The parameter uncertainty inflation fallacy

Pascal Pernot

The parameters uncertainty inflation fallacy

Pascal PERNOT[1]

Laboratoire de Chimie Physique, UMR8000, CNRS / Univ. Paris-Sud, F-91405 Orsay, France[a)]

Statistical estimation of the prediction uncertainty of physical models is typically hindered by the inadequacy of these models due to various approximations they are built upon. The prediction errors caused by model inadequacy can be handled either by correcting the model's results, or by adapting the model's parameters uncertainty to generate prediction uncertainties representative, in a way to be defined, of model inadequacy errors. The main advantage of the latter approach (thereafter called PUI, for Parameters Uncertainty Inflation) is its transferability to the prediction of other quantities of interest based on the same parameters. A critical review of implementations of PUI in several areas of computational chemistry shows that it is biased, in the sense that it does not produce prediction uncertainty bands conforming with model inadequacy errors.

Keywords: *computational chemistry; uncertainty quantification; prediction uncertainty; model inadequacy.*

[a)]Electronic mail: pascal.pernot@u-psud.fr

I. INTRODUCTION

Prediction uncertainty of physical models or simulations is difficult to estimate[1]. Yet, it is a necessary step to produce virtual measurements[2], *i.e.*, to enable simulations or models to replace experiments.

Estimation of model prediction uncertainty requires a thorough analysis of three major error sources: (i) systematic errors due to the model formulation and approximations (model inadequacy); (ii) numerical errors (notably for stochastic models); and (iii) parameter uncertainty. Numerical errors are expected to be kept to a negligible or well controlled level (except maybe for chaotic model)[2–4], while parameter uncertainty is estimated by well established calibration methods, notably bayesian inference[5–7]. The most challenging part of the uncertainty quantification process remains model inadequacy[8], which takes often a major fraction of the uncertainty budget[9].

Model inadequacy is characterized by the inability of a model to produce results in statistical agreement with reference data, within their uncertainty range. Even empirical physical models, having adjustable parameters, cannot always achieve a statistically valid representation of the reference data used for their calibration. As model improvement is often impractical or impossible, it is important to be able to deal with the limitations of existing models. Model inadequacy should not be seen as a failure of physical models, but more as an intrinsic component of their predictions that has to be taken care of. Two examples are provided and commented in Fig 1.

Prediction errors due to model inadequacy can be handled either *internally*, by model improvement in the spirit of Jacob's ladder for DFT[12] and composite methods of quantum chemistry[13], or *externally*, by statistical correction of model predictions. The focus of the present study is on the latter approach, which consists in designing a statistical model representing the unexplained part of the model residuals on a set of reference data[14–20].

A major drawback of the statistical correction of model predictions is its lack of transferability to other observables[21,22], which is an issue with generalist models, such as atomistic/molecular simulation or electronic structure computing. As model parameters and their uncertainties are in principle transferable, a solution is to assign them the residual dispersion due to model inadequacy, by a controlled increase in parameters uncertainty. This has been implemented, for instance, through ensemble methods in the calibration of density functionals approximations[17,23–29], or through the concept of embedded models[30].

However, this *parameter uncertainty inflation* (PUI) approach suffers from intrinsic limi-

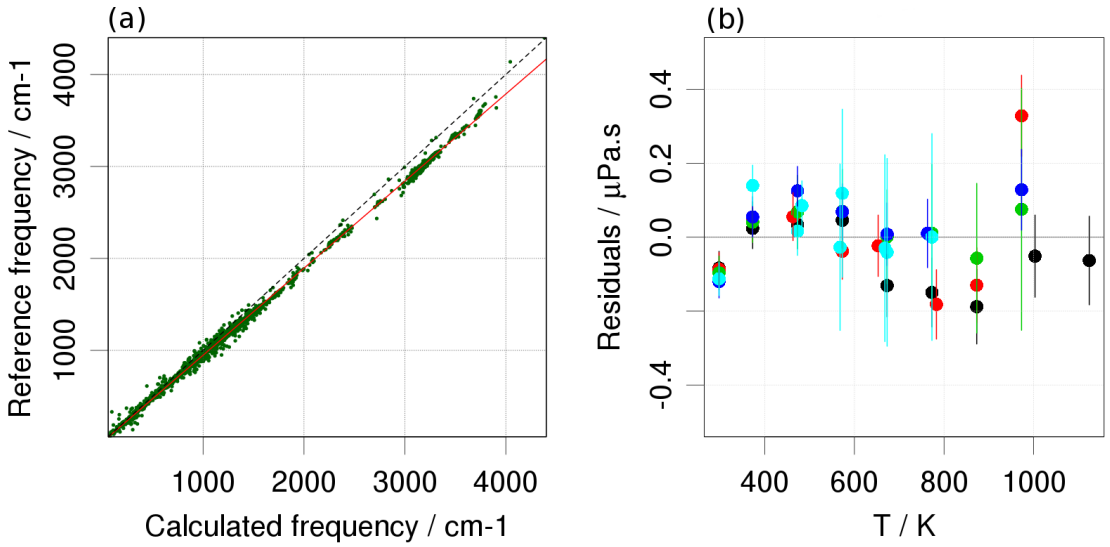

Figure 1. Examples of model inadequacy:
(a) scatterplot of reference fundamental vibrational frequencies with respect to harmonic *ab initio* frequencies calculated at the CCD/6-31G* theory/basis-set level. The red line depicts the linear tendency in the data cloud, which is not the unit line. The error bars on the reference data are invisible at this scale. The data are extracted from the CCCBDB[10];
(b) residuals of the fit of Argon viscosity data by a Chapman-Enskog model. The error bars represent 2-σ experimental confidence intervals. Even if the empirical model achieves well centered residuals, the T-dependent oscillation of the latter reveals an unsatisfactory fit, notably at low temperature. This dataset is described by Cailliez and Pernot[11].

tations which have to be carefully considered:

1. Due to the geometry of the problem in data space[31,32], enlarging the uncertainty patch on the model manifold around the optimal parameters does not contribute to improve the validity of an inadequate model. Besides, even if model adequacy were recovered by PUI for a calibration property, no guarantee exists on the transferability of adequacy to other properties of interest, which have different model manifolds.

2. Considering a model $M(x;\boldsymbol{\vartheta})$,[33] depending on a control variable x (*e.g.* temperature, pressure...), and parameters $\boldsymbol{\vartheta}$, propagation of parameter uncertainty is governed by the functional shape of the model sensitivity coefficients $(\partial M(x;\boldsymbol{\vartheta})/\partial\vartheta_i)$ as functions of x[34]. This means that the shape of the prediction uncertainty bands over the control space does not necessarily conform with the shape of the model inadequacy errors. As will be shown below, this might lead to uncontrolled under- or over-estimation of prediction uncertainty, depending on the value of the control variables.

This short study focuses on the second problem and considers a series of examples inspired from the computational chemistry literature. It focuses on deterministic models, or stochastic models with negligible numerical errors. The next section introduces three methods implementing the PUI approach in a common bayesian framework. Section III treats three examples: (1) a simple linear model involving the statistical correction of *ab initio* molecular vibrational frequencies; (2) a meta-analysis of the prediction uncertainty for formation heats of solids calculated by the mBEEF density functional; and (3) an original application to the calibration of a Lennard-Jones potential on temperature-dependent viscosity data. A discussion of the encountered problems and recommendations to users of these PUI methods serve as conclusion in Section IV.

II. METHODS

Bayesian data analysis is a convenient framework to develop calibration-prediction methods, and it has been used here to present and develop PUI methods. A brief introduction to bayesian analysis is provided in the next section. More details can be found in several excellent textbooks[5–7].

