

## METHOD

## Towards robust statistical inference for complex computer models

Johannes Oberpriller,<sup>1\*</sup>   
 David R. Cameron,<sup>2</sup>  
 Michael C. Dietze<sup>3</sup>  and  
 Florian Hartig<sup>1</sup> 

**Abstract**

Ecologists increasingly rely on complex computer simulations to forecast ecological systems. To make such forecasts precise, uncertainties in model parameters and structure must be reduced and correctly propagated to model outputs. Naively using standard statistical techniques for this task, however, can lead to bias and underestimation of uncertainties in parameters and predictions. Here, we explain why these problems occur and propose a framework for robust inference with complex computer simulations. After having identified that model error is more consequential in complex computer simulations, due to their more pronounced nonlinearity and interconnectedness, we discuss as possible solutions data rebalancing and adding bias corrections on model outputs or processes during or after the calibration procedure. We illustrate the methods in a case study, using a dynamic vegetation model. We conclude that developing better methods for robust inference of complex computer simulations is vital for generating reliable predictions of ecosystem responses.

**Keywords**

Bayesian Inference, bias correction, biased models, data imbalance, robust inference.

*Ecology Letters* (2021) **24**: 1251–1261

**INTRODUCTION**

Ecological systems are often complex and interdependent (Levin, 1998). To understand these systems, and to forecast their dynamics under changing conditions, ecologists rely increasingly on complex computer simulations (CCS, near synonymous terms include: process-based models, mechanistic models, system models; see e.g. Evans *et al.*, 2012; Briscoe *et al.*, 2019; Thompson *et al.*, 2020), for example to predict ecosystem responses to climate change (e.g. Cheaib *et al.*, 2012; Rahn *et al.*, 2018). The trend towards an increasing use of complex computer simulations mirrors similar developments in other scientific fields, for example galaxy formation (Somerville & Davé, 2015), macroevolutionary dynamics (Rangel *et al.*, 2018) or epidemiological disease control (Drake *et al.*, 2015).

For any of these models, precise forecasts and correct estimates of predictive uncertainty are paramount, both for their scientific interpretation (Petchey *et al.*, 2015), and for decision making and governmental actions (Dietze *et al.*, 2018). The IPCC report, for example, uses a combination of different earth system models to simulate future behaviour of the atmosphere, ocean, land surface and fluxes (Bindoff *et al.*, 2013). Using computer simulations for decision making is only sensible, however, if their predictions are sufficiently precise, and if their uncertainties are correctly communicated (Budescu *et al.*, 2009).

Achieving these goals depends on correctly determining model structure, parameters and their uncertainties. Where parameters and model structure cannot be determined directly by measurement or theory, they have to be estimated by comparing model predictions to data (model calibration and selection, e.g. Hartig *et al.*, 2012; Dietze, 2017). In recent years, the field has moved from informal methods for model calibration to established statistical methods such as maximum likelihood estimation (MLE, e.g. Castiglioni *et al.*, 2010) or Bayesian inference (e.g. Harrison *et al.*, 2012; Luke *et al.*, 2017). Superficially, it would seem that parameter calibration and uncertainty propagation in CCS are no different from the statistical regression models familiar to most ecologists, and that no special statistical theory is needed for these models (at least as long as model outputs are approximately deterministic, for stochastic simulation models see Hartig *et al.*, 2011).

In practice, however, there are important differences between calibrating simple statistical models and CCS. One trivial difference is the sheer computational challenge of constraining large models to big data (e.g. Fer *et al.*, 2018). Another, more fundamental disparity arises through the model structure. Compared to statistical models, CCS are characterised by having a higher level of interconnectedness and nonlinearity, as well as multiple variables and outputs. Moreover, CCS typically make a large number of structural assumptions based on prior knowledge (Dormann *et al.*, 2012). As a consequence, they are often less flexible in terms

<sup>1</sup>Theoretical Ecology, University of Regensburg, Universitätsstraße 31, Regensburg 93053, Germany

<sup>2</sup>UK Centre for Ecology & Hydrology, Bush Estate, Penicuik, Midlothian EH260QB, UK

<sup>3</sup>Department of Earth & Environment, Boston University, Boston, MA, USA

\*Correspondence

Johannes Oberpriller, Faculty of Biology and Pre-Clinical Medicine, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany.

Email: johannes.oberpriller@ur.de

of what outputs or patterns can be produced, despite having a large number of parameters (Fatichi *et al.*, 2016).

These traits lead to certain problems when calibrating CCS that are less common in statistical models. A particularly important example is trade-offs when calibrating to multiple data streams. It has been argued that using multiple data streams is desirable because information from different biological levels of organisation (e.g. daily carbon fluxes and yearly inventory data) contains more complementary information than a single data stream (e.g. Grimm, 2005; Medlyn *et al.*, 2015). However, the combination of internal constraints (e.g. mass- or energy balance) with structural error will often make it impossible for a CCS to fit all data streams simultaneously (for a list of examples, see MacBean *et al.*, 2016). Moreover, the information or observation density of data at different organisational levels can differ substantially, leading to unbalanced data (substantial differences in the number of observations of different data streams) for the calibration. This means that the calibration cannot avoid a systematic misfit (bias) in some of the model outputs, and additionally faces a conflict between the information provided by different, possibly unbalanced data streams, both situations that are less common in statistical models.

The goal of this paper was to explore these problems in more detail and provide an overview of strategies for robust statistical inference with CCS. In the remainder of the text, we first explain the problems that may occur when calibrating CCS with structural error, illustrated with the example of a complex forest ecosystem model. Based on our results, we test a range of suggested remedies, and finally provide practical recommendations for using statistical inference with CCS in ecology and evolution.

### WHY DOES MODEL ERROR AFFECT STATISTICS DIFFERENTLY IN COMPLEX COMPUTER SIMULATIONS?

To start our discussion, it will be helpful to further clarify how conventional statistical models differ from CCS. Models exist on a continuum between these two classes (Dormann *et al.*, 2012), but considering the ends of this spectrum, we see clear distinctions between models typically used for statistical data analysis (e.g. GLMMs, see Bolker *et al.*, 2009) and CCS (e.g. Trotsiuk *et al.*, 2020). One key difference is that CCS usually connect a sizeable number of state variables via processes that aim to represent our scientific understanding of the natural system, often with submodels that are calculated at different time steps (e.g. daily, weekly and annual, see as an example the LPJ-GUESS model Smith *et al.*, 2001). It has often been argued that their mechanistic nature makes CCS more appropriate than regression models for forecasting far into the future, because, at least in principle, they should be able to predict into domains for which no previous data exists (e.g. Kearney *et al.*, 2010; Rastetter, 2017; Radchuk *et al.*, 2019).

These benefits of CCS, however, come along with larger structural complexity, which exacerbates challenges in identifying the correct model structure and correcting possible model-data discrepancies (Peng *et al.*, 2011). For example

their typically high interconnectedness hampers the localisation of structural errors. Moreover, while their mechanistic underpinning grants better inclusion of prior knowledge regarding the processes driving system dynamics (Dietze *et al.*, 2013), it can become a liability when mechanisms or parameters are unknown and have to be guessed. A final point is that CCS have to apply certain simplifications and discretisations for computational reasons (e.g. discrete soil layers Tiktak & Bouten, 1992). As a result of these and many more challenges, most CCS display certain structural errors, which are difficult to fix immediately (e.g. Richardson *et al.*, 2012).

