A Dynamic Baseline Calibration Procedure for CGE models

Johannes Ziesmer[1] · Ding Jin[2] · Sneha D Thube[3] · Christian Henning[2]

Abstract

Baseline assumptions play a crucial role in conducting consistent quantitative policy assessments for dynamic Computable General Equilibrium (CGE) models. Two essential factors that influence the determination of the baselines are the data sources of projections and the applied calibration methods. We propose a general, Bayesian approach that can be employed to build a baseline for any recursive-dynamic CGE model. We use metamodeling techniques to transform the calibration problem into a tractable optimization problem while simultaneously reducing the computational costs. This transformation allows us to derive the exogenous model parameters that are needed to match the projections. We demonstrate how to apply the approach using a simple CGE and supply the full code. Additionally, we apply our method to a multi-region, multi-sector model and show that calibrated parameters matter as policy implications derived from simulations differ significantly between them.

Keywords Dynamic baseline calibration · Model uncertainty · Bayesian approach · Metamodeling · Simulation optimization · Quantitative policy analysis

✉ Johannes Ziesmer
  johannes.ziesmer@lifbi.de

1   Leibniz Institute for Educational Trajectories, Bamberg, Germany

2   Institute of Agricultural Economics, University of Kiel, Kiel, Germany

3   Kiel Institute for World Economy, Kiel, Germany

1 Introduction

Nowadays, Computable General Equilibrium (CGE) models are considered the workhorse models of policy analysis focusing on economy-wide effects induced by exogenous economic shocks or policy interventions (de Melo 1988; Shoven and Whalley 1984; Dixon and Jorgenson 2013). For example, CGE models have been widely used to assess the impacts of policies in the area of international trade (Hertel 1997; Hertel et al. 2007; Caliendo and Parro 2014), migration (Stifel and Thorbecke 2003; Fan et al. 2018), agricultural policies (Milczarek-Andrzejewska et al. 2018; Taylor et al. 1999), and energy(Burniaux and Chateau 2014) or climate policies (Böhringer et al. 2014; Fujimori et al. 2016; Böhringer et al. 2021).

However, longstanding criticisms of CGE models include that these models have weak econometric foundations (McKitrick 1998; Jorgenson 1984). This criticism arises from the fact that CGE models are complex, and the available empirical data is limited, implying that it is often impossible to estimate all model parameters econometrically (McKitrick 1998; Jorgenson 1984; Hansen and Heckman 1996). Thus, typically model parameters are either assumed ad hoc or weakly derived from empirical data to calibrate CGE models.

In static CGE models, a classical baseline calibration corresponds to calculating model exogenous variables, such that model output in the equilibrium replicates the economic structure defined by a given social accounting matrix (SAM) empirically observed in a specific base year. This approach can be problematic because generally, an infinite number of parameter set-ups can be generated such that all of them can perfectly replicate the SAM. A good case in point is the often-used practice of assuming ad hoc values for relevant elasticities of substitution or transformation to determine remaining parameters of corresponding Constant Elasticity of Substitution (CES)- and Constant Elasticity of Transformation (CET)-functions. Thus, calibration of static CGE models corresponds to a more or less arbitrary parameter specification. In response to this criticism, Systematic Sensitivity Analysis (SSA) is being increasingly used in CGE model applications. With this approach endogenous CGE output variables are simulated based on sampled CGE model parameters that are derived from estimated or assumed distributions (Olekseyuk and Schürenberg-Frosch (2016)).

However, though SSA is a good method to reveal induced uncertainty of model outputs explicitly as it derives the means and variances of all relevant endogenous variables based on an assumed distribution of parameters, it does not provide a way to use external information to "update" the assumed distribution. Hence, it is not an appropriate procedure to reduce model uncertainty. Furthermore, SSA does not provide a method to integrate observed model uncertainty into the derivation of optimal policy choices.

For dynamic CGE models, in addition to the base year calibration, modellers also need to calibrate the long-term behaviour (Sánchez 2004). Typically, while constructing the baseline of a dynamic model, modellers use historical data and forecasts to specify relevant model parameters minimizing forecasts errors comparing simulated model output with observed or forecasted data over the entire

forecast period from the base year to the end year. Calibration of a dynamic baseline has the advantage of focusing on the long-term behaviour of the model, i.e., delivering information about the key parameters that drive the model response to an exogenous shock. Understanding the response of the economy to exogenous shocks is not only crucial to predict correctly future economic development, such as sectoral growth, employment, capital stock, trade or Greenhouse Gas (GHG) emissions in the baseline but also to assess counterfactual economic developments implied by different policy scenarios (Nong and Simshauser (2020); Pothen and Hübler (2021), see also Sect. 5.3). Moreover, the different approaches applied to derive model baselines and calibrating model parameters are far from being innocuous but rather crucially impact on corresponding policy analyses. Accordingly, parameter calibration and baseline derivation has received wide-attention for several types of numerical models that are used for policy analysis (e.g., Dynamic Stochastic General Equilibrium (DSGE) models in Gomme and Lkhagvasuren (2013), New Quantitative Trade (NQT) models in Pothen and Hübler (2021) and spatial trade models in Paris et al. (2011)).

Technically, baseline calibration of dynamic equilibrium models corresponds to a high dimensional optimization problem, i.e., a set of model parameters must be identified such that the equilibrium values of the variables of interest should match the forecasts from external sources of data. Additionally, a set of theoretical restrictions like closure rules also need to hold valid on the model parameters. Given the high dimensionality of parameters and output variables, solving this problem is very challenging. Moreover, beyond the theoretical parameter restrictions, a priori expert information regarding the empirical range of model parameters is generally available, and prior parameter distributions can formally represent this. Hence, given all these constraints, a Bayesian estimation approach appears to be an appropriate methodological framework for baseline calibration of dynamic equilibrium CGE models. Moreover, Bayesian estimation techniques are applied to estimate parameters of DSGE models (Smets and Wouters 2003; Hashimzade and Thornton 2013). In particular, like CGEs also DSGE models correspond to sound micro-founded and theoretically consistent model structures, where applying Bayesian estimation techniques DSGE models combine these theoretical properties with good forecasting performance. For example, Smets and Wouters (2003) have demonstrated that the estimated DSGE models perform quite well in forecasting compared to standard and Bayesian vector autoregressions. Not only does the system-wide estimation procedure deliver a more efficient estimate of the structural model parameters, but it also provides a consistent estimate of the structural shock processes driving recent economic developments. Understanding the contribution of the various structural shocks to recent economic developments is an important input in the monetary policy decision process. Accordingly, they conclude especially that applying Bayesian estimation techniques to calibrate DSGE models the latter can become a useful tool for projection and policy analysis in central banking. However, compared to CGE models DSGEs are still highly aggregated models, which assume highly aggregated economic agents, e.g., do not include heterogenous sectoral structures. Therefore, in contrast to DSGE models applying Bayesian estimation

techniques to estimate parameters of a CGE model is still challenging due to the much more complex structure and much high number of parameters of the latter.

Some researchers have used an informal Bayesian estimation procedure by which they run a model forward over a historical period and compare model results with historical data (Gehlhar 1994; Kehoe et al. 1995; Dixon et al. 1997). Though such approaches can be helpful to revise parameter estimates and recalibrate the model (Tarp et al. 2002), they are nevertheless ad hoc and do not yet offer a systematic Bayesian procedure that can be applied baseline calibration of a dynamic model. Alternatively, Arndt et al. (2002) and more recently Go et al. (2016) propose a maximum entropy approach for parameter estimation of CGE models by using information theory to estimate CGE parameters based on a sequence of observed SAMs. Their approach can be interpreted as a special case of the Bayesian approach and hence, already provides several advantages as compared to the standard calibration methods. However, the approach still has limitations that can be addressed in a general Bayesian framework. First, it is based on specific assumptions regarding a priori parameter distributions, which can be significantly relaxed in a general Bayesian framework (Heckelei and Mittelhammer 2008). Second, this approach is focused on an ex-post analysis, and thus, it requires empirical observations of corresponding SAMs. Furthermore, so far, the Cross Entropy (CE)-method has only been applied to single country and static CGE models, and this approach cannot be easily extended to baseline calibration of dynamic CGE models.

Our paper proposes a Bayesian estimation approach that generalizes the CE-method of Arndt et al. (2002); Go et al. (2016) thereby allowing a more direct and straightforward formulation of available prior information (see Sect. 2). An important innovative step in our approach is that we complement the Bayesian estimation with metamodeling techniques (Kleijnen and Sargent 2000) to replace the CGE model with a simplified surrogate model to reduce complexity and thereby, significantly reduce the computational effort (Sect. 2.1). Our Bayesian approach also enables us to simulate endogenous CGE outputs based on sampled model parameters that are derived from the corresponding posterior distribution, where technically metamodeling also facilitates Metropolis-Hasting sampling from this posterior distribution. We show how to apply the approach with a toy model and provide the complete code for this demonstration (Sect. 3). Since our proposed approach can also be applied to calibrate complex multi-region and multi-sector CGEs subsequently in (Sect. 4) we demonstrate our approach by calibrating the dynamic baseline of the Dynamic Applied Regional Trade (DART-CLIM) model. Lastly, we also show that the chosen calibration approach has a noticeable impact on policy assessments by using our method to construct two baselines for DART-CLIM albeit with differences in the extent of calibration (Sect. 5).

2 Methods

Simplified, the application of CGE models in policy analysis follows three main steps in general (Fig. 1).

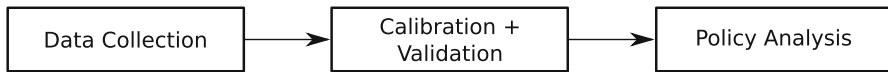| Data Collection | → | Calibration + Validation | → | Policy Analysis |

Fig. 1 Step-wise CGE-modeling approach. (adapted and simplified from Shoven and Whalley (1992))

The first step is collecting statistical data, like SAMs or Input/Output (I/O)-tables besides information about exogenous parameters (like elasticities) and forecasts for key outputs from other studies. The second step is model calibration, where the parameters of the CGE are derived to replicate observed data and forecasts to a certain degree. When the calibration gives a satisfying specification, the model is applied in policy analysis by implementing policy shocks, and the policy impact is measured by comparing outputs from the policy scenario to the baseline scenario. In the rest of this section, we present the foundations of calibration and derive a consistent and comprehensive approach to the calibration step.

Formally, let F denote a CGE model, which implicitly determines outputs, y as a function of a set of model parameters, θ:

$$F(y, \theta) \equiv 0. \tag{1}$$

F is an I-dimensional vector valued function, y an I-dimensional vector of endogenous output variables and θ a K-dimensional vector of exogenous model parameters. Model parameters can be further disaggregated into different subsets, e.g. behavioural parameters or exogenous variables. The latter also includes policy variables that are controlled by the government, e.g. taxes or tariffs, etc. Further, exogenous variables include demographic or economic variables uncontrolled by the government, e.g. world market prices, population growth, etc.... Changes in the values of exogenous variables correspond to exogenous shocks, which induce changes in the endogenous variables. The latter corresponds to the response of the economic system to exogenous shocks. Behavioural parameters define the response of the economic system to exogenous shocks. Hence, assuming behavioural parameters correspond to their true values implies that the model replicates the true response of the economy to exogenous shocks. In this context, we define a dynamic model as a model for which variables and parameters are defined for different periods, while a static model corresponds to a model for which variables and parameters are only defined for one specific time (base) period. Further, we define baseline by a complete parameter specification which corresponds to a clearly defined base run scenario. Obviously, since the model $F(y, \theta)$ determines for any complete parameter specification (θ) the corresponding endogenous output variables a baseline also delivers specific values for all endogenous variables (y).

An advantage of CGE models is that one can simulate counterfactual scenarios, i.e. one can calculate the values of the endogenous variables that result from assuming parameter values that differ from their corresponding baseline values. Hence, one can simulate the impact of different policies by changing the subset of policy variables compared to their baseline values, or one can simulate the impact of different framework conditions on policy impacts by changing the subset of

exogenous variables uncontrolled by the government. Finally, one can also simulate the impact of characteristic properties of the economic system on the impact of specific policies via simulating policy impacts, assuming different values for the subset of behavioural variables and comparing these to policy impacts derived for the baseline values. However, important for any model-based policy analysis is that the baseline parameters can be empirically validated. In particular, this implies that the subset of behavioural parameters corresponds to the true structural properties of the economic system.

Given the fact that in contrast to endogenous and exogenous variables, behavioural parameters can generally not be directly observed, empirical model validation is based on a comparison of observed values of the endogenous model variables with corresponding values derived from model simulations. Accordingly, all model validation methods basically correspond to the identification of behavioural model parameters that maximize the likelihood of the model specification. Given enough observed data, y^o, of sufficient quality, θ can be estimated econometrically. In the context of CGEs, data is usually scarce, and therefore, the identification of θ becomes a calibration problem. A standard calibration method for CGEs is based on given empirical observations of endogenous and exogenous variables in a given base year, where behavioural parameters are determined in a way that the observed variables are replicated in this base year. Since for this procedure, data is only observed in a single (base) year, we call this a static baseline calibration method. In contrast, assuming parameters are determined based on variable values observed for a sequence of time periods, we use the term dynamic baseline calibration method.

