DATA-AWARE CALIBRATION OF PHYSICS-BASED COMPUTER MODELS OF ENGINEERING SYSTEMS: QUANTIFYING FLEXIBILITY OF CALIBRATION CAMPAIGN

Zhiyuan Qin, Clemson University, USA

ABSTRACT

Model calibration entails inferring the uncertain parameters of a computer model of an engineering or a natural system from available measurements of system response. In this context, flexibility can be defined as formulating an overly flexible model calibration campaign would lead to models with inferior predictive capability, meaning that while model calibration seemingly reduces the *fitting error* (error in model predictions at the settings of the calibration experiments), it increases the *prediction error* (error in model predictions at untested settings). In this paper, we argue that for a given set of available experiments, an optimally flexible model calibration campaign would yield maximum generalizability in calibrated model predictions. Factors that affect flexibility due to model itself are identified in the literature as the number of calibration parameters, their range, functional form of the model, experimental uncertainty. The availability of experiments, such as the number of experiments and their distribution within the domain of applicability, also affect the flexibility of the calibration campaign. An additional factor that can be considered is the independent discrepancy model that is empirically trained to represent systematic bias in model predictions due to model incompleteness. Taking all the these factors into account, we present a generally applicable approach for quantifying flexibility of model calibration, and demonstrate the proposed approach along with its accompanying metric on several academic examples.

Keywords: Predictive Modeling, Goodness-of-fit, Model Selection, Prediction Maturity, Model Updating, Model Validation, Uncertainty Quantification.

1.0 INTRODUCTION

Physics-based computer models of engineering systems are instruments of prediction that define a functional relationship between parameters that control the operational conditions of a system (input) and the system response of interest (output). In this context, these input parameters, known as *control variables*, define the domain of applicability, in which the system operates. Here, we will assume that experimentalists have control over and full knowledge of these control variables. While developing a model that links control variables to the output responses of interest (for instance, loads acting on a structure to deformations), other input parameters that define the characteristics of the system (for instance, material properties or boundary conditions of the structure) are also introduced to the model. Oftentimes, a number of these input parameters that define the system characteristics are poorly-known, and thus, must be inferred from experimental measurements[1]; while the rest of the parameters are accepted as well known. We refer to this subset of poorly-known input parameters that are selected for such inference as *calibration parameters*. Hence, given the values for well-defined control variables and poorly-known calibration parameters, simulation models are conceived to predict unknown output responses within a predefined domain of applicability.

Model calibration then entails estimating the best-fit values for a few calibration parameters (which are believed to be identifiable) from experiments conducted at various control parameter settings within this domain of applicability. Of course, we should not expect to get the complete 'truth' from the model, for they are mere approximations, and hence, incomplete (i.e. systematically biased) representations of the underlying behavior of the system. Such incompleteness may originate from, for instance omission of input parameters from the model, omission of interactions between the model input parameters and/or control variables, or assigning incorrect values to model input parameters that are considered to be known. Thus, computer models invariably have systematic discrepancy biases in the way they predict the true behavior of the systems. This inherent discrepancy bias may be identified during model calibration by inferring an independent error model from the experimental data (Kennedy and O'Hagan 2001; Bayarri et al. 2002; Higdon et al. 2007; Farajpour and Atamturktur 2013) or by

[1] Common approach for selecting calibration parameters entails not only evaluating their uncertainty but also their influence (i.e. sensitivity) on the model output of interest (Atamturktur et al. 2012)

blending emulators with mechanistic models to explain the omitted relationships between model input parameters (Atamturktur and Brown 2015).

Here, we define the *flexibility* of the model calibration as the freedom[2] assigned to the model to conform to the experiments at the tested settings within the domain of applicability. According to this definition, calibration campaign formulated for the same model may be very different depending on the choices made, for instance, in the selection of calibration parameters and their plausible ranges, as well as the representation of the model discrepancy bias with an independent error model. With this interpretation, a physics-based model with no calibration parameters and no error model to correct for discrepancy bias would have zero flexibility. Only when we acknowledge the uncertainties in the input parameters and the presence of discrepancy bias in model predictions and allow these parameters to be bias-calibrated and model predictions to be corrected against experimental measurements does the issue of flexibility of model calibration emerges. Hence, similar to Sober (1975), we can interpret the flexibility of a model calibration campaign to be related to the amount of information that needs to be inferred from the experimental data to be able to use the model in a predictive manner.

In model calibration, the goodness-of-fit of a model to experiments reflects how well a model fits a particular set of observed data. A good fit is a necessary but not a sufficient condition (Bard 1974; Roberts and Pashner 2000), as it is possible to calibrate physics based models to different sets of calibration parameter values that can fit a finite set of experiments reasonably well due to the inevitable compensations between various sources of errors and uncertainties (Brooks & Tobias, 1996, Li et al, 1996; Atamturktur et al., 2014a). In contrast to goodness of fit, generalizability is defined as the ability of a model to represent the reality of interest at all settings of the domain, including the settings where experiments are not available (MacKay 1992). Generalizability of a calibrated model is important as computer models are most often calibrated with the ultimate objective of predicting at settings for which experiments are unavailable.

[2] Cutting et al. (1992) called this freedom “scope” of a model.

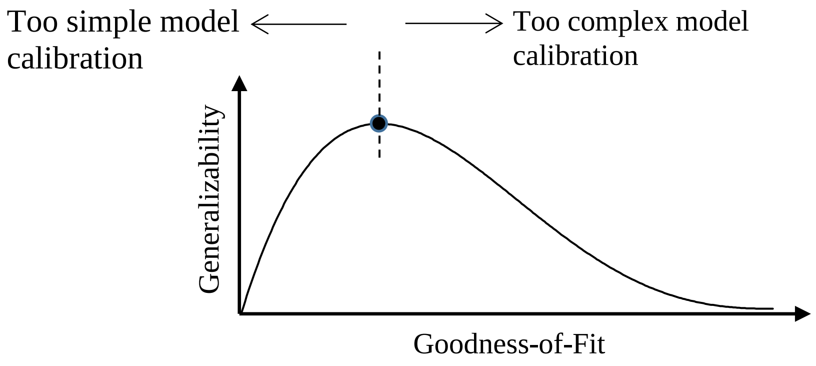


Fig. 1: Notional relationship between goodness-of-fit and generalizability (referred to as Ockham's Hill by MacKay 1992)

The complexity of model calibration campaign leads to a hill-like relationship between good fitness to a finite amount of noisy measurements (at the tested settings) and generalizability of the model predictions (at the untested settings) (MacKay 1992, Pitt & Myung 2002, Schoups et al., 2008) (See Fig. 1). A model calibration campaign with too little flexibility would lose valuable information that could have otherwise been inferred fom the data (Gauch 1993, MacKay 1992, Ruano 2007). On the other hand, a calibration campaign with too much flexibility would encourage the model to fit to noise in the measurements seemingly improving the goodness-of-fit while degrading the generalizability (see Fig. 1)[3]. Hypothetically, in the most extreme case, a calibration campaign that can produce a model capable of matching practically any possible outcome (infinite flexibility) yields an uninformative tool that is impossible to falsify (Popper 1959) and that has no generalizability (tail end of Fig. 1). For physics-based models, the inherent functional structure of the model would impose a differential ability to fit patterned data and would prevent us from reaching this hypothetical infinite flexibility. Such differential ability of a model was referred to as "selectivity" by Cutting et al. (1992).

In the published literature, five major contributors of flexibility of calibration campaign were identified. The first three of these factors contribute to the flexibility of the model itself: (i) number of calibration parameters to be inferred from the available measurements; (ii) the range within which the parameters were allowed to vary during model calibration; and (iii) functional

[3] This concept, referred to as *degenerative* by Forster and Sober (1994), is widely known as overfitting.

form of the model's mathematical expression (or model form) (Myung & Pitt, 1997; Bozdogan & Haughton, 1998; Zucchini, 2000; Dunn 2000; Pitt et al., 2002; So, 2003). The last two factors relate to the amount of available calibration experiments[4]: (iv) experimental uncertainty, which reduces the restraining effect measurements place on model calibration and (v) the number of available experimental measurement points (Forster and Sober 1994; Myung & Pitt, 1997; Pitt et al., 2002; Moriasi 2007). In addition to these five factors, we also consider the coverage of the experiments (i.e. how well the experiments explore the domain of applicability) (Atamturktur et al. 2015) and the role of independent, empirical error models inferred during model calibration process to correct for discrepancy bias.

Taking these aforementioned factors into account, in this paper, we propose a quantitative and generally applicable metric to quantify the flexibility of a model calibration campaign. Specifically, we evaluate the extreme values a model's output can assume all while confirming to available experiments. Furthermore, we demonstrate the effect of the aforementioned factors on the proposed metric using academic examples and compare the performance of the metric against existing criteria developed for model selection.

This paper is organized as follows. Section 2 provides a background review of the established literature on complexity and its role in model selection. Section 3 makes a description of the quantitative metric for determining complexity of the model calibration. In Section 4, it demonstrates the effect of five aforementioned factors on the proposed calibration complexity and compares against existing model selection criteria through controlled examples. Next, using the proposed metric, the Section 5 introduces a concept for data-aware calibration and demonstrates its application with a case study. Finally, in section 6, it makes a discussion of the results and concludes remarks.

[4] Here, with the term '*amount of measurements,'* we imply all relevant aspects of the calibration measurements, namely the number of measurements, how well they explore the domain of applicability (also known as coverage) and experimental measurements repeatability (lack of uncert*ainty*) (Atamturktur et al., 2015) .

2.0 BACKGROUND

2.1 Complexity and Model Selection

In the context of physics-based modeling and simulation where the underlying model structure is dictated by first-principle physics, the term *complex*[5] was often used to mean *detailed* (Webster et al., 1984; Smith & Vaughan, 1980; Ward, 1989) and determining the appropriate level and type of detail was considered to be one of the most important steps in the formulation of a simulation model (Law 1991; Salt 1993). It was advised, for instance, to start from a simple model, progressively add details until sufficient accuracy is obtained and select the least detailed model that meets the modeling objectives (Brooks & Tobias, 1996; Pidd 1996; Hill 1998).

In the context of empirical curve fitting, Sober (1975) argued that models that are more informative are less complex. Kuhn (1977) stated that everything else being equal, it is rational to prefer a simpler model over a more complex one. Turney (1990) showed that simpler models tend towards a greater stability (or robustness) in face of experimental uncertainty. Many approaches have been developed for purposes of comparing alternative models built with varying levels of detail, which in turn has resulted in a new branch of mathematical statistics known as *model selection*. Although originally conceived to aid the model formulation process, the model selection criteria originated from this field also supply means for comparing alternative model calibration campaigns (see Table 1 for a list of common model selection criteria). These criteria typically consider both goodness-of-fit and complexity and differ in their representation of the latter.

Table 1 Widely used model selection metrics

Selection Metric	Criterion Equation
Akaike information criterion (AIC)	$AIC = -2 * \ln\left(f(y\middle|\widehat{\boldsymbol{\theta}})\right) + 2k$
Bayesian information criterion (BIC)	$BIC = -2 * \ln\left(f(y\middle|\widehat{\boldsymbol{\theta}})\right) + 2k * \ln(n)$
Deviance information criterion (DIC)	$DIC = \overline{D(\boldsymbol{\theta})} + p_D$
Information-theoretic measure of complexity (ICOMP)	$ICOMP = -\ln\left(f(y\middle|\widehat{\boldsymbol{\theta}})\right) + \frac{k}{2}\ln\left(\frac{trace\left[\boldsymbol{\Omega}\left(\widehat{\boldsymbol{\theta}}\right)\right]}{k}\right) - \frac{1}{2}\ln\left(\det\left[\boldsymbol{\Omega}\left(\widehat{\boldsymbol{\theta}}\right)\right]\right)$

[5] See Brooks & Tobias, 1996 for a thorough discussion on various meaning complexity has taken in engineering and science literature.