A. Statistical calibration and prediction

One considers a model represented by the function $M(x;\boldsymbol{\vartheta})$, which parameters $\boldsymbol{\vartheta}$ have to be identified, *i.e.* characterized by their probability density function (pdf) or, in the gaussian hypothesis, their "best" value and covariance matrix.

Calibration. Parameters inference is done by calibration of the model on a set of reference data $\boldsymbol{D} = \{x_i, y_i\}_{i=1}^{N}$, accompanied by uncertainties $\{u_{y_i}\}_{i=1}^{N}$. In the general case, the full covariance matrix, $\boldsymbol{V}_D$, might be available in addition to the usual diagonal elements (u_{y_i}). All the knowledge about the parameters is encoded in the *posterior* pdf $p(\boldsymbol{\vartheta}|\boldsymbol{D})$ for $\boldsymbol{\vartheta}$, conditional on $\boldsymbol{D}$ (and M). The posterior pdf is obtained by Bayes theorem

$$p(\boldsymbol{\vartheta}|\boldsymbol{D}) \propto p(\boldsymbol{D}|\boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta}), \tag{1}$$

where $p(\boldsymbol{\vartheta})$ is the prior pdf of the parameters and $p(\boldsymbol{D}|\boldsymbol{\vartheta})$ is the likelihood. Assuming normal data error distribution, the likelihood can be written as

$$p(\boldsymbol{D}|\boldsymbol{\vartheta}) \propto |\boldsymbol{V}_D|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{R}^T \boldsymbol{V}_D^{-1} \boldsymbol{R}\right), \tag{2}$$

where $\boldsymbol{R}$ is the column vector of residuals

$$R_i(\boldsymbol{\vartheta}) = y_i - M(x_i; \boldsymbol{\vartheta}). \tag{3}$$

The maximum a posteriori (MAP)

$$\hat{\boldsymbol{\vartheta}} = \text{argmax}_{\boldsymbol{\vartheta}} p(\boldsymbol{\vartheta}|\boldsymbol{D}) \tag{4}$$

is a point estimate of the set of parameters providing the best fit to the data, constrained by the prior pdf. The mean value of the parameters $\boldsymbol{\mu}_{\vartheta|D}$ and their covariance matrix $\boldsymbol{V}_{\vartheta|D}$ are often used to summarize the posterior pdf. It is important to note that, except if the model is not identifiable, the variance of the parameters is a decreasing function of the calibration dataset cardinal. Moreover, the covariance matrix of the parameters should not be used if the model calibration is not statistically valid.

Validation of a calibration can be done by posterior predictive assessment (see below)[6,22,35], but simple statistics, such as the Birge ratio[36,37] can be very useful. It is defined as

$$R_B = \frac{1}{N-n} \boldsymbol{R}^T(\hat{\boldsymbol{\vartheta}}) \boldsymbol{V}_D^{-1} \boldsymbol{R}(\hat{\boldsymbol{\vartheta}}), \tag{5}$$

where n is the number of parameters in the model, and should be close to 1 for satisfactory fits. Values smaller than 1 point to over-estimated data variance, while too high values can be due to under-estimated data variance or, most often, to model inadequacy.

Prediction. For deterministic models, the mean value of a prediction at a new control value $\tilde{x}$ and its variance can be approximated by linear uncertainty propagation[34]

$$\mu_{M|D}(\tilde{x}) = M(\tilde{x}; \boldsymbol{\mu}_{\vartheta|D}) \tag{6}$$

$$u^2_{M|D}(\tilde{x}) = \boldsymbol{J}^T(\tilde{x}; \boldsymbol{\mu}_{\vartheta|D}) \boldsymbol{V}_{\vartheta|D} \boldsymbol{J}(\tilde{x}; \boldsymbol{\mu}_{\vartheta|D}), \tag{7}$$

where $\boldsymbol{J}$ is a vector of *sensitivity coefficients*

$$\boldsymbol{J}_k(x; \boldsymbol{\mu}_{\vartheta|D}) = \left. \frac{\partial M(x; \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}_k} \right|_{\boldsymbol{\mu}_{\vartheta|D}}. \tag{8}$$

The linear approximation is exposed here mostly for didactic reasons. If it is not appropriate, one has to estimate $\mu_{M|D}$ and $u^2_{M|D}$ by higher order Taylor expansions[34], or by numerical integration (Monte Carlo method)[38].

Various prediction statistics can be used for model validation[6,22,35]. Posterior predictive assessment compares model predictions with reference data and/or validation data. Visual inspection of prediction probability intervals (prediction bands) is generally very useful.

In the following, one will mostly refer to the mean prediction variance on the calibration set

$$MPV = \frac{1}{N}\sum_{i=1}^{N} u^2_{M|D}(x_i), \tag{9}$$

and note the mean prediction uncertainty as

$$\overline{u}_{M|D} = \sqrt{MPV}. \tag{10}$$

The mean squared errors of the model at the MAP

$$MSE = \frac{1}{N}\sum_{i=1}^{N} R_i^2(\hat{\boldsymbol{\vartheta}}) \tag{11}$$

will be used as a reference point for the validation of model predictions.

Inadequate models. If the covariance matrix of the reference dataset used for model validation is known, a Birge ratio value higher than 1 is a good indicator of model inadequacy. Otherwise, inspection of the residuals and comparison with typical reference data uncertainties is often used. A notable trend in the residuals, possibly quantified by their correlation length, is also a feature to be checked.

B. The Parameters Uncertainty Inflation strategy

The aim of PUI is to adjust a model's parameters uncertainty in order to produce enough model output variance to encompass the part of the variance in the residuals due to model inadequacy. This is achieved in Eq. 7 by adapting the parameters covariance matrix $\boldsymbol{V}_{\vartheta|D}$. Two methods are considered: an *indirect* one, based on a scaling of the data covariance matrix $\boldsymbol{V}_D$ (Eq. 2); and a *direct* one, based on the optimization of the elements of $\boldsymbol{V}_{\vartheta|D}$, a more complex option with several variants.

1. The indirect approach

A statistical approach, inspired from bayesian statistics, developed by Brown and Sethna[39], and adapted by Frederiksen *et al.*[23,40] identifies "parameters ensembles" from which prediction statistics are estimated. To relieve the problem of model inadequacy, a scaling factor, T, is introduced in the pdf describing the ensemble.

Translating this in the bayesian framework, an *empirical* likelihood is used

$$p(\boldsymbol{D}|\boldsymbol{\vartheta}, T) \propto |T\boldsymbol{V}_D|^{-1/2} \exp\left(-\frac{1}{2T}\boldsymbol{R}^T\boldsymbol{V}_D^{-1}\boldsymbol{R}\right), \tag{12}$$

which can be seen as the standard likelihood (Eq. 2) with a scaled data covariance matrix $T\boldsymbol{V}_D$.

Jacobsen and collaborators choose T so that the mean variance of model predictions reproduces the mean squared error for the best parameters[26], *i.e.*

$$MPV(T) \simeq MSE. \tag{13}$$

This equation assumes that model inadequacy is a strongly dominant part of the residuals, otherwise, data uncertainty should be explicitly considered. In the ensemble method, T is chosen using a statistical mechanics analogy with a temperature, leading to[40]

$$T = \frac{2C_0}{n}, \tag{14}$$

where $C_0 = \frac{1}{2}\boldsymbol{R}^T(\hat{\boldsymbol{\vartheta}})\boldsymbol{V}_D^{-1}\boldsymbol{R}(\hat{\boldsymbol{\vartheta}})$, and n is the number of parameters.