These structural errors (including observational bias as part of the statistical model) and their associated uncertainties increase the uncertainties in the calibration process (Bayarri *et al.*, 2007; see also Beven 2005; Trucano *et al.*, 2006). To address this issue, the field has moved towards using formal, statistical methods for model calibration and uncertainty propagation. These methods, however, infer parameters and uncertainties conditional on the assumed model structure being correct. Statistical modellers are usually not overly concerned about these assumptions, because their models flexibly adjust to data, and thus their main concerns are distributional assumptions (e.g. Warton *et al.*, 2015). In CCS, however, this assumption will not hold, and structural errors will interact with the inference, in particular when nonlinearities are large, and when the model is fit to imbalanced data (Abramowitz *et al.*, 2008), i.e. when one data stream has much more observations than another.

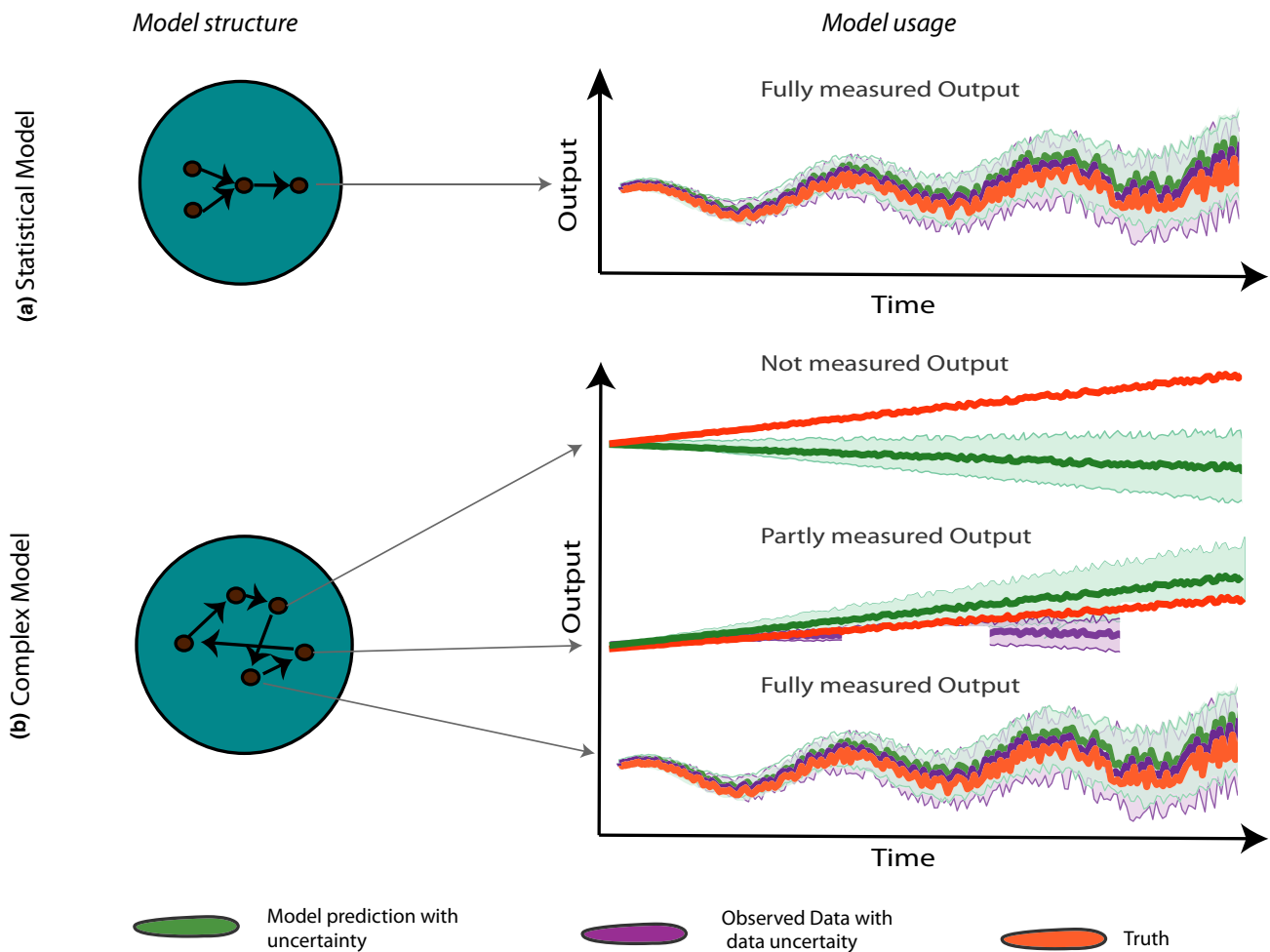
A statistical calibration will respond to this problem by compensating structural error through adjusting parameters to values that differ from the true values of the underlying process (Bell & Schlaepfer, 2016). The resulting model may still display acceptable performance in the domain for which data are available, but parameter estimates may be biased, and their uncertainties may be underestimated. Moreover, when extrapolating beyond the data domain, which is considered an important strength of CCS, biases and underestimation of uncertainty can become substantial (He *et al.*, 2014), especially when the model is calibrated to multiple unbalanced data streams (an example dealing with these issues is Richardson *et al.*, 2010). If the model is not able to fit both data streams at the same time, the calibration algorithm will face a conflict (MacBean *et al.*, 2016). In this situation, the calibration will tend to use parameters adjustments to compensate the error in the more data-rich outputs, at the cost of increased error and too narrow confidence intervals (Sargsyan *et al.*, 2019) particularly in the data-poor model outputs (Fig. 1).

### Case study

To provide a practical example of these problems, we examine the influence of structural model error on predictions, parameters and uncertainty estimation by calibrating the Basic forest model (BASFOR) to multiple balanced or unbalanced data streams.

#### *Model structure and introduced structural error*

BASFOR simulates horizontal homogeneous forest stands by representing three biogeochemical cycles (carbon, nitrogen



**Figure 1** A visualisation of differences between complex computer simulations and statistical models. While statistical models are generally fit to only one response variable, complex computer simulations often predict multiple response variables and thus can be fit to multiple data sources, which may vary in sample size, and can be used to extrapolate to unobserved variables. Moreover, complex computer simulations typically have more variables that are in a more nonlinear and connected dependence structure. From these differences, we hypothesise that 1) biased complex models will lead to biased parameter estimates and wrong predictions, 2) standard calibration underestimates uncertainty and 3) both of these problems increase when calibrating against unbalanced data sets.

and water) as well as soil environment interaction. It is driven by environmental data (atmospheric  $CO_2$  concentration, solar radiation, air temperature, precipitation, wind speed and humidity) and describes the forest stand by 17 state variables (nine tree-related and eight soil-related).

To examine the implications of structural error, we modified several key processes in BASFOR. First, we changed the temperature dependence of NPP allocation (higher optimal temperature, fewer allowed deviations). Second, we made decomposition of litter temperature-dependent. Third, we changed dependence of water runoff to leaf-area-index (exponential quadratic instead of exponential linear). Fourthly, we weighted nitrogen allocation to tree components with their nitrogen use efficiency. Lastly, we made nitrogen leaching root-depth dependent. Although the exact location and nature of these modifications were somewhat arbitrary, we think of those modifications as realistic for structural errors that could also occur in real ecosystem models.

#### Statistical inference

We then used the original BASFOR model (henceforth called the ‘true’ model) to simulate synthetic data with random observation errors (0.2) for daily observations of Gross Primary Production (GPP) and daily (balanced data streams) or 10-day (imbalanced data streams, so called because of an unbalance between the number of observations of GPP and ET) measurements of evapotranspiration (ET). Drivers for the simulation were climate data from 1920 to 2005 from Hyytiälä, Finland (Reyer *et al.*, 2020).