However, as we already explained above, a problem of a static baseline calibration corresponds to the fact that there exists an infinite set of different behavioural parameter specifications that imply that simulated outputs perfectly replicate a given set of observed variables, while simulated responses of the economic system to given exogenous shocks differ in contrast significantly for different behavioural parameter values. Therefore, model-based policy analysis becomes arbitrary. We suggest a Bayesian framework for an advanced parameter calibration method to avoid this arbitrariness. The latter can be applied as a static as well as a dynamic baseline calibration method.

In the general Bayesian framework, observed variables are noisy, i.e., data $y^o = \{Y^o_1, \ldots, Y^o_N\}$ correspond to true variable values, $y = \{Y_1, \ldots, Y_n\}$ and noises $\epsilon = \{\epsilon_1, \ldots, \epsilon_N, \}$. Assuming ϵ_n is *iid* normally distributed, $\epsilon_n \propto N(0, I)$ implies that the posterior results as:

$$h(\theta \mid y^o) \propto p(\theta) \prod_{n=1}^{N} p(\epsilon_n)$$
$$\text{with}(\theta, \epsilon) \in \Psi := \{(\theta, \epsilon) \mid \epsilon = y^o - y \text{ and } F(y, \theta) \equiv 0\}$$

(2)

Building upon this general Bayesian framework, we develop a calibration procedure for quasi-dynamic CGE models. To this end we define different subsets of parameters, i.e., $\theta = (\theta_0, \theta_t)$. θ_0 denotes parameter values in the base year while θ_t denotes parameter values in a period $t = 1, \ldots, T$ different to the base year. Given a

forecast for a subset of output variables, $z^o \in y$, and $y^o_0 \in y$ empirical observations of a subset of CGE-variables in the base year t_0, we define $\epsilon_0 = y^o_0 - y_0$ and $\epsilon_z = z^o - z$, and $\epsilon = (\epsilon_0, \epsilon_z)$. Additionally, assuming normal distributions for $\epsilon \propto N(0, \Sigma_\epsilon)$ and $\theta \propto N(\bar{\theta}, \Sigma_\theta)$ with the co-variance matrices Σ_ϵ, $\Sigma_\theta = (\Sigma_0, \Sigma_T)$ as diagonal matrices with elements σ_ϵ^2, σ_θ^2, we can derive the following optimization problem for the Highest Posterior Density (HPD) -estimator:

$$\theta^* = \arg \min_\theta \underbrace{(\theta - \bar{\theta})' \Sigma^{-1} (\theta - \bar{\theta})}_{\propto p(\theta)} + \underbrace{\epsilon' \Sigma_\epsilon^{-1} \epsilon}_{\propto \mathcal{L}(y^o|\theta)}$$
$$\epsilon_0 = y^o_0 - y_0$$
$$\epsilon_z = z^o - z \tag{3}$$
$$0 \equiv F(y_0, z, y \setminus (y_0, z), \theta_0, \theta_T)$$
$$0 \equiv H(\theta_0, \theta_T).$$

$H(\theta_0, \theta_T)$ takes additional parameter constraints into account that are induced by economic theory or apriori expert information. In general, given observations y^o_0 and forecasts z^o, HPD estimation follows an optimization problem described by the system 3. It is also possible to choose other distributions, i.e., other extremum metrics, while keeping the general idea the same. Moreover, please note that in general, the approach 3 includes the option that some variables or parameters are known with certainty, e.g. governmental policies like taxes or tariffs are normally known with certainty. In this case, these variables or parameters are fixed to their prior values and excluded from the prior density function.

However, outputs z are only defined as an implicit function by theCGE, and therefore, the solution of the optimization problem is very tedious. Hence, complex methods of simulation optimization (SO) have to be applied to solve (3). We apply metamodeling techniques to reduce complexity and computational effort (see 2.1 for a short introduction). The advantage of using metamodels is that they may result in more efficient SO methods. As a result, F is substituted with a metamodel M, which approximates the mapping between model parameters and a subset of output variables, (y_0, z), derived from the original model F.

$$\theta^* = \arg \min_\theta (\theta - \bar{\theta})' \Sigma^{-1} (\theta - \bar{\theta}) + \epsilon' \Sigma_\epsilon^{-1} \epsilon$$
$$\epsilon_0 = y^o_0 - y_0$$
$$\epsilon_z = z^o - z \tag{4}$$
$$(y_0, z) = M(\theta_0, \theta_T)$$
$$0 \equiv H(\theta_0, \theta_T)$$

The formulation in (3 and 4) allows a straightforward interpretation of the assumed variances σ_ϵ^2 and σ_θ^2. In particular, we can interpret σ_ϵ^2 as weights that indicate the importance of matching the corresponding forecast Z^o relative to the other variables. A lower value of σ_ϵ^2 suggests higher importance relative to $\sigma_{\epsilon'}^2$. The variance σ_θ^2 captures the information we know about a parameter θ with a lower value for σ_θ^2

compared to $\sigma_{\theta'}^2$ implying more certainty or higher knowledge about θ than θ'.

2.1 Metamodeling

Metamodeling techniques are widely used in a variety of research fields such as design evaluation and optimization in many engineering applications (Simpson et al. 1997; Barthelemy and Haftka 1993; Sobieszczanski-Sobieski and Haftka 1997), as well as in natural science (Razavi et al. 2012; Gong et al. 2015; Mareš et al. 2016). In recent years, metamodeling is increasingly being applied to economic research. For example, Ruben and van Ruijven (2001) have applied the approach to bio-economic farm household models to analyze the potential impact of agricultural policies on changes in land use, sustainable resource management, and farmers' welfare; Villa-Vialaneix et al. (2012) have compared eight metamodels for the simulation of N_2O fluxes and N leaching from corn crops; Yildizoglu et al. (2012) have applied the technique to two well-known economic models, Nelson and Winter's industrial dynamics model and Cournot oligopoly with learning firms, to conduct sensitivity analysis and optimization respectively. Regardless of the research fields, the metamodeling technique simplifies the underlying simulation model, leading to a more in-depth understanding. The technique also brings the possibility of embedding simulation models into other analysis environments to solve more complex problems, such as the previously described calibration process.

The use of metamodeling entails three steps: selection of metamodel types, Design of Experiments (DOE), and model validation (Kleijnen and Sargent 2000).

2.1.1 Metamodel Types

Metamodels are classified into parametric and non-parametric models (Rango et al. 2013). Parametric models, such as polynomial models (Forrester et al. 2008; Myers et al. 2016), have explicit structure and specification. On the other hand, non-parametric models do not depend on assumptions of model specification and determine the (I/O) relationship of the underlying simulation model using experimental data. Examples of non-parametric models consist of Kriging models (Cressie 1993; Yildizoglu et al. 2012; Kleijnen 2015), support vector regression models (Vapnik 2013), random forest regression models (Breiman 2001), artificial neural networks (Smith 1993), and multivariate adaptive regression splines (Friedman et al. 1991).

In this paper, we focus on the polynomial models that are defined by their order. For example, a second-order polynomial model is given as follows:

$$Y = \beta_0 + \sum_{h=1}^{k} \beta_h \theta_h + \sum_{h=1}^{k} \sum_{g \geq h}^{k} \beta_{h,g} \theta_h \theta_g + \eta, \tag{5}$$

where $\theta_1, ..., \theta_k$ are the k independent variables, Y is the dependent variable and η is the error term. The corresponding coefficients β are usually estimated through a linear regression based on least squares estimation (Chen et al. 2006). Polynomial metamodels generally provide only local approximations. Theoretically, certain

types of metamodels like Kriging models (Kleijnen 2015) can also provide a global fit. Additionally, different metamodels may also be combined into an ensemble (Bartz-Beielstein and Zaefferer 2017; Friese et al. 2016). Some of the advantages of the polynomial models are:

– They have simple forms that are easy to understand and manipulate.
– They require low computational efforts.
– They can be easily integrated into other research frameworks.

For a more thorough introduction to other types of metamodels, see for example Dey et al. (2017); Simpson et al. (2001).

2.1.2 Design of Experiments

To utilize the metamodels, we need to estimate the corresponding coefficients. We generate the simulation sample by DOE, which is a statistical method of drawing samples in computer experiments (Dey et al. 2017) and perform the estimation by entering the simulation sample into the simulation model. DOE could be set-up in two ways: the classical experimental design and the space-filling experimental design (see Fig. 2). The former places the sample points at the boundaries and the centre of the parameter space to minimize the influence of the random errors from the stochastic simulation models. However, Sacks et al. (1989) have argued that this is not the case for deterministic simulation models where systematic errors prevail. Therefore, the space-filling experimental designs should be employed to replace the classical ones. Among popular space-filling designs, Latin Hypercube design enjoys great popularity due to its ability to generate uniformly distributed sample points with ideal coverage of the parameter space as well as the flexibility with the number of the sample points (Sacks et al. 1989).

2.1.3 Model Validation

Validation refers to assessing whether the prediction performances of the metamodels hold an acceptable level of quality (Kleijnen 2015; Villa-Vialaneix
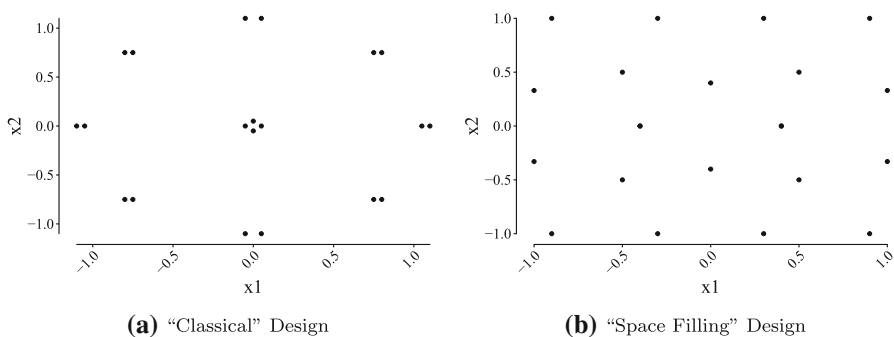

(a) "Classical" Design   **(b)** "Space Filling" Design

Fig. 2 Classical and space-filling design. (adapted from Simpson et al. (2001))

et al. 2012; Dey et al. 2017). Normally, two samples are needed to assess the quality of a derived metamodel: the training sample and the test sample. The training sample is used to fit the parameters of the metamodel, whereas the test sample is used to validate the trained metamodel, and the test sample must include data points that are not part of the training sample. It is important that the metamodels have good predictions while maintaining generality. For this reason, a test sample is essential because it helps us evaluate if the metamodels can be generalized and whether the simulation model can be replaced with them.

The following textbook statistics are often considered to assess the validation results.

$$
\begin{aligned}
R^2 &= 1 - \frac{\sum_{i=1}^{n}(Y_i - Y^o{}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y^o})^2} \\
RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - Y^o{}_i)^2}
\end{aligned}
\tag{6}
$$

where Y_i and $Y^o{}_i$ are the predicted values and true values for the test sample at sample point i, and $\overline{Y^o}$ is the mean of Y^o in the test sample. In regression analysis, R^2 is a statistical measure of how close the data are to the fitted regression line. The root mean squared error (RMSE) is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data i.e., how close the model's predicted values are to the true values.

To compare the prediction performances for dependent variables that have different scales, we introduce the absolute error ratio (AER), which is calculated by taking the absolute value of RMSE divided by the corresponding mean:

$$
\text{AER} = \left| \frac{RMSE}{\overline{Y^o}} \right| = \left| \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - Y^o{}_i)^2}}{\overline{Y^o}} \right|.
\tag{7}
$$

The AER metric gives us an idea of how large the prediction errors are in comparison to the true simulated values on average, i.e., the lower the AER values, the better the prediction performances.

3 Demonstration

In this section, we first demonstrate how the theoretically derived approach (see Sect. 2) can be applied using a simple toy model (Hosoe et al. 2010, 2016). We use the toy model to provide technical insights to the reader on how our approach can be applied. Therefore, we also refer explicitly to the corresponding programming code of our calibration procedure, giving pointers on where each step is implemented. The complete implementation code can be found as a supplement to this paper[1]. The DART-CLIM application in Sect. 4 is used to demonstrate that our approach can

---

[1] https://github.com/universal/cgecalibration/.

also handle the calibration of huge complex models with a very large number of parameters ($\gg 1000$).

The toy model we use is from the General Algebraic Modeling System (GAMS) model library (GAMS Development Corporation 2020), and is a simple four sectors, two factors, one household recursive dynamic CGE model (Hosoe et al. 2010, 2016).