Minimum description length (MDL)	$MDL = -\ln\left(f(y\|\hat{\boldsymbol{\theta}})\right) + \frac{k}{2}\ln\left(\frac{n}{2}\right) + \ln\left(\int\ d\hat{\theta}\det\left[\boldsymbol{I}(\hat{\boldsymbol{\theta}})\right]\right)$

Note: y = data function; n =sample of size; $\hat{\boldsymbol{\theta}}$ = parameter value that maximizes the likelihood function $f(y|\hat{\boldsymbol{\theta}});$ k = number of parameters; D is the deviance of the likelihood, $D(\hat{\boldsymbol{\theta}}) = -2 * log(f(y|\hat{\boldsymbol{\theta}});$ $p_D = \overline{D(\hat{\boldsymbol{\theta}})} - D(\bar{\theta}),$ $\overline{D(\theta)}$ is the expectation of $D(\hat{\boldsymbol{\theta}})$ and $\bar{\theta}$ is the expectation of $\hat{\boldsymbol{\theta}}; \Omega$ = covariance matrix of the parameter estimates; ln= the natural logarithm of base e.

For instance, the Akaike Information Criterion (AIC) (Akaike, 1973) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) represent a model's complexity considering only the number of calibration parameters. Both methods use likelihood to assess the model goodness-of-fit and then penalize it with model complexity represented by the number of calibration parameters. Although user-friendly, the number of calibration parameters alone is not a sufficient definition of complexity. This can be demonstrated by considering two models $y_1 = \theta * x$ and $y_2 = \theta + x$ which have the same number of calibration parameters θ but different functional forms (multiplicative versus additive). Despite both having one calibration parameter, these models have vastly different data-fitting abilities. AIC and BIC also fail to discount parameters which are not constrained by the data. A Bayesian generalization of AIC, the Deviance information criterion (DIC) overcomes the problem (Spiegelhalter, Best, Carlin, & van der Linde, 2002). The complexity term of DIC is calculated by assessing the number of parameters that can be constrained by experiments (a concept referred to as the effective number of parameters) (Liddle, 2007). DIC is calculated in a straight forward manner using Monte Carlo posterior samples (Parkinson et al. 2006). The calculation is easier performed with posterior samples generated by nested sampling, which have non-integer weights, than AIC and BIC. However, no efficient method has been developed for calculating reasonably accurate MC standard errors of DIC.

The three criteria, AIC, BIC and DIC are all sensitive only to one aspect of complexity: the number of parameters (the effective number for DIC. Furthermore, all calibration parameters in these criteria are considered to have equal contribution to the complexity of model as functional form and range of parameters are not considered. Information-Theoretic Measure of Complexity (ICOMP) criterion (Bozdogan, 2000) overcomes these shortcomings by considering not only the number of calibration parameters but also the effects of their sensitivity and interdependence. From the table 2, the second and third terms together represent a complexity measure that takes into account the effects of parameter sensitivity through the trace and parameter interdependence through the determinant which are two principal components of the functional form that contribute to model complexity (Li, Lewandowski, & DeBrunner, 1996).

However, Pitt et al. (2002) emphasized the importance of model selection metrics being invariant under reparameterization and recognized ICOMP criterion's inability to remain invariant. Being invariant under reparameterization means that when parameters of the model are transformed without loss of information, and the new model with transformed parameters (that behaves equivalently with the original model) should have the same complexity value as the original. For instance, if the π^a in $y = sin(\pi^{ax} + b)$ is transformed into a new parameter, c, the new model $y = sin(c^x + b)$ should be identified as equivalent by the model selection metric. AIC, BIC and ICOMP would however consider these two models to have different complexities.

Cutting et al. (1992) also recognized that the number of parameters alone is an insufficient indicator of model complexity and advocated for evaluating the fitting power (i.e. what they refer to as 'scope') of a model to random data. They suggested using binomial tests to compare the fitting ability of a model to the data from actual system with the fitting ability to random data. Similarly, complexity has been defined as the range of data patterns that a model can fit (Myung et al. 2000 and Pitt et al. 2002) quantified by a geometric complexity measure known as the Minimum Description Length (MDL) (Grunwald, 2000; Rissanen, 1983,1996). This metric considers the experimental data as a code or description to be compressed by the model and evaluates the models according to their ability to compress a data set by extracting necessary information from the data without random noise. MDL is based on the understanding that the more data is compressed, the more information about the underlying regularities governing the process of interest would be learnt (Pitt et al., 2002). Therefore, MDL would chooses a model which has the shortest description code (length) of the data (Pitt et al., 2006). MDL has been criticized for not fully considering the parameter interdependencies in model fitting and selection process (Bozdogan, 2003).

In our study, these aforementioned metrics will be used as baseline for comparisons.

3.0 A NEW QUANTITATIVE METRIC FOR CALIBRATION COMPLEXITY

For a fixed set of experiments, increasing the flexibility of a model calibration campaign in general improves the goodness-of-fit to available experiments. Up to a certain point, this added complexity might indeed be benefitting the model[6] and improving the generalizability; beyond that point, the model starts to fit to the noise in the measurements (referred to as Ockham's hill by McKay 1992). This concept was demonstrated earlier in Figure 1. For a given model, at what point over fitting starts depends strongly on the calibration experiments: (1) How many experiments available? (2) What are their uncertainties? (3) How well do these experiments explore the domain? Figure 2 demonstrate that a model calibration too flexible for a given set of experiments may lead to a model that can predict vastly incorrect values away from the experiments: the more complex the model calibration campaign, the worse the predictions may become away from the experiments.

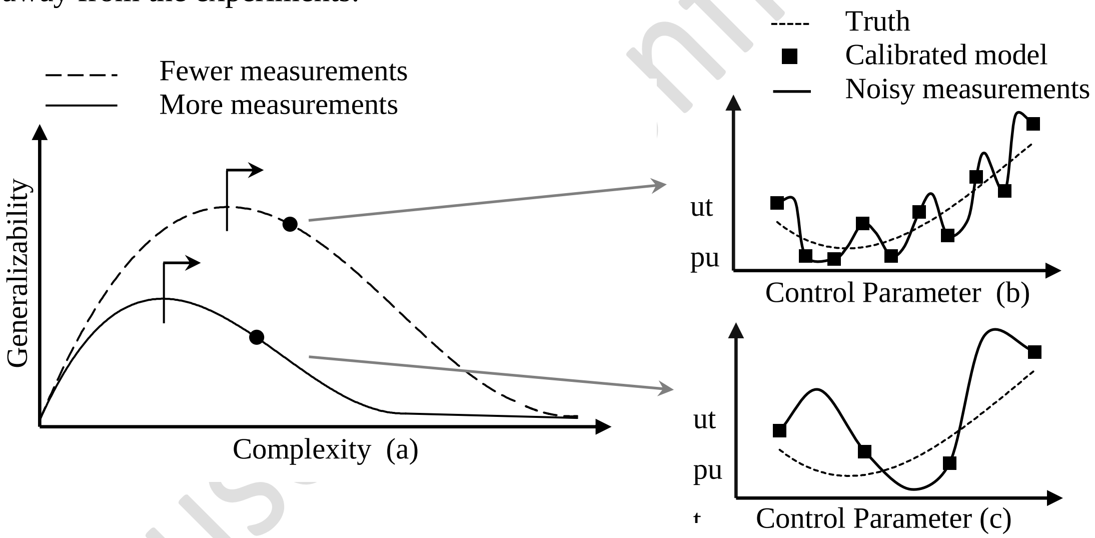


Fig. 2 (a) Notional relationship between complexity and generalizability for differing amount of experiments (b) Calibration with more measurements (c) Calibration with fewer measurements.

[6] Here, we use the term "model" to mean a specific functional form with a predefined set of parameters.

3.1 Indicator for Calibration Flexibility (ICaF)

We articulate that the range of possible predictions that a model can produce at untested settings of the domain of applicability while conforming to the available measurements[7] reflects the flexibility of model calibration campaign. Thus, we calculate the *extreme values* a calibrated model can assume at untested settings while complying with the experiments at the tested setting. Here, these extreme values indicate the lowest and highest possible values the model can be *forced* to assume given the bounded uncertainties assigned to calibration parameters. These lowest and highest model predictions are obtained for the entirety of the domain of applicability.

In the process of obtaining the ICaF, θ is the calibration parameter, and n is the total number of available measurements. The upper and lower bounds for the experimental values are represented by u_j and l_j, respectively. At each grid point within the domain applicability of the control variables, the algorithm searches for the θ value that enables the model to produce the lower and upper extreme values, while conforming to the bounds defined by the experimental uncertainty. To assure that the model predictions stay within the experimental limits, a penalization factor of P (100 in this study) is included in the algorithm. The aforementioned process can be visualized as stretching an elastic band constrained at a number of points where experiments are conducted throughout its length. The algorithm determines how far this band can be stretched if it is pulled up (highest extreme values) or down (lowest extreme values) at various points throughout its length. Figure 3a provides a notional representation of this concept for a one-dimensional problem (i.e. with a single control parameter, *x*). Here, the domain of applicability is bounded $a <= x <= b$, and the flexibility of calibration is determined between these settings. Of course, the flexibility of model calibration would be different definitions of domains of applicability.

At a given control parameter setting, the difference between the upper and lower flexibility limits reflects the ability of the model to mimic a wide range of possible outputs. This ability would vary throughout the domain of applicability. The area that is bounded between the

[7] Here, the emphasis on "available" measurements is important. We are referring to the measurements that are used in the calibration process.

flexibility limits for a given domain of applicability is highlighted in Figure 3(b). This area will herein be used as an indicator for calibration flexibility (ICaF), and is calculated as the area over the entire domain of applicability,

$$ICa\mathrm{F} = \int_a^b [f(x,\theta)_{max} - f(x,\theta)_{min}]\, dx = \sum_{i=1}^{N} \frac{f(x_i,\theta)_{max} - f(x_i,\theta)_{min}}{b-a} \tag{1}$$

where N is the number of grid points within the domain at which flexibility limits be calculated. It is of course important to evaluate a high enough number of grid points to accurately determine complexity.

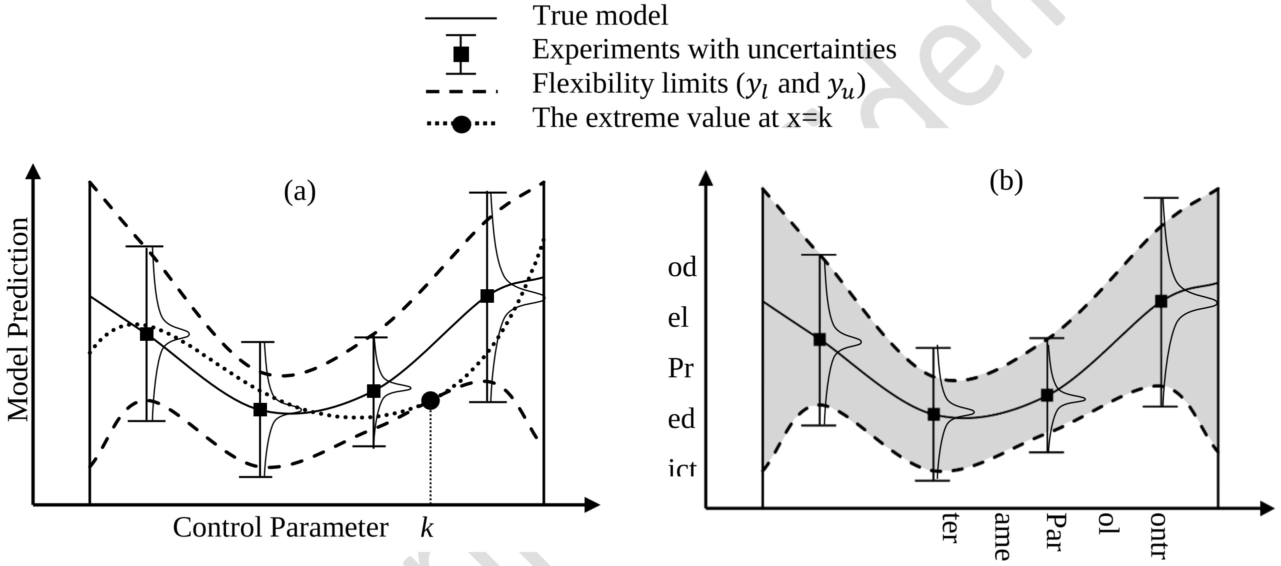


Fig.3 (a) Flexibility limits representing the extreme values calculated individually for each control parameter setting (b) The area between the upper and lower flexibility limits

