Received: 11 April 2022 | Accepted: 3 September 2022

DOI: 10.1111/2041-210X.14002

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Issues in calibrating models with multiple unbalanced constraints: the significance of systematic model and data errors

David Cameron[1] | Florian Hartig[2] | Francesco Minnuno[3] | Johannes Oberpriller[2] | Björn Reineking[4] | Marcel Van Oijen[5] | Michael Dietze[6]

[1]UK Centre for Ecology and Hydrology Bush Estate, Penicuik, UK; [2]Theoretical Ecology, University of Regensburg, Regensburg, Germany; [3]Department of Forest Sciences, University of Helsinki, Helsinki, Finland; [4]Univ. Grenoble Alpes, INRAE, LESSEM, France; [5]Independent Researcher, Edinburgh, UK and [6]Department of Earth & Environment, Boston University, Boston, Massachusetts, USA

Correspondence
David Cameron
Email: dcam@ceh.ac.uk

Funding information
European Space Agency, Grant/Award Number: 4000135015/21/I-NB; Academy of Finland, Grant/Award Number: 312559

Handling Editor: Carl Boettiger

Abstract

1. Calibrating process-based models using multiple constraints often improves the identifiability of model parameters, helps to avoid several errors compensating each other and produces model predictions that are more consistent with underlying processes. However, using multiple constraints can lead to predictions for some variables getting worse. This is particularly common when combining data sources with very different sample sizes. Such unbalanced model-data fusion efforts are becoming increasingly common, for example when combining manual and automated measurements.
2. Here we use a series of simulated virtual data experiments that aim to demonstrate and disentangle the underlying cause of issues that can occur when calibrating models with multiple unbalanced constraints in combination with systematic errors in models and data. We propose a diagnostic tool to help identify whether a calibration is failing due to these factors. We also test the utility of adding terms representing uncertainty in systematic model/data systematic error in calibrations.
3. We show that unbalanced data by itself is not the problem—when fitting simulated data to the 'true' model, we can correctly recover model parameters and the true dynamics of latent variables. However, when there are systematic errors in the model or the data, we cannot recover the correct parameters. Consequently, the modelled dynamics of the low data volume variables departs significantly from the true values. We demonstrate the utility of the diagnostic tool and show that it can also be used to identify the extent of the imbalance before the calibration starts to ignore the more sparse data. Finally, we show that representing uncertainty in model structural errors and data biases in the

calibration can greatly improve the model fit to low-volume data, and improve coverage of uncertainty estimates.

4. We conclude that the underlying issue is not one of sample size or information content per se, despite the popularity of ad hoc approaches that focus on 'weighting' datasets to achieve balance. Our results emphasize the importance of considering model structural deficiencies and data systematic biases in the calibration of process-based models.

KEYWORDS
Bayesian inference, inverse modelling, model calibration, model discrepancy, multiple constraints, predictive uncertainty, structural model error, systematic data bias

1 | INTRODUCTION

Calibrating a model with multiple constraints means that we use data on several model outputs, often at similar organizational levels of the modelled system, to constrain uncertainties about model structure or parameters. The value of this approach has long been recognized: from a theoretical perspective, models that make multiple predictions are considered to be 'efficient' as they are often supported by multiple lines of evidence and can be tested against different types of observations (Grimm & Railsback, 2012; Marquet et al., 2014). From a practical perspective, scientists are increasingly reliant on complex process-based models (Fisher et al., 2018; Fisher & Koven, 2020), together with methods to combine models and data, to generate precise forecasts and improve system understanding (Dietze et al., 2018; Van Oijen, 2020). In both cases, the use of multiple constraints is important when alternative competing hypotheses or models are compatible with a single set of observations. While it is often not hard for complex models to get a single 'right' answer for the wrong reasons, it is much harder to hit multiple benchmarks at the same time, and careful comparisons to multiple data constraints can help isolate incorrect assumptions (Grimm & Railsback, 2012; Medlyn et al., 2015).

The value of multiple data constraints is not limited to model testing, but extends equally, if not more so, to model calibration. Issues of equifinality (i.e. multiple alternative parameter combinations producing the same model output) and parameter identifiability are common when complex models are constrained by a single type of data, making it easy for models to get the right answer for the wrong reason (Williams et al., 2009). In principle, the process of constraining model uncertainties via calibration (a.k.a. inverse modelling or model-data fusion, e.g. Hartig et al., 2012) is relatively straightforward. The idea is to infer model parameters that produce outputs that agree with the observed data. This can be achieved via informal calibration or optimization procedures (e.g. Aber & Federer, 1992; Parton et al., 1993), but as increasingly more data have become available in the recent years (Farley et al., 2018; Hampton et al., 2013; LaDeau et al., 2017), the field has moved towards formal statistical calibration methods based on likelihood or Bayesian statistics (Fer et al., 2018; Hilborn & Mangel, 1997; Van Oijen et al., 2005).