In order to apply our dynamic calibration procedure, five main steps are necessary, which are summarized in Algorithm 1. We implemented the individual steps in a mix of R (R Core Team 2020) and GAMS (GAMS Development Corporation 2020), where the GAMS part only requires a free demo license[2]. The R code drives the full example, where two scripts have to be run to replicate the steps and results described in the following (first R/sample_generation.R and second R/metamodels.R). The computation should be done in about five minutes on a current system.

Algorithm 1 Steps

1) DOE: Θ
2) Computing simulations
3) Estimating and validating metamodels
4) Bayesian calibration $\Rightarrow \boldsymbol{\theta}^*$
5) Evaluating calibration point, $\boldsymbol{z}^*$, with $F(\boldsymbol{z}^*, \boldsymbol{\theta}^*) \equiv 0$

Following the standard calibration procedure of CGE models, an observed SAM is used to calibrate behavioural model parameters. However, based on an observed SAM using equilibrium conditions of the CGE the behavioural parameters are underdetermined, i.e. for any observed SAM there exists infinite parameter values for which equilibrium conditions are fulfilled. Accordingly, a subset of parameters is usually exogenously fixed, and the remaining parameters are calculated using equilibrium conditions based on an observed complete SAM.

Similar to Hosoe et al. (2010, 2016) we exogenously determine the parameters listed in Table 1, i.e. the four elasticities of substitution for each of the Armington function, *sigma*, the four elasticity of transformation for each of the CET-function, *psi*, as well as the exogenous variables population growth, *pop*, rate of return for capital, *ror*, depreciation rate, *dep*. Further, exogenously determined parameters include a parameter determining sectoral investments, i.e. the elasticity of investment, *zeta*[3].

---

[2] https://www.gams.com/try_gams/.

[3] The complete model of Hosoe et al. (2010, 2016) comprises of 62 behavioral parameters, $4 \times 5 = 20$ input coefficients for the upper Leontief nest of the sectoral production functions, $4 \times 2 = 8$ parameters for the four lower Cobb-Douglas (CD) production functions, $4 \times 3 = 12$ parameters for the four Armington functions aggregating domestic supply and imports to domestic supply as well as $4 \times 3 = 12$ parameters for the four CET-functions transforming domestic production into domestic supply and exports for each commodity and finally 4 parameters for the CD-demand systems determining domestic demand of the representative domestic household. Finally, we have 6 parameters determining sectoral investments (*iota*, *zeta*, and *lambda(i)*). At this point we neglect exogenous variables, e.g. world market prices, tax rates and tariffs, etc...

Table 1 Parameters and targets used in demonstration of approach using the toy model of Hosoe et al. (2010, 2016)

Label	Description	Dimension
ror	Rate of return for capital	1
dep	Depreciation rate of capital	1
pop	Population growth rate	1
sigma	Elasticity of substitution (Armington)	4x1
psi	Elasticity of transformation (CET)	4x1
zeta	Elasticity of investment	1
ω	Weight of ideal-typical SAM	1
Z1	Sectoral output	4x1
Z2	Total sectoral output	1
Z3	Sectoral exports	4x1
Z4	Total exports	1
Z5	Exchange rate	1
Z6	Sectoral imports	4x1
Z7	Total imports	1

While the standard calibration procedure implicitly assumes that the true SAM can be completely and perfectly observed, empirically observed SAM data is often noisy or incomplete. Hence, like the behavioural parameters, also the observed SAM data correspond to stochastic variables. In general, a Bayesian approach can deal with partly noisy and incomplete SAM data (see also Go et al. (2016)). For example, forward and backward linkages of a sector or detailed data on specific factor inputs might be missing, or (I/O) relations might be noisy, i.e. they do not perfectly match with the true SAM. Dealing with this kind of incomplete SAM data CGE-modellers normally apply a data imputation procedure like, for example, the entropy approach suggested by Robinson et al. (2001). Following the latter missing data is estimated based on a prior SAM. The latter is constructed using available statistical data taken from SAMs available for prior periods, or data available from other countries with a similar economic structure like the country to be analyzed or data taken from expert estimations. However, imputed SAMs crucially depend on the constructed prior SAM where generally different options exist to construct this prior. Hence, constructed SAM data is noisy. As a way to deal with this, we suggest the following: We assume that based on available data one can construct a finite set of different ideal-typical SAMs, $\{SAM_r\}$, where the true SAM follows as a convex combination of these ideal-typical SAMs. Let ω^{true} denote the vector of weights of the ideal-typical SAMs, i.e. it holds: $SAM^{\text{true}} = \sum_r \omega_r SAM_r$. Thus, once the set of ideal-typical SAMs have been constructed, we simply include the weights ω into $\boldsymbol{\theta}$.

For the following demonstration, we constructed two SAMs, where each corresponds to a specific economic structure characterized by specific forward and backward linkages of the agricultural sector as well as by a specific labour intensity of agricultural production (see `gams/sam.gms#L*`). The true SAM corresponds to a linear combination of these two SAMs, where the SAM combination parameter ω is a priori unknown and hence randomly drawn between zero and one.

Further, to be able to evaluate the goodness of fit of our calibration procedure we generate a benchmark parameter specification θ^{true} (see `R/sample_genera-tion.R`#L32-33. Based on this benchmark specification, we run the model for $T = 30$ time periods generating a time path for all endogenous variables. However, given the fact that in a true empirical application of our approach, exogenous forecasts will be available only for a limited subset of endogenous variables let $z_t = \{Z_{k,t}\}$ denote a matrix of selected endogenous variables of the CGE, where $Z_{k,t}$ denotes the forecasted equilibrium path of the endogenous variable k observed for time periods t.

3.1 Design of Experiments

The first step in the process is the generation of a suitable sample to estimate and validate the metamodels.

After having decided on the parameters to be used for the calibration, the allowed range for the parameters is defined based on prior experience, expert opinions, and literature review. For *ror*, *dep*, *pop*, *zeta* we picked a small range around the original values (± 0.01 or ± 0.02). For the two elasticity parameters *sigma* and *psi* we picked a large range of [1, 18] (see `R/sample_generation.R`#L9- 17). This information is then transformed into a suitable format for the Latin Hypercube Sample (LHS) generation function, and a larger estimation sample and a validation sample is generated.

The number of necessary simulations depends on the kind and form of the metamodel to be estimated. A rule of thumb for the required simulations is the number of coefficients times ten for polynomial models. As we will apply a second-order polynomial model, this is $\approx |\theta|^2 \times 10 = 13^2 \times 10 = 1690$. For the validation sample, which is separate to the estimation sample, we generate $|\theta|^2 \times 2 = 338$ parameter settings (`R/sample_generation.R`#L28-31).

The R code exports the sample to a `GDX`-file(see #L46) and calls the code for the next step (see #L41-60).

3.2 Simulations

Following the sample generation, one needs to calculate the individual simulations and collect the results. We abstain from calculating it in a cluster environment given the simple structure and, therefore, nearly instantaneous solves of the toy model. In order to calculate the results for the samples, minimal changes are necessary to the model: The parameter values need to be loaded from a `GDX`-file (as is often the case for non-toy models) and unload the simulation results in the end. In our case, we also replaced the SAM, that was entered in text-form into the model (see `gams/cge/model.gms`#L44-45,208). A second part needed is some wrapping code that loops over the individual parameter settings and collects the results, which we also implemented in GAMS (see `gams/mm_simulations.gms`#L*). This code defines the necessary parameters(see #L4-69), loops over all parameter settings(see

#L71-90), scales results for numerical stability(see #L92-97) and finally unloads the aggregated results for further consumption.

The loop in lines $\#71 - 90$ also provides the basis for splitting the work across multiple machines. Instead of looping over the full set of specifications, only loop over a subset of those, save the partial results and aggregate those together in a final step. This is possible since the individual specifications do not depend on each other.

3.3 Metamodels

Using the results from the previous two steps, we can now estimate the metamodels, that replace the original CGE $F(y, \theta) \equiv 0$ with a set of metamodels $z = M(\theta)$ (see R/metamodels.R#L*). In our simple toy application we assume $y = (y_0, z)$, where z corresponds to the outputs being matched in the calibration procedure (see Table 1), and $y_0 = y \setminus z$ to other endogenous model outputs including SAM entries in the base year. To keep our demonstration example simple, we estimate metamodels for each target $Z_j \in z$ only for the base year t_0 and the end year t_{30}[4].

Given the relatively simple structure of the underlying CGE and the relatively small number of parameters, we can fit a second-order polynomial model that captures the relationships quite well. In a more realistic setting with a larger model that requires much more computational effort, it will become expensive to run the required number of simulations (see Sect. 4 for a way to deal with this). The required number of simulations roughly scales quadratically with the number of calibration parameters. We use linear regression to estimate the coefficients of the metamodels. With the help of some loops and data wrangling of the estimation and validation samples, we can estimate the metamodels and calculate the validation measures (see R/steps/estimate_models.R#L*). The estimations and validations use helper functions from the R packages DiceDesign / DiceEval (Dupuy et al. 2015).

As evidenced by the low AER, we achieve a nearly perfect fit for all of the selected endogenous output variables, z, in the final year t_{30}. For comparison we also show the mean, the RMSE and R^2 values in Table 2. For brevity, we skip the results for the base year t_0, where the achieved goodness of fit for base year variables is similar to the presented results.

3.4 Bayesian Calibration

With the help of the estimated metamodels, we can now solve the calibration problem (see Equation 4). We implemented the calibration procedure in GAMS, but R (or any other language with the possibility to solve NLPs problems) would also be fine (see gams/calibration.gms#L*). The model for the optimization problem is implemented in GAMS (see gams/calibra-tion/model.gms#L*), which consists of the objective function ((see #L90-

---

[4] Please note that generally, we could estimate metamodels for each time period or we could assume a specific growth function for every forecasted variable, Y, and estimate relevant growth rates as a metamodel.

Table 2 Validation results: metamodels for t_{30}

Target	AER	Mean	RMSE	R^2
$Z1_{AGR}$	0.0102	1.2575	0.0128	0.9969
$Z1_{HMN}$	0.0023	34.4134	0.0787	0.9997
$Z1_{LMN}$	0.0054	6.0468	0.0327	0.9985
$Z1_{SRV}$	0.0007	85.6500	0.0560	1.0000
$Z2$	0.0007	127.3677	0.0844	1.0000
$Z3_{AGR}$	0.0754	0.0030	0.0002	0.9835
$Z3_{HMN}$	0.0069	7.6843	0.0530	0.9978
$Z3_{LMN}$	0.0137	0.1195	0.0016	0.9943
$Z3_{SRV}$	0.0112	1.8367	0.0205	0.9966
$Z4$	0.0045	9.6435	0.0438	0.9989
$Z5$	0.0006	0.9850	0.0006	0.9901
$Z6_{AGR}$	0.0317	0.5118	0.0162	0.9870
$Z6_{HMN}$	0.0065	3.7753	0.0244	0.9981
$Z6_{LMN}$	0.0081	3.5032	0.0284	0.9970
$Z6_{SRV}$	0.0136	1.8532	0.0252	0.9950
$Z7$	0.0045	9.6435	0.0438	0.9989

Metamodels are able to replicate the CGE nearly perfectly, as evidenced by the low AER

118) and the metamodels (see #L120-160)[5]. The required data, metamodel coefficients, parameter bounds, variances and priors, and forecasts are exported in the metamodel estimation code (see `R/metamodels.R#L15-16`.) As shown in Table 5 we assumed large values for the variances, with their values being set directly in the code (see `gams/calibration.gms#L23-31`). Please note that we include only a subset of time periods (only the base year and the end year) into the Bayesian estimation to account for the fact that often forecasts for relevant variables are only available for selected time periods. The GAMS code for the calibration is called automatically from R (see `R/metamodels.R#L18-40`).

Please note further that in our simple toy model application, we generally assume that behavioural parameters are constant over time, i.e. $\theta_t = \theta_0$[6]. Second, please note that while $\epsilon_z = z^o - M(\theta)$, it holds $\epsilon_0 = SAM^o - \sum_r \omega_r^* SAM_r$. Hence, in general observed values, $SAM_{i,j}^o$ may differ from the estimated true SAM-values, $\sum_r \omega_r^* SAM_{r,i,j}$.

Overall, the Bayesian approach allows to combine multiple sources of information: (1) empirical observation of SAM-values in the base run, SAM_{ij}^0, (2) dynamic forecasts of endogenous variables, z as well as expert information on

---

[5] Please note that to directly match variables in the GAMS and R-code we kept the variable notation used in R also in GAMS, i.e. instead of defining in GAMS one variable $\sigma^2(J)$ where the set J includes all four production sectors, we defined four variables, $\sigma_j^2, J = \{AGR, LMN, HMN, SRV\}$.

[6] This follows by construction as we generated the true forecast via simulation of the CGE based on the fixed true parameters, θ^{true}.