It is thorough to establish the link with the Birge ratio, using Eq. 5, as

$$T = \frac{N-n}{n}\, R_B. \tag{15}$$

Note that this *indirect* PUI method is akin to the Birge ratio method used in metrological inter-laboratory comparisons to reconcile inconsistent data[37,41]. The Birge ratio method, assuming an adequate model and misestimated data variances, rescales the latter in order to get a valid statistical estimation of the data mean, whereas, in the hypothesis of reliable data variances, T is chosen here to compensate for model inadequacy and obtain valid prediction statistics.

An alternative estimation of T can be based on Eqns. 7 and 13, assuming a near-linear dependence of the model on its parameters in their uncertainty range and negligible data uncertainty:

$$T \simeq \frac{MSE}{MPV(T_0)}, \tag{16}$$

using the mean prediction variance from a reference calibration with $T = T_0 \equiv 1$.

2. The direct approach

In the direct approach, the model's parameters are considered as random variables, with a pdf conditioned by a set of hyperparameters, typically their mean values $\boldsymbol{\mu}_\vartheta$ and a covariance matrix $\boldsymbol{V}_\vartheta$, defining a multivariate normal distribution $p(\boldsymbol{\vartheta}|\boldsymbol{\mu}_\vartheta, \boldsymbol{V}_\vartheta)$.

Such stochastic parameters can be handled in the bayesian inference problem, either at the model level, leading to use a stochastic model within the standard likelihood framework (Eq. 2), or at the likelihood level.

Model level. At the model level, one estimates the impact of stochastic parameters on model predictions by uncertainty propagation[38]

$$f_M(\boldsymbol{\xi};\boldsymbol{x},\boldsymbol{\mu}_\vartheta,\boldsymbol{V}_\vartheta) = \int \prod_{i=1}^{N} \delta\left(\xi_i - M(x_i;\boldsymbol{\vartheta})\right)\, p(\boldsymbol{\vartheta}|\boldsymbol{\mu}_\vartheta,\boldsymbol{V}_\vartheta)\, d\boldsymbol{\vartheta}, \tag{17}$$

where $f_M(.;\boldsymbol{x},\boldsymbol{\mu}_\vartheta,\boldsymbol{V}_\vartheta)$ is the multivariate pdf of the model's predictions at the vector of control points $\boldsymbol{x}$. Inserting this stochastic model in Eq. 2 can be done by replacing $M(x_i;\boldsymbol{\vartheta})$ by the mean predictions (Eq. 6) and their covariance matrix $\boldsymbol{V}_M$

$$p(\boldsymbol{D}|\boldsymbol{\mu}_\vartheta,\boldsymbol{V}_\vartheta) \propto |\boldsymbol{V}_D + \boldsymbol{V}_M|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{R}^T\left(\boldsymbol{V}_D + \boldsymbol{V}_M\right)^{-1}\boldsymbol{R}\right), \tag{18}$$

where

$$\boldsymbol{V}_{M,ij} \equiv u_M^2(x_i,x_j) = \boldsymbol{J}^T(x_i;\boldsymbol{\mu}_\vartheta)\boldsymbol{V}_\vartheta\boldsymbol{J}(x_j;\boldsymbol{\mu}_\vartheta), \tag{19}$$

$$R_i = y_i - \mu_{M|D}(x_i). \tag{20}$$

Note that using the full variance matrix of Eq. 18 in the calculation of the Birge ratio (Eq. 5), by increasing the variance without affecting the residuals, should enable to validate the model with $R_B \simeq 1$.

For a deterministic model M, when the number of parameters is smaller than the number of data points, $\boldsymbol{V}_M$ is singular (non positive-definite), causing the likelihood to be degenerate, and the calibration to be intractable[30]. By definition, for inadequate models, the data covariance matrix is too small to alleviate the degeneracy problem.

As all data points cannot be reproduced *simultaneously* by the model, one has to replace the multivariate problem by a set of univariate problems (marginal likelihoods[30]), *i.e.*, one ignores the covariance structure of model predictions by taking

$$\boldsymbol{V}_{M,ij} = u_M^2(x_i,x_j)\delta(i-j). \tag{21}$$

Likelihood level. A new likelihood, conditioned on the hyperparameters to be inferred[22,30], is obtained by integration of the standard likelihood (Eq. 2) over the possible values of the parameters (marginalization)

$$p(\boldsymbol{D}|\boldsymbol{\mu}_\vartheta,\boldsymbol{V}_\vartheta) = \int p(\boldsymbol{D}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|\boldsymbol{\mu}_\vartheta,\boldsymbol{V}_\vartheta)\, d\boldsymbol{\vartheta}. \tag{22}$$

As in the previous case, it is pointed out by Sargsyan *et al.*[30] that this likelihood is in general degenerate, so that the inference problem has to be solved by alternative methods, such as Approximate Bayesian Computation (ABC)[42,43]. In this case, the full likelihood (Eq. 22) is replaced by a tractable expression, involving summary statistics of the model

predictions, to be compared to similar statistics of the data. An example is provided in Sargsyan *et al.*[30], where the mean value of the model and its prediction uncertainty are used. A version adapted to the present problem, with an explicit treatment of experimental uncertainty is

$$p_{ABC}(\boldsymbol{D}|\boldsymbol{\mu}_\vartheta, \boldsymbol{V}_\vartheta) \propto \exp\left(-\frac{1}{2}\boldsymbol{R}^T\boldsymbol{V}_D^{-1}\boldsymbol{R}\right) \times p_{reg}(\boldsymbol{D}|\boldsymbol{\mu}_\vartheta, \boldsymbol{V}_\vartheta) \tag{23}$$

where the first term has the same expression as the standard likelihood (Eq. 2) using residuals evaluated at the mean of the model prediction (Eq. 20), and the second term ensures that the predicted model uncertainty $u_M(x_i)$, combined with experimental uncertainty u_{y_i}, is of a magnitude compatible with the residuals

$$p_{reg}(\boldsymbol{D}|\boldsymbol{\mu}_\vartheta, \boldsymbol{V}_\vartheta) = \exp\left(-\sum_{i=1}^{N} \frac{\left(\sqrt{u_M^2(x_i) + u_{y_i}^2} - |R_i|\right)^2}{2u_{y_i}^2}\right). \tag{24}$$

As evidenced in our notation, this term can also be seen as a regularization function, necessary to constrain the parameters covariance matrix $\boldsymbol{V}_\vartheta$ in the inference process. The constraint imposed here is a statistical variant of Eq. 13, but aims at the same effect.

III. APPLICATIONS

A. Harmonic vibrational scaling factors

Various approximations in the *ab initio* calculation of harmonic vibrational frequencies of molecules lead to a systematic bias with respect to fundamental experimental frequencies (Fig. 1(a)), which can be statistically corrected by a simple scaling of the calculated values[44–49]. This *a posteriori* scaling corrects empirically for the approximations involved in the *ab initio* calculation. After scaling, the residual errors are typically still much larger than the reference data uncertainties[50], and the corrected model is still inadequate ($R_B \gg 1$).

