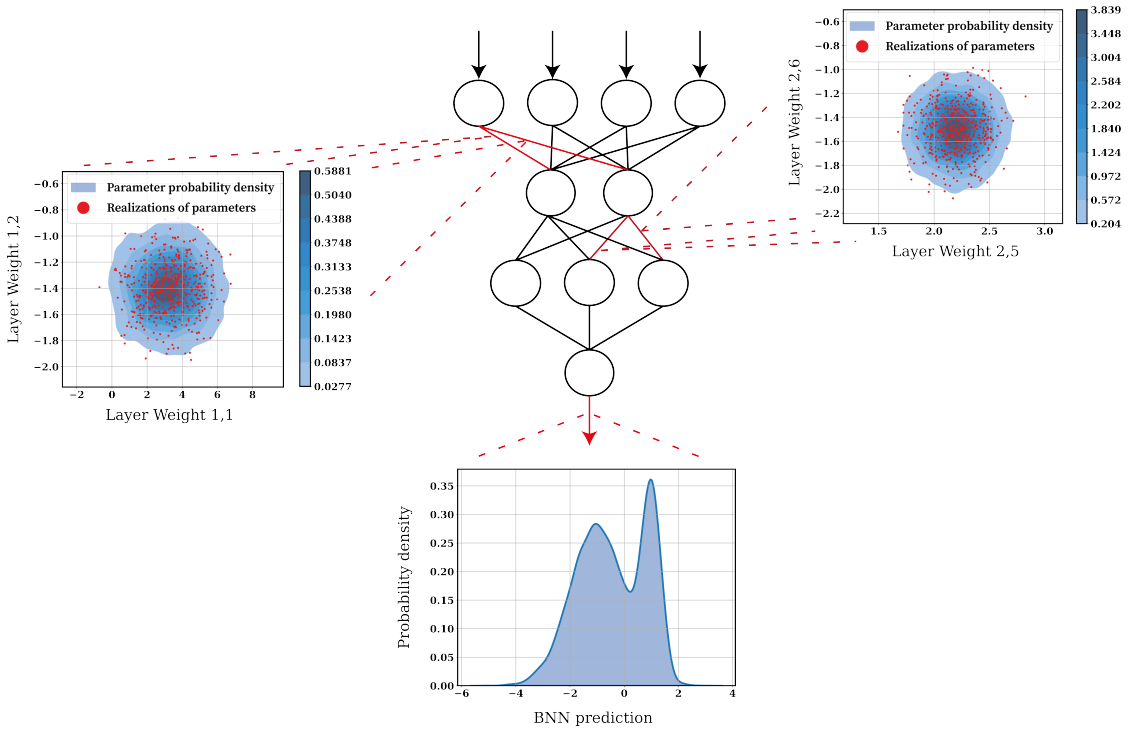

Figure 4: Description of explicit approaches to handle uncertainty in neural networks. a) Heteroscedastic ANN, b) Bayesian Neural Networks

The high-level idea of parametric modeling is the identification of some finite set of parameters (subject to a model structure) to provide a mapping between a set of inputs and outputs. The assumption asserts a constraint on the flexibility of the ultimate model one can construct. There also exists a nonparametric modeling paradigm, where instead of assuming a finite-dimensional parameter vector for model construction, one can instead assume an infinite-dimensional vector i.e. (a function). This provides greater flexibility in model construction and enables the information expressed by the model to grow as more data is acquired. One popular class of (bayesian) nonparametric models is the stochastic process (SP). Formally, SPs define a probability model over an infinite collection of random variables (i.e. functions), any finite subset of which have a joint distribution. This often leads to the interpretation of SPs as a probability distribution over functions, such that a realization of an SP is equivalent to obtaining a sample from a function space. When the distribution over the function space is assumed Gaussian, one obtains a Gaussian process (GP). GPs are a particularly appealing form of SP, primarily due to the fact that multivariate Gaussians are closed under both marginalization and conditioning. This means that given a model input (a value of screw speed, die melt temperature, feed rates, and pressures in our end-product quality prediction problem) one can construct a predictive distribution (i.e. a distribution over melt viscosity) analytically via Bayesian inference. This is particularly convenient for computation. Further convenience lies in the fact that being nonparametric models, hyperparameter selection need only consider the selection of an appropriate mean and covariance function (and its associated parameters), as this selection is primarily responsible for the properties of the distribution over the function space. For more detailed information on the mathematics underlying GPs, we direct to [20], and for the introductory tutorial, we recommend [21].

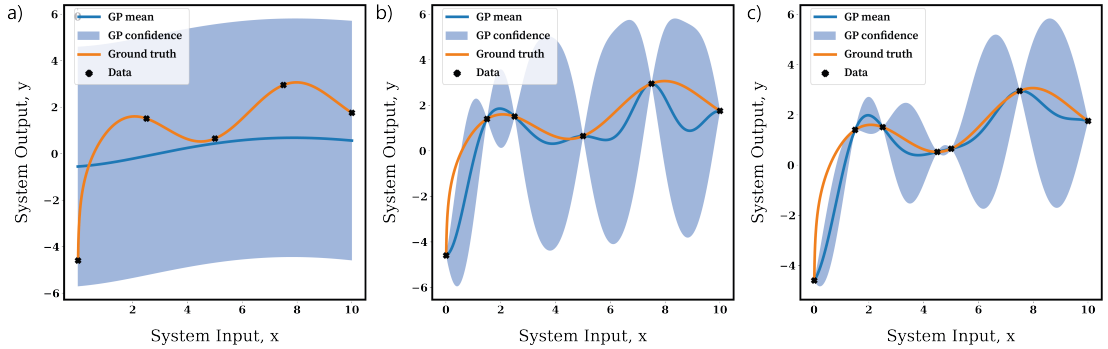

Figure 5: Expression of Gaussian process uncertainty in different data regimes. a) low-data regime, high epistemic uncertainty, b) medium-data regime, reduced epistemic uncertainty, c) high data regime, low epistemic uncertainty