Table 3 Goodness-of-fit (*gof*) of predicted values of metamodels and of calibrated CGE

Target	σ^2	True $F(\mathbf{y}^{\text{true}}, \boldsymbol{\theta}^{\text{true}}) \equiv 0$	Metamodel $z^M = M(\boldsymbol{\theta}^*)$	Simulation $F(\mathbf{y}^{F,*}, \boldsymbol{\theta}^*) \equiv 0$	$\frac{(Z_j^M - Z_j^{\text{true}})^2}{\sigma^2}$	$\frac{(Z_j^{F,*} - Z_j^{\text{true}})^2}{\sigma^2}$
$Z1_{AGR,0}$	0.1000	1.2652	1.2636	1.2636	0.00	0.00
$Z1_{AGR,30}$	0.0100	1.0569	1.0766	1.0475	0.04	0.01
$Z2_{30}$	0.1000	143.5337	143.5164	143.6120	0.00	0.06
$Z3_{LMN,30}$	0.0100	0.1393	0.1130	0.1122	0.07	0.07
$Z4_{30}$	0.0001	10.9218	10.9221	10.9548	0.00	10.88
$Z5_{30}$	0.0100	0.9812	0.9781	0.9774	0.00	0.00
$Z6_{AGR,30}$	0.0100	0.7807	0.7807	0.8098	0.00	0.09
$Z7_{30}$	0.0100	10.9218	10.9221	10.9548	0.00	0.11
Tot	–	–	–	–	0.11	11.22

The calibration procedure is able to match the different targets across time periods as evidenced by the small differences

economic structures, captured in derived ideal-typical SAMs SAM_r and their corresponding prior weights, ω_r. Finally, please note that beyond dealing with noisy SAM data, the Bayesian approach can easily deal with missing SAM-data, where $\sum_r \omega_r^* SAM_{r,i,j}$ correspond to the Bayesian estimation of missing SAM-values.

3.5 Evaluation

Using the found parameter values $\boldsymbol{\theta}^*$ and given that we know the "truth" in this example, we can evaluate the model predictions in three aspects: (1) The first one is using a (*gof*) measure, which is proportional to the (negative) likelihood part $(\mathcal{L}(\mathbf{y}^o \mid \boldsymbol{\theta}))$ of our target function (see Equation 3). This captures how modelers view the importance of the different targets in the calibration procedure, which is reflected by the relative value of $\sigma_{Z_j}^2$ to $\sigma_{Z_{j'}}^2$, but also $\sigma_{Z_j}^2$ to Z_j. A lower value for goodness-of-fit *gof* means a higher likelihood and hence a better fit, with a value of zero indicating a perfect fit. (2) We can compare model outputs of the metamodel with simulation outputs of the true model. (3) We can compare estimated parameters $\boldsymbol{\theta}^*$ with their corresponding true parameter values $\boldsymbol{\theta}^{truth}$. In the Tables 3, 4 and 5 we report different *gof* measures. These Tables are created as the final step in the R code (see R/metamodels.R#L43).

As evidenced by the results in Table 3 comparing predicted outcomes of the metamodel and the calibrated CGE with the true outcomes for the forecasted variables we achieve an excellent fit[7] for the calculated *gof* measure presented in Table 3. Further, we note a difference between the *gof* from the calibration step (calculated comparing the true and the predicted outcomes of the metamodel), and the *gof* calculated for the prediction of the calibrated CGE ($\boldsymbol{\theta}^*$).

---

[7] Please note that assuming a 1% deviation on all target variables implies a value of ≈ 141.

Table 4 Model output comparison

Target	True	Metamodel	Simulation	$\% \frac{Z^M_j - Z^{\text{true}}_j}{Z^{\text{true}}_j}$	$\% \frac{Z^{F,*}_j - Z^{\text{true}}_j}{Z^{\text{true}}_j}$
	$F(y^{\text{true}}, \theta^{\text{true}}) \equiv 0$	$z^M = M(\theta^*)$	$F(y^{F,*}, \theta^*) \equiv 0$		
$Z1_{AGR,0}$	1.2652	1.2636	1.2636	0.00	−0.13
$Z1_{AGR,30}$	1.0569	1.0766	1.0475	2.78	−0.89
$Z2_{30}$	143.5337	143.5164	143.6120	−0.07	0.05
$Z3_{LMN,30}$	0.1393	0.1130	0.1122	0.64	−19.45
$Z4_{30}$	10.9218	10.9221	10.9548	−0.30	0.30
$Z5_{30}$	0.9812	0.9781	0.9774	0.07	−0.39
$Z6_{AGR,30}$	0.7807	0.7807	0.8098	−3.60	3.74
$Z7_{30}$	10.9218	10.9221	10.9548	−0.30	0.30
$Z1_{HMN,0}$	24.3287	24.3298	24.3298	0.00	0.00
$Z1_{HMN,30}$	39.6825	39.3790	39.5151	−0.34	−0.42
$Z1_{LMN,0}$	4.9905	4.9909	4.9909	0.00	0.01
$Z1_{LMN,30}$	6.2138	6.3873	6.3696	0.28	2.51
$Z1_{SRV,0}$	63.2146	63.2148	63.2148	0.00	0.00
$Z1_{SRV,30}$	96.5805	96.6736	96.6799	−0.01	0.10
$Z2_0$	93.7990	93.7990	93.7990	0.00	0.00
$Z3_{AGR,30}$	0.0009	0.0013	0.0013	2.75	40.78
$Z3_{HMN,30}$	8.9388	9.0068	9.0797	−0.80	1.58
$Z3_{SRV,30}$	1.8428	1.8010	1.7617	2.23	−4.40
$Z4_0$	6.7709	6.7709	6.7709	0.00	0.00
$Z5_0$	1.0000	1.0000	1.0000	0.00	0.00
$Z6_{AGR,0}$	0.2158	0.2174	0.2174	0.00	0.73
$Z6_{HMN,0}$	3.0769	3.0758	3.0758	0.00	−0.04
$Z6_{HMN,30}$	3.6622	3.8645	3.8543	0.26	5.24
$Z6_{LMN,0}$	2.3922	2.3919	2.3919	0.00	−0.01
$Z6_{LMN,30}$	4.2514	4.0416	4.0638	−0.55	−4.41
$Z6_{SRV,0}$	1.0860	1.0858	1.0858	0.00	−0.01
$Z6_{SRV,30}$	2.2276	2.2353	2.2269	0.38	−0.03
$Z7_0$	6.7709	6.7709	6.7709	0.00	0.00

Given the relative simple structure of the applied CGE, not only the targets used in the calibration are matched well, but also other endogenous outputs are mostly matched well

So far, we only looked at the targets that were active in the previous step. Next, we look at the percentage differences for all targets reported in Table 4 and we notice only small differences between the values predicted by the metamodels (M) and the values calculated using the true simulation model (F). The differences are below one per cent, with few exceptions and a maximum difference of 3.6 per cent. For the comparison between the true specification (y^{true}, with $F(y^{\text{true}}, \theta^{\text{true}}) \equiv 0$ and the values from the simulation model ($y^{F,*}$ with $F(y^{F,*}, \theta^*) \equiv 0$ we see few larger

Table 5 Parameter (θ) comparison

Coefficients	.lo	.up	σ^2	Prior	True (θ^{true})	Calibration (θ^*)	$\% \frac{\theta^*-\theta^{\text{true}}}{\theta^{\text{true}}}$
dep	0.03	0.05	0.02	0.0400	0.0490	0.0500	2.04
pop	0.01	0.03	0.02	0.0200	0.0292	0.0300	2.74
psi_{AGR}	1.00	18.00	17.00	9.5000	12.7386	9.4850	−25.54
psi_{HMN}	1.00	18.00	17.00	9.5000	5.2547	6.8734	30.80
psi_{LMN}	1.00	18.00	17.00	9.5000	2.6228	9.3955	258.22
psi_{SRV}	1.00	18.00	17.00	9.5000	10.3935	10.3813	−0.12
ror	0.04	0.06	0.02	0.0500	0.0537	0.0530	−1.39
$sigma_{AGR}$	1.00	18.00	17.00	9.5000	10.6068	10.1165	−4.62
$sigma_{HMN}$	1.00	18.00	17.00	9.5000	12.5040	10.6556	−14.78
$sigma_{LMN}$	1.00	18.00	17.00	9.5000	14.2158	8.4732	−40.40
$sigma_{SRV}$	1.00	18.00	17.00	9.5000	9.0117	7.8304	−13.11
zeta	0.98	1.02	0.04	1.0000	1.0158	1.0167	0.09
ω	0.00	1.00	1.00	0.5000	0.3804	0.4125	8.43

The table shows the used parameter values and the comparison between true and calibrated parameters. The calibration procedure is generally able to re-identify the true parameter values

differences ($Z3_{LMN,30}$, $Z3_{AGR,30}$), but overall we see a very good match of values, even for those targets, that were not active in the calibration step.

Beyond prediction errors, it is interesting to analyze the goodness of fit regarding the estimation of the CGE parameters. As can be seen from Table 5 for most parameters, the true value could be nearly identified, with over 60% of the parameters having only a deviation below 15%. The largest deviations can be found for the elasticity of substitution and elasticity of transformation for the sector *LMN*, with estimation errors of −40% and 258%. Especially the latter corresponds to the high forecast error found for exports of the *LMN* sector. Interestingly, based on our approach, we can identify the true underlying economic structure, i.e. the Bayesian estimation delivers a posterior value for ω of 0.41, which moves the uninformative prior value of 0.5 significantly towards the true value of 0.38, implying only an estimation error of slightly over 8%.

Overall, based on achieved *gof* we can conclude that our dynamic calibration procedure is an efficient approach to estimate relevant CGE-parameters, which corresponds to a superior method when compared to standard CGE-calibration. The latter follows especially given the fact that our suggested procedure allows an explicit evaluation of the goodness of fit of derived parameter estimations.

4 DART Application

In a second step, we will apply our approach to calibrate the DART-CLIM model to demonstrate that our approach is suitable to calibrate large multi-regions-multi-sector recursive-dynamic CGE models.

4.1 Dynamic Applied Regional Trade Model

DART-CLIM is a recursive multi-region, multi-sector CGE model that is developed at the Kiel Institute for the World Economy and is being used to assess the effects of climate policies on the global economy (Winkler et al. 2021; Thube et al. 2022) The model is based on the Global Trade Analysis Project (GTAP)-9 database (Aguiar et al. 2016), with an aggregation to 20 regions and 24 sectors. Detailed definitions on regional and sector aggregation are shown in Tables 9 and 10, respectively.

For the application of our calibration method, as shown in Table 6, we selected six region-specific endogenous variables z (total of 120 output variables) that need to be matched to their forecasts. The data for the projection of Gross Domestic Product (GDP) is taken from the OECD macroeconomic forecasts (OECD 2019)

Table 6 Description of DART-CLIM parameters and output variables used in calibration

Outputs	Description	Data Source
gdp	Real regional Gross Domestic Product	OECD (2019)
esolar	Regional electricity production from solar PV	World Energy Outlook (International Energy Agency 2018): Current Policies Scenario
ewind	Regional electricity production from wind	
ffu	Total electricity production from fossil sources (coal, oil and natural gas)	
eother	Rest of electricity technologies excluding nuclear and hydroelectric sources	
semis	Regional CO_2 emissions from burning fossil fuels	

Inputs	Description	Dimension
TFP	Factor productivity parameter for labor and capital	24 x 20
EEI	Autonomous energy efficiency	18 x 20
ESUB	Elasticity of substitution parameter needed to calculate elasticity of fuel supply	3 x 1
dep	Depreciation	20 x 1
eagg_ele	Elasticity of substitution between the eight electricity technologies	20 x 1
preleexp	Exponent for increase of fixed resource for solar PV, wind, and other renewable electricity	3 x 20
esub_res	Elasticity of substitution between fixed resource and capital-labor-energy aggregate	3 x 20
esub_kle	Elasticity of substitution between capital-labor and energy aggregate	24 x 20
armel	Armington elasticity of substitution between imported goods	24 x 1

and the data for the electricity production and CO_2 emissions comes from the World Energy Outlook (International Energy Agency 2018). Specifically, we will match the Compound Annual Growth Rate (CAGR) of the model outputs to those calculated from the forecasts for better numerical stability, assuming an exponential growth path.

Table 6 also shows the nine input parameters of DART-CLIM that we will use to replicate the forecasts of the above said endogenous variables in the baseline scenario. All model parameters are either region-specific and/or sector-specific and thus, account for a total of about 1500 input variables. These nine input parameters correspond to the typical set of model parameters chosen by modellers to calibrate the dynamics of macroeconomic variables, energy sector outputs, and GHG emissions, especially in models that focus on the analysis of energy and climate policies (Foure et al. 2020; Faehn et al. 2020). They affect the development of different electricity technologies and impact their production pathways, portfolio of regional electricity technologies, and total CO_2 emissions. Lastly, since results from CGE models are very sensitive to assumed values for trade elasticities as well as for elasticities of substitution (Mc Daniel and Balistreri 2003; Turner 2009, 2008), we also include two sets of elasticity parameters, namely, *armel* and *esub_kle*. Additional details of the DART-CLIM model are provided in the "Appendix".