Prior to the calibration, we conducted a sensitivity analysis of BASFOR. Based on the results, we removed insensitive parameters and three parameters that showed very high trade-offs with other parameters from the calibration by fixing them to their true values (the goal of this procedure is to speed up MCMC computations; see, e.g. Minunno *et al.*, 2013). Because the true parameter values were known, no model error was introduced by this procedure, and the validity of

our further results is thus not affected by the parameter screening. In a real application, where “true” parameter values would be unknown, this procedure could introduce additional model error, which would further motivate the need to find methods to compensate for model error, such as the ones we present in this study.

We applied Bayesian inference (e.g. van Oijen *et al.*, 2011) to infer the values and uncertainties of the remaining six model parameters and the two standard deviation parameters of the observation model from the synthetic data. We specified flat (uniform) priors on the model parameters and vague gamma priors for the standard deviation parameters. We estimated posteriors with the Differential-Evolution Markov-Chain Monte-Carlo (ter Braak & Vrugt, 2008) algorithm, implemented in the R package BayesianTools, Hartig *et al.* (2019). To speed up computations, we generated initial values and the Z matrix with a differential evolution optimiser (DEoptim, Ardia *et al.*, 2016). We applied this procedure to both the ‘true’ model and the model with structural error.

*Quantification of the error in inference*

To assess the effect of model error on the inference, we calculated the average error of parameter estimates by averaging the percentage difference between the ‘true’ parameter ( $p^*$ ) and the calibrated parameter over the posterior, averaged over  $N = 10000$  samples from the posterior, the different parameters ( $P$ ) and the five replicates ( $M$ ).

$$\text{Parameter error} = \frac{1}{P} \sum_i^P \left| \frac{1}{M} \sum_j^M \frac{1}{N} \sum_k^N \frac{p_{i,j,k} - p_i^*}{p_i^*} \right| \quad (1)$$

Moreover, to assess the error of model predictions (also called time-series error), we calculated the mean absolute error of data  $d_i$  and model prediction  $m_i(x, \theta_j)$  (driven with climatic drivers  $x$  and parameters  $\theta_j$ ) averaged over time ( $T$ ), the

posterior distribution (through  $N = 120$  samples from the posterior) and five calibration replicates ( $M$ ).

$$\text{Error} = \frac{1}{M} \sum_j^M \frac{1}{N} \sum_k^N \frac{1}{T} \sum_i^T |d_i - m_i(x, \theta_{j,k})| \quad (2)$$

Note that in most cases with structural model error, the error in the parameters and predictions was systematic, meaning that it can be interpreted as bias.

To relate the error to the estimated uncertainties, and thus examine if uncertainty estimates were reliable, we calculated error scaled to estimated uncertainty (ESEU) by dividing the mean error per day by the posterior standard deviation  $\sigma_i(m_i(x, \theta_j))$ , averaged over time, the posterior distribution and the five replicates.

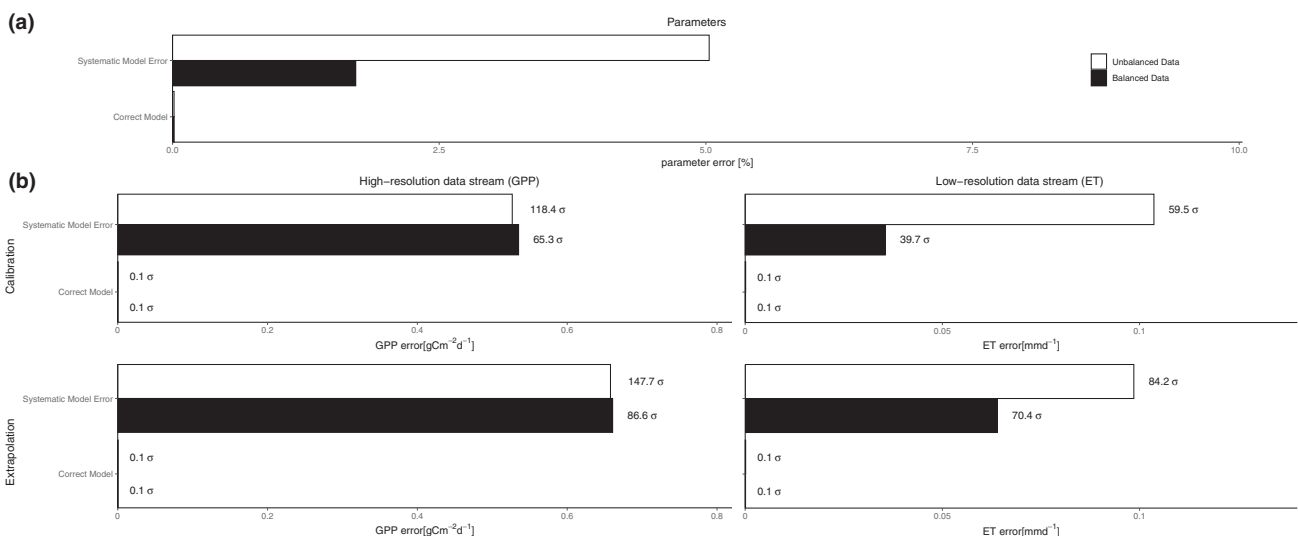
$$\text{ESEU} = \frac{1}{T} \sum_i^T \frac{|d_i - \frac{1}{M} \sum_j^M \frac{1}{N} \sum_k^N m_i(x, \theta_j)|}{\sigma_i(m_i(x, \theta_j))} \quad (3)$$

A mean absolute error the same magnitude as the estimated uncertainty (standard deviation) will result in an ESEU of 1. Values substantially larger than one suggest that the estimation or prediction error is larger than the estimated uncertainty. For the model outputs and uncertainties, we differentiated between calibration and extrapolation domain.

*Comparison between calibrating a ‘true’ model and a model with structural error*

The results of the calibration with the ‘true’ model (without structural error) show that the error of the inferred parameters was virtually zero (<0.02%) for balanced and unbalanced data sets (Fig. 2a). In both of these cases, extrapolation and calibration error were small with narrow uncertainties (ESEU = 0.1) (Fig. 2b).

For the model with structural error, inferential errors were much larger (Fig. 2a). In particular, the parameter error was



**Figure 2** Performance of the model with and without structural model error for balanced and unbalanced data. The bars reflect error in absolute values and numbers reflect the error scaled to estimated uncertainty (i.e. the error of the model which can be explained due to a high estimated uncertainty). The case study indicates that structural model bias leads to (a) parameters with serious errors, (b) erroneous model outcomes and high error scaled to estimated uncertainty.

three times larger for the unbalanced data (*c.* 5%) compared to the balanced data (*c.* 1.7%) (Fig. 2a). Higher parameter error for the model with structural error led in all cases to higher time-series errors compared to the correct model (Fig. 2b). For the balanced data set, the error for calibration was smaller than for extrapolation, whereas for the unbalanced data set this only was true for the high-resolution data (GPP). Moreover, GPP error was slightly smaller for the unbalanced than the balanced data set, but ET error otherwise. These errors led to a very high ESEU (Fig. 2b). This effect was stronger for the unbalanced data, especially for the undersampled data (ET) in the calibration domain (Fig. 2b).

These results support our theoretical expectations that calibrating with a correct structural model leads to unbiased parameter estimates, correct predictions and reliable uncertainty estimates, regardless whether data streams are balanced or unbalanced. Introducing structural model error, however, led to erroneous parameter estimations (Fig. 2a), caused erroneous time-series predictions and high ESEU (Fig. 2b), and these effects are intensified by unbalanced data sets (Fig. 2a and b).