One considers here the scaling model $M(x; s) = s * x$, where s is the scaling factor and x a calculated frequency. Irikura *et al.*[50] proposed a method to evaluate the prediction uncertainty of vibrational frequencies corrected by scaling factors. Their approach assumes a multiplicative uncertainty model

$$\nu_i = s * x_i, \tag{25}$$

$$u_{\nu_i} = u_s * x_i, \tag{26}$$

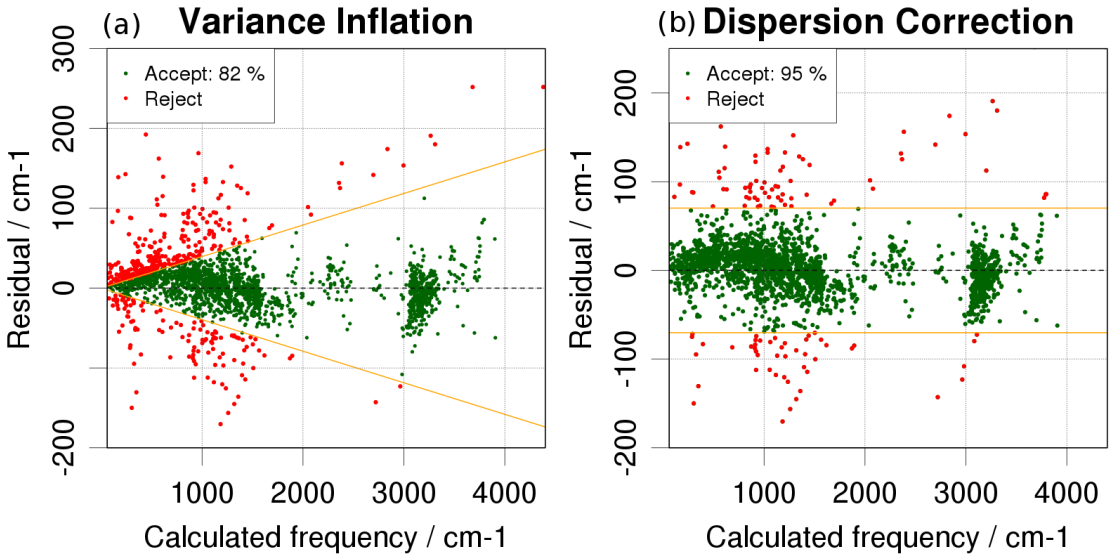

Figure 2. Scatter plots of the residuals of 2279 scaled vibrational frequencies for the CCD/6-31G* theory/basis-set level. The orange lines show the 95 % confidence prediction range for two methods: (a) parameters uncertainty inflation; (b) dispersion correction. The red points lie outside of the 95 % prediction range.

where ν_i a scaled frequency and u_s is the scaling factor uncertainty. An expression of u_s has been derived (Eq. 21 in Irikura *et al.*[50]) as

$$u_s^2 \simeq \frac{\sum_i (y_i - s * x_i)^2}{\sum_i x_i^2}, \tag{27}$$

which is different from the uncertainty that would result from an ordinary least squares calibration model[51], *i.e.*, for large data sets,

$$u_s^2 \simeq \frac{MSE}{N \sum_i x_i^2}. \tag{28}$$

We want to emphasize here that it is possible to recover Eq. 27 by the indirect PUI approach. Namely, equating the mean prediction variance with the mean squared errors (Eq. 13) leads to

$$\frac{1}{N} \sum_i u_{\nu_i}^2 = \frac{1}{N} \sum_i (y_i - s * x_i)^2, \tag{29}$$

while Eq. 26 gives

$$\sum_i u_{\nu_i}^2 = u_s^2 \sum_i x_i^2, \tag{30}$$

from which one derives Eq. 27.

This shows clearly that the derivation of u_s by Irikura *et al.*[50] does not provide the uncertainty on the scale parameter resulting from the calibration procedure, but the parameter

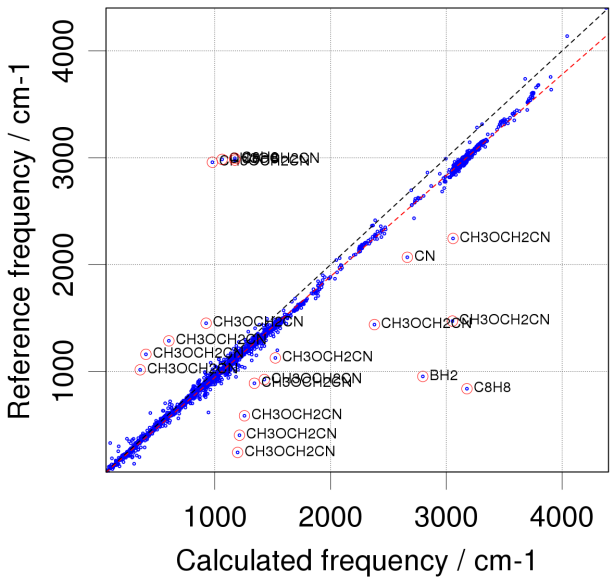

Figure 3. Scatterplot of reference fundamental vibrational frequencies with respect to harmonic *ab initio* frequencies calculated at the CCD/6-31G* theory/basis-set level, as directly extracted from the CCCBDB[10]; the most outlying points have been circled and labeled.

uncertainty necessary to recover a prediction variance over the calibration set compatible with the model errors, in the hypothesis of a multiplicative uncertainty model.

To illustrate the implication of this choice on prediction uncertainty bands, let us consider a set of vibrational frequencies extracted from the CCCBDB[52]. A link to the R[53] scripts used for data extraction, cleanup and treatment is provided in the Supporting Information section.

For the CCD/6-31G* theory/basis-set combination, a data set containing 2323 frequencies is recovered (7 records with incomplete data have been removed). A sanity check, based on a plot of the reference frequencies *vs.* the calculated frequencies, enables to detect several aberrant points, mainly due to incorrect symmetry assignment for CH_3OCH_2CN and C_8H_8 (Fig. 3). Also, the CN frequency and one BH_2 frequency are outstanding and marked as outliers. These data were removed, and the final data set contains 2279 frequencies.[54] Let us note that a more rigorous data curation procedure would be required to generate reference scaling factors, which is not the aim of the present paper.

The statistical analysis of this set gives $s = 0.947$, in conformity with the CCCBDB value[10], and $u_s = 0.020$ (Eq. 27), smaller than the value of 0.046 reported in the CCCBDB, which reflects the impact of aberrant points in the original dataset on u_s.

One can check on Fig. 2(a) that the linear dependence of the prediction uncertainty implied by Eq. 26 is not representative of the residuals cloud. It underestimates the dispersion

at low frequency and overestimates it at high frequency. Furthermore, the probability for a calibration data point to lie in a 95 percent confidence band $[-2u_s x_i,\, 2u_s x_i]$ is only 82 %.

It has been shown that in this case, model inadequacy should not be accounted for by Eq. 26[51,55]. Instead, the completion of the model by a stochastic variable $\delta \sim N(0, \sigma^2)$ representing model inadequacy,

$$\nu_i = s * x_i + \delta \tag{31}$$

$$u_{\nu_i}^2 = u_s^2 * x_i^2 + \sigma^2, \tag{32}$$

enables a more consistent estimation of prediction uncertainty bands (Fig. 2(b)). The use of the stochastic correction δ recognizes that the model errors have a random and uniform distribution with respect to the control variable. In this case, $u_s = 4.1\,10^{-4}$ is the standard uncertainty of the scale factor resulting from ordinary least-squares regression[51]. For large calibration datasets like the present one, the first term of the prediction variance is negligible, and one finds that $u_{\nu_i}^2 \simeq \sigma^2 \simeq MSE$[17,51].

This example shows how the one-parameter scaling model, and the implied sensitivity coefficient, prevents the indirect PUI strategy to achieve reasonable prediction uncertainty bands. It is now acknowledged that the uncertainty factor defined by Eq. 27 should not be used for prediction uncertainty[56,57], although this is not clearly stated in the CCCBDB where the corresponding values of u_s are still reported[10].

B. Calibration of density functional approximations

Jacobsen and coworkers[23,24,27,40,58,59] have elaborated an ensemble method to account for the uncertainty in the parameters of their calibrated mBEEF density functional. Considering that the prediction errors are typically much larger than the reference data uncertainty (model inadequacy), they scale the parameter covariance matrix to get a mean prediction uncertainty in agreement with the MSE (Eq. 13).