4.2 Iterative Approach

The main idea is to follow the same approach as in Sect. 3, using equation 3 to calibrate the selected DART-CLIM parameters. In general, Bayesian estimation could be performed using the original DART-CLIM model. However, since DART-CLIM is a recursive dynamic model, the optimization problem (3) is tedious to solve numerically even when powerful solvers are applied. Given that we want to calibrate over 1500 parameters in this application, the corresponding optimization problem requires a relatively large number of simulation runs, making it costly. Therefore, the efficiency gained by metamodel-based SO is important[8].

The required number of simulations for a polynomial metamodel with interaction effects would also need a large number of simulations, and therefore, we only use local main effects. First-order polynomial metamodels are efficient and effective, provided they are "adequate" approximations (Kleijnen 2020). However, polynomial metamodels generally provide only local approximations, so a series of metamodels are estimated iteratively, optimizing goodness of fit. The developed iterative procedure follows the hill-climbing technique and wraps the steps from Algorithm 1 in a loop. Algorithm 2 gives an overview of the individual steps.

---

[8] Note that in the metamodeling literature, metamodels dealing with more than ten parameters are considered as high dimensional (For example, Shan and Wang (2010); Wang et al. (2011); Sanchez (2006)).

Algorithm 2 Steps

Start iteration $k \in \{1, 2, \dots\}$
    1) DOE: Θ_k
    2) Computing simulations
    3) Estimating and validating metamodels
    4) Bayesian calibration: $\Rightarrow \boldsymbol{\theta}_k^*$
    5) Evaluating calibration point: $\boldsymbol{z}_k^F = F(\boldsymbol{\theta}_k^*)$
    6) Exit or start next iteration $k + 1$ at beginning
End

First, we draw a sample of parameters following a specific DOE. In particular, at each step k, we sample each parameter from an interval $[\theta_{i,k}^{lo}, \theta_{i,k}^{up}]$, with the bounds being iteration specific[9]. Given that we are looking for local first-order main effects, we generate a *metamodel* sample Θ^M where each parameter θ_i is sampled individually within its range, and all other parameters remain unchanged. The sampling strategy is based on the "one factor at a time" sampling method (Kleijnen 2015). In particular, we sample four points for each parameter θ_i with two points close to the current value and two points further away since some parameter values might result in infeasible simulations. The sample points Θ_k^M, a $|\boldsymbol{\theta}| * 4x|\boldsymbol{\theta}|$ matrix, are calculated as follows:

$$
\begin{aligned}
\boldsymbol{\lambda} \quad &= [0.05, 0.45, 0.55, 0.95] \\
\boldsymbol{\theta}_{k,i} \quad &= \theta_i^{lo} + \boldsymbol{\lambda}(\theta_i^{up} - \theta_i^{lo}) \\
\Theta_k^M \quad &=
\begin{bmatrix}
\boldsymbol{\theta}_{k,1} & \theta_{k-1,2}^* & \theta_{k-1,3}^* & \cdots & \theta_{k-1,|\boldsymbol{\theta}|}^* \\
\theta_{k-1,1}^* & \boldsymbol{\theta}_{k,2} & \theta_{k-1,3}^* & \cdots & \theta_{k-1,|\boldsymbol{\theta}|}^* \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\theta_{k-1,1}^* & \theta_{k-1,2}^* & \theta_{k-1,3}^* & \cdots & \boldsymbol{\theta}_{k,|\boldsymbol{\theta}|}
\end{bmatrix}.
\end{aligned}
\tag{8}
$$

In addition, a smaller *validation* sample Θ_k^V (500 parameter specifications) is generated using LHS (Sacks et al. 1989), where parameters $\boldsymbol{\theta}_k$ are sampled simultaneously in the same ranges to test the predictive ability of the estimated meta-models. Computation of the individual runs of the *metamodel* and *validation* samples can easily be split into multiple parallel computations (Herlihy et al. 2020). The parallel computation is time-saving because a single simulation run with DART-CLIM takes from 15 to 60 minutes of computation time and we need to compute $\approx 4 * 1500 + 500 = 6500$ different runs.

In the second step, we conduct simulation with the DART-CLIM model for each sampled parameter vector.

Next, we estimate and validate relevant metamodels M_k (see Sect. 2). In particular, we estimate a metamodel $M_{k,j}$ for each output Z_j and perform the validation.

---

[9] For notational convenience we now denote $\boldsymbol{\theta}_T$ with $\boldsymbol{\theta}$, as we keep all other parameters fixed.

In the fourth step, we conduct the Bayesian estimation of the model parameters θ_k^* based on the metamodels M_k. Using these metamodels we can solve the slightly adapted problem 9 resulting in θ_k^*:

$$
\begin{aligned}
\theta_k^* &= \arg\min_{\theta}(\theta - \theta_{k-1})'\Sigma_{\theta}^{-1}(\theta - \theta_{k-1}) + \epsilon_z'\Sigma_{\epsilon_z}^{-1}\epsilon_z \\
&\text{s.t.} \\
\epsilon_z &= z^o - z \\
z &= M_k(\theta) \\
0 &\equiv H(\theta),
\end{aligned}
\tag{9}
$$

where $0 \equiv H(\theta)$ includes theoretical and empirical restrictions on the parameters including the iteration specific bounds $(\theta_k^{lo}, \theta_k^{up})$.

In the fifth step, we use estimated parameter values θ_k^* to derive trend predictions of relevant output variables, z_k^F, from simulation runs of the original DART-CLIM model. If the prediction error (z^o to z_k^F) is below a critical threshold, then the algorithm stops, and the estimated parameters, θ_k^*, are taken to calibrate DART-CLIM[10]. Otherwise, the process starts again at step 1, while we sample from the interval $[\theta_{k+1}^{lo}, \theta_k^{up}]$ defined around the estimated parameters θ_k^*.

We repeatedly search for a local optimum in a restricted parameter space in this process. The subsequent parameter space for each parameter is set around the best solution found in the previous iteration. This iterative approach builds upon our described methodology (see Sect. 2), though it does not guarantee a global optimum since the iterative process might converge to a local optimum. However, the iterative process guarantees that goodness of fit predicting exogenous forecast variables is improved moving from one iteration to the next[11].

The above-described process is implemented in a combination of R (R Core Team 2020), GAMS (GAMS Development Corporation 2020) and Ansible scripts (Red Hat Software 2020). The Ansible scripts are used to automate the parallel computation in the cluster environment. The cluster environment uses the Slurm workload manager (Yoo et al. 2003). The DART-CLIM model and the Bayesian calibration are implemented in GAMS, while the sampling procedure and analysis are implemented in R (Wickham et al. 2019; Wickham 2016; Dupuy et al. 2015; Dirkse et al. 2020).

5 Results

In Sect. 5.1, we briefly evaluate the validation results by assessing the prediction performances of the metamodels. Subsequently, we assess the calibration results using a *gof* measure and compare the simulated trends with their forecast trends.

---

[10] We set the critical threshold corresponding to an average prediction error of 5%. Approximatively, this critical threshold reflects the relative information value of the forecasts vis-a-vis information encapsulated in the prior distribution of the calibrated DART-CLIM parameters.

[11] It should be further noted that as long as the data on exogenous forecasts is itself noisy, parameter estimates implying perfect predictions are generally not desirable.

Finally, we conclude the section with a short demonstration of the importance of baseline calibration on policy implications by simulating the impacts of fulfilling the emission reduction targets as committed under the Paris Agreement in the DART-CLIM model.

5.1 Validation

We use the *metamodel* sample to derive metamodels which are then used to make predictions of the *validation* sample. The iterative approach moves the model parameter space recurrently towards a parameter space, where the simulated trends are close to the forecast trends. We check whether the local first-order polynomial metamodels for each iteration can approximate the underlying (I/O) relationships sufficiently well by assessing the prediction performance by means of the AER (see Equation 7):

$$
\text{AER}_{k,j} = \left| \frac{RMSE_{k,j}}{\bar{Z}_{k,j}^F} \right| = \left| \frac{\sqrt{\frac{1}{v} \sum_{u=1}^{v} \left(Z_{k,j,u}^F - Z_{k,j,u}^M \right)^2}}{\frac{1}{v} \sum_{u=1}^{v} Z_{k,j,u}^F} \right|,
\tag{10}
$$

where $Z_{k,j,u}^F$ and $Z_{k,j,u}^M$ denote the uth simulated and predicted value for the jth output variable in the kth iteration. Moreover, we can see the evolution of the prediction performances across the iterations.

The results are shown in Fig. 3 for each output variable in each region and each iteration, which is denoted by the different colors (see Table 11 for detailed results).

The different scales on the y-axes provide a rough indication of the difficulty of using local first-order polynomial metamodels to approximate the underlying (I/O) relationships of the DART-CLIM model, particularly in earlier iterations. Specifically, from Fig. 3 we observe that the (I/O) relationships for gdp, esolar and ewind are relatively easier to approximate than those for ffu, eother and semis as is evident from the lower AER values of the former three outputs relative to the latter three, especially in earlier iterations. A possible explanation for this could be that for ffu, eother and semis the underlying relationships are more complex than what the metamodels can capture, particularly when the model parameter space is large. Within each of these three variables, we also see regional differences in the magnitude of the prediction errors. For instance, the linear metamodels perform better at capturing the (I/O) relationships in eother in *India* than in eother in *USA*. This regional variation can be partly attributed to the linearity of the metamodels and, thus, their limitation in capturing the underlying relationship driven by factors like interaction effects across regions and sectors.

However, in the majority of the cases, the first-order polynomials give satisfactory prediction performance and therefore, we still consider this assumption robust. Overall, from Fig. 3 it can be seen that for each output variable and across regions, the AER values tend to decrease with a higher number of iterations and
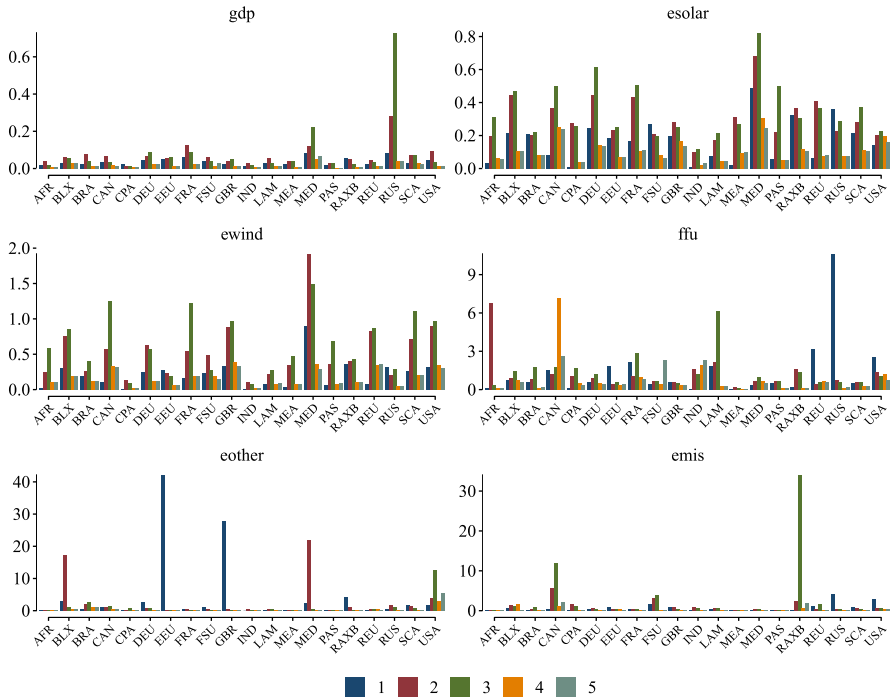
Fig. 3 The prediction performance (AER) for each output variable in each region and each iteration. Lower AER values indicate better prediction performance of metamodels. Performance of metamodels improves with more iterations

finally reach relatively low values, thereby indicating fine prediction performances of the metamodels.

5.2 Baseline Calibration

The primary goal of our approach is to calibrate the baseline of dynamic models wherein the simulated trends from model outputs in the baseline scenario are close to their forecast trends. To this end, we use two measures to assess the calibration results. The first measure is again (*gof*) which is proportional to the (negative) likelihood part ($\mathcal{L}(\boldsymbol{y}^o \mid \boldsymbol{\theta})$) of the target function (see Equations 3 and 9). The second measure κ follows a more intuitive idea. We compare the simulated values z_k^F (with $F(\boldsymbol{y}, \boldsymbol{\theta}_k^*) \equiv 0$) with the corresponding forecasts z^o and define an indicator function to count the number of output variables that fall in the pre-defined range:

$$\kappa_{k,j}(Z_j, Z^o_j) = \begin{cases} 1 & \text{if } 0.9 * f(Z^o_j) \le f(Z^F_{k,j}) \le 1.1 * f(Z^o_j) \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

with f a transformation function. In the simplest case f is just the identity function but it can also transform the CAGR into real values like GDP in *US*$.

To assess the results from our application, we first take a look at the *gof* measure. The results are given in Table 7 where "predicted" represents the *gof* computed by using $z_k^M = M_k(z, \theta_k^*)$ while "simulated" represents the *gof* computed by using z_k^F.