## A TOOLBOX FOR STATISTICAL INFERENCE IN COMPLEX COMPUTER SIMULATIONS

After having confirmed our intuition that statistical calibrations of CSS are highly susceptible to structural error, we turn our attention to possible solutions. Few general treatments of the problem exist in literature, but there are certain strategies and suggestions that are frequently used in practice. To deal with the problem of imbalanced data, many studies rebalance or reweight data streams. The remaining model-data discrepancies (bias) have sometimes been addressed by introducing data-driven models to the process-model after or during the calibration. In the following, we will discuss these potential solutions and test their applicability in our case study.

## Weighting of data streams

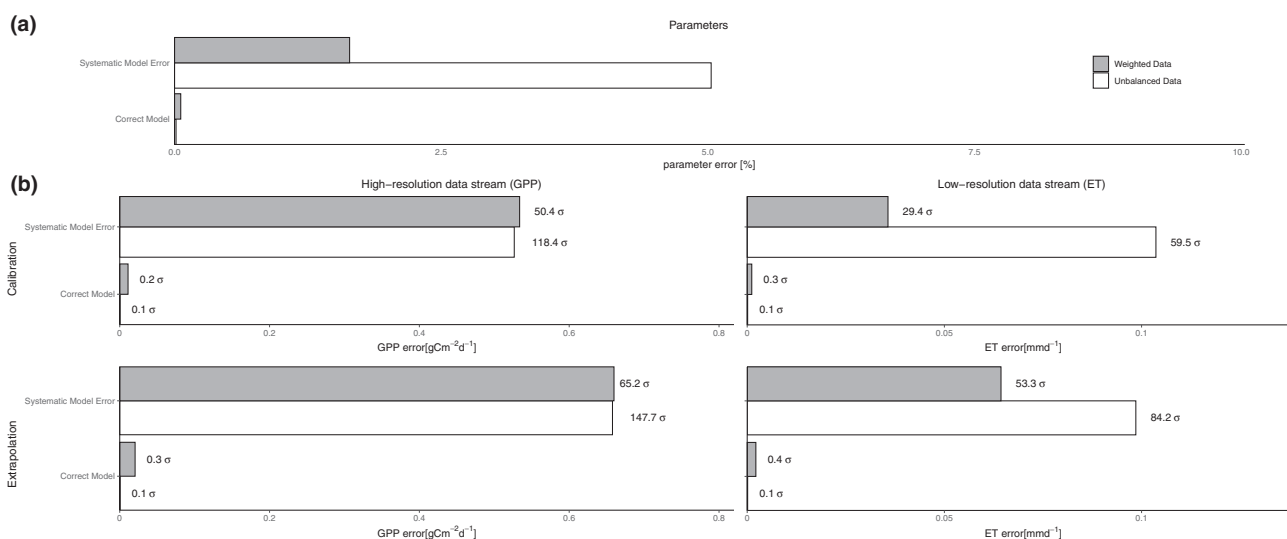
The strategy of rebalancing and reweighting data addresses the issue that standard statistical methods weight, the importance of each data stream principally by its content of independent observations. While the latter is perfectly sensible for a correct model, it will lead to distortions towards the model output with more data when structural error makes it impossible to fit both data streams at the same time.

### Case study – weighting of data streams

To examine the possible benefits of weighting for our case study, we down-weighted the likelihood for the GPP data with 1/10, the ratio of ET to GPP observations, thus giving both data streams the same weight (for details, see Supporting Information S1, section 1.1). Weighting the data streams increased the error for the estimated parameters of the correct model (Fig. 3a) by a small amount, which propagates through the model into a small error in predictions and a higher ESEU (Fig. 3b). For the model with structural error, introducing weights in the likelihood decreased parameter error leading to smaller ET error, but slightly increased GPP error (Fig. 3b). Moreover, the ESEU of ET in the calibration domain is smaller due to a reduction of ET error. Overall, we can thus conclude that weighting slightly decreased the inferential performance for the correct model, but dramatically improved the performance for the model with structural error.

### Bias correction after calibration

Another option to deal with model error is statistical bias correction. The simplest approach is to fit flexible statistical or machine learning models post hoc (*i.e.* after the CCS has been calibrated) to the residual errors (but see Beyer *et al.*, 2019). The logic here is that if the model makes the same error under similar conditions (called ‘time invariance’ by Ehret *et al.*,



**Figure 3** Comparison of the performance of the model with structural model error and the correct model for weighted and unbalanced data. The bars reflect error in absolute values and numbers reflect the error scaled to estimated uncertainty. The case study indicates that weighting the data streams decreases a) parameters error, b) shifts error in model outcomes and improves ESEU



2012), we can learn this error and apply corrections to future predictions. Obviously, this method only corrects predictions and not the parameter estimates, as the actual inference remains unchanged.

#### Case study – bias correction after calibration

To test this method, we used a flexible Gaussian process (GP) model from the kernlab package (Karatzoglou et al., 2004) with a distance-based covariance structure (for details see Supporting Information 1, section 1.2). We fitted the model to approximately 6 years of residual errors as a response, and the corresponding model drivers (e.g. temperature and humidity) and CCS output as predictors, and extrapolated the error to future predictions. Our results show that this approach decreased the predictive GPP error of the model with structural error by similar amounts in the calibration and extrapolation periods (Fig. 4). ET error was approximately the same between the corrected and uncorrected versions of the model with a structural error, but there was a large decrease in ESEU (Fig. 4), not only caused by reduced error, but mostly by the variance coming from the explicitly modelled model error. Applying the same method to the true model introduced a slightly larger error in the time series and increased ESEU (Fig. 4). We speculate that this is due to the GP overfitting on random error.

#### Bias correction during calibration

A second option is to perform the bias correction within the calibration. A common example of this is the Kennedy-O'Hagan (KOH) approach (Kennedy & O'Hagan, 2001). In this approach, we fit again a GP for the bias together with the other model parameters in the same likelihood:

$$L(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[d - (m(\theta, x) + GP(x, m))]^2\right\} \quad (4)$$

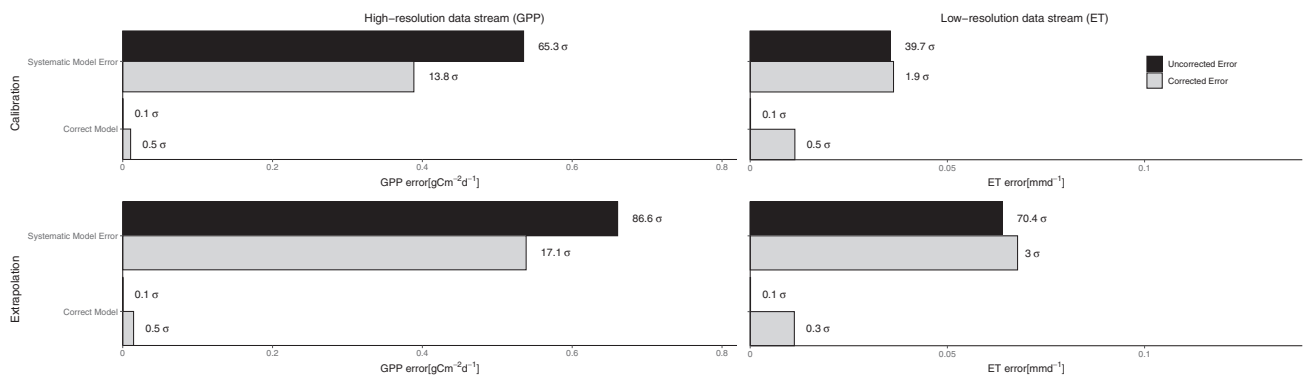
Here,  $\sigma$  is the standard deviation of the observational error,  $GP$  a Gaussian process. While the advantage of this approach is that the bias correction can also improve the inference on the model's parameters, the drawback is that it may suffer from an identifiability issue between parameters and model

error. Whether this problem occurs depends on how distinct the structure of the process and the error model are. Note also that multiple data streams can be helpful in this regard, because they would typically impose independent constraints on the process model. Moreover, it has been shown, that incorporating suitable prior knowledge about the model error (e.g. smooth with respect to some predictor variables) allows the KOH method to separate between parameters and model error (Brynjarsdóttir & O'Hagan, 2014). Because of these attractive properties, there are a sizeable number of studies which have tested and modified this approach (e.g. Higdon et al., 2004, 2008; Goldstein & Rougier, 2009; Tuo & Wu, 2016; Tuo, 2017).