To summarise the potential contribution to the process industries, GPs provide a class of highly flexible models that, through operation within a Bayesian nonparametric framework for inference, express both epistemic and aleatoric model uncertainties. This is particularly useful for quantifying when one is extrapolating and when one is interpolating. Typically, additional, data-dependent mechanisms are required to quantify when this is happening [22], however, in GPs this is expressed automatically. Further, the information expressed by GP models can grow as more data becomes available for model construction (leading to a reduction in epistemic uncertainty). However, for practical use it should be noted:

- The computational complexity of GPs grows cubically with the number of data points, providing either a computational barrier or basis for the use of approximate methods for GP modeling with large datasets.
- The assumption of a homoscedastic Gaussian noise model is computationally convenient, but given that GPs are interpolation models, this can be restrictive in some physical processes where only a small amount of data is available. More robust noise models can be used, (such as Student's T and Laplace distributions) but this typically implies the use of approximate methods for the generation of the predictive distribution.

For more information and insight into the high-level mathematical basis of all the methods described in this section, we refer the reader to provided references.

References

[1] C. Abeykoon, "Design and applications of soft sensors in polymer processing: A review," *IEEE Sensors Journal*, vol. 19, pp. 2801–2813, April 2019.

[2] D. Bonvin and G. François, "Control and optimization of batch chemical processes," tech. rep., Butterworth-Heinemann, 2017.

[3] J. T. Oden, "Adaptive multiscale predictive modelling," *Acta Numerica*, vol. 27, pp. 353–450, 2018.

[4] D. Zhang, T. R. Savage, and B. A. Cho, "Combining model structure identification and hybrid modelling for photo-production process predictive simulation and optimisation," *Biotechnology and Bioengineering*, vol. 117, no. 11, pp. 3356–3367, 2020.

[5] A. Memarian, S. K. Varanasi, and B. Huang, "Mixture robust semi-supervised probabilistic principal component regression with missing input data," *Chemometrics and Intelligent Laboratory Systems*, vol. 214, p. 104315, 2021.

[6] W. C. Rooney and L. T. Biegler, "Design for model parameter uncertainty using nonlinear confidence regions," *AIChE Journal*, vol. 47, no. 8, pp. 1794–1804, 2001.

[7] D. Luengo, L. Martino, M. Bugallo, V. Elvira, and S. Särkkä, "A survey of monte carlo methods for parameter estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2020, pp. 1–62, 2020.

[8] K. P. Kusumo, L. Gomoescu, R. Paulen, S. Garci´a Mun˜oz, C. C. Pantelides, N. Shah, and B. Chachuat, "Bayesian approach to probabilistic design space characterization: A nested sampling strategy," *Industrial & Engineering Chemistry Research*, vol. 59, no. 6, pp. 2396–2408, 2019.

[9] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer school on machine learning*, pp. 63–71, Springer, 2003.

[10] N. D. Perić, R. Paulen, M. E. Villanueva, and B. Chachuat, "Set-membership nonlinear regression approach to parameter estimation," *Journal of Process Control*, vol. 70, pp. 80–95, 2018.

[11] J. Friedman, T. Hastie, R. Tibshirani, *et al.*, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001.

[12] M. Von Stosch, R. Oliveira, J. Peres, and S. F. de Azevedo, "Hybrid semi-parametric modeling in process systems engineering: Past, present and future," *Computers & Chemical Engineering*, vol. 60, pp. 86–101, 2014.

[13] J. W. Coulston, C. E. Blinn, V. A. Thomas, and R. H. Wynne, "Approximating prediction uncertainty for random forest regression models," *Photogrammetric Engineering & Remote Sensing*, vol. 82, no. 3, pp. 189–197, 2016.

[14] J. Pinto, C. R. de Azevedo, R. Oliveira, and M. von Stosch, "A bootstrap-aggregated hybrid semi-parametric modeling framework for bioprocess development," *Bioprocess and biosystems engineering*, vol. 42, no. 11, pp. 1853–1865, 2019.

[15] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, p. 859–877, Apr 2017.

[16] S. Sun, G. Zhang, J. Shi, and R. B. Grosse, "Functional variational bayesian neural networks," *CoRR*, vol. abs/1903.05779, 2019.

[17] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *CoRR*, vol. abs/2011.06225, 2020.

[18] W. Chu, S. S. Keerthi, and C. J. Ong, "Bayesian support vector regression using a unified loss function," *IEEE transactions on neural networks*, vol. 15, no. 1, pp. 29–44, 2004.

[19] D. J. MacKay, "Bayesian model comparison and backprop nets," 1992.

[20] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, vol. 2. MIT press Cambridge, MA, 2006.

[21] R. Turner and M. P. Deisenroth, "Ml tutorial: Gaussian processes (richard turner),"

[22] J. Ash, L. Lancaster, and C. Gotwalt, "A method for controlling extrapolation when visualizing and optimizing the prediction profiles of statistical and machine learning models," 2022.