The "predicted" *gof* values in the *tot* column measure the overall deviations of z_k^M from z^o. The "predicted" *gof* values decreases with each iteration reflecting improvements in the prediction performance of metamodels. This pattern is also similar to the results seen in Sect. 5.1. Moreover, we can discern differences between the *gof* values for "predicted" and "simulated". The differences are large in earlier iterations but subsequently reduce. Thus, the prediction performance of metamodels has a significant influence on the calibration process, and the improvement of the prediction performance leads to the refinement of the calibration point. After five iterations, we achieve an average prediction error below 5% of the calibrated baseline compared to the forecast trends.

Second, in order to have a more straightforward understanding of the results in the calibrated baseline scenario, we take a look at the results of the κ measure in Table 7. We transform the CAGR into real values for the year 2030 for each output variable in each region. This is important since small differences in CAGRs can lead to large differences in real values after accumulation over the years. From the last column in Table 7 we see that in the fifth iteration, 108 out of 120 simulated trends fall into the corresponding $\pm 10\%$ ranges around the forecast trends, while in the first iteration, only 67 satisfy this criterion. Again, this demonstrates the advantages of the iterative approach and the improvement gains with additional iterations.

These results show that the iterative calibration method is capable of calibrating the baseline scenario of dynamic CGE model DART-CLIM in a structured and systematic way. The approach provides satisfactory calibration results that can be quantified, and it can be easily applied to a variety of simulation models. It should be noted that since the goal here is to illustrate the application of our method, we only performed five rounds of iteration. However, the error margins of the baseline calibration can be further reduced with further iterations.

5.3 Implications of an Improved Calibration Procedure on Policy Analysis

The main aim of parameter calibration is to use all available data that is informative regarding the specification of relevant model parameters. Our explanations demonstrate that using existing exogenous forecast data for parameter calibration is a rather complex and laborious process. Thus, one could finally ask whether it is worth the additional effort. This is an empirical question since parameters resulting from different calibration processes (ad-hoc static versus Bayesian dynamic methods) can imply significantly different economic responses to given policy shocks derived from the same CGE model.

An important response of economic systems to simulated climate policies is encapsulated in the induced change of marginal abatement costs. Hence, we use this variable to assess whether our dynamic calibration method induces significantly different economic responses than a standard calibration procedure. To this end, we impose in a simulated climate policy scenario regional unilateral CO_2 reduction

Table 7 Goodness-of-fit (*gof*) across iterations

i	Comparison	Emis		EOther		ESolar		EWind		ffu		gdp		tot	
		gof	κ	gof	κ	gof	κ	gof	κ	gof	κ	gof	κ	gof	κ
1	Predicted	120.61	11	695.54	18	205.43	16	2.36	19	282.42	11	255.99	10	1562.36	85
	Simulated	88.48	14	6632.54	13	1139.42	8	113.93	13	559.22	7	181.37	12	8714.95	67
2	Predicted	82.25	15	1.16	20	1.95	18	1.01	19	136.89	15	95.82	16	319.09	103
	Simulated	81.37	15	1227.83	18	122.05	12	213.74	13	206.53	13	95.81	16	1947.32	87
3	Predicted	84.65	15	13.30	19	0.74	18	0.70	19	141.45	16	79.38	17	320.23	104
	Simulated	91.38	14	170.27	15	562.78	8	203.95	16	220.21	14	68.28	17	1316.88	84
4	Predicted	98.92	15	6.78	19	0.51	18	0.40	19	98.50	15	45.22	18	250.33	104
	Simulated	97.97	15	81.49	19	3.71	20	70.64	19	103.27	15	44.69	18	401.76	106
5	Predicted	84.05	15	1.61	20	0.35	18	0.05	19	75.33	18	20.30	18	181.69	108
	Simulated	77.27	15	111.74	19	9.46	19	1.37	20	85.44	17	21.15	18	306.44	108

Lower values of *gof* or higher values of κ indicate better matching of the forecasts. Matching improves with more iterations

targets on all of the regions represented in DART-CLIM. The regional targets correspond to the conditional Nationally Determined Contributions (NDC)[12] that countries have pledged within the proceedings of the Paris Agreement under conditional emission reduction pledges. The emission reduction targets are defined as percentage reductions relative to the baseline.

We simulate the Paris Agreement policy scenario with DART-CLIM applying three different calibration methods implying three different parameter specifications of the model. First, we specify parameters applying our dynamic calibration method, labelled as *full*. Next, we use the standard static calibration method based on observed national SAMs in the base year with ad hoc assumed values for parameters driving future response. We refer to this specification method as *static*. Finally, we applied our dynamic calibration method again. However, in this case, we only matched three trends, namely regional GDP, total fossil-fueled electricity, and CO_2 emissions, and stopped the iterations after two rounds. In this case, we matched 48 out of 60 targets in a $\pm 10\%$ range and 54 in a $\pm 20\%$ range. This specification is labeled as *subset*.

As can be seen from Table 8 estimated carbon prices (marginal abatement costs) differ significantly across parameter specifications. Since the CO_2 reduction targets are defined as percentage reductions relative to the baseline, the absolute emission reductions that need to be undertaken by regions differ across the three scenarios. Owing to the differences across baselines, the inherent economic trade-offs for regions and sectors differ under the policy scenario. This difference gives to different estimates of abatement costs across regions. On average, percentage differences in national carbon prices equals $\approx 21\%$ comparing standard parameter calibration (*static*) with our advanced dynamic calibration method (*full*), while on average, this difference amounts to $\approx 16\%$ for the *subset* specification. Moreover, in some regions like *PAS* or *BRA* simulated carbon prices vary by $\approx 50\%$ depending on the applied calibration procedures. We refrain from diving into further details about the results since the goal of this exercise is only to showcase that choices made in baseline calibration of CGE models are of crucial importance for further policy analysis.

6 Conclusion

Nowadays, evidence-based policies correspond to a standard approach of good governance. However, policy analysis is plagued by model uncertainty (Manski 2018). This holds especially true for CGE-applications as these models usually have weak econometric foundations. Accordingly, given that available data allowing a sophisticated econometric parameter specification remain rather limited, this paper develops an innovative, dynamic calibration method applying a Bayesian estimation framework. Bayesian estimation techniques are already successfully applied to estimate parameters of DSGE models. Like CGEs DSGE models correspond to sound micro-founded and theoretically consistent model structures. Applying

---

[12] The method for calculating the emission reduction targets is based on Böhringer et al. (2021).

Table 8 Carbon price by region derived from different calibration approaches

r	Full	Subset	Static
AFR	10.59	12.65	11.95
BLX	195.14	172.90	158.31
BRA	125.10	82.49	65.11
CAN	38.40	35.56	33.49
CPA	12.10	15.31	14.67
DEU	47.37	51.14	51.40
EEU	27.01	25.20	27.25
FRA	122.71	113.14	103.67
FSU	51.44	51.19	45.00
GBR	41.42	44.73	39.76
IND	16.28	17.43	15.33
LAM	27.72	26.41	22.86
MEA	33.10	30.70	26.76
MED	67.78	85.26	98.18
PAS	53.61	85.46	80.51
RAXB	35.93	41.12	36.42
REU	365.74	272.76	258.10
RUS	11.88	8.44	8.01
SCA	55.32	59.78	47.37
USA	19.99	20.49	28.75

Differences in baseline calibration choices impact characteristics of the baseline. Consequently, implementation of the same policy with different baselines leads to different economic trade-offs in the model and thus, different abatement cost estimates

Bayesian estimation techniques DSGE combine these theoretical properties with good forecasting performance and hence have become an additional useful tool in the forecasting kit of central banks (Smets and Wouters 2003; Hashimzade and Thornton 2013). However, compared to CGE models DSGEs are still highly aggregated models, which assume highly aggregated economic agents, e.g., do not include heterogenous sectoral structures. Therefore, in contrast to DSGE models, Bayesian estimation techniques have rarely been applied to estimate parameters of a CGE model and even nowadays are still challenging. Technically, the Bayesian framework corresponds to an optimization problem. Given the complexity of most CGE models solution of the latter optimization problem cannot be solved with standard numerical solution algorithms. For example, the calibration of the DART-CLIM model includes over 1500 parameters to be specified.

In this context, this paper develops a metamodel-based simulation optimization (SO) approach. Given the fact that SO approaches generally require many simulation runs, and each run is quite expensive, the advantage of using metamodels is that they significantly reduce computation time, implying more efficient SO methods.

In particular, we first demonstrate how our approach can be applied using a toy model. Then, to make our approach accessible, especially to CGE-modellers, we explicitly supply all necessary code for replication. Second, we show the effectiveness of our approach by applying it to perform the calibration of a dynamic baseline of the DART-CLIM model based on forecasts of six central outputs. Our iterative calibration approach delivers convincing results. For example, for fitting values for about 1500 parameters using a total of 120 outputs, the prediction errors could be significantly reduced from iteration to iteration until the average prediction error led below an acceptable range of 5%. Based on derived metamodels, Bayesian estimation of model parameters can be done quite effectively, requiring significant lower computational effort when compared to estimation procedures using the original CGE, i.e. it takes less than a minute. The derivation of the metamodels requires a significant computational effort, though, but that can be easily split across multiple computers or can be computed in a cluster environment.

Nevertheless, it is fair to conclude that compared to simple static calibration methods typically used in the literature, our advanced approach is definitely technically more demanding. Thus, the question arises: Is it worth the additional effort? In this regard, we demonstrate in this paper that economic response to climate policies encapsulated in regional emission prices derived from DART-CLIM differ by up to 21% depending on the applied calibration method. Hence, we conclude that investing additional resources into comprehensive parameter calibration is definitely worth the effort.

Furthermore, we illustrate that applying our Bayesian framework, the posterior distribution of the model can be used to calculate the complete distribution of the forecast and various risk measures proposed in the literature. In particular, the structural nature of the model allows computing the posterior distribution forecast conditional on a policy path. Moreover, it also allows examining the structural sources of the forecast errors and their policy implications. Technically, metamodeling also facilitates Metropolis-Hasting sampling from the posterior distribution. Thus, our method is not only a good method to explicitly reveal induced uncertainty of model like SSA, but also an appropriate procedure to reduce model uncertainty. Finally, it also allows for better reproducibility of policy research, as it replaces the ad-hoc (often trial and error) approaches with a theoretically founded and algorithmic approach.

Of course, many challenges still lie ahead. Here we particularly like to highlight three of those challenges that we consider interesting topics of future research. First, as Manski (2018) highlighted in his seminal paper, integrating fundamental model uncertainty is crucial for conducting comprehensive model-based policy analysis. We certainly consider our Bayesian approach as a promising methodological starting point in this context. However, much more work is needed to both deepen and broaden the scope of existing CGE-based policy analysis. For example, when forecasting in a policy context, realistic assumptions about how practical politicians form their beliefs on how alternative policies impact relevant policy outcomes are crucial. In particular, political reality constraints on information processing, pervasive model uncertainty and the associated need for perpetual learning

highlight the limits of the rational expectations assumption. Another topic for future research is to extend our calibration method to other complex models like interlinked ecological-economic model frameworks. Furthermore, at a methodological level, it will be interesting to investigate other metamodeling techniques in the context of this approach, especially regarding computational intensity and prediction accuracy. In the iterative procedure used for large models, it appears worthwhile to investigate how to reuse information from previous iterations, like samples or even the estimated metamodels. The latter might be done by building a metamodel ensemble (Bartz-Beielstein and Zaefferer 2017; Friese et al. 2016), investigating if weighted linear combinations of metamodels improve the calibration procedure. The weights might be determined by a distance measure to the sample space of the respective metamodel. Given the advances in machine learning and automatic differentiation, in particular, it will also be interesting to investigate if those techniques can be used in our approach (Baydin et al. 2018).

Appendix

A Dart

DART-CLIM is a recurisve multi-region, multi-sector CGE model. The model version used in the application is based on the GTAP9 database. The primary goal of this model version is analysis of climate and energy policies. Representation of sectoral details affects the policy impact anaylsis in CGE model(Fischer and Morgenstern 2006; Kuik et al. 2009). Therefore, in order to have a rich representation of the energy sectors, data about electricity production by technology is additionally included from the GTAP - Power database (Peters 2016). In the GTAP9 - Power database the supply of some of these technologies is differentiated by base-load and peak-load. However, in DART-CLIM we have aggregated the base-load and peak-load sectors to a single homogeneous sector.

Overall DART-CLIM has eight different technologies that produce electricity: solar PV, wind, nuclear, hydro, coal, gas, oil, and other renewable technologies. We assume that the development of nuclear and hydroelectric technologies is significantly determined by political will-power and/or geographical framework conditions, and less by price effects in electricity markets. Accordingly, the production pathways of these two technologies is defined exogenously in the model based on data from International Energy Agency (2018). The production activites are defined as CES functions in the model. The details about the nesting structures for different sectors are available in (Thube et al. 2022).

See Tables 9 and 10.