#### Case study – bias correction during calibration

In its original version, the KOH method fits the  $GP$  against all calibration data with all drivers and state variables as predictors. However, as the computational cost of GP fitting and evaluation scale unfavourable with the number of data points, this makes it more difficult for typical environmental model calibrations. The computational problems occur because the calculation of the  $GP$  requires an inversion of a large covariance matrix. Moreover, the KOH method assumes having enough observational data of model determining variables (model state and external drivers) to fully constrain the Gaussian process (Kennedy & O'Hagan, 2001), which for typical ecological models is not a realistic assumption (in our case study, we do not have virtual measurements of any state variables, we measured only the fluxes GPP and ET).

For our case study, we propose an alternative variant of the KOH method, which makes three changes to decrease computational cost. First, we only use the drivers and the observed values as predictors. Secondly, we calibrate against a subsample of data (in our case we subsample to 10% of the data, the last 8 years of data and drivers as best proxies for future drivers). We do so because, typically models systematically predict GPP that is too small on warm summer days and ET that is too high when humidity is low. Thirdly, we avoid the costly inversion of the covariance matrix that is only needed to match GP parameters to their prior by approximating the inverse covariance by its diagonal, while still inferring the full



**Figure 4** Comparison of the performance of the model with structural error and the correct model fitting a correction term after calibration. The bars reflect error in absolute values and numbers reflect the error scaled to estimated uncertainty. The case study indicates that correcting the data streams decreases error in model outcomes and decreases ESEU.

covariance matrix (rbfdot kernel) in the likelihood. To code a preference for explaining the data by the process-model, we apply a regularising  $\gamma(2,0.1)$  prior with a high probability weight near zero on the diagonal. Based on the GP predictions, we calculate model-data discrepancies for the rest of the time series (a detailed tutorial is given in Supporting Information S1, section 1.2).

When applying bias correction during calibration, parameter error stayed near zero for the correct model, and decreased for the model with structural error (Fig. 5a). However, whereas time series error decreased in both outputs, for the model with structural error, for the true model, error increased (Fig. 5b), with an almost identical pattern to the post hoc GP (Fig. 5b). For the model with structural error, the calibration resulted in higher estimated uncertainty and thus lower ESEU compared to a calibration without an explicit model error term (Fig. 5b). Overall, the method improves parameters, predictions and ESEU for the model with structural error, but decreases the performance for the correct model.

### Correcting processes rather than outputs

We have seen so far that correcting bias on the model outputs can improve predictions and inference. The true error, however, is not on the outputs, but in the model processes themselves. It, therefore, seems obvious to explore if the processes themselves could be bias-corrected. For simple population models, this idea has been suggested under the name ‘partially specified ecological models’ (Wood, 2001). The drawback of this approach for CCS is that the complexity of the error term and therefore the issue of identifiability increases significantly if errors in all possible subprocesses are considered. For our case study, we attempted to correct process-errors directly via a state-space approach (details see supporting information S1, section 1.3), but did not succeed in improving the statistical

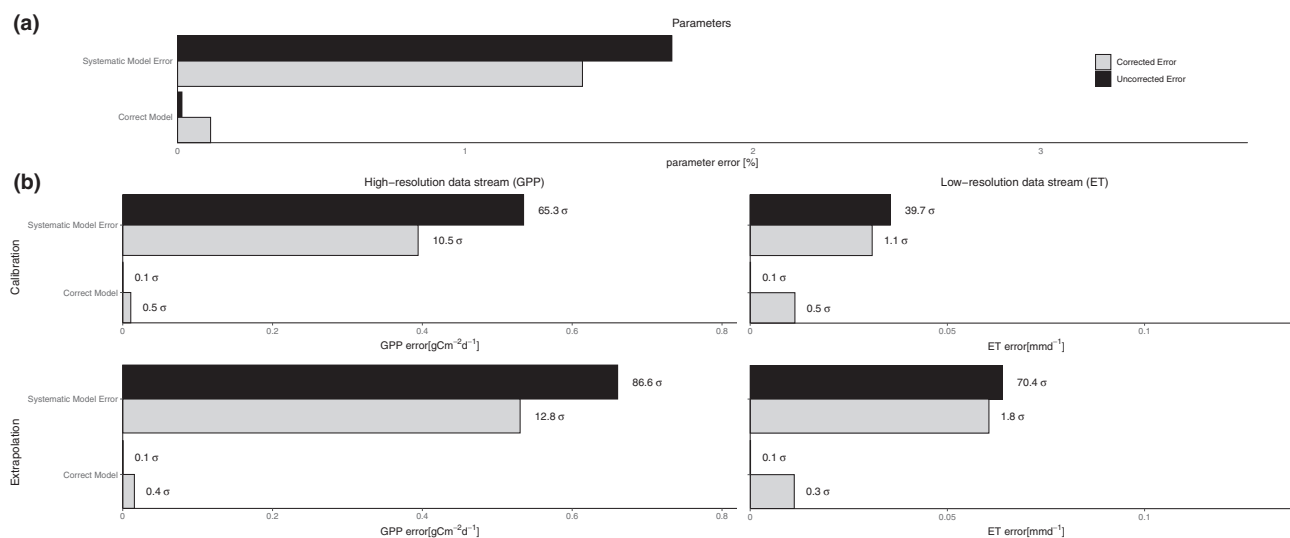
inference in this way. Nevertheless, we believe that this is worthwhile for further research, in particular because it would not only correct errors, but also allow to identify their location.

### DISCUSSION

CCS are increasingly used in ecology, evolution and earth system sciences. Our ability to confront these models with data and to estimate uncertainties in parameters and predictions is critical for their utility.

In this contribution, we highlighted that certain issues emerge when using standard statistical methods to calibrate CCS. Most importantly, our theoretical explanations as well as our case study demonstrated that naive applications of standard calibration methods to imperfect computer simulations can lead to biased parameter estimates and predictions, and to underestimated uncertainties (Fig. 2), and that these biases are more pronounced than in flexible statistical models. These issues are particularly severe when calibrating against unbalanced data (Fig. 2). Weighting of data streams can reduce the aggravating effect of unbalanced data (Fig. 3). Data-driven models can be used to describe and remove the remaining bias after or during the calibration. In our case study, fitting model bias with a GP after calibration improved time series predictions (Fig. 4). Thus, our results show that robust methods exist for ameliorating negative consequences of structural model for making predictions with calibrated CCS.

Using a GP during calibration can additionally improve parameter inference (Fig. 5). We acknowledge that the interpretation of parameter values across structurally different models is tricky, because those parameters have different meanings in the respective models, and thus, one could argue that both the true and the model with structural error have parameters that are correct under their respective



**Figure 5** Comparison of the performance of the model with structural model error and the correct model for a correction during calibration. The bars reflect error in absolute values and numbers reflect the uncertainty in units of standard deviation. The case study indicates that correcting error during the calibration decreases (a) parameters error for the wrong model, (b) reduces error in the model outcomes and improves uncertainty estimation.

assumptions. This view, however, neglects, that researchers will tend to interpret parameter values as if their models were structurally unbiased, and representation of the true process. Our comparison of the estimated and the true parameter, therefore, measures to what extent this interpretation is justified and shows that explicitly modelling structural error increases the chances of model parameters representing their real values (Goldstein & Rougier, 2009).