For a given value of x_i, the Pseudo algorithm for the determination process of ICaF can be obtained as follows:

Given

x: grid point

i: the index of *x*

k: location of experiment

a: starting point of *x*

b: ending point of *x*

u_k*:* upper bounds for experiments value

l_k*:* lower bounds for experiments value

$mval$*:* model prediction value

$fval$*:* function value that optimized by PSO

d_1: absolute difference value between $mval_i$ and u_k

d_2: absolute difference value between $mval_i$ and l_k

P: penalization factor ($P_1=100$, $P_2=10000$)

θ*:* calibration parameter

Ub: upper bounds for calibration parameter

Lb: lower bounds for calibration parameter

$f(x_i, \theta)$*:* objective function

Find

$f(x_i, \theta)_{max}$ & $f(x_i, \theta)_{min}$

Satisfy

Bounds: $a<x_i<b$

System objective function

```
Do
    Calculate the $f(x_i, \theta)_{max}$
    Initialize $fval(\theta) = 0$
    For each grid point $x_i$
        For each experiments $k$
            If $mval_i > u_k$, then $fval(\theta) = fval(\theta)+P_1*d1$
            End If
            If $mval_k < l_k$, then $fval(\theta)= fval(\theta)+P_1*d2$
            End If
        End For
    $fval(\theta) = fval(\theta)+ P2- mval_i$
    $f(x_i, \theta)_{max}=10000-PSO(fval(\theta), Ub, Lb)$
    End For
End Do
Do
```

```
        Calculate the $f(x_i, \theta)_{min}$
        Initialize $fval(\theta) = mval_i$
        For each grid point $x_i$
            For each experiments $k$
                If $mval_i > u_k$, then $fval(\theta) = fval(\theta)$+P$_1$*d1
                End If
                If $mval_i < l_k$, then $fval(\theta) = fval(\theta)$+P$_1$*d2
                End If
            End For
        $f(x_i, \theta)_{min}$=PSO( $fval(\theta), Ub, Lb$)
        End For
End Do
System goals:
$ICaF = \sum_{i=1}^{N} \frac{f(x_i,\theta)_{max} - f(x_i,\theta)_{min}}{b-a}$
```

Algorithm 1: Pseudo algorithm for the determination process of $ICaF$

ICaF is useful in the process of determining the flexibility limits allows compensations between various terms uncertainties and errors to occur. Such compensations occur when the correlated parameters are calibrated simultaneously, the available experimental observations are insufficient, an incorrectly assigned parameter value compensates for another incorrect parameter value or acts to offset model bias, or errors in predictions of multiple outputs are combined into a single fidelity metric. In addition, *ICaF* fully considers the parameter interdependencies and refrain from the shortcomings of being not invariant to reparameterization.

3.2 Optimization

Throughout this study, the objective function is minimized using the Particle Swarm Optimization (PSO) (Eberhart and Kennedy, 1995), a probabilistic search algorithm inspired by the bird flocks seeking food. In this analogy, each bird (*particle)* in a population (*swarm*) tries to find the best accessible food resource (optimum solution) (Venter & Sobieszczanski-Sobieski, 2003). Hence, in out implementation, each particle within the swarm is assigned different starting values for the calibration parameters at a given control variable setting. Particle then searches for the value that will minimize the objective function. Each particle changes the parameter value according to its own recorded minimum value for the objective function (*local optimum*) and for the swarm’s recorded minimum value for the objective function (global optimum).

```
Given
  t: swarm size, 25
  V[]: the velocity of particle
  present[]: the current location of particle
  p-best[]: the local optimum value;
  g-best[]: the global optimum value;
  rand(): random value between 0 and 1;
  $c_1$: social acceleration coefficient, 1.3
  $c_2$: cognitive acceleration coefficient, 2.8
Find  fval(θ)
Optimization Loop
  Do
      For each particle
          Calculate fitness value
          If present [] is better than p-best[] , then p-best[] =present[]
          End If
      End For
  Choose the particle with the best fitness value of all the particles as the g-best[]
      For each particle
          Calculate particle velocity:
           $V[]=V[]+c_1*rand()*(p\text{-}best[]-present[])+c_2*rand()*(g\text{-}best[]-present[])$
          Update particle position:
           $fval(\theta) = present[] = present[] + V[]$
      End For
  End Do
```

Algorithm 2: Pseudo algorithm for optimization process (Note: p-best represents the local best, and g-best represents the global best)

In a single iteration of the PSO, a movement of particle is controlled by swarm size, social acceleration coefficient, and cognitive acceleration coefficient (Shi & Eberhart, 1998). Herein, the swarm size represents the total number of particles, the cognitive acceleration coefficient represents the private thinking of the particle itself, and the social acceleration coefficient represents the collaboration among the particles. In this study, the swarm size, the social acceleration coefficient and the cognitive acceleration coefficient are set at 25, 1.3 and 2.8, respectively (Farajpour et al. 2012). The movement of the particle is terminated when the difference between the best and worst objective function values is less than a predefined threshold value (10^{-6}). After termination, the values for the calibration parameters corresponding to the extreme values are assigned as the final value for the parameters at a certain grid point of the control parameter.

4.0 INFLUENTIAL ATTRIBUTES OF FLEXIBILITY

4.1 PARAMETER SET FOR CALIBRATION

As discussed earlier in the section 3, the number of parameters has been widely recognized as an indicator of complexity (Akaike 1970; Hockly and Murdock 1987; Raaismakers and Shiffrin 1992; Snowling et al. 2001; Hemez et al. 2010). Since there is little justification for calibrating a parameter that has negligible uncertainty or a negligible influence on the model output of interest (Li et al. 1996; Atamturktur et al. 2012), the calibration parameters must be typically selected considering both the uncertainty and the sensitivity of the parameters. It is also important that the selected parameters are independent in that they do not cancel out their respective effects on the model outputs, an aspect known as collinearity, which prevents successful identification of model parameters from the data (Brun et al. 2002). Generally speaking, reducing the number of calibration parameters by calibrating only a subset of uncertain parameters reduces complexity and improves the identifiability (i.e. the ability to find a unique solution in a deterministic sense). This would however mean that rest of the poorly known parameters are assumed to be fixed, and are potentially assigned incorrect values. On the contrary, increasing the number of calibration tend to allow the model produce larger number of patterns of output response leading to an increase in flexibility and therefore, in complexity of model calibration campaign. However, the number of parameters is an incomplete indicator of model complexity (Bamber 1985; Jaafar & Han 2012). The subset of parameters chosen to be calibrated is also very important (Weijers and Vanrolleghem, 1999; Brun et al. 2002; Ruano et al. 2007; Atamturktur et al. 2012). Two different parameter subsets with the same number of parameters may yield model calibration campaigns with significantly different flexibilities. Consider, for instance, a polynomial with four orthogonal calibration parameters:

$$y = ax^3 + bx^2 + cx + d, \tag{2}$$

where x represents control variable defined within [-1 5] and a, b, c and d are calibration parameters with exact, but unknown, values of $[0.5 - 1 - 5\ 20]$. One way to choose to calibrate as few as one or as many as all four calibration parameters from any one of the 15 combinations as shown in Table 1. Let us focus on the effect of the number of parameters on the complexity of calibration by minimizing the effect of other factors by (i) setting the range of calibration parameters to a constant 20% of their nominal (best initial estimate) for all parameters, (ii) keeping the model form constant so that parameter sensitivities do not change and (iii) keeping the experiments identical with four experiments distributed uniformly over the domain of applicability with degree of confidence 0.95 for the confidence intervals. (Fig. 5).

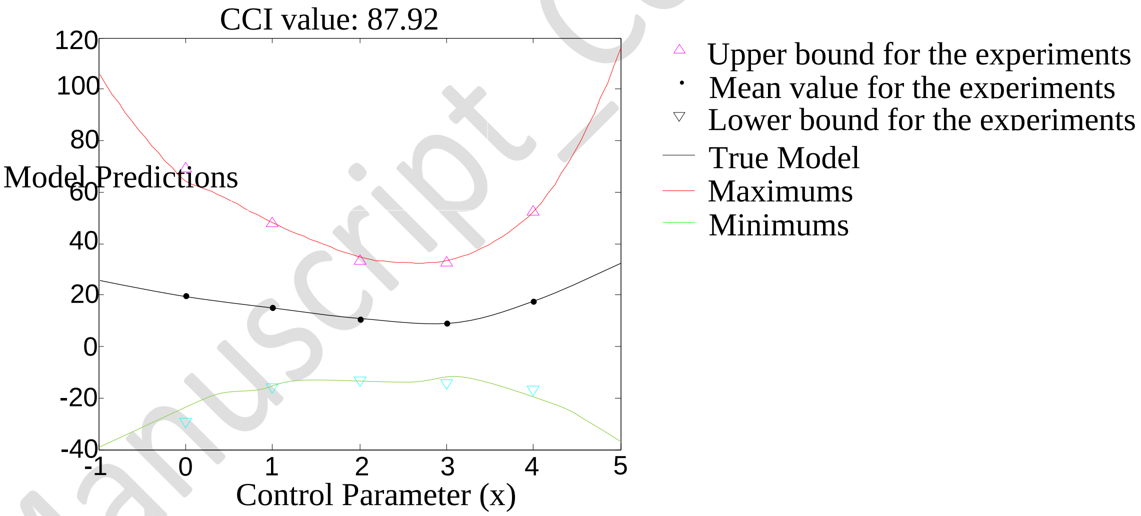


Fig. 5 The complexity of the combination with four parameters [a b c d]

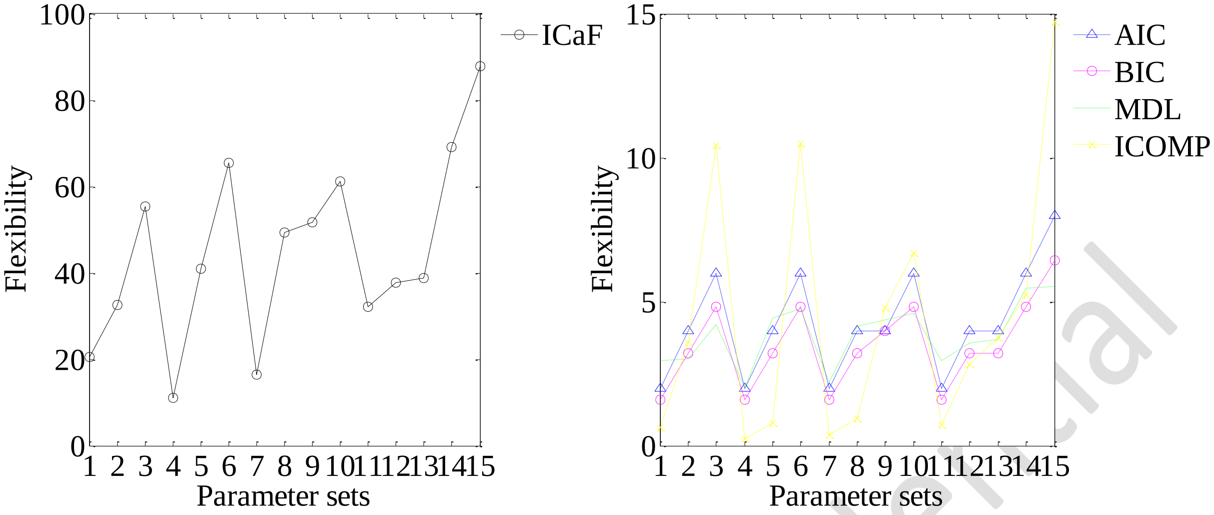


Fig. 6 Complexity Comparisons of Different Numbers of Calibration Parameters (1(a),2(a,b),3(a,b,c) ,4(b),5(b,d),6(a,b,d) ,7(c),8(c,d),9(b,c),10(b,c,d), 11(d),12(a,c),13(a,d),14(a,c,d), 15(a,b,c,d))