In the publications on mBEEF, one has only access to histograms of the scaled errors, which do not enable us to appreciate the structure of the prediction uncertainty for this method. Thanks to the formation heat data provided in a recent article[27], one can now compare the prediction uncertainty with the residual errors of the calibrated method and test their conformity. The dataset of residual errors and prediction uncertainty used below has been extracted from Table I of this article. The measurement uncertainty of the reference data is not provided, but the typical experimental uncertainty on formation heats has been reported to be well below 0.1 eV/atom[60].

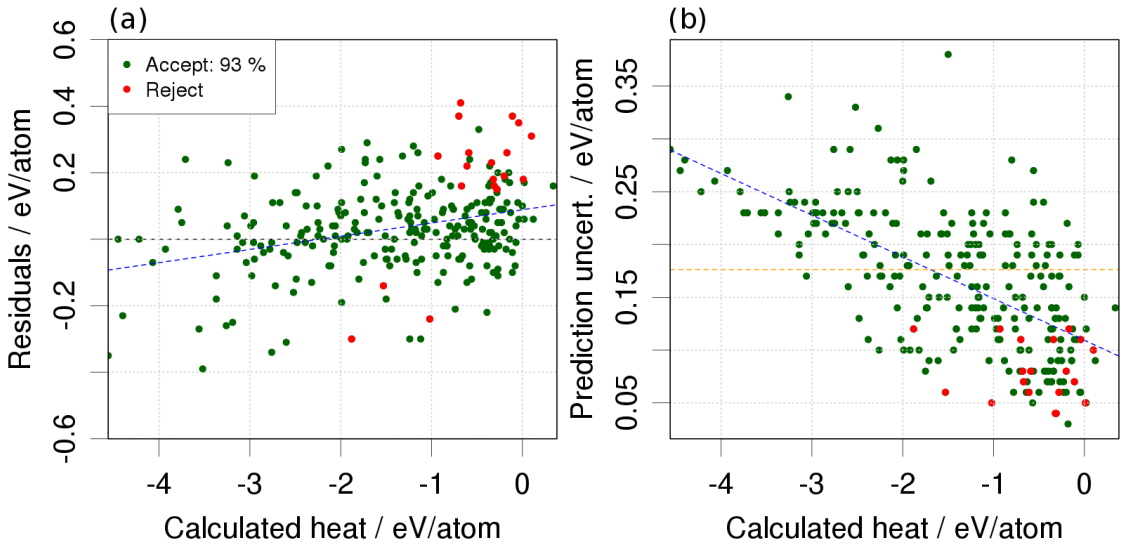

Figure 4. Residuals (a) and prediction uncertainties (b) of the mBEEF density functional on a dataset of formation heats. The red points are those for which a 95% confidence interval around the calculated value does not contain the reference value.

The residual errors are plotted in Fig. 4(a) as a function of calculated heats (the control variable): their distribution presents a small positive linear trend, but the amplitude is weak, and one would not gain much by an additional a posteriori correction. One can therefore assume that the method is well calibrated and enables to make predictions without significant bias (smaller than 0.1 eV/atom) within the calibration range. One can also see that the residual errors are often much larger than the typical reference data uncertainty, revealing model inadequacy.

Prediction uncertainties generated by the mBEEF method are plotted in Fig. 4(b): they display a marked negative linear dependency with the control variable, with a correlation coefficient of -0.63 and a ratio of about 3 between the extreme average values (blue dashed line). This trend is not observable on the absolute values of the residuals.

The mean prediction uncertainty $\overline{u}_{M|D}$ (0.18 eV/atom) is slightly higher than the $RMSE$ (0.14 eV/atom). As a consistency check, one calculates for each residual a 95 % confidence interval using the prediction uncertainty provided in the original article and checks if this interval contains 0. This is verified in 93 % of the cases, confirming the good average properties of the estimated prediction errors. However, instead of being uniformly distributed over the heat range, all the intervals failing the test appear only for formation heats above -2 eV/atom (red points in Fig. 4), *i.e.*, 100 % of the intervals below -2 eV/atom include the null value.

This leads us to conclude that the uncertainty of the lower heats is overestimated (Fig. 4(b)), while the uncertainty of a fraction of the higher heats is underestimated. The $RMSE$ for the heats below $-2\,\mathrm{eV/atom}$ is about $0.14\,\mathrm{eV/atom}$, while the mean prediction uncertainty for this group is about $0.23\,\mathrm{eV/atom}$ ($\sim$65 % overestimation).

Even if the effect is less striking than in the vibrational frequencies case (Section III A), this example shows also that indirect PUI produces prediction uncertainties that are not distributed like the model errors they are supposed to represent.

C. Lennard-Jones parameters

The third example concerns the estimation of the σ, ϵ Lennard-Jones (LJ) potential parameters from the analysis of temperature-dependent viscosity data. The data and Chapman-Enskog viscosity model are described in Cailliez and Pernot[11]. The reference set contains 41 points $\{x_i,\, y_i,\, u_{y_i}\}$, where x is the temperature, y the viscosity and u_y is the viscosity measurement uncertainty. They result from 5 measurement series $\left\{D^{(i)}\right\}_{i=1}^{5}$, but one did not attempt here to model inter-series discrepancy. Therefore the data covariance matrix is diagonal, with $\boldsymbol{V}_{D,ij} = u_{y_i}^2 \delta(i-j)$[11,32].

The indirect and direct PUI methods presented above have been implemented in `Stan`[61], using the `rstan`[62] interface package for `R`[53]. `Stan` is a very flexible and efficient probabilistic programming language to implement bayesian statistical models. A link to the codes to reproduce the results of this example is provided in the Supporting Information section. The indirect method (Sect. II B 1), implementing Eqns. 12 and 15, is named VarInf_Rb; the direct method based on marginal likelihoods (Sect. II B 2) is named Margin; and the approximate bayesian method (Sect. II B 2) ABC. The covariance matrix of the parameters is parameterized by u_ϵ, u_σ and ρ, the uncertainty on ϵ, σ, and their correlation coefficient, respectively.

A `Stan` code provides a sample of the posterior pdf of the model's parameters $p(\boldsymbol{\mu}_\vartheta, \boldsymbol{V}_\vartheta|\boldsymbol{D})$, from which statistics are calculated. The No-U-Turn sampler[63] was used, and convergence of the sampling was assessed by examining the parameters samples and the *split Rhat* statistics provided by `rstan`[62]. Uniform prior pdfs have been used for location parameters, and log-uniform for scaling parameters, unless stated explicitly. All models were run with 4 parallel Markov Chains of 5000 iterations each, 1000 of which are used as warm-up for the No-U-Turn sampler. The convergence criteria and parameters statistics are thus estimated on four samples of 4000 points.

Table I. Parameters of the posterior pdf of the Lennard-Jones parameters recovered by different PUI methods. The RMSE for all fits is 0.10 μPa.s.

Method	μ_ϵ (K)	μ_σ (Å)	u_ϵ (K)	u_σ (Å)	ρ	R_B	$\overline{u}_{M\|D}$ (μPa.s)
WLS	146.1(4)	3.315(1)	-	-	-0.97	6.60	0.01
VarInf_Rb	146(5)	3.32(1)	-	-	-0.97	0.05	0.13
Margin	146(1)	3.316(3)	0.6(8)	0.004(2)	0.0(6)	0.98	0.11
ABC	146.2(4)	3.315(1)	0.7(7)	0.003(2)	0.0(6)	1.20	0.09
Margin1	144(2)	3.321(4)	5.0(10)	0.015(3)	-0.98(2)	0.88	0.13
Margin2	146(1)	3.315(3)	0.01(2)	0.0043(6)	0.0(6)	1.00	0.11
Margin3	145(1)	3.318(3)	1.6(2)	0.0001(2)	0.0(6)	0.98	0.10

The mean values of the parameters and hyperparameters estimated for all methods are reported in Table I, along with their Birge ratio (R_B), and mean prediction uncertainty $\overline{u}_{M|D}$. The $RMSE$ for all methods is 0.1 μPa.s.