Table 9 Regions disaggregation in DART-CLIM

r	Full name
AFR	Sub Saharan Africa
BLX	Belgium, Netherlands and Luxembourg
BRA	Brazil
CAN	Canada
CPA	China and Hong Kong
DEU	Germany
EEU	Czech Republic, Slovakia, Slovenia, Hungary, Estonia, Latvia, Lithuania, Bulgaria, Romania, Croatia, Austria, Poland
FRA	France
FSU	Kazakhstan, Kyrgyzstan, Ukraine, Albania, Belarus, Armenia, Azerbaijan, Tajikistan, Turkmenistan, Uzbekistan, Georgia, Rest of Europe
GBR	United Kingdom, Ireland
IND	India
LAM	Central- and South America
MEA	Middle East, Northern Africa and Turkey
MED	Mediterranean Europe: Italy, Spain, Portugal, Malta, Greece, Cyprus
PAS	Pacific Asia
ANJ	Australia, New Zealand and Japan
REU	EFTA and rest of the World: Norway, Iceland, Liechtenstein, Switzerland, Overseas Territories and Antarctica
RUS	Russia
SCA	Sweden, Denmark and Finland
USA	United States of America

Table 10 Sectoral disaggregation in DART-CLIM

i	Full name
Non-Energy Products (12)	
CRP	Chemical Products (rubber, plastic)
ETS	Energy-intensive production
MOB	Mobility
OLI	Other light industries
OHI	Other heavy industries
SVCS	Services
TND	Transmission and distribution
ANI	Animal Products
GRN	grains
OSD	oilseeds
CRO	rest of crops
RAGR	Rest agriculture and other processed food
Energy Products (12)	
ENuclear	Nuclear electricity
ESolar	Solar PV electricity
EWind	Wind electricity
EHydro	Hydro-electricity
ECoal	Coal electricity
EGas	Gas for electricity
EOil	Petroleum and coal products for electricity
EOther	Biofuels, waste, geothermal and tidal technologies
COL	Coal
OIL	Petroleum and coal products
GAS	Gas
CRU	Oil

B Results

See Table 11.

Table 11 AER by target and iteration

Target	i	AFR	BLX	BRA	CAN	CPA	DEU	EEU	FRA	FSU	GBR
gdp	1	0.02	0.03	0.02	0.03	0.02	0.04	0.05	0.06	0.04	0.02
	2	0.04	0.06	0.08	0.07	0.01	0.07	0.06	0.13	0.06	0.04
	3	0.02	0.06	0.04	0.03	0.01	0.09	0.06	0.09	0.04	0.05
	4	0.01	0.03	0.01	0.02	0.01	0.02	0.01	0.02	0.01	0.01
	5	0.01	0.03	0.01	0.02	0.01	0.02	0.01	0.03	0.03	0.01
Esolar	1	0.03	0.22	0.21	0.08	0.01	0.25	0.19	0.17	0.27	0.19
	2	0.20	0.45	0.20	0.37	0.28	0.44	0.23	0.43	0.21	0.28
	3	0.31	0.47	0.22	0.50	0.26	0.61	0.25	0.50	0.20	0.25
	4	0.06	0.11	0.08	0.25	0.04	0.14	0.07	0.11	0.08	0.17
	5	0.06	0.11	0.08	0.24	0.04	0.14	0.07	0.11	0.06	0.14
Ewind	1	0.02	0.31	0.19	0.11	0.00	0.24	0.27	0.16	0.24	0.33
	2	0.25	0.76	0.27	0.57	0.13	0.63	0.23	0.55	0.48	0.88
	3	0.59	0.85	0.40	1.25	0.09	0.58	0.19	1.22	0.28	0.96
	4	0.11	0.20	0.12	0.33	0.02	0.12	0.07	0.19	0.18	0.38
	5	0.11	0.19	0.13	0.32	0.02	0.11	0.06	0.19	0.15	0.33
ffu	1	0.10	0.72	0.58	1.56	0.12	0.55	1.81	2.18	0.40	0.57
	2	6.78	0.93	0.84	1.24	1.07	0.87	0.40	1.09	0.65	0.62
	3	0.31	1.46	1.76	1.79	1.64	1.21	0.62	2.84	0.66	0.50
	4	0.08	0.73	0.15	7.13	0.54	0.49	0.33	1.00	0.45	0.36
	5	0.08	0.57	0.16	2.64	0.34	0.44	0.41	0.83	2.29	0.34
Eother	1	0.06	2.87	0.34	1.22	0.00	2.64	42.20	0.33	1.22	27.85
	2	0.21	17.30	1.94	1.24	0.26	0.79	0.29	0.37	0.54	0.50
	3	0.11	1.07	2.61	1.35	0.75	0.66	0.13	0.26	0.29	0.27
	4	0.07	0.40	1.07	0.49	0.05	0.18	0.02	0.08	0.11	0.03
	5	0.08	0.55	1.15	0.51	0.05	0.19	0.02	0.07	0.15	0.03
Emis	1	0.04	0.58	0.13	0.42	0.05	0.28	0.90	0.30	1.60	0.81
	2	0.24	1.41	0.47	5.54	1.64	0.55	0.34	0.38	3.02	0.86
	3	0.14	1.07	0.94	11.96	1.20	0.37	0.28	0.47	3.76	0.42
	4	0.04	1.56	0.08	1.03	0.10	0.16	0.27	0.07	0.19	0.22
	5	0.04	0.14	0.08	2.19	0.08	0.15	0.22	0.07	0.23	0.20
Target	**i**	**IND**	**LAM**	**MEA**	**MED**	**PAS**	**RAXB**	**REU**	**RUS**	**SCA**	**USA**
gdp	1	0.01	0.03	0.02	0.09	0.02	0.06	0.03	0.08	0.03	0.05
	2	0.03	0.06	0.04	0.12	0.03	0.05	0.05	0.28	0.07	0.10
	3	0.02	0.03	0.04	0.22	0.03	0.02	0.04	0.73	0.07	0.04
	4	0.01	0.01	0.01	0.05	0.01	0.01	0.01	0.04	0.03	0.01
	5	0.01	0.01	0.01	0.07	0.01	0.01	0.01	0.04	0.03	0.01
Esolar	1	0.01	0.08	0.02	0.49	0.06	0.32	0.06	0.36	0.22	0.14
	2	0.10	0.17	0.31	0.68	0.22	0.36	0.41	0.23	0.28	0.20
	3	0.12	0.22	0.27	0.82	0.50	0.31	0.37	0.29	0.37	0.23
	4	0.02	0.04	0.10	0.31	0.05	0.12	0.08	0.07	0.11	0.20
	5	0.03	0.05	0.10	0.24	0.05	0.11	0.08	0.08	0.11	0.16

Table 11 continued

Target	i	IND	LAM	MEA	MED	PAS	RAXB	REU	RUS	SCA	USA
Ewind	1	0.00	0.08	0.04	0.90	0.07	0.36	0.08	0.33	0.26	0.31
	2	0.10	0.22	0.35	1.92	0.36	0.41	0.83	0.20	0.71	0.90
	3	0.07	0.27	0.47	1.49	0.68	0.43	0.88	0.29	1.12	0.97
	4	0.03	0.08	0.07	0.36	0.08	0.11	0.35	0.05	0.20	0.34
	5	0.03	0.09	0.08	0.30	0.09	0.10	0.36	0.04	0.20	0.30
ffu	1	0.02	1.85	0.05	0.32	0.50	0.21	3.19	10.64	0.51	2.57
	2	1.58	2.13	0.20	0.69	0.66	1.64	0.45	0.71	0.58	1.35
	3	1.20	6.15	0.15	0.98	0.67	1.41	0.59	0.60	0.60	1.06
	4	1.90	0.24	0.04	0.68	0.15	0.15	0.64	0.14	0.26	1.20
	5	2.30	0.25	0.04	0.50	0.13	0.14	0.61	0.18	0.24	0.72
Eother	1	0.00	0.15	0.04	2.36	0.03	4.16	0.26	0.49	1.79	1.79
	2	0.43	0.41	0.07	21.78	0.13	0.99	0.36	1.80	1.50	3.97
	3	0.26	0.56	0.06	0.49	0.19	0.26	0.53	0.97	0.62	12.45
	4	0.08	0.19	0.03	0.25	0.08	0.04	0.33	0.16	0.27	3.02
	5	0.07	0.17	0.03	0.23	0.10	0.06	0.30	0.14	0.22	5.37
Emis	1	0.02	0.32	0.03	0.21	0.23	0.18	1.05	4.21	0.83	2.94
	2	0.76	0.56	0.09	0.50	0.09	2.48	0.40	0.41	0.59	0.61
	3	0.60	0.54	0.13	0.43	0.12	34.08	1.60	0.35	0.40	0.72
	4	0.09	0.11	0.03	0.24	0.06	0.59	0.12	0.11	0.14	0.33
	5	0.11	0.13	0.03	0.18	0.06	1.79	0.10	0.15	0.14	0.31

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Aguiar, A., Narayanan, B., & McDougall, R. (2016). An overview of the GTAP 9 data base. *Journal of Global Economic Analysis, 1*, 181–208.

Arndt, C., Robinson, S., & Tarp, F. (2002). Parameter estimation for a computable general equilibrium model: A maximum entropy approach. *Economic Modelling, 19*(3), 375–398.

Barthelemy, J. F., & Haftka, R. T. (1993). Approximation concepts for optimum structural design—A review. *Structural optimization, 5*(3), 129–144.

Bartz-Beielstein, T., & Zaefferer, M. (2017). Model-based methods for continuous and discrete global optimization. *Applied Soft Computing, 55*, 154–167. https://doi.org/10.1016/j.asoc.2017.01.039

Baydin, A.G., Pearlmutter, B.A., Radul, A.A., Siskind, J.M. (2018). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research 18(153), 1–43. http://jmlr.org/papers/v18/17-468.html*

Böhringer, C., Dijkstra, B., & Rosendahl, K. E. (2014). Sectoral and regional expansion of emissions trading. *Resource and Energy Economics, 37*, 201–225. https://doi.org/10.1016/j.reseneeco.2013.12.003

Böhringer, C., Peterson, S., Rutherford, T. F., Schneider, J., & Winkler, M. B. J. (2021). Climate policies after paris: Pledge, trade and recycle: Insights from the 36th energy modeling forum study (emf36). *Energy Economics, 103*

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Burniaux, J. M., & Chateau, J. (2014). Greenhouse gases mitigation potential and economic efficiency of phasing-out fossil fuel subsidies. *International Economics, 140*, 71–88. https://doi.org/10.1016/j.inteco.2014.05.002

Caliendo, L., & Parro, F. (2014). Estimates of the trade and welfare effects of NAFTA. *The Review of Economic Studies, 82*(1), 1–44. https://doi.org/10.1093/restud/rdu035

Chen, V. C., Tsui, K. L., Barton, R. R., & Meckesheimer, M. (2006). A review on design, modeling and applications of computer experiments. *IIE Transactions, 38*(4), 273–291.

Cressie, N.A.C. (1993). Statistics for Spatial Data. Wiley. 10.1002/9781119115151

Dey, S., Mukhopadhyay, T., & Adhikari, S. (2017). Metamodel based high-fidelity stochastic analysis of composite laminates: A concise review with critical comparative assessment. *Composite Structures, 171*, 227–250.

Dirkse, S., Ferris, M., Jain, R. (2020). gdxrrw: An Interface Between 'GAMS' and R. http://www.gams.com, r package version 1.0.6

Dixon, P. B., Jorgenson, D. W. (eds) (2013). Handbook of Computable General Equilibrium Modeling. NORTH HOLLAND

Dixon, P.B., Rimmer, M., Parmenter, B.R. (1997). The Australian textiles, clothing and footwear sector from 1986-87 to 2013-14: analysis using the Monash model. Monash University

Dupuy, D., Helbert, C., Franco, J. (2015). DiceDesign and DiceEval: Two R packages for design and analysis of computer experiments. Journal of Statistical Software 65(11):1–38, http://www.jstatsoft.org/v65/i11/

Faehn, T., Bachner, G., Beach, R.H., Chateau, J., Fujimori, S., Ghosh, M., Hamdi-Cherif, M., Lanzi, E., Paltsev, S., Vandyck, T., et al. (2020). Capturing key energy and emission trends in cge models: Assessment of status and remaining challenges. Tech. rep.

Fan, Q., Fisher-Vanden, K., & Klaiber, H. A. (2018). Climate change, migration, and regional economic impacts in the united states. *Journal of the Association of Environmental and Resource Economists, 5*(3), 643–671. https://doi.org/10.1086/697168

Fischer, C., Morgenstern, R.D. (2006). Carbon abatement costs: Why the wide range of estimates? The Energy Journal 27(2):73–86, http://www.jstor.org/stable/23297020

Forrester, A., Sobester, A., & Keane, A. (2008). *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons.