Table 2: Ranking of different metrics for complexity

Ranking	**Complexity**					**Generalizability**[9]
	AIC	**BIC**	**MDL**	**ICOMP**	**ICaF**	
1	b (2.00)	b (1.61)	b (2.01)	b (0.25)	b (11.05)	b (83.36%)
2	c (2.00)	c (1.61)	c (2.21)	c (0.38)	C (16.44)	c (84.35%)
3	a (2.00)	a (1.61)	a (2.94)	a (0.62)	a (20.44)	a (84.55%)
4	d (2.00)	d (1.61)	d (2.96)	d (0.73)	d (32.11)	d (85.10%)
5	a,c (4.00)	a,c (3.22)	a,b (3.01)	b,d (0.79)	a,b (32.51)	a,b (85.20%)
6	a,d (4.00)	a,d (3.22)	a,c (3.55)	c,d (0.95)	a,c (37.65)	a,c (85.42%)
7	a,b (4.00)	b,d (3.22)	a,d (3.69)	a,c (2.83)	a,d (38.77)	a,d (85.47%)
8	b,c (4.00)	b,c (4.00)	c,d (4.15)	a,b (3.60)	b,d (40.89)	a,b,c,d (85.53%)
9	b,d (4.00)	c,d (3.22)	b,c (4.38)	a,d (3.76)	c,d (49.29)	b,d (85.89%)
10	c,d (4.00)	a,b (3.22)	b,d (4.43)	b,c (4.80)	b,c (51.63)	c,d (86.06%)
11	a,c,d (6.00)	a,c,d (4.83)	a,b,c (4.21)	a,c,d (5.25)	a,b,c (55.45)	b,c (86.17%)
12	a,b,c (6.00)	a,b,d (4.83)	b,c,d (4.60)	b,c,d (6.70)	b,c,d (61.28)	a,b,c (86.28%)

13	a,b,d (6.00)	a,b,c (4.83)	a,b,d (4.77)	a,b,c (10.42)	a,b,d (65.51)	b,c,d (86.53%)
14	b,c,d (6.00)	b,c,d (4.83)	a,c,d (5.48)	a,b,d (10.49)	a,c,d (69.23)	a,c,d (87.17%)
15	a,b,c,d (8.00)	a,b,c,d (6.44)	a,b,c,d (5.52)	a,b,c,d (14.73)	a,b,c,d (87.92)	a,b,d (87.69%)

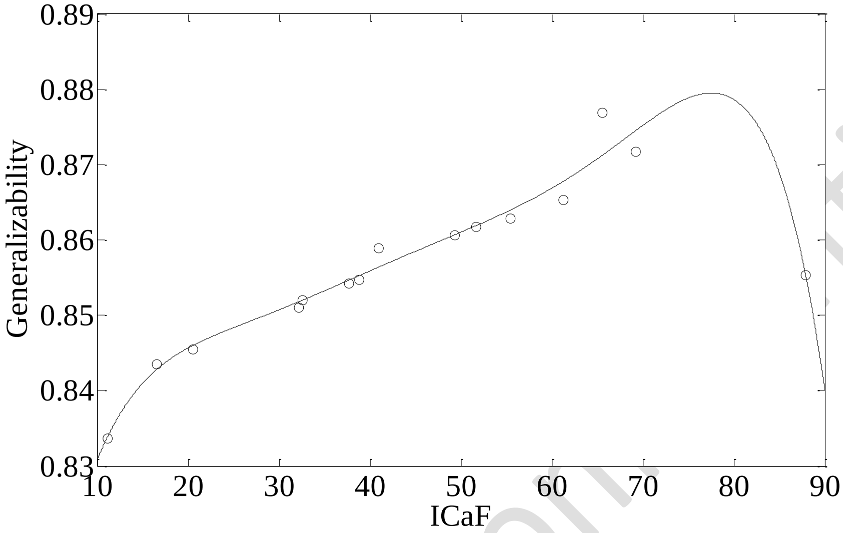


Fig. 7 The relationship between *ICaF* and Generalizability

4.2 RANGE OF THE CALIBRATION PARAMETERS

The prior knowledge of calibration parameter is another factor that affects the complexity of calibration. In the most general case, allowing the model parameters to vary in a wider range increases the flexibility of the model. However, determining the prior knowledge in the parameters selected for calibration is identified to be one of the most challenging aspects (Brun 2002). Here, we represent the parametric uncertainty with bounded uncertainty where the parameter value is enveloped within predefined lower and upper limits (Herbert Kay et al. 2000). This representation is particularly suitable in situations where reliable probabilistic information

[9] Generalizability (Gr) is used to express the ability of the model calibration to predict the true model. Generalizability is defined by the following equation (Pitt, Myung, & Zhang, Toward a method of selecting among computational models of cognition, 2002):

$$Gr(\%) = \left\{ 1 - \frac{\sum_{i=1}^{m} \left| y(x_i, \theta) - y_{truth}(x_i, \theta_n) \right|}{\sum_{i=1}^{m} \left| y_{truth}(x_i, \theta_n) \right|} \right\} \times 100$$

Where m is the number of discretized points that are used to evaluate Gr, and $y_{truth}(x_i, \theta_n)$ is the response of the true model.

about the parameters is not available. The higher the uncertainty about a parameter, the wider the assigned range would be, which in turn would allow greater flexibility to the calibration campaign.

Once again, we will refer to Equation 2 to evaluate the influence of the predefined bounds on the calibration parameter on complexity. We calculated complexity for 10 different parameter uncertainty levels keeping the experiments and model form constant at all times. Four calibration measurements are distributed uniformly over the domain of applicability with bounded uncertainties of 500% their mean value. The asymptotical relationship between the complexity of model calibration and the range of calibration parameters is shown in Fig. 8.

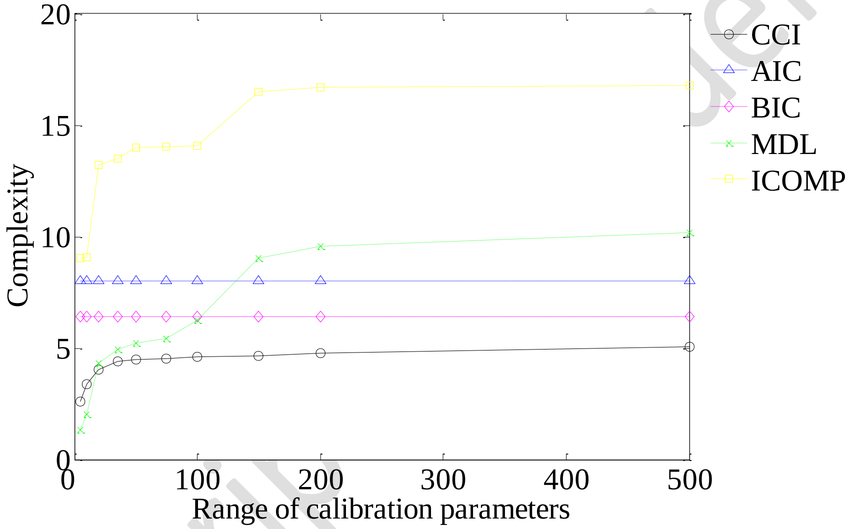


Fig. 8: Relationship between complexity and range of the calibration parameters

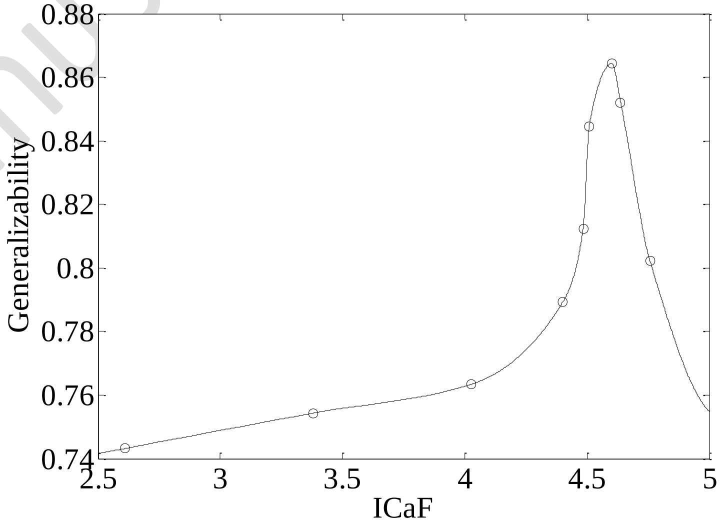


Fig. 9 The relationship between *ICaF* and Generalizability

4.3 EXPERIMENTAL COVERAGE

Experimental coverage refers to the ability of a set experiments represent the entire domain (Atamturktur et al., 2015). Experimental coverage therefore entails two different characteristics (the number of experiments and the distribution of experiments) both of which impact the complexity of a calibration campaign. Generally speaking more complex models tend to have a greater number of uncertain parameters and hence require a greater amount experiments (Snowling et al. 2001). Conducting new experiments at untested settings within the domain then forces the model to fit to these additional constraints, which in turn decreases the flexibility of the calibration campaign.

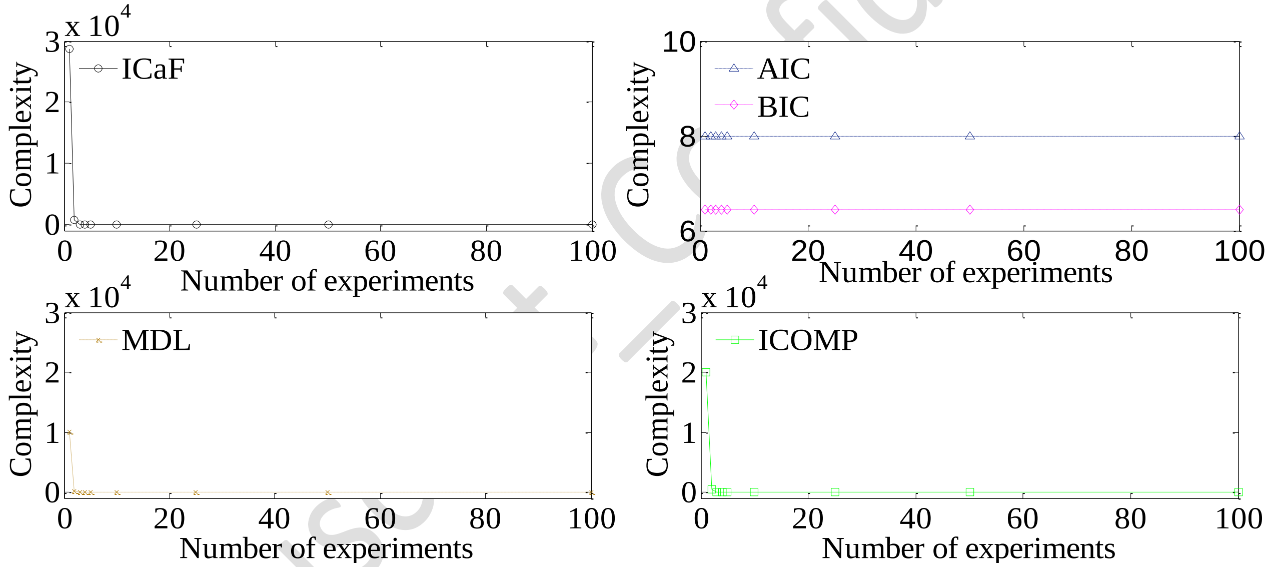


Fig. 10 The relationship between complexity and the number of experiments for different metrics

Herein, we will refer to Equation 2 to demonstrate the influence of the number of experiments has on complexity. We investigate nine different calibration campaign with varying number of experiments which are distributed uniformly over the domain of applicability with bounded uncertainties of 50% their mean value. For all experiments, the parameter set (with 20% parametric uncertainty) and model form are kept constant.