Our results regarding the consequences of model error are qualitatively supported by the few earlier studies that have looked at the problem (for balanced data by White *et al.*, 2014 and for unbalanced data sets by Abramowitz *et al.*, 2008). In general, however, this topic seems surprisingly underappreciated in the statistical literature. We speculate that most statisticians do not operate with large system models, and the modellers that do are not primarily interested in statistical methods. Nevertheless, a good understanding of these issues is urgently needed, as many important forecasts rely on the correct identification of parameters and their uncertainties. In the next subsections, we summarise our conclusions from existing literature and our new simulations, provide practical guidance for their use, and delineate a statistical research program to develop a theory of robust inference for CCS.

#### **Which methods work to improve inference for biased system models?**

To achieve a more balanced impact of the different data on the calibration, many modelling studies weigh data streams. Despite its popularity, few studies have examined the justification for this practice. Contrary to Wutzler & Carvalhais (2014), who only found minor improvements, we found that weighting improved all considered performance measures (Fig. 3). Different CCS and a different severity of model error may explain the differences in the two studies. In general, benefits from weighting likely depend on the statistical context, the weighting strategy, and the model error. Overall, however, we believe that weighting is a useful and conservative strategy if structure model error is suspected. One open question that would profit from more research is how the weighting of different data streams should be performed. Creating balance by upweighting the less abundant data stream, which essentially corresponds to the common practice of oversampling in machine learning, could lead to a serious underestimation of uncertainties as it is equivalent to using the same data multiple times. Downweighting, the far more common approach in studies calibrating CCS, is more conservative, but it also artificially decreases the information in the more abundant data stream to the level of the less abundant stream, which can hardly be optimal to get realistic uncertainties. In general, these two options represent the extremes of a broad spectrum of possibilities, and more research is required to understand how an optimal weighting could be justified. An option to avoid the problem would be to calibrate against patterns, as suggested by the POM (Grimm, 2005), to independently update subsets of parameters against different data streams (Wutzler & Carvalhais, 2014), or to set up subjective likelihoods (White *et al.*, 2014), as in the GLUE approach (Beven & Binley, 1992). The downside, however, is that these

approaches could be considered even more subjective than weights on the data streams.

A complementary class of methods directly addresses the issue of model error, by identifying and correcting structural biases from model's predictions. In our case study, this approach (via the KOH method) improved parameters, predictions and uncertainty quantification (in line with Brynjarsdóttir & O'Hagan 2014). However, the standard KOH method has two main challenges – high computational complexity (Conti & O'Hagan, 2010) and possible identifiability issues between model parameters and model error (Brynjarsdóttir & O'Hagan, 2014). We addressed the first problem by only using a fraction of the available data to fit the GP and extrapolated to the remaining calibration domain. We speculate subsampling works for models with mechanistic structure, as long as the learned discrepancy will behave similarly in the future. We appreciate that using a fraction of the calibration data potentially disregards useful information, and that our additional numerical approximations could further reduce the method's performance. The fact that we reduced the model error, however, suggests that these problems are probably mild. Still, in situations where computational costs are not limiting, it would be better to use the original method suggested in Kennedy & O'Hagan (2001). The issue of identifiability is important, but arises in many statistical situations, and several strategies exist to deal with it, for example regularisation or informative priors (Brynjarsdóttir & O'Hagan, 2014). Thus, we think these methods can lead to better predictions for ecological CCS and modellers should be using them.

A limitation of our case study is that it tested validity and effectiveness for one specific model, with one specific error structure. While we do think that the chosen example is typical and representative for the field, it would be useful to explore the generality of our results in future studies and their robustness to observation errors and uncertainties, which can be expected to exacerbate statistical problems.

Finally, all our successful examples used bias corrections on model outputs. In particular, when making predictions, these implicitly assume that the model error is stationary, which is unlikely to be true (Chen *et al.*, 2015). It would therefore be preferable to move bias corrections directly inside the modelled processes. In our case study, we attempted such a correction with a state-space approach, but could not achieve an increase in inferential performance. It is possible that idiosyncrasies of our setup were responsible for this negative result, but it seems equally plausible that corrections on the outputs are already at the limit of what can be sensibly inferred from data. Either way, these considerations suggest that bias corrections are currently no panacea, and that careful improvements of the model structure, if possible, are still the preferable solution.

#### **Practical suggestions**

As famously noted by Box (1976): 'All models are wrong, but some are useful'. Accepting this fact, the question for CCS is what type of error is dominant. If statistical error dominates the structural error (this can be checked by an analysis of residuals, see Supporting information section 2), all standard



Cases	Statistical error dominates, balanced or unbalanced data	Structural error dominates, balanced data	Structural error dominates, unbalanced data
Naive use of standard methods	Standard methods are sufficient for parameters, predictions and uncertainty quantification	Standard methods lead to biased predictions and parameters.	Standard methods lead to a higher bias in parameters, predictions and uncertainty estimation
Our recommendations	1. Standard methods are sufficient	1. Bias correction after calibration improves predictions 2. Bias correction during calibration additionally improves parameters	1. Weighting reduces bias by a lot 2. Bias correction after calibration improves predictions 3. Bias correction during calibration additionally improves parameters
Remarks	-	Bias correction has high computational costs	Bias correction has high computational costs

**Figure 6** The different situations in environmental model calibration and our suggestions for improving model performance. The two main factors, which need to be taken into account are the data situation (balanced or unbalanced) and the sources of error (random or structural). This general advice can slightly change in different situations as model complexity and computational demand strongly depend on the CCS, domain of extrapolation and number of data streams. Overall performance will become worse with increasing observational error for all methods including standard calibration.

statistical techniques work fine, regardless of the balance of data. In this case, using methods that accounting for possible structural model errors tends to somewhat increase uncertainties (Fig. 4 and 5, see recommendations Fig. 6). When structural error dominates, however, severe statistical problems can arise, in particular for imbalanced data. In this case, weighting of data streams or adding bias correction to the CCS can improve the outcomes of a model calibration dramatically. Our recommendation for modellers with little statistical background is that downweighting imbalanced data is a simple, conservative approach that can alleviate some of issues created by structural error. Although it is somewhat ad-hoc, it improved results in our case study, and it makes uncertainty estimates (e.g. confidence intervals) more conservative. For more experienced modellers, we propose to consider additional bias corrections after or during calibration, or even consider if bias corrections can be moved inside the processes, which would not only improve the inference, but also model understanding. For all these purposes we provide sample code (<https://github.com/JohannesOberpriller/Oberpriller-et-al-2021>).

#### Towards a statistical theory for robust inference in complex computer simulations

More broadly, our paper highlights that structural model error raises specific problems for statistical inference with complex computer simulations. This should alert the ecological community that model error is a real problem for the calibration of CCS, and naively applying standard statistical methodologies does not always lead to the desired results.

Although we did a step into the direction of robust inference in CCS by reviewing proposed solutions, explaining their theoretical justification and providing practical guidance for their application, further work is required to arrive at a general solution for robust statistical inference. For example, we

have no good theory about how to set weights for different data streams. When considering a data stream with only one observation, it becomes clear that downweighting to the least common data stream is likely not always optimal. Moreover, it would be interesting to extend bias corrections also to methods that use simulation-based inference, such as Approximate Bayesian Computing (ABC) or synthetic likelihood (Csilléry *et al.*, 2010; Hartig *et al.*, 2011).