The Birge ratio for a model implementing the standard likelihood (Eq. 2) is $R_B \simeq 6.6$ (method WLS in Table I), indicating clearly that the Chapman-Enskog model is unable to fit the data within their uncertainty range. Model inadequacy is also apparent through the trend/oscillation in the residuals (Fig. 1(b)). As the VarInf_Rb method has inflated data uncertainty (the scale factor T has been estimated from the Birge ratio of the WLS method by Eq. 15, giving $T \simeq 129$), its Birge ratio (0.05) is too small. The Margin method achieves a near-unit Birge ratio, but the ABC method cannot reach this value, because of the constraints introduced in the likelihood (Eq. 24).

The residuals (points) and prediction band (gray area) for the indirect method (VarInf_Rb) are shown in Fig. 5. To be comparable with the residuals, the prediction bands are corrected from the mean prediction value, $\mu_{M|D}(T)$. One sees that the prediction band adopts a diabolo structure, with a pronounced waist around 500 K. By contrast, optimization of the covariance matrix of the parameters by the direct methods, Margin and ABC, leads to prediction bands with more regular shapes (Fig. 5). All methods achieve a mean prediction uncertainty in fair agreement with their $RMSE$ of 0.1 μPa.s, although VarInf_Rb returns a value slightly in excess, with $\overline{u}_{M|D} = 0.13\,\mu$Pa.s. It is difficult at this stage to pick a best prediction uncertainty model: the Margin method might be favored due to its better Birge ratio.

Inspection of a sample of the posterior pdf for the Margin (Fig. 6) and ABC methods (not

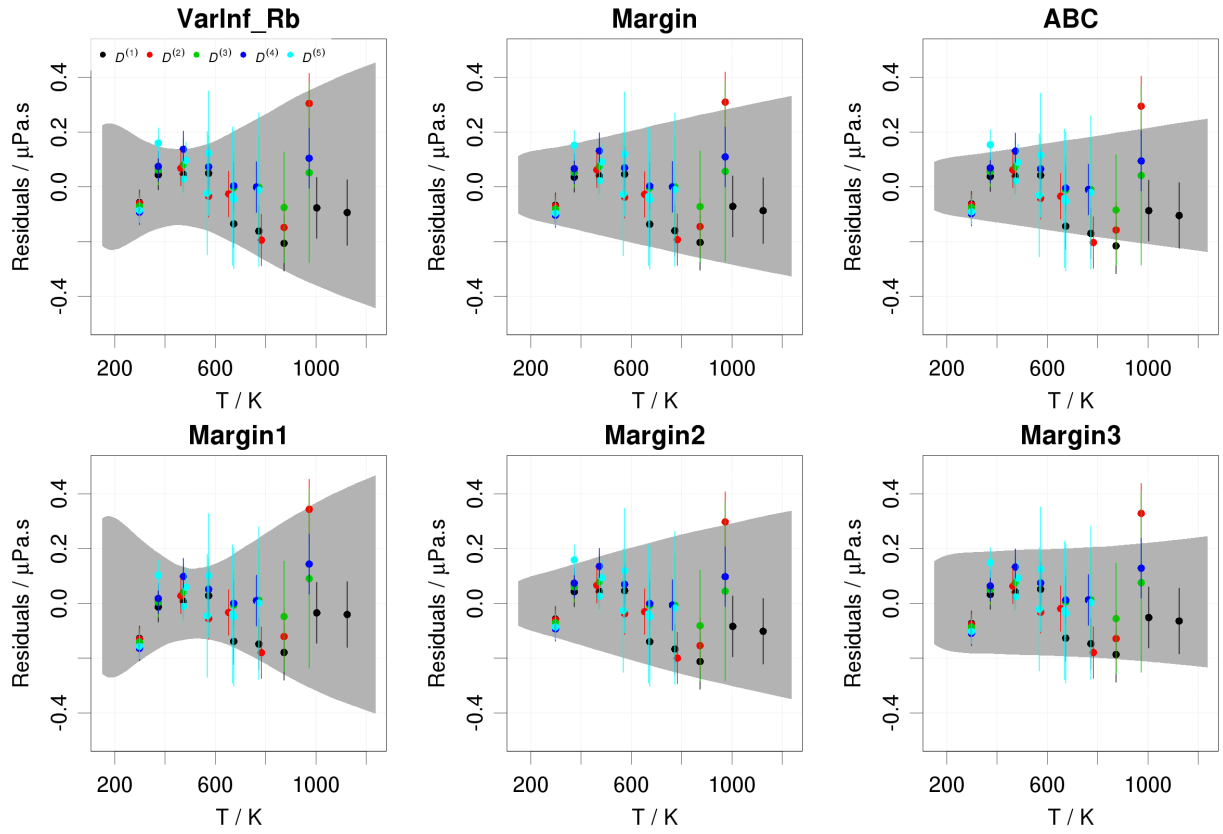

Figure 5. Residuals and centered prediction bands of a Chapman-Enskog model of Ar viscosity for the VarInf_Rb, Margin and ABC methods (top row), and for the three degenerate solutions of the Margin method (bottom row). The dark-gray bands represent model prediction confidence interval at the 2-σ level, corrected from the mean prediction.

shown) reveals the presence of three modes (high-density areas), each one corresponding to a minimum value of a parameter of the $\boldsymbol{V}_\vartheta$ covariance matrix. In the present case, the mode corresponding to $\rho \simeq -1$ is less outstanding than the modes at $u_\epsilon \simeq 0$ and $u_\sigma \simeq 0$.

By constraining the support of the parameters through their prior pdf, the three modes have been sampled independently for the Margin method. They are reported as Margin1 to Margin3 in Table I, and in Fig. 5. The three samples produce very slightly different estimates of the LJ parameters, achieve good Birge ratios near unity, but present marked differences on the variance matrix parameters:

- Margin1 corresponds to an extreme negative correlation of ϵ and σ , and to large values of both u_ϵ and u_σ. This solution gives a prediction band very similar to the one of VarInf_Rb (Fig. 5), and leads to the same excess in mean prediction uncertainty (Table I). Its Birge ratio value (0.88) is the smallest of the 3 modes.

- Margin2 corresponds to a minimal value of u_ϵ and an undetermined value of ρ. Con-