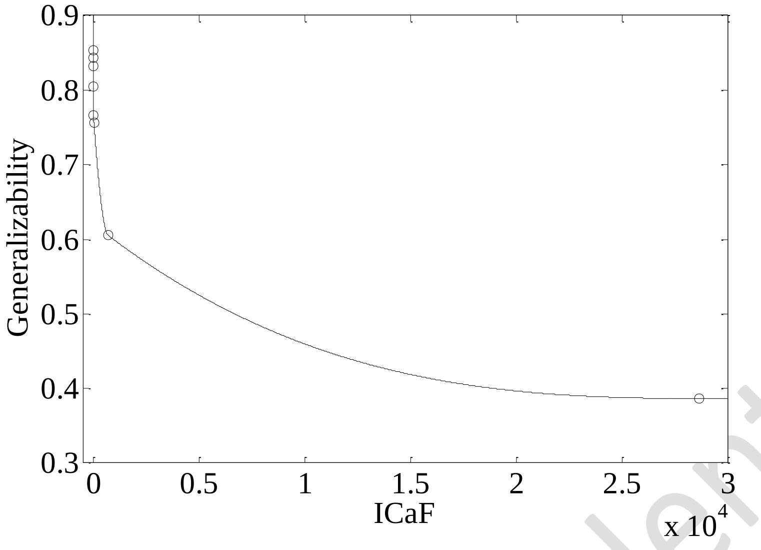


Fig. 11 The relationship between *ICaF* and Generalizability

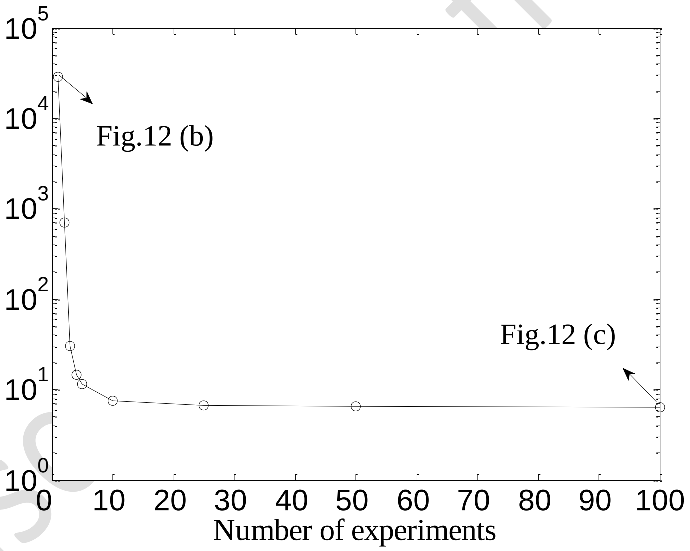


Fig. 12 (a) The relationship between the *ICaF* and the number of experiments

As seen in Fig.12 (a), an increase in the number of experiments decreases the complexity as the domain of applicability is better explored and calibrated is further restricted. However, after 9 experiments, adding further experiments only marginally affect the exploration of the domain of applicability, resulting in diminishing returns in the reduction of the complexity of calibration campaign.

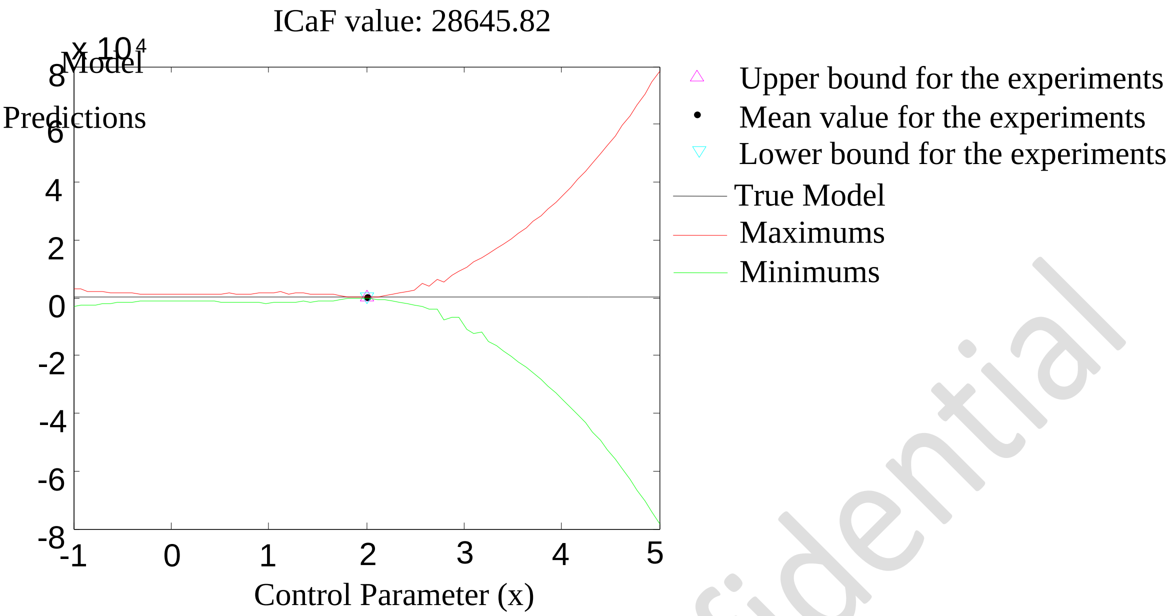


Fig.12 (b) The number of experiments is 1

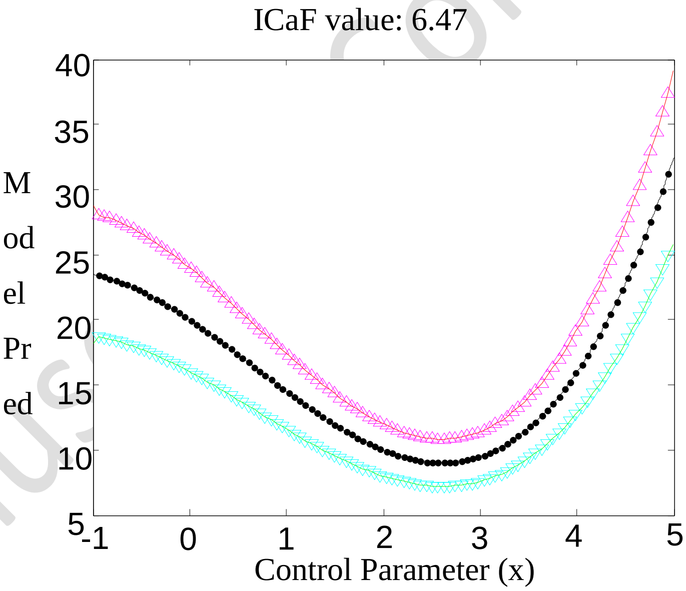


Fig. 12 (c) The number of experiments is 100

Experimental coverage must also focus on how well the domain is explored by experiments into account. (Hemez et al. 2010, Atamturktur et al. 2015). To analyze the effect of distribution of experiments on complexity, we refer back to equation 2, and calibrate the uncertain parameters of the model against nine different distributions of ten experiments with uncertainties that are 50% of their mean value. The model form and parameter set are kept constant for all distributions, with the range of calibration parameters set at 20% for all distributions.

Table 3: The distribution of experiments versus the complexity (*ICaF*)

Experiment settings within domain [-1 5]	**Complexity values**	**Nearest-Neighbor Metric**
[-1; -0.33; 0.33; 1; 1.66; 2.33; 3; 3.66; 4.33; 5]	3.44	0.38
[-1; -0.44; 0.11; 0.66; 1.22; 1.78; 2.33; 2.89; 3.44; 5]	3.52	0.42
[-1; -0.42; 0.33; 0.77; 1.22; 1.74; 2.24; 2.95; 3.02; 5]	3.59	0.44
[-1; -0.55; -0.11; 0.33; 0.88; 1.22; 1.67; 2.11; 2.55; 5]	3.64	0.46
[-1; -0.74 -0.41; 0.12; 0.45; 0.72; 1.3; 1.42; 1.64; 5]	4.35	0.48
[-1; -0.66; -0.33; 0; 0.33; 0.66; 1; 1.33; 1.66; 5]	4.83	0.50
[-1; -0.78; -0.55; -0.33; -0.11; 0.11; 0.33; 0.55; 0.77; 5]	8.43	0.56
[-1; -0.78; -0.56; -0.61; -0.52; -0.41; -0.36; -0.57; 0.32; 5]	10.21	0.59
[-1; -0.89; -0.78; -0.67; -0.56; -0.45; -0.34; -0.23; -0.12; 5]	11.99	0.61

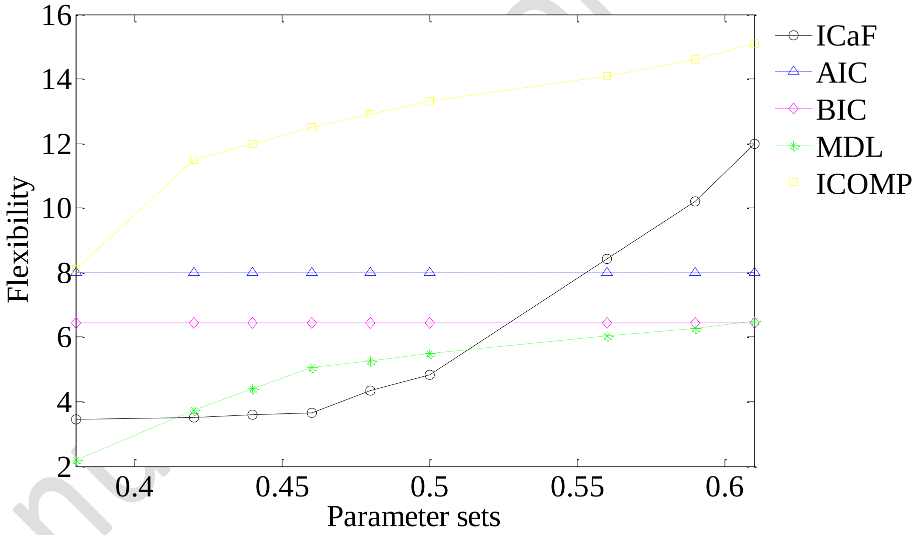


Fig. 13 Influence of distribution of experiments on complexity

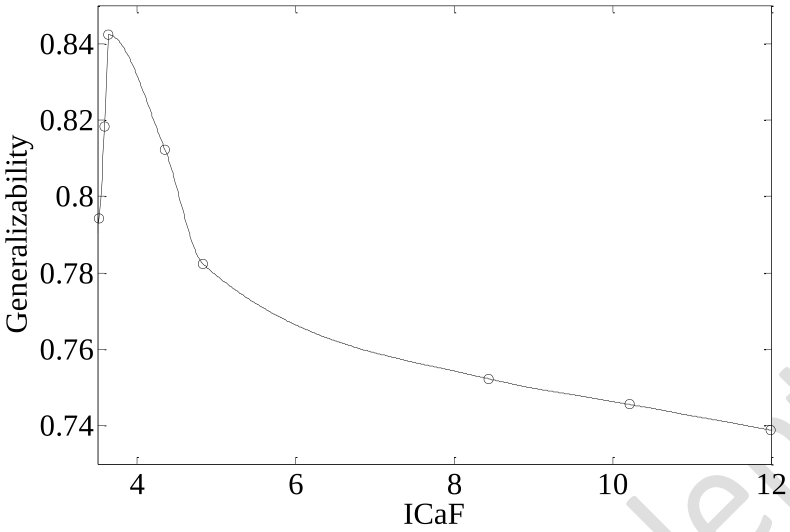


Fig. 14 The relationship between *ICaF* and Generalizability

4.4 EXPERIMENTAL UNCERTAINTIES

Experimental uncertainties reflect the restraining effect experiments place on model calibration (Moriasi 2007). Large experimental uncertainties introduce flexibility to the calibration increasing the complexity (see Fig. 13). Once again, we refer to Equation 2 and investigate the effect of this experimental uncertainty on complexity for eight bounded experimental uncertainty levels. The parameter set, whose range is set to 20%, as well as model form are kept constant for all scenarios; five calibration experiments are distributed uniformly over the domain of applicability.