Foure, J., Aguiar, A., Bibas, R., Chateau, J., Fujimori, S., Lefevre, J., Leimbach, M., Rey-Los-Santos, L., & Valin, H. (2020). Macroeconomic drivers of baseline scenarios in dynamic cge models: review and guidelines proposal. *Journal of Global Economic Analysis, 5*(1), 28–62.

Friedman, J. H., et al. (1991). Multivariate adaptive regression splines. *The annals of statistics, 19*(1), 1–67.

Friese, M., Bartz-Beielstein, T., Emmerich, M. (2016). Building ensembles of surrogate models by optimal convex combination. Tech. rep., http://nbn-resolving.de/urn:nbn:de:hbz:832-cos4-3480

Fujimori, S., Kubota, I., Dai, H., Takahashi, K., Hasegawa, T., Liu, J. Y., Hijioka, Y., Masui, T., & Takimi, M. (2016). Will international emissions trading help achieve the objectives of the paris agreement? *Environmental Research Letters, 11*(10), 104001. https://doi.org/10.1088/1748-9326/11/10/104001

GAMS Development Corporation (2020) General Algebraic Modeling System (GAMS) Release 33.1. Washington, DC, USA, http://www.gams.com/

Gehlhar, M.J. (1994). Economic growth and trade in the pacific rim: An analysis of trade patterns. PhD thesis, Purdue University

Go, D. S., Lofgren, H., Ramos, F. M., & Robinson, S. (2016). Estimating parameters and structural change in CGE models using a bayesian cross-entropy estimation approach. *Economic Modelling, 52*, 790–811. https://doi.org/10.1016/j.econmod.2015.10.017

Gomme, P., Lkhagvasuren, D. (2013). Calibration and simulation of DSGE models. In: Hashimzade N, Thornton MA (eds) Handbook of Research Methods and Applications in Empirical Macroeconomics, Chapters, Edward Elgar Publishing, chap 24, pp 575–592, https://ideas.repec.org/h/elg/eechap/14327_24.html

Gong, W., Duan, Q., Li, J., Wang, C., Di, Z., Dai, Y., Ye, A., & Miao, C. (2015). Multi-objective parameter optimization of common land model using adaptive surrogate modeling. *Hydrology and Earth System Sciences, 19*(5), 2409–2425.

Hansen, L. P., & Heckman, J. J. (1996). The empirical foundations of calibration. *Journal of Economic Perspectives, 10*(1), 87–104. https://doi.org/10.1257/jep.10.1.87

Hashimzade, N., & Thornton, M. A. (2013). *Handbook of research methods and applications in empirical macroeconomics*. Cheltenham, UK: Edward Elgar.

Heckelei, T., Mittelhammer, R. (2008). A bayesian alternative to generalized cross entropy solutions for underdetermined econometric models. Discussion Paper 2, Institute for Food and Resource Economics University of Bonn

Herlihy, M., Shavit, N., & Luchangco, V. (2020). *The Art of Multiprocessor Programming*. Oxford: Elsevier LTD.

Hertel, T., Hummels, D., Ivanic, M., & Keeney, R. (2007). How confident can we be of CGE-based assessments of Free Trade Agreements? *Economic Modelling, 24*, 611–635.

Hertel, T.W. (1997). Global Trade Analysis: Modeling and Applications. Cambridge Univ. Press, Cambridge [u.a.]

Hosoe, N., Gasawa, K., & Hashimoto, H. (2010). Textbook of Computable General Equilibrium Modelling. *Palgrave Macmillan UK*. https://doi.org/10.1057/9780230281653

Hosoe, N., Gasawa, K., Hashimoto, H. (2016). Textbook of Computable General Equilibrium Modeling: Programming and Simulations 2nd Edition. University of Tokyo Press, 10.1057/9780230281653

International Energy Agency. (2018). World Energy Outlook 2018. *OECD*

Jorgenson, D. (1984). Econometric Methods for Applied General Equilibrium Analysis, Cambridge University Press, Cambridge, pp 139–203. Growth 2, ch. 2, pp. 89-155.

Kehoe, T. J., Polo, C., & Sancho, F. (1995). An evaluation of the performance of an applied general equilibrium model of the Spanish economy. *Economic Theory, 6*(1), 115–141. https://doi.org/10.1007/bf01213943

Kleijnen JP (2015) Design and analysis of simulation experiments. In: International Workshop on Simulation, Springer, pp 3–22

Kleijnen, J. P., & Sargent, R. G. (2000). A methodology for fitting and validating metamodels in simulation. *European Journal of Operational Research, 120*(1), 14–29.

Kleijnen JPC (2020) Simulation optimization through regression or kriging metamodels. In: High-Performance Simulation-Based Optimization, Springer International Publishing, pp 115–135, 10.1007/978-3-030-18764-4_6

Kuik, O., Brander, L., & Tol, R. S. (2009). Marginal abatement costs of greenhouse gas emissions: A meta-analysis. *Energy Policy, 37*, 1395–1403.

Manski, C. F. (2018). Communicating uncertainty in policy analysis. *Proceedings of the National Academy of Sciences, 116*(16), 7634–7641. https://doi.org/10.1073/pnas.1722389115

Mareš, T., Janouchová, E., & Kučerová, A. (2016). Artificial neural networks in the calibration of nonlinear mechanical models. *Advances in Engineering Software, 95*, 68–81.

Mc Daniel, C. A., & Balistreri, E. J. (2003). A review of armington trade substitution elasticities. *Economie internationale, 2*, 301–313.

McKitrick, R. R. (1998). The econometric critique of Computable General Equilibrium modeling: The role of functional forms. *Economic Modelling, 15*, 543–573. https://doi.org/10.1016/S0264-9993(98)00028-5

de Melo, J. (1988). Computable general equilibrium models for trade policy analysis in developing countries: A survey. *Journal of Policy Modeling, 10*(4), 469–503. https://doi.org/10.1016/0161-8938(88)90017-8

Milczarek-Andrzejewska, D., Zawalińska, K., & Czarnecki, A. (2018). Land-use conflicts and the common agricultural policy: Evidence from poland. *Land Use Policy, 73*, 423–433. https://doi.org/10.1016/j.landusepol.2018.02.016

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). *Response surface methodology: process and product optimization using designed experiments.* John Wiley & Sons.

Nong, D., & Simshauser, P. (2020). On energy and climate change policies: The impact of baseline projections. *Applied Energy, 269*, 115062. https://doi.org/10.1016/j.apenergy.2020.115062

OECD (2019) OECD Economic Outlook, Volume 2019 Issue 2. Paris, https://doi.org/https://doi.org/10.1787/9b89401b-en, https://www.oecd-ilibrary.org/content/publication/9b89401b-en

Olekseyuk, Z., & Schürenberg-Frosch, H. (2016). Are Armington elasticities different across countries and sectors? *A European study. Economic Modelling, 55*, 328–342.

Paris, Q., Drogué, S., & Anania, G. (2011). Calibrating spatial models of trade. *Economic Modelling, 28*(6), 2509–2516. https://doi.org/10.1016/j.econmod.2011.07.008

Peters, J.C. (2016). The gtap-power data base: Disaggregating the electricity sector in the gtap data base. Journal of global economic analysis, Global Trade Analysis Project (GTAP), Department of Agricultural Economics, Purdue University, West Lafayette, IN, https://www.gtap.agecon.purdue.edu/resources/ res_display.asp?RecordID=4619

Pothen, F., & Hübler, M. (2021). A forward calibration method for analyzing energy policy in new quantitative trade models. *Energy Economics, 100*,

R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/

Rango, J., Schnorbus, T., Kwee, H., Beck, R., Kinoo, B., Arthozoul, S., Zhang, M. (2013). Comparison of different approaches for global modeling of combustion engines. Design of Experiments (DoE) in engine development pp 70–91

Razavi S, Tolson BA, Burn DH (2012) Review of surrogate modeling in water resources. Water Resources Research 48(7)

Red Hat Software (2020) Ansible. https://www.ansible.com/

Robinson, S., Cattaneo, A., & El-Said, M. (2001). Updating and estimating a social accounting matrix using cross entropy methods. *Economic Systems Research, 13*(1), 47–64. https://doi.org/10.1080/09535310120026247

Ruben, R., & van Ruijven, A. (2001). Technical coefficients for bio-economic farm household models: a meta-modelling approach with applications for southern mali. *Ecological Economics, 36*(3), 427–441. https://doi.org/10.1016/s0921-8009(00)00240-8

Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989) Design and analysis of computer experiments. Statistical science pp 409–423

Sánchez, M. V. (2004). Rising inequality and falling poverty in Costa Rica's agriculture during trade reform: a macro-micro general equilibrium analysis. *Shaker Publishing, Maastricht, chap Chapter, 7*, 189–226.

Sanchez, S.M. (2006). Work smarter, not harder: Guidelines for designing simulation experiments. In: Proceedings of the 2006 Winter Simulation Conference, IEEE, 10.1109/wsc.2006.323037

Shan, S., Wang, G.G. (2010). Metamodeling for high dimensional simulation-based design problems. Journal of Mechanical Design 132(5), 10.1115/1.4001597

Shoven, J. B., & Whalley, J. (1984). Applied general-equilibrium models of taxation and international trade: An introduction and survey. *Journal of Economic Literature, 22*(3), 1007–1051. https://doi.org/10.2307/2725306

Shoven, J.B., Whalley, J. (1992). Applying General Equilibrium. Cambridge University Press, https://EconPapers.repec.org/RePEc:cup:cbooks: 9780521266550

Simpson TW, Peplinski J, Koch PN, Allen JK (1997) On the use of statistics in design and the implications for deterministic computer experiments. Design Theory and Methodology-DTM'97 pp 14–17

Simpson, T. W., Lin, D. K., & Chen, W. (2001). Sampling strategies for computer experiments: design and analysis. *International Journal of Reliability and Applications, 2*(3), 209–240.

Smets, F., & Wouters, R. (2003). An estimated dynamic stochastic general equilibrium model of the euro area. *Journal of the European Economic Association, 1*(5), 1123–1175. https://doi.org/10.1162/154247603770383415

Smith, M. (1993). Neural networks for statistical modeling. Thomson Learning

Sobieszczanski-Sobieski, J., & Haftka, R. T. (1997). Multidisciplinary aerospace design optimization: survey of recent developments. *Structural Optimization, 14*(1), 1–23. https://doi.org/10.1007/bf01197554

Stifel, D. C., & Thorbecke, E. (2003). A dual-dual CGE model of an archetype african economy: trade reform, migration and poverty. *Journal of Policy Modeling, 25*(3), 207–235. https://doi.org/10.1016/s0161-8938(03)00014-0

Tarp, F., Arndt, C., Jensen, H.T., Robinson, S., Heltberg, R. (2002). Facing the development challenge in Mozambique: an economywide perspective. resreport 126, International Food Policy Research Institute (IFPRI)

Taylor, J. E., Yunez-Naude, A., & Dyer, G. (1999). Agricultural price policy, employment, and migration in a diversified rural economy: A village-town CGE analysis from mexico. *American Journal of Agricultural Economics, 81*(3), 653–662. https://doi.org/10.2307/1244030

Thube, S. D., Delzeit, R., & Henning, C. H. (2022). Economic gains from global cooperation in fulfilling climate pledges. *Energy Policy, 160,*. https://doi.org/10.1016/j.enpol.2021.112673

Turner, K. (2008). A computable general equilibrium analysis of the relative price sensitivity required to induce rebound effects in response to an improvement in energy efficiency in the uk economy. SIRE Discussion Papers 2008-20, Scottish Institute for Research in Economics (SIRE), https://ideas.repec.org/p/edn/sirdps/33.html

Turner, K. (2009). Negative rebound and disinvestment effects in response to an improvement in energy efficiency in the UK economy. *Energy Economics, 31*(5), 648–666. https://doi.org/10.1016/j.eneco.2009.01.008

Vapnik, V. (2013). The nature of statistical learning theory. Springer science & business media

Villa-Vialaneix, N., Follador, M., Ratto, M., & Leip, A. (2012). A comparison of eight metamodeling techniques for the simulation of n2o fluxes and n leaching from corn crops. *Environmental Modelling & Software, 34*, 51–66.

Wang, H., Tang, L., & Li, G. (2011). Adaptive MLS-HDMR metamodeling techniques for high dimensional problems. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2011.04.220

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, https://ggplot2.tidyverse.org

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019) Welcome to the tidyverse. Journal of Open Source Software 4(43):1686, 10.21105/joss.01686

Winkler, M. B. J., Peterson, S., & Thube, S. (2021). Gains associated with linking the eu and chinese ets under different assumptions on restrictions, allowance endowments, and international trade. *Energy Economics, 104,*

Yildizoglu, M., Salle, I., et al. (2012). *Efficient sampling and metamodeling for computational economic models*. Groupe de Recherche en Economie Théorique et Appliquée: Tech. rep.

Yoo, A.B., Jette, M.A., Grondona, M. (2003). SLURM: Simple linux utility for resource management. In: Job Scheduling Strategies for Parallel Processing, Springer Berlin Heidelberg, pp 44–60, 10.1007/10968987_3

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.