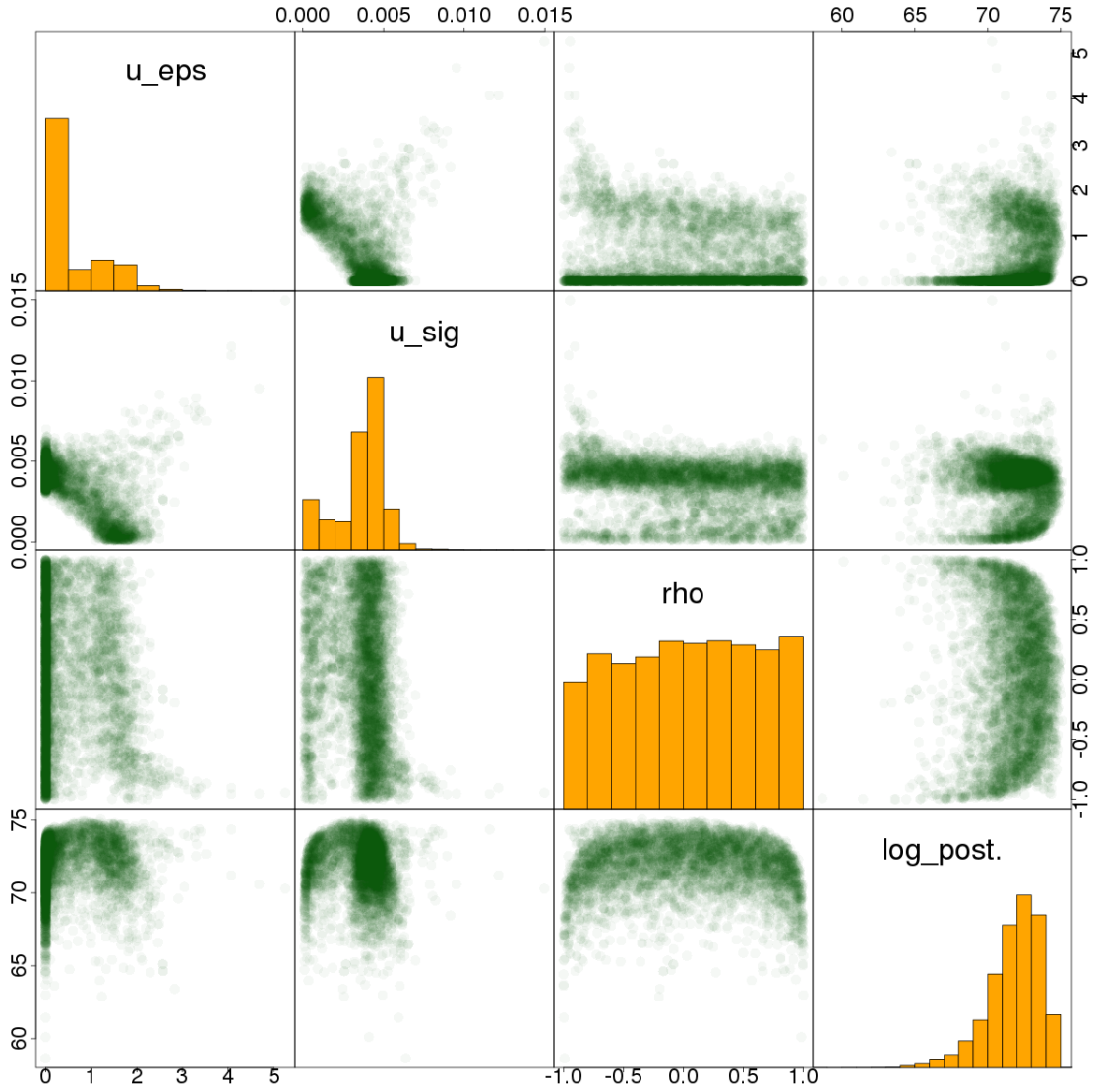

Figure 6. Scatterplots matrix of the posterior sample for the Margin method's covariance matrix parameters, showing its multimodality; the diagonal provides histograms of the parameters posterior distribution, while the out-of-diagonal plots represent a projection of the sample in the 2D space of the corresponding parameters pair (the upper and lower matrices are redundant); "log_post." is the logarithm of the posterior pdf.

sidering the prediction bands the Margin2 mode appears to have a major contribution to the Margin sample.

- Margin3 is the symmetrical of Margin2, with a minimal value of u_σ, and corresponds to an almost uniform prediction band.

The direct PUI methods Margin and ABC are therefore subject to a degeneracy in the optimal hyperparameters describing the stochastic LJ parameters, the implications of which are presented in the next section.

IV. DISCUSSION AND CONCLUSION

We have presented PUI methods in a unified formalism and established links between these methods, and with other methods in uncertainty quantification. We have shown through a series of examples in different contexts that existing methods attempting to capture model inadequacy errors in the covariance matrix of the model parameters present a series of problematic properties.

The prediction uncertainty bands are constrained by the functional form of the sensitivity coefficients of the model (Eq. 8), notably when using a simple inflation of the data variance (indirect method, Sect. II B 1). This has been apparent in the three examples above, and it might lead to areas of the control parameter with systematic under- or over-estimation of the prediction uncertainty. Unfortunately, this information is hard to gather directly from the literature, as the authors report typically the mean prediction uncertainty, or histograms of prediction uncertainties, which mask trends or systematic effects along the control variable. We can only recommend that authors working with these methods provide more informative/detailed representations of prediction uncertainties, and discuss the impact of under- or over-estimated prediction uncertainties on their intended use.

The influence of the model sensitivity coefficients can be modulated by the covariance matrix of the parameters, when the latter is optimized (direct methods, Sect. II B 2). We have seen in Sect. III C, that direct PUI methods based on a stochastic representation of the model's parameters might present degenerate modes leading to very different shapes of the prediction uncertainty bands, and that one has no *a priori* criterion to choose among them.[64] However, it is interesting that one of the modes of the Margin method is similar to the solution obtained by the scaling of data variance, and that this mode achieves sub-optimal statistics, both for its Birge ratio and for its mean prediction uncertainty. This would suggest that the indirect PUI method does not provide the best solution to the prediction uncertainty estimation problem. As discussed by Pernot and Cailliez[32], the posterior pdf multimodality/degeneracy might imply an undesirable high sensitivity of the prediction band shape to changes in the calibration dataset. Besides, the multimodality problem of the posterior pdf can be expected to increase with the number of parameters.

Considering the direct PUI method, the Margin method has no tuning option that would enable a performance improvement. At the opposite, the empirical likelihood on which the ABC method is based enables to envision additional constraints which might help to relieve the multimodality problem. This is in our opinion the most promising route to a design a

satisfying PUI method. But, we have also seen that these constraints tend to produce sub-optimal residuals, leading to a compromise between fit quality and prediction uncertainty quality.

Linear uncertainty propagation (LUP) has been used to estimate the mean value and covariance matrix of the model predictions in the Margin and ABC methods (Eqns. 19, 20 and 24). In the present application (Ar viscosity), the uncertainty on the model's parameters is small (less than 1%) and the viscosity model is monotonous and continuous in the LJ parameters variation range. There is no reason to be worried about uncontrolled non-linearity effects. However, this is not necessarily the case for other models, and the use of LUP has to be handled with care. For instance, Pernot and Cailliez[32] validated the use of LUP in a similar application by estimating the relative errors between an LJ parameter-wise linear approximation of the viscosity model over the whole T range and a sample of model values for LJ parameters drawn from their posterior pdf.

The main conclusion of this study is that methods to estimate prediction uncertainty of inadequate models based on parameters uncertainty inflation have to be used with great care and subjected to careful inspection, both of parameter space, and of prediction uncertainty trends. There is no sense in using prediction uncertainties if they are not reliable.

SUPPLEMENTARY MATERIAL

See supplementary material for the data and codes used in Sections III A and III C.

REFERENCES

[1]S. Glotzer, S. Kim, P. Cummings, A. Deshmukh, M. Head-Gordon, G. Karniadakis, L. Petzold, C. Sagui, and M. Shinozuka, "International assessment of research and development in simulation-based engineering and science," Tech. Rep. (World Technology Evaluation Center, Inc. (WTEC), 2009).

[2]K. K. Irikura, R. D. Johnson, and R. N. Kacker, Metrologia **41**, 369 (2004).

[3]C. I. Williams and M. Feher, J. Comput.-Aided Mol. Des. **22**, 39 (2008).

[4]M. Feher and C. I. Williams, J. Chem. Inf. Model. **52**, 724 (2012).

[5]P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences* (Cambridge University Press, 2005).

[6]A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. (Chapman and Hall/CRC, 2013).

[7]R. McElreath, *Statistical Rethinking*, Texts in Statistical Science (CRC Press, 2015).