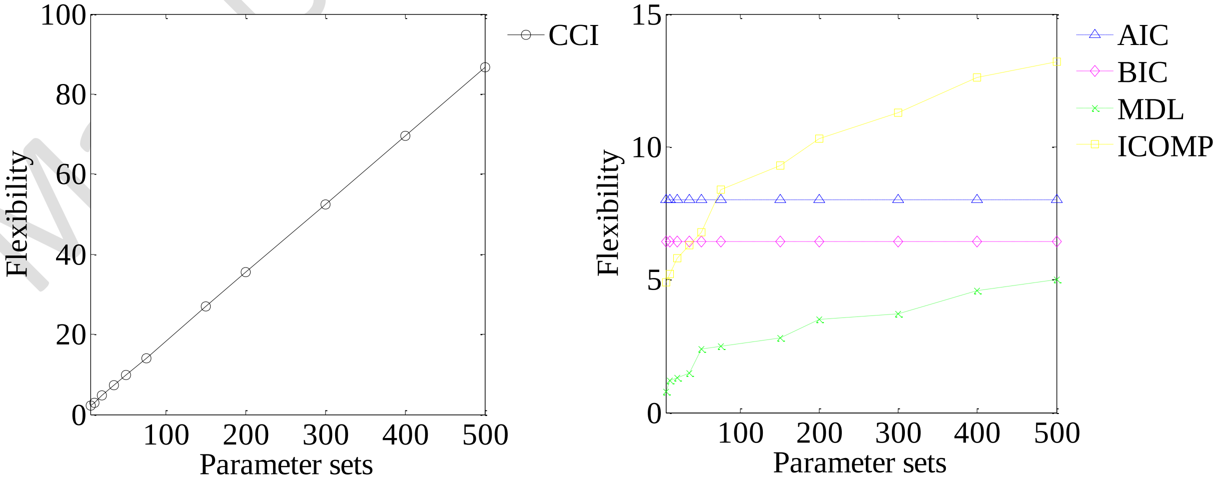


Fig. 15 Relationship between flexibility and the experimental uncertainties

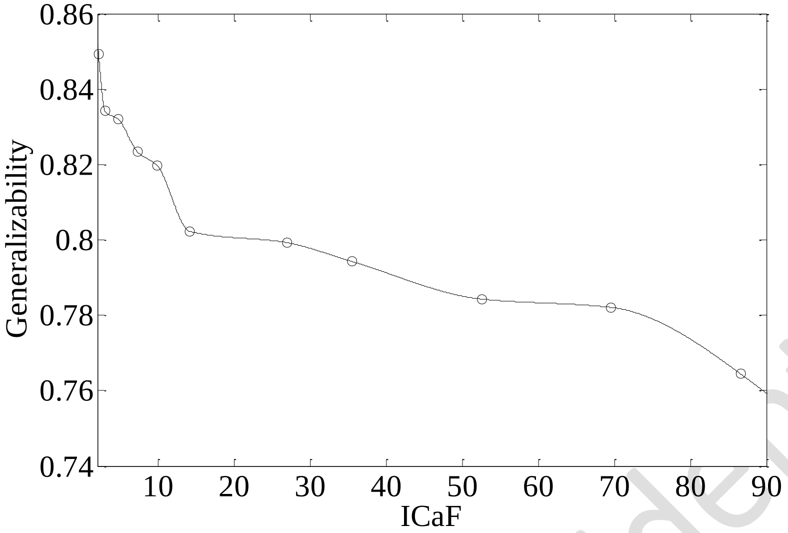


Fig. 16 The relationship between *ICaF* and Generalizability

4.5 TRAINING OF INDEPENDENT BIAS MODEL

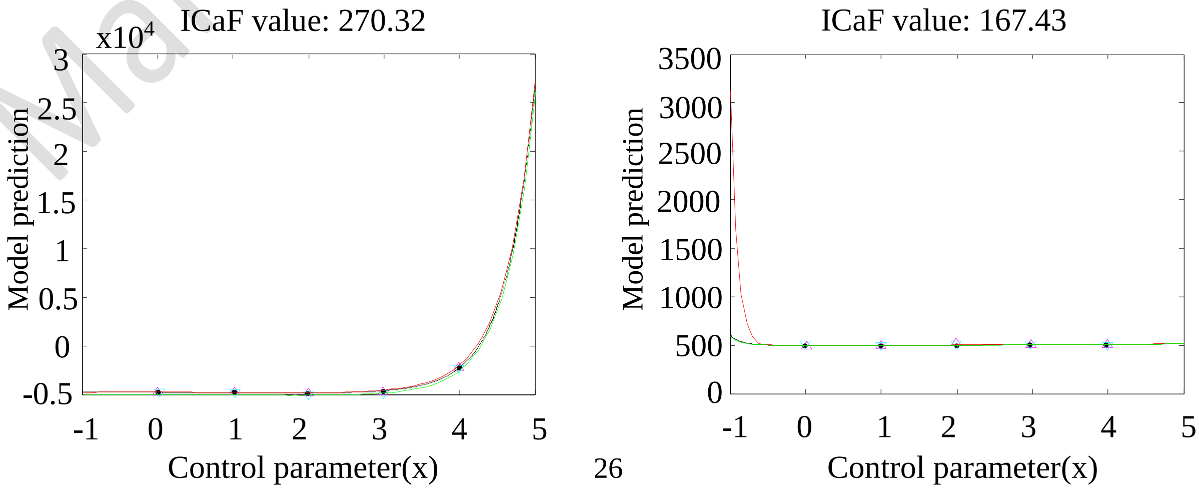
Aside from the parameters of a model, the model form that is intrinsic to the model's mathematical expression (linear dynamics representation versus nonlinear dynamics) plays an important role in the model complexity (Cutting et al. 1992). Although, in this study, our emphasis is on different calibration schemes as applied to a given model, a discussion on model calibration complexity would have been incomplete without consideration of the model form itself.

In most engineering applications, alternative models result from the additional mechanistic principles being added to existing models, resulting in a nested set of models. Often times, the additional details result in the introduction of new parameters. For instance, Halfon (1983) added details into models in a logical order and noted that after the model reaches a certain level of maturity adding more details was not justifiable as they did not affect the model behavior significantly and increased the predictions uncertainty due to the uncertainty in the added parameters. Similarly, Atamturktur et al. (2015) has shown that an increase in model detail often leads to an increase in model form complexity. In their study, a more detailed model is shown to have less discrepancy bias but increase prediction uncertainty. Geman et al. (1992) referred to these two conflicting objectives reducing as the bias-variance dilemma. In these studies, the quest for simplicity in model form is recommended not as a guide to the "truth," but to predictive accuracy given the available experimental data.

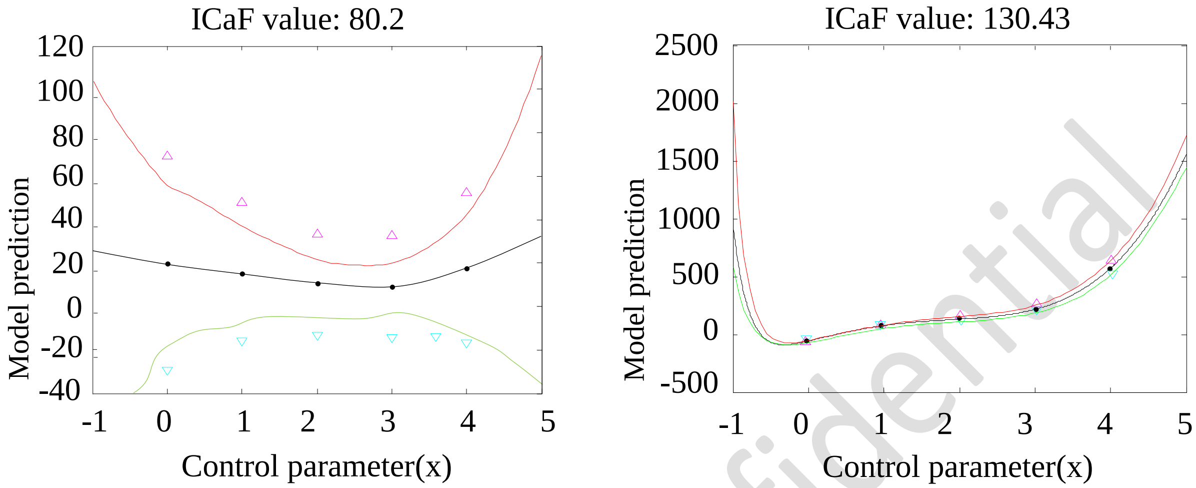
Unlike these earlier studies, in this paper our interest is not the model complexity as we envision that the optimum calibration campaign is sought for a given model. However, we are interested in the ability of the independent error model (discrepancy bias) to capture irrelevant patterns of data (Myung & Pitt, 1997). Our interest relates to the 21st century interpretation of methodological principle attributed to William of Ockham (widely known as Ockhan's Razor) which is based on the maxim "entities are not to be multiplied beyond necessity." This axiom has been interpreted within the context of anti-superfluity principle (Barnes, 2000). In our application, this principle calls for eliminating posits from models that are not necessary to explain the underlying behavior of the system. This concept is demonstrated by adding an additional functional term to Equation 2 (Equation 3).

$$\boldsymbol{y} = \mathrm{a}\boldsymbol{x}^3 + \mathrm{b}\boldsymbol{x}^2 + \mathrm{c}\boldsymbol{x} + \mathrm{d} + \varphi(\boldsymbol{x}, \boldsymbol{\theta}) \tag{3}$$

Herein, $\varphi(\boldsymbol{x}, \boldsymbol{\theta})$ represents the additional functional term, which makes the model form more complex without adding any new calibration parameters to the model. Here, three different functional forms are evaluated: sin (x), exp(x), $2^{(\mathrm{ex})}$ and x^{x}. Consistent with the section 1, the same five original calibration parameters are evaluated with their range set to 500% for all model form to eliminate a constraining effect due to parameters. The original function (without the addition of, $e(\boldsymbol{x})$) is used to generate the calibration experiments are kept constant and are distributed uniformly over the domain of applicability with bounded uncertainties of 500% of their mean value.



(a)$\boldsymbol{y} = a\boldsymbol{x}^3 + b\boldsymbol{x}^2 + c\boldsymbol{x} + d + \boldsymbol{x}^{\boldsymbol{x}}$ (b) $\boldsymbol{y} = a\boldsymbol{x}^3 + b\boldsymbol{x}^2 + c\boldsymbol{x} + d + \exp(\mathbf{x})$



(c)$\boldsymbol{y} = a\boldsymbol{x}^3 + b\boldsymbol{x}^2 + c\boldsymbol{x} + d + sin(\boldsymbol{x})$ (d) $\boldsymbol{y} = \boldsymbol{a}\boldsymbol{x}^3 + \boldsymbol{b}\boldsymbol{x}^2 + \boldsymbol{c}\boldsymbol{x} + \boldsymbol{d} + 2^{\boldsymbol{ex}}$

Fig. 17 Complexity calculation of different forms

Table 6: Influence of model form on complexity

		Complexity				
		AIC	**BIC**	**MDL**	**ICOMP**	**ICaF**
1	**No additional term**	8.00	6.44	5.02	13.2	84.69
2	**exp(x)**	16.60	16.38	13.83	21.8	270.32
3	**$\mathbf{x}^{\mathbf{x}}$**	10.82	10.61	8.06	16.02	167.43
4	**sin(x)**	7.56	7.35	4.8	12.76	84.20
5	**$\mathbf{2}^{\mathbf{(ex)}}$**	8.10	7.88	5.33	13.3	130.43

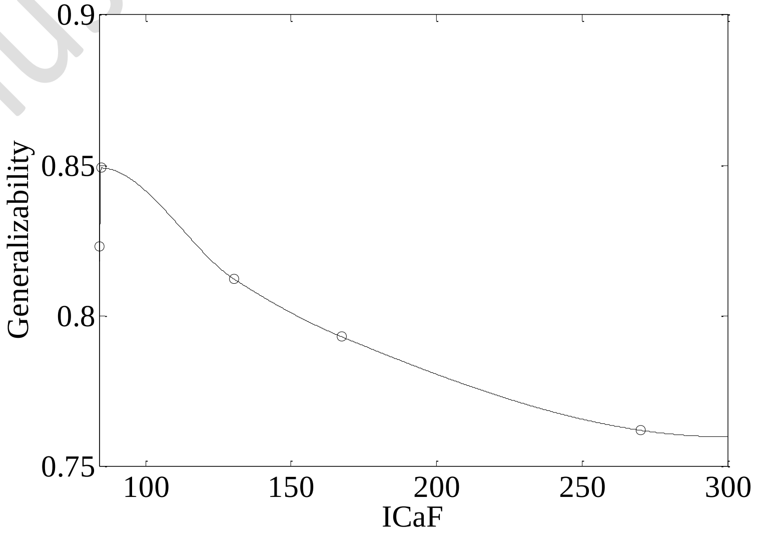


Fig. 18 The relationship between *ICaF* and Generalizability

What is important to note in Table 6 is that each of the three cases where an additional term is added to the original model yields increased complexity than the original model that is used to generate data. However, no comparison can be made among the three alternative, $sin(\boldsymbol{x})$ functions listed in Table 6 as it would lead to subjectivity as models that may "look" complex in their mathematical functional may not necessarily exhibit higher flexibility (Dunn 2000). The Ockham's razor principle is easy to agree on when it is considered to shave off superficial aspects of the model as demonstrated in the example above. The problem is often times; one may not know what is superficial – especially when an independent error model is being trained to account for the bias.