A last point is that statistical bias corrections are important for improving the inference, but the correct model still consistently performed best in our case study, and we should thus also think about how to develop methods to track down the location of the error. To localise errors, one could start by analysing model discrepancies for patterns, and use those to attempt a rough localisation of the structural error. Moreover, we speculate that when a dramatic change of a parameter value between KOH and standard calibration happens, this gives a hint that model error affects this specific parameter and thus that model error is ‘near’ to this parameter. Then using time-dependent parameters (instead of constant) (Reichert & Mieleitner, 2009) could be an option to get a better localisation of the error. Another idea (Wood, 2001) goes a step further, by saying that flexible models (generalised additive models) should account for the processes, or by (Reichstein *et al.*, 2019), which propose to learn entire submodels. These approaches should be tested in practice to finally improve model performance.

#### AUTHORSHIP

FH and JO conceived and designed the study. JO implemented the case studies, ran the experiments and analysed the results. DC advised regarding implementing errors in the BASFOR model. All authors contributed equally to discussing and interpreting the results, and to the preparation of the manuscript.

## ACKNOWLEDGEMENTS

The idea for this work originated from discussions within COST Action FP1304 PROFOUND. We thank Maximilian Pichler and Lukas Heiland as well as three anonymous reviewers for their valuable comments and suggestions. JO was funded by the Bavarian Ministry of Science and the Arts in the context of Bavarian Climate Research Network (bayklif). Open Access funding enabled and organized by Projekt DEAL.

## PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/ele.13728>.

## DATA AVAILABILITY STATEMENT

Code and examples are available under <https://github.com/JohannesOberpriller/Oberpriller-et-al-2021>. <https://zenodo.org/badge/latestdoi/272397284>

## REFERENCES

- Abramowitz, G., Leuning, R., Clark, M. & Pitman, A. (2008). Evaluating the performance of land surface models. *J. Clim.*, 21, 5468–5481.
- Ardia, D., Mullen, K.M., Peterson, B.G. & Ulrich, J. (2016). Differential evolution instead of Dif593 ferential evolution.
- Bayarri, M.J., Berger, J.O., Paulo, R., Sacks, J., Cafeo, J.A., Cavendish, J. et al. (2007). A framework for validation of computer models. *Technometrics*, 49, 138–154.
- Bell, D.M. & Schlaepfer, D.R. (2016). On the dangers of model complexity without ecological justification in species distribution modeling. *Ecol. Model.*, 330, 50–59.
- Beven, K. (2005). On the concept of model structural error. *Water Sci. Technol.*, 52, 167–175.
- Beven, K. & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Process.*, 6, 279–298.
- Beyer, R., Krapp, M. & Manica, A. (2020). An empirical evaluation of bias correction methods for palaeoclimate simulations. *Clim. Past.*, 16, 1493–1508. <https://doi.org/10.5194/cp-16-1493-2020>
- Bindoff, N.L., Stott, P.A., AchutaRao, K.M., Allen, M.R., Gillett, N., Gutzler, D., Hansingo, K., Hegerl, G., Hu, Y., Jain, S., Mokhov, I.I., Overland, J., Perlwitz, J., Sebbari, R. & Zhang, X. (2013). Detection and Attribution of Climate Change: From Global to Regional. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V. & Midgley, P.M.) Cambridge University Press. Cambridge, UK and New York, NY, USA, pp. 867–952.
- Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. et al. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends Ecol. Evol.*, 24, 127–135.
- Box, G.E.P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71, 791–799.
- ter Braak, C.J.F. & Vrugt, J.A. (2008). Differential Evolution Markov Chain with snooker updater and fewer chains. *Statistics and Computing*, 18, 435–446.
- Briscoe, N.J., Elith, J., Salguero-Gómez, R., Lahoz-Monfort, J.J., Camac, J.S., Giljohann, K.M. et al. (2019). Forecasting species range dynamics with process-explicit models: Matching methods to applications. *Ecol. Lett.*, 22, 1940–1956.
- Brynjarsdóttir, J. & O'Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, 30, 114007.
- Budescu, D.V., Broomell, S. & Por, H.H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science*, 20, 299–308.
- Castiglioni, S., Lombardi, L., Toth, E., Castellarin, A. & Montanari, A. (2010). Calibration of rainfall-runoff models in ungauged basins: A regional maximum likelihood approach. *Advances in Water Resources*, 33, 1235–1242.
- Cheabib, A., Badeau, V., Boe, J., Chuine, I., Delire, C., Dufrière, E. et al. (2012). Climate change impacts on tree ranges: Model intercomparison facilitates understanding and quantification of uncertainty. *Ecol. Lett.*, 15, 533–544.
- Chen, J., Brissette, F. & Lucas-Picher, P. (2015). Assessing the limits of bias correcting climate model outputs for climate change impact studies. *J. Geophys. Res. Atmos.*, 120, 1123–1136.
- Conti, S. & O'Hagan, A. (2010). Bayesian emulation of complex multi-output and dynamic computer models. *Journal of Statistical Planning and Inference*, 140, 640–651.
- Csilléry, K., Blum, M.G.B., Gaggiotti, O.E. & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends Ecol. Evol.*, 25, 410–418.
- Dietze, M.C. (2017). *Ecological Forecasting*. Princeton University Press, Princeton.
- Dietze, M.C., Fox, A., Beck-Johnson, L.M., Betancourt, J.L., Hooten, M.B., Jarnevich, C.S. et al. (2018). Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proc. Natl Acad. Sci.*, 115, 1424–1432.
- Dietze, M.C., LeBauer, D.S. & Kooper, R. (2013). On improving the communication between models and data. *Plant, Cell Environ.*, 36, 1575–1585.
- Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C., Hartig, F. et al. (2012). Correlation and process in species distribution models: Bridging a dichotomy. *J. Biogeogr.*, 39, 2119–2131.
- Drake, J.M., Kaul, R.B., Alexander, L.W., O'Regan, S.M., Kramer, A.M., Pulliam, J.T. et al. (2015). Ebola cases and health system demand in Liberia. *PLoS Biol.*, 13, e1002056.
- Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K. & Liebert, J. (2012). Should we apply bias correction to global and regional climate model data? *Hydrology and Earth System Sciences Discussions*, 9, 5355–5387.
- Evans, M.R., Norris, K.J. & Benton, T.G. (2012). Predictive ecology: Systems approaches. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367, 163–169.
- Fatichi, S., Vivoni, E.R., Ogden, F.L., Ivanov, V.Y., Mirus, B., Gochis, D. et al. (2016). An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. *J. Hydrol.*, 537, 45–60.
- Fer, I., Kelly, R., Moorcroft, P.R., Richardson, A.D., Cowdery, E.M. & Dietze, M.C. (2018). Linking big models to big data: Efficient ecosystem model calibration through Bayesian model emulation. *Biogeosciences*, 15, 5801–5830.
- Goldstein, M. & Rougier, J. (2009). Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference*, 139, 1221–1239.
- Grimm, V. (2005). Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science*, 310, 987–991.
- Harrison, K.W., Kumar, S.V., Peters-Lidard, C.D. & Santanello, J.A. (2012). Quantifying the change in soil moisture modeling uncertainty from remote sensing observations using Bayesian inference techniques. *Water Resour. Res.*, 48. <https://doi.org/10.1029/2012wr012337>
- Hartig, F., Calabrese, J.M., Reineking, B., Wiegand, T. & Huth, A. (2011). Statistical inference for stochastic simulation models – theory and application. *Ecol. Lett.*, 14, 816–827.
- Hartig, F., Dyke, J., Hickler, T., Higgins, S.I., O'Hara, R.B., Scheiter, S. et al. (2012). Connecting dynamic vegetation models to data – an inverse perspective. *J. Biogeogr.*, 39, 2240–2252.
- Hartig, F., Minunno, F. & Stefan, P. (2019). BayesianTools: General-Purpose MCMC and SMC Samplers and Tools for Bayesian Statistics.