[8]A. O'Hagan, J. Stat. Plan. Inference **143**, 1643 (2013).

[9]K. K. Irikura, J. Phys. Chem. Ref. Data **36**, 389 (2007).

[10]R. Johnson III, "Vibrational frequency scaling factors for CCD/6-31G* ." NIST Computational Chemistry Comparison and Benchmark Database, Release 17b; NIST Standard Reference Database Number 101 (2006).

[11]F. Cailliez and P. Pernot, J. Chem. Phys. **134**, 054124 (2011).

[12]J. P. Perdew, AIP Conference Proceedings **577**, 1 (2001).

[13]K. Raghavachari and A. Saha, Chem. Rev. **115**, 5643 (2015).

[14]M. C. Kennedy and A. O'Hagan, J. R. Stat. Soc. B **63**, 425 (2001).

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

[16]K. Lejaeghere, J. Jaeken, V. V. Speybroeck, and S. Cottenier, Phys. Rev. B **89**, 014304 (2014).

[17]P. Pernot, B. Civalleri, D. Presti, and A. Savin, J. Phys. Chem. A **119**, 5288 (2015).

[18]S. De Waele, K. Lejaeghere, M. Sluydts, and S. Cottenier, Phys. Rev. B **94**, 235418 (2016).

[19]J. Proppe, T. Husch, G. N. Simm, and M. Reiher, Faraday Discuss. **195**, 497 (2016).

[20]J. Proppe and M. Reiher, J. Chem. Theory Comput. **13**, 3297 (2017).

[21]K. Campbell, *The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004) - SAMO 2004*, Reliab. Eng. Syst. Safe. **91**, 1358 (2006).

[22]T. A. Oliver, G. Terejanu, C. S. Simmons, and R. D. Moser, Comput. Methods in Appl. Mech. Eng. **283**, 1310 (2015).

[23]J. J. Mortensen, K. Kaasberg, S. L. Frederiksen, J. K. Norksov, J. P. Sethna, and K. W. Jacobsen, Phys. Rev. Lett. **95**, 216401 (2005).

[24]J. Wellendorff, K. T. Lundgaard, A. Møgelhøj, V. Petzold, D. D. Landis, J. K. Nørskov, T. Bligaard, and K. W. Jacobsen, Phys. Rev. B **85**, 235149 (2012).

[25]K. Lejaeghere, V. V. Speybroeck, G. V. Oost, and S. Cottenier, Crit. Rev. Solid State Mater. Sci. **39**, 1 (2013).

[26]J. Wellendorff, K. T. Lundgaard, K. W. Jacobsen, and T. Bligaard, J. Chem. Phys. **140**, 144107 (2014).

[27]M. Pandey and K. W. Jacobsen, Phys. Rev. B **91**, 235201 (2015).

[28]K. Lejaeghere, L. Vanduyfhuys, T. Verstraelen, V. V. Speybroeck, and S. Cottenier, Comput. Mater. Sci. **117**, 390 (2016).

[29]G. N. Simm and M. Reiher, J. Chem. Theory Comput. **12**, 2762 (2016).

[30]K. Sargsyan, H. N. Najm, and R. Ghanem, Int. J. Chem. Kinet. **47**, 246 (2015).

[31]M. K. Transtrum, B. B. Machta, and J. P. Sethna, Phys. Rev. E **83**, 036701 (2011).

[32]P. Pernot and F. Cailliez, AIChE Journal , n/a (2017).

[33]Boldface type refers to vectors or matrices.

[34]BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML, "Evaluation of measurement data - Guide to the expression of uncertainty in measurement (GUM)," Tech. Rep. 100:2008 (Joint Committee for Guides in Metrology, JCGM, 2008).

[35]A. Vehtari and J. Ojanen, Statist. Surv. **6**, 142 (2012).

[36]R. T. Birge, Phys. Rev. **40**, 207 (1932).

[37]R. N. Kacker, A. Forbes, R. Kessel, and K.-D. Sommer, Metrologia **45**, 257 (2008).

[38]BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML, "Evaluation of measurement data - Supplement 1 to the "Guide to the expression of uncertainty in measurement" - Propagation of distributions using a Monte Carlo method," Tech. Rep. 101:2008 (Joint Committee for Guides in Metrology, JCGM, 2008).

[39]K. S. Brown and J. P. Sethna, Phys. Rev. E **68**, 021904 (2003).

[40]S. L. Frederiksen, K. W. Jacobsen, K. S. Brown, and J. P. Sethna, Phys. Rev. Lett. **93**, 165501 (2004).

[41]O. Bodnar and C. Elster, Metrologia **51**, 516 (2014).

[42]K. Csilléry, M. G. Blum, O. E. Gaggiotti, and O. François, Trends Ecol. Evol. **25**, 410 (2010).

[43]M. Sunnaker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz, PLoS Comput. Biol. **9**, e1002803 (2013).

[44]J. A. Pople, H. B. Schlegel, R. Krishnan, D. J. Defrees, J. S. Binkley, M. J. Frisch, R. A. Whiteside, R. F. Hout, and W. J. Hehre, Int. J. Quantum Chem. **20**, 269 (1981).

[45]D. J. DeFrees and A. D. McLean, J. Chem. Phys. **82**, 333 (1985).

[46]G. Rauhut and P. Pulay, J. Phys. Chem. **99**, 3093 (1995).

[47]A. P. Scott and L. Radom, J. Phys. Chem. **100**, 16502 (1996).

[48]M. D. Halls, J. Velkovski, and H. B. Schlegel, Theor. Chem. Acc. **105**, 413 (2001).

[49]J. Neugebauer and B. A. Hess, J. Chem. Phys. **118**, 7215 (2003).

[50]K. K. Irikura, R. D. Johnson, and R. N. Kacker, J. Phys. Chem. A **109**, 8430 (2005).

[51]P. Pernot and F. Cailliez, J. Chem. Phys. **134**, 167101 (2011), arXiv:arXiv:1010.5669.

[52]R. Johnson III, "NIST Computational Chemistry Comparison and Benchmark Database, Release 14; NIST Standard Reference Database Number 101," (2006).

[53]R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2015).

[54]For the improvement of this invaluable database, I strongly encourage to report any observed data problem to the CCCBDB curator, through the error form at http://cccbdb.nist.gov/errorformx.asp.

[55]P. Pernot and F. Cailliez, ArXiv e-prints **1010.5669** (2010), arXiv:1010.5669 [physics.chem-ph].

[56]K. K. Irikura, R. D. Johnson, R. N. Kacker, and R. Kessel, J. Chem. Phys. **134**, 167102 (2011).

[57]R. L. Jacobsen, R. D. Johnson, K. K. Irikura, and R. N. Kacker, J. Chem. Theory Comput. **9**, 951 (2013).

[58]V. Petzold, T. Bligaard, and K. W. Jacobsen, Top. Catal. **55**, 402 (2012).

[59]A. J. Medford, J. Wellendorff, A. Vojvodic, F. Studt, F. Abild-Pedersen, K. W. Jacobsen, T. Bligaard, and J. K. Norskov, Science **345**, 197 (2014).

[60]V. Stevanović, S. Lany, X. Zhang, and A. Zunger, Phys. Rev. B **85**, 115104 (2012).

[61]A. Gelman, D. Lee, and J. Guo, J. Educ. Behav. Stat. **40**, 530 (2015).

[62]Stan Development Team, *RStan: the R interface to Stan* (2016), R package version 2.14.1.

[63]M. D. Hoffman and A. Gelman, Journal of Machine Learning Research **15**, 1593 (2014).

[64]Note that when the reference data are abundant, which was not the case in the chosen example, keeping aside a validation set might help in this regard.