5.0 CASE STUDY APPLICATION

The proposed methodology is demonstrated on a 9 degree of freedom lumped mass cantilever tapered beam model. Mass of the beam has been considered as the control parameter. Stiffness of elements of beam has been considered as calibration parameters. To be able to calculate the accuracy, exact values for stiffness of elements has been assumed and accordingly the true model has been generated. According to this true model, a set of hypothetical experiments has been generated for calibration.

Alternative calibration campaigns have been implemented to calibrate the model. In the first calibration campaign (See Fig.19), the stiffness of all elements has been calibrated with one calibration parameter. In other words, all the parameters have been assumed to be equal, and in the second calibration campaign, all the parameters have been divided into two groups and calibrated with 2 different calibration parameters. Finally for the last calibration process, all 9 stiffness parameters have been considered separately as a calibration parameter. In consequence, by using the presented definition for complexity of model calibration, the complexity of each calibration process has been quantified.

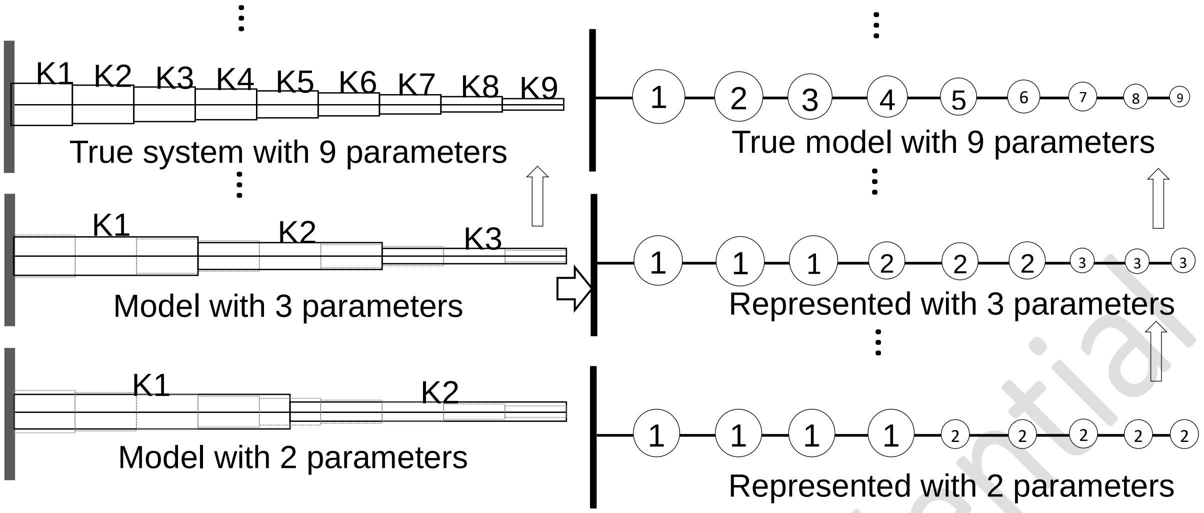

K1
K2
K3
K4
K5
K6
K7
K8
K9
True system with 9 parameters
1
2
3
4
5
6
7
8
9
True model with 9 parameters
K1
K2
K3
Model with 3 parameters
1
1
1
2
2
2
3
3
3
Represented with 3 parameters
K1
K2
Model with 2 parameters
1
1
1
1
2
2
2
2
2
Represented with 2 parameters

Fig. 19 Calibration process of the example

The total energy of a beam under a free vibration caused by an initial displacement is considered as the true model, as modeled by a nine-degree of freedom system. To obtain the total energy of the system, time history response of the beam has been used (Blevins 1979):

$$Y(x,t) = \sum_{i=1}^{9} A_i \tilde{y}_i(x) sin\,(w_{n_i}t + \Phi_i) \qquad (3)$$

In equation (3), Y(x, t) is the response of the beam at point x where x starts at the clamped end towards the free end. In this paper the deflection at the end of the beam will be considered:

$$x = l, Y(l,t) = \sum_{i=1}^{9} A_i \tilde{y}_i(l) sin\,(w_{n_i}t + \Phi_i) \qquad (4)$$

In equation (4), $Y(l,t)$ is the response at the free end of the beam and t is time. For each i there is a mode shape and natural frequency. $\tilde{y}_i(x)$ is the mode shape vector and accordingly $\tilde{y}_i(l)$ is the mode shape vector at the end of the beam. A_i is a constant with the units of length and Φ_i is a phase angle. w_{n_i} is the natural circular frequency in units of rad/sec and can be expressed by the following (Blevins 1979):

$$\omega_{n_i} = \frac{\lambda_i^2}{l^2}\left(\frac{EI}{\mu A}\right)^{\frac{1}{2}} \qquad (5)$$

In equation (5), I and A are the moment of inertia, and the area of the cross section of the beam. E is the elastic modulus and μ is the material density. λ is a dimensionless constant which is a function of boundary conditions. Related values are presented in Table 4. The equivalent stiffness for all elements of the beam are considered in this paper, also is presented in Table 5.

Table 4: Values for the true model

Mode Number	$A_i\tilde{y}_i(l)$	w_n	Φ_i
1	6.56E-01	5.99	$\pi/2$
2	2.07E-01	17.80	$\pi/2$
3	4.19E-02	29.06	$\pi/2$
4	3.45E-02	39.47	$\pi/2$
5	1.66E-02	48.63	$\pi/2$
6	1.59E-02	56.29	$\pi/2$
7	1.00E-02	62.35	$\pi/2$
8	8.73E-03	66.44	$\pi/2$
9	9.55E-03	68.61	$\pi/2$

Table 5: Equivalent stiffness of beam elements

Element Number	K_i
1	123.46
2	121.38
3	125.11
4	124.57
5	120.95
6	125.39
7	125.23
8	123.01
9	124.81

Thus, from the figure below, regardless of the number of experiments we have, the least error is always provided by the 15th calibration process, however, from the 9th to the 15th calibration, the fitting error has very small changes, whose ranges are between 0 and 10^{-3}.

Next, a different number of calibration parameters were used to calibrate each process, the complexity of which was then quantified using the presented definition of the complexity of model calibration. The results of those experiments are shown in Fig. 21.

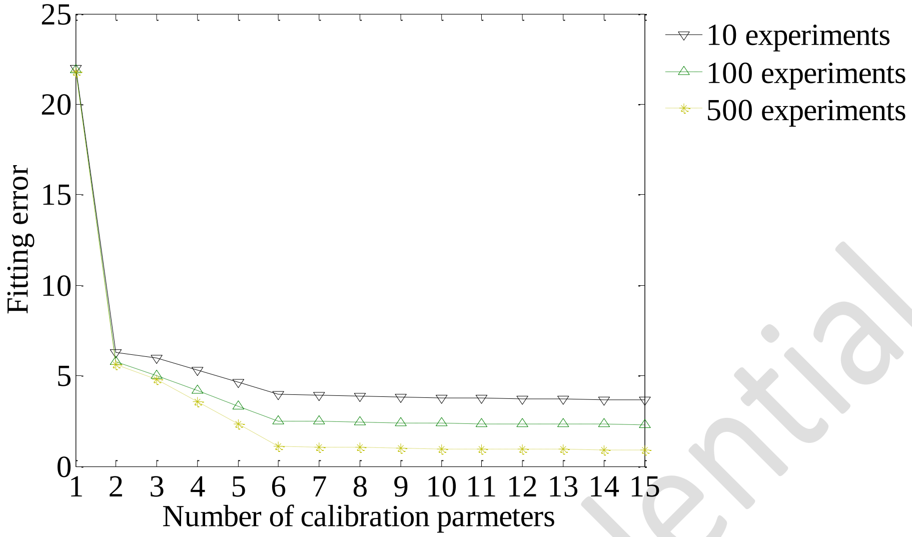


Fig. 20 Normalized error for each calibration process at different experiments

As expected, complexity of model calibration increases with increasing the number of calibration parameters. Determining the complexity that provides the maximum generalizability for each graph is used to determine the optimum complexity for different amounts of available data (See Fig.19). According to Fig. 22, a graph for optimum complexity versus available data can be extracted that is presented in Fig. 23.

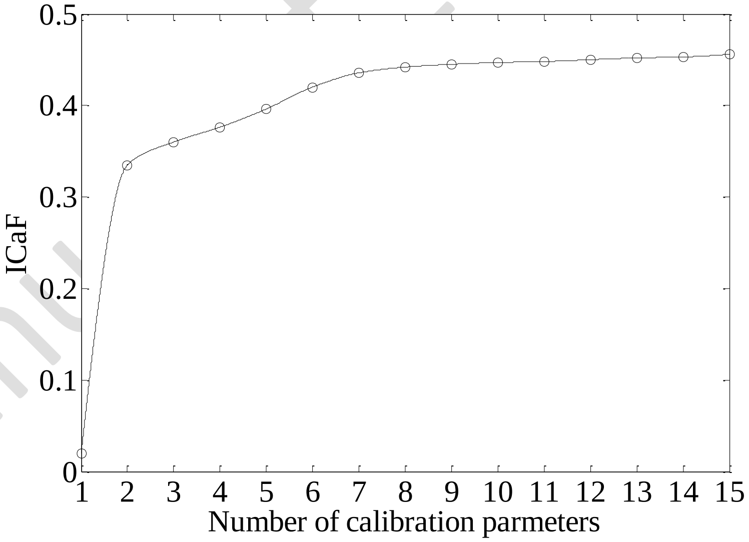


Fig. 21: Complexity of Model Calibration

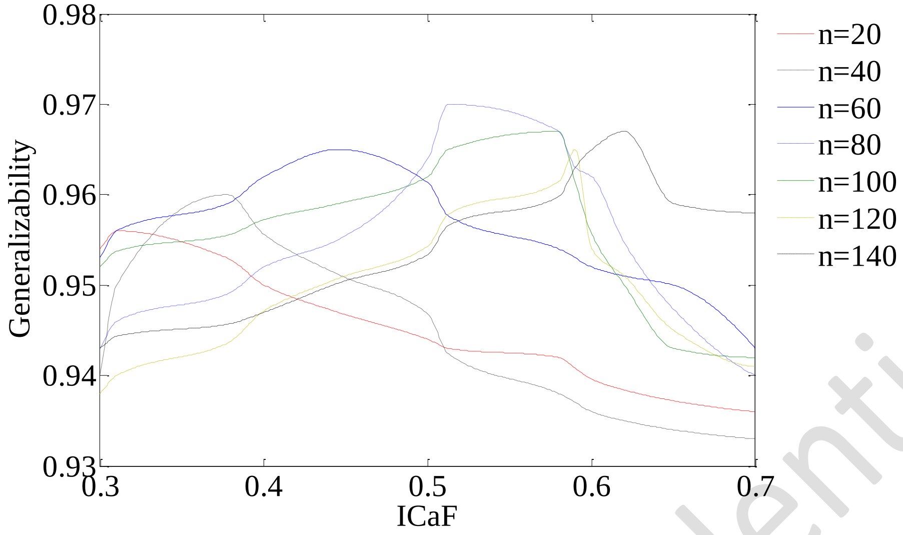


Fig. 22: Generalizability for different ICaF at different experiments

According to Fig. 23, with fewer experiments, fewer number of calibration parameters gives better predictions accuracy, for more experiments, more number of calibrations gives better predictions accuracy.