- He, Y., Yang, J., Zhuang, Q., McGuire, A.D., Zhu, Q., Liu, Y. *et al.* (2014). Uncertainty in the fate of soil organic carbon: A comparison of three conceptually different decomposition models at a larch plantation. *J. Geophys. Res. Biogeosci.*, 119, 1892–1905.
- Higdon, D., Gattiker, J., Williams, B. & Rightley, M. (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103, 570–583.
- Higdon, D., Kennedy, M., Cavendish, J.C., Cafeo, J.A. & Ryne, R.D. (2004). Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26, 448–466.
- Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. (2004). Kernlab - An S4 Package for Kernel Methods in R. *J. Stat. Softw.*, 11, 1–20.
- Kearney, M.R., Wintle, B.A. & Porter, W.P. (2010). Correlative and mechanistic models of species distribution provide congruent forecasts under climate change. *Conserv. Lett.*, 3, 203–213.
- Kennedy, M.C. & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 425–464.
- Levin, S.A. (1998). Ecosystems and the biosphere as complex adaptive systems. *Ecosystems*, 1, 431–436.
- Luke, A., Vrugt, J.A., AghaKouchak, A., Matthew, R. & Sanders, B.F. (2017). Predicting nonstationary flood frequencies: Evidence supports an updated stationarity thesis in the United States. *Water Resour. Res.*, 53, 5469–5494.
- MacBean, N., Peylin, P., Chevallier, F., Scholze, M. & Schürmann, G. (2016). Consistent assimilation of multiple data streams in a carbon cycle dataassimilation system. *Geoscientific Model Development*, 9, 3569–3588.
- Medlyn, B.E., Zaehle, S., De Kauwe, M.G., Walker, A.P., Dietze, M.C., Hanson, P.J. *et al.* (2015). Using ecosystem experiments to improve vegetation models. *Nat. Clim. Chang.*, 5, 528–534.
- Minunno, F., van Oijen, M., Cameron, D.R. & Pereira, J.S. (2013). Selecting parameters for bayesian calibration of a process-based model: A methodology based on canonical correlation analysis. *SIAM/ASA Journal on Uncertainty Quantification*, 1, 370–385.
- van Oijen, M., Cameron, D.R., Butterbach-Bahl, K., Farahbakhshazad, N., Jansson, P.E., Kiese, R. *et al.* (2011). A Bayesian framework for model calibration, comparison and analysis: Application to four models for the biogeochemistry of a Norway spruce forest. *Agric. For. Meteorol.*, 151, 1609–1621.
- Peng, C., Guiot, J., Wu, H., Jiang, H. & Luo, Y. (2011). Integrating models with data in ecology and palaeoecology: Advances towards a model– data fusion approach. *Ecol. Lett.*, 14, 522–536.
- Petchey, O.L., Pontarp, M., Massie, T.M., Kéfi, S., Ozgul, A., Weilenmann, M. *et al.* (2015). The ecological forecast horizon, and examples of its uses and determinants. *Ecol. Lett.*, 18, 597–611.
- Radchuk, V., Kramer-Schadt, S. & Grimm, V. (2019). Transferability of mechanistic ecological models is about emergence. *Trends Ecol. Evol.*, 34, 487–488.
- Rahn, E., Vaast, P., Läderach, P., van Asten, P., Jassogne, L. & Ghazoul, J. (2018). Exploring adaptation strategies of coffee production to climate change using a process-based model. *Ecol. Model.*, 371, 76–89.
- Rangel, T.F., Edwards, N.R., Holden, P.B., Diniz-Filho, J.A.F., Gosling, W.D., Coelho, M.T.P. *et al.* (2018). Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves. *Science*, 361, eaar5452.
- Rastetter, E.B. (2017). Modeling for understanding v. *Modeling for Numbers. Ecosystems*, 20, 215–221.
- Reichert, P. & Mieleitner, J. (2009). Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. *Water Resour. Res.*, 45. <https://doi.org/10.1029/2009wr007814>
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. *et al.* (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566, 195–204.
- Reyer, C.P.O., Silveyra Gonzalez, R., Dolos, K., Hartig, F., Hauf, Y., Noack, M. *et al.* (2020). The PROFOUND Database for evaluating vegetation models and simulating climate impacts on European forests. *Earth System Science Data*, 12, 1295–1320.
- Richardson, A.D., Anderson, R.S., Arain, M.A., Barr, A.G., Bohrer, G., Chen, G. *et al.* (2012). Terrestrial biosphere models need better representation of vegetation phenology: Results from the North American Carbon Program Site Synthesis. *Glob. Change Biol.*, 18, 566–584.
- Richardson, A.D., Williams, M., Hollinger, D.Y., Moore, D.J.P., Dail, D.B., Davidson, E.A. *et al.* (2010). Estimating parameters of a forest ecosystem C model with measurements of stocks and fluxes as joint constraints. *Oecologia*, 164, 25–40.
- Sargsyan, K., Huan, X. & Najm, H. (2019). Embedded model error representation for bayesian model calibration. *International Journal for Uncertainty Quantification*, 9, 365–394.
- Smith, B., Prentice, I.C. & Sykes, M.T. (2001). Representation of vegetation dynamics in the modelling of terrestrial ecosystems: Comparing two contrasting approaches within European climate space. *Glob. Ecol. Biogeogr.*, 10, 621–637.
- Somerville, R.S. & Davé, R. (2015). Physical models of galaxy formation in a cosmological framework. *Ann. Rev. Astron. Astrophys.*, 53, 51–113.
- Thompson, P.L., Guzman, L.M., Meester, L.D., Horváth, Z., Ptačnik, R., Vanschoenwinkel, B. *et al.* (2020). A process-based metacommunity framework linking local and regional scale community ecology. *Ecol. Lett.*, 23(9), 1314–1329.
- Tiktak, A. & Bouten, W. (1992). Modelling soil water dynamics in a forested ecosystem. III: Model description and evaluation of discretization. *Hydrol. Process.*, 6, 455–465.
- Trotsiuk, V., Hartig, F., Cailleret, M., Babst, F., Forrester, D.I., Baltensweiler, A. *et al.* (2020). Assessing the response of forest productivity to climate extremes in Switzerland using model– data fusion. *Glob. Change Biol.*, 26, 2463–2476.
- Trucano, T.G., Swiler, L.P., Igusa, T., Oberkampf, W.L. & Pilch, M. (2006). Calibration, validation, and sensitivity analysis: What's what. *Reliability Engineering & System Safety*, 91, 1331–1357.
- Tuo, R. (2017). Adjustments to Computer Models via Projected Kernel Calibration. *arXiv:1705.03422 [stat]*.
- Tuo, R. & Wu, C.F.J. (2016). A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification*, 4, 767–795.
- Warton, D.I., Foster, S.D., De'ath, G., Stoklosa, J. & Dunstan, P.K. (2015). Model-based thinking for community ecology. *Plant Ecol.*, 216, 669–682.
- White, J.T., Doherty, J.E. & Hughes, J.D. (2014). Quantifying the predictive consequences of model error with linear subspace analysis. *Water Resour. Res.*, 50, 1152–1173.
- Wood, S.N. (2001). Partially specified ecological models. *Ecol. Monogr.*, 71, 1–25.
- Wutzler, T. & Carvalhais, N. (2014). Balancing multiple constraints in model-data integration: Weights and the parameter block approach. *J. Geophys. Res. Biogeosci.*, 119, 2112–2129.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Editor, Tim Coulson

Manuscript received 28 August 2020

First decision made 16 October 2020

Second decision made 26 January 2021

Manuscript accepted 11 February 2021