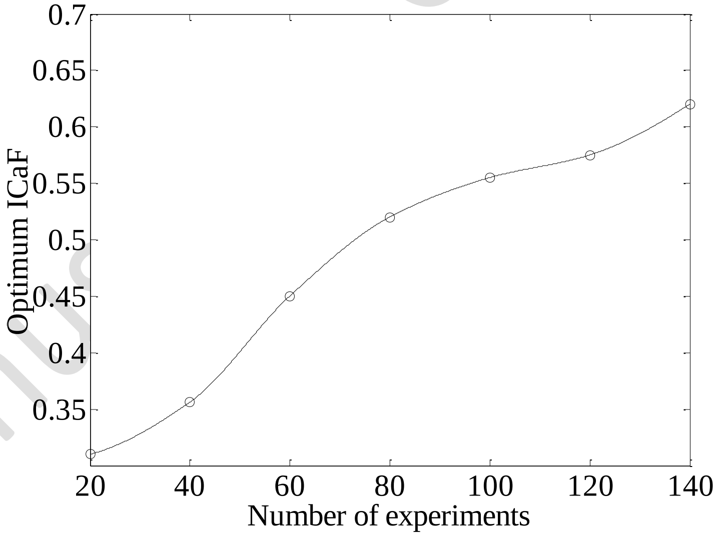


Fig. 23: Data Aware Optimum ICaF of Model Calibration

6.0 CONCLUDING REMARKS

The authors present a thorough discussion on the quantifying calibration complexity of numerical models of engineering in a data-aware manner. Data-aware calibration enables the analyst to obtain the maximum generalizability in calibrated model predictions by using an optimally complex model calibration campaign for a given set of available experiments. The authors strongly propose that complexity quantification of a model calibration campaign must go beyond the complexity of the model and also consider the availability of experiments in their quantity, quality and coverage of the domain of applicability.

The number of calibration parameters, their range and the functional form of the model were recognized as factors that affecting complexity of a numerical model (Myung & Pitt, 1997; Bozdogan & Haughton, 1998; Zucchini, 2000; Pitt et al., 2002; So, 2003). Although the association between such flexibility and the complexity of a model has been recognized previously by using these three factors (Myung & Pitt, 1997), the concept has not yet been implemented to quantify the complexity of model calibration taking the inevitable limitations of experimental resources into account. In this article, the authors present a quantitative and generally applicable approach for determining the complexity of model calibration by taking into account the experimental coverage and experimental uncertainties in addition to the number of calibration parameters, their range and the functional form of the model.

The authors associate the complexity of a model calibration campaign to the flexibility of the model to produce a range of values at untested settings while conforming to the experiments at the tested settings. The extremes of this range at a given control parameter setting are represented as the upper and lower flexibility limits in this paper. These limits bound an area over the entire domain of applicability, which is defined as an index for the complexity of model calibration. The average of this total area over the entire domain applicability is used to calculate the complexity of model calibration (Calibration Complexity Indicator, *ICaF*).

The five aforementioned factors and their influence on model calibration complexity are demonstrated on a generic polynomial function (the third-degree polynomial model). To investigate solely each factor, the effects of other aforementioned factors on complexity are minimized by designing a corresponding calibration campaign accordingly. It was determine that the additional calibration parameters expanded the predictive ability of the model at given control parameter setting to produce results of greater complexity, the influence of which was validated on the generic polynomial model by altering the number of calibration parameters.

The authors also used a wider range of the calibration parameters to expand the variance of the calibration parameter to produce a broader range of model predictions at a given control parameter setting, which in turn increased the complexity to model. The relationship between a range of calibration parameters and complexity was clearly established through the use of different ranges of the calibration parameters to the model form. These different model forms in turn affected how well the model to fit many data patterns, which in turn yielded a greater or lesser model complexity. The complexity results of various model forms were correlated with their model form indicators (defined as derivative of the model expression with respect to control parameter). Subsequent correlations found an increase in complexity with an increase in the model form indicator.

Most importantly, in these experiments the authors executed model calibration campaigns in a data-aware manner, through which they derived the final two factors of calibration complexity (experimental coverage and experimental uncertainties). In that both restrict the model calibration both affect the model flexibility. The experimental coverage reflects how well the experiments explored the domain of applicability, which was found to relate to both the number of available experimental data points and their distribution within that domain. Increasing the number of uniformly distributed experiments forces the model to fit to additional data points, which in turn lessens the flexibility of the model.

To demonstrate this effect on complexity, the authors implemented 6 various uniformly distributed experiments on the model, which was also a distribution over the domain of applicability. Unlike uniformly distributed experiments, the model becomes flexible when the concentrated experiments on a certain portion of the domain of applicability fail to restrict the calibration at the uncovered portion of the domain. By gradually pulling the experimental locations over to that uncovered portion, however, it was possible to derive a relationship between the experimental distribution and complexity. These experimental uncertainties were found to determine the level of restriction placed on the calibration by each experiment. Consequently, a more complex model was obtained with an increase in these uncertainties.

This novel methodology detailed in this analysis is expected to enhance the quantification of the model calibration complexity by considering the inevitable limitations of experimental resources, which has never been attempted. Most importantly, by using an optimally complex model calibration campaign for a given set of available experiments, the authors provide a method for obtaining the maximum generalizability in calibrated model predictions

REFERENCES

Akaike. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika, 60(2), 255-265.

Akaike. (1974). A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control, AC-19(6), 716-723.

Atamturktur, S., Hemez, F., and Laman, J., (2012), "Uncertainty Quantification in Model Verification and Validation as Applied to Large Scale Historic Masonry Monuments," Engineering Structures (Elsevier), Vol. 43, pp. 221-234.

Atamturktur, S., Hegenderfer, J., Williams, B., Egeberg, M., Lebensohn, R. A., and Unal, C. (2015), "A resource allocation framework for experiment-based validation of numerical models," Mechanics of Advanced Materials and Structures, 22, 641–654.

Atamturktur, S., Egeberg, M.*, Hemez, F., and Stevens, G.*, (2015), "Defining Coverage of an Operational Domain Using a Modified Nearest-Neighbor Metric," Mechanical Systems and Signal Processing (Elsevier), Vol. 50-51, pp. 349-361.

Atamturktur, S. and Brown, A. (2015), "State-Aware Calibration for Inferring Systematic Bias in Computer Models of Complex Systems," NAFEMS World Congress, June 21-25, Manchester Grand Hyatt, San Diego Market Place, San Diego, California, USA, ISBN 978-1-910643-24-2.

Bard, Y., 1974. Nonlinear Parameter Estimation. Academic Press, New York/London.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C.-H., and Tu, J. (2007), "A framework for validation of computer models," Technometrics, 49, 138–154.

Bozdogan, H., & Haughton, D. M. (1998). Informational complexity criteria for regression models. Computational Statistics & Data Analysis, 28, 51-76.

Bozdogan H. (2000). Akaike's Information Criterion and Recent Developments in Information Complexity. Journal of Mathematical Psychology, 44(1), 62-91.

Brooks, R., & Tobias, A. (1996). Choosing the best model: Level of detail, complexity, and model performance. Mathematical and computer modelling, 24(4), 1-14.

Casti, J. (1979). Connectivity, Complexity and Catastrophe in Large-Scale Systems. New-York: John Wiley and Sons.

Farajpour, I. and Atamturktur, S., (2013), "Error and Uncertainty Analysis of Inexact and Imprecise Computer Models," Journal of Computing in Civil Engineering (ASCE), Vol. 27, No. 4, pp. 407-418.

Friedman, D., Massaro, D. W., Kitzis, S. N., & Cohen, M. M. (1995). A Comparison of Learning Models. Journal of Mathematical Psychology, 39, 164-178.

Gell-Mann, M. (1995). What is Complexity? Complexity 1(1), 16-19.

Hemez, F., Atamturktur, H., & Unal, C. (2010). Defining predictive maturity for validated numerical simulations. Computer & Structures, 88(7-8), 497-505.

Henneman, R., & Rouse, W. (1986). On measuring the complexity of monitoring and controlling large-scale systems. IEEE Transactions on Systems, Man and Cybernetics, SMC-16, 193-207.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008), "Computer model calibration using high-dimensional output," Journal of the American Statistical Association, 103, 570–583.

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J., and Ryne, R. D. (2004), "Combining field data and computer simulations for calibration and prediction," SIAM Journal on Scientific Com- puting, 26, 448–466.

Hill, M. C. (1998). Methods and guidelines for effective model calibration. Denver, CO, USA: US Geological Survey.

Jaafar, W. Z., & Han, D. (2012). Uncertainty in index flood modelling due to calibration data sizes. Hydrological Processes, 26(2), 189-201.

Kennedy, M. C. and O'Hagan, A. (2001), "Bayesian calibration of computer models," Journal of the Royal Statistical Society, Series B, 63, 425–464.

Kunz, M., Trotta, R., & Parkinson,, D. R. (2006), Measuring the effective complexity of cosmological models. Physical Review D, 74(2).

Law, A.M. (1991), Simulation model's level of detail determines effectiveness, Industrial Engineering 23 (10), 16-18.

Liddle, A. R. (2007), Information criteria for astrophysical model selection. The Author Journal Compilation, 74-78.

MacKay, D.J.C. (1992), “Bayesian Interpolation,” NeuralComput. 4:415–447.

Myung, I., & Pitt, M. (1997). Applying Occam's razor in modelling cognition: A Bayesian approach. Psychonomic Bulletin & Review, 4(1), 79-95.

Myung, I., Balasubramanian, V., & Pitt, M. (2000). Counting probability distributions: Differential geometry and model selection. Proceedings of the National Academy of Sciences, 97(21), 11170-11175.

Oreskes, N., Shrader-Frechette, K., Belitz, K., (1994), Verification, validation, and confirmation of numerical models in the earth sciences. Science 263, 641–646.

Pidd, M. 1996. Five simple principles of modelling, In Proceedings of the 1996, Winter Simulation Conference, ed. J. M. Charnes, D.M. Morrice, D. T. Brunner, and J. J. Swain, p. 721-728. Institute of Electrical and Electronics Engineers, Piscataway, N.J.

Pitt, M. A., & Myung, J. (2002). When a good fit can be bad. TRENDS in Cognitive Sciences, 6(10), 421-425.

Pitt, M., Myung, I., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. Psychologial Review, 109(3), 472-491.

Rosen, R. (1977). Complexity as a system property. International Journal of General Systems, 3, 227-232.

Salt, J.D. (1993), “Keynote address: Simulation should be easy and fun!,” In Proceedings of the 1993 Winter Simulation Conference, (Edited by G.W. Evans et al.), pp. 1-5, IEEE, New York.

Schoups, G., van de Giesen, N. C., & Savenije, H. H. (2008). Model complexity control for hydrologic prediction. Water Resources Research, 44(12).

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics (6), 461-464.

Seth Roberts, Harold Pashler (2000). How Persuasive Is a Good Fit? A Comment on Theory Testing. Psychological Review, Vol.107, No.2, 358-367.

Simon, H. (1964). The architecture of complexity. General Systems Yearbook, 10, 63-76.

So, B. (2003). Maximized log-likelihood updating and model selection. Statistics & Probability Letters, 64(3), 293-303.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, B, 64, 583-639.

Van Buren, K. L., & Atamturktur, S. (2012). A comparative study: predictive modeling of wind turbine blades. Wind Engineering, 36(3), 235-250.

Ward, S. (1989). Arguments for Constructively Simple Models. The Journal of the Operational Reseach Society, 40(2), 141-153.

Webster, D., Padgett, M., Hines, G., & Sirois, D. (1984). Determining the level of detail in a simulation model - a case study. Computers & Industrial Engineering, 8(3-4), 215-225.

Zhu, L., and Carlin, B. P. (2000), “Comparing Hierarchical Models for SpatioTemporally Misaligned Data Using the Deviance Information Criterion,”Statistics in Medicine, 19, 2265–2278.

Zucchini, W. (2000). An Introduction to Model Selection. Journal of Mathematical Psychology, 44, 41-61.