Technically, combining multiple, heterogeneous data sources within a statistical calibration is straightforward. Provided that measurement errors associated with the data are uncorrelated and hence independent, we can combine them by multiplying the statistical likelihoods (the probability of observing a dataset under any particular set of proposed model parameters) for the individual data streams (Van Oijen, 2020). In practice however, the statistical calibration of complex models can be challenging, especially when data sources differ greatly in volume (e.g. Medvigy et al., 2009; Ricciuto et al., 2011; Richardson et al., 2010). Unbalanced calibration datasets are now common as low volumes of manually collected field data are frequently combined with high volumes of automatically collected data from in situ sensors or remote sensing. Since each data point is usually modelled as an independent piece of information in a statistical likelihood, sparse observations can often be overwhelmed by the higher volume of data. This is undesirable as the low-volume data often constrain parts of the system with high uncertainties that are crucial for future projections (e.g. soil carbon and nitrogen), and require higher labour costs to collect. As increasingly more data become available, this issue of unbalanced datasets is likely to worsen significantly. For example, NASA's earth observation system is expected to grow by an order of magnitude, from an already overwhelming ~5 PB/year in 2018–2020 to a staggering ~50 PB/year, as soon as 2022 (https://earthdata.nasa.gov/eosdis/cloud-evolution).

Since the apparent issue is the imbalance in data volume, existing approaches often try to correct that balance by thinning-out, aggregating or reweighing the calibration datasets so that they have a more balanced influence on the calibration. Common examples include, reweighting different datasets so they count equally (Cailleret et al., 2020; Keenan et al., 2013; Medvigy et al., 2009; Richardson et al., 2010) or weighting by inverse sample size (Thum et al., 2017). In fisheries, Maunder et al. (2017) and Carvalho et al. (2017) have suggested that in likelihood-based statistical procedures used to assess stock measurements, the down weighting or elimination of data is often used (e.g. Kell et al., 2014; Siddeek et al., 2017) to deal with data conflicts arising from model misspecification. Maunder et al. (2017) suggest that model misspecification is a main cause of sensitivity of calibration results to data weighting, and that down-weighting data are not necessarily appropriate because it may not

resolve model misspecification (Wang et al., 2015). The main purpose of these ad hoc approaches is to down-weigh the high-volume data so that its influence on the calibration is more balanced.

Unfortunately, such ad hoc approaches have no basis in probability theory; indeed, it makes no logical sense that the information content of a dataset in the calibration should be determined by the presence of another more sparse dataset. The significance of a dataset in the calibration should be determined by the reliability of that dataset alone. By arbitrarily changing the reliability of the calibration data, we are also throwing away potentially useful information that can be used to improve models. In reweighting the data, we introduce subjective control over the calibration by some measure of how close we want the model to fit the different data streams after calibration. A better option would be to develop solutions based on the underlying causes that lead to poor outcomes when calibrating models with unbalanced data.

Oberpriller et al. (2021) showed that the calibration problem with unbalanced data streams in not due the imbalance per se, but because the model cannot fit both data sources when structural error is present. The calibration will favour the high-volume data, at the expense of worse model predictions for the low-volume data, because the former has a higher weight in the likelihood.

Here we investigate this problem in more detail, using several virtual experiments to illuminate the underlying reasons for the issues discussed. Second, we propose a diagnostic tool to help researchers identify whether issues that they are facing during calibration can be attributed to the interaction of imbalanced calibration data with model/data error rather than some other cause. Finally, we illustrate, as simply as possible, that including uncertainty in model structural error and data systematic bias in the likelihood improves model predictions and provides a quantification of uncertainty that has greater utility than using ad-hoc methods such as reweighting.

2 | MATERIALS AND METHODS

2.1 | Very simple ecosystem model

To illustrate the issues of model calibration to multiple constraints, we developed the very simple ecosystem model (VSEM). The model was designed to be as simple as possible, yet resemble more complicated, process-based ecosystem models that are commonly used in terrestrial ecosystem modelling.

In essence, the model calculates the daily accumulation of carbon in the plant and soil from the growth of the plant via photosynthesis and senescence to the soil, which respires carbon back to the atmosphere. While we rely on a terrestrial carbon budget model for this example, the underlying issues are general to any model that predicts multiple outputs (e.g. species, life-history stages, biogeochemical pools and fluxes). These issues apply to wide classes of models in routine use across marine, freshwater and terrestrial systems that are used to describe physiological, population, community, ecosystem and evolutionary processes.

The VSEM requires only one input dataset to drive the model, daily photosynthetically active radiation (PAR, MJ m^{-2} day^{-1}).

Since we are interested in virtual experiments, we simulated PAR input data using a sinusoidal function, $\mathrm{PAR} = (|\sin(\mathrm{Days}/365 \times \pi) + \epsilon \times 0.25|) \times 10$

$$\epsilon \sim N(0, 1),$$

where ϵ represents Gaussian random noise and Days is a vector of integers from 1 to the number of days in the simulation (2048 in this case).

The model calculates gross primary productivity (GPP) using a very simple light-use efficiency (LUE) formulation multiplied by light interception. Light interception is calculated via Beer's law with a constant light extinction coefficient, KEXT, operating on Leaf Area Index, which itself is calculated based on vegetation foliar carbon (C_v) and leaf area ratio (LAR). A respiration parameter (GAMMA) determines the fraction of GPP that is autotrophic respiration, giving the net primary productivity (NPP).

$$\begin{aligned} \mathrm{GPP} &= \mathrm{PAR} \times \mathrm{LUE} \times \left(1 - \exp^{(-\mathrm{KEXT} \times \mathrm{LAR} \times C_v)}\right) \\ \mathrm{NPP} &= (1 - \mathrm{GAMMA}) \times \mathrm{GPP} \end{aligned}$$

There are three state equations representing the change in vegetation (C_v), root (C_r) and soil (C_s) carbon pools over time. The NPP is allocated to above (vegetation) and below (root) ground carbon pools via a fixed allocation fraction (A_v). Carbon is lost from the plant pools to a single soil pool via fixed vegetation and root turnover rates (τ_v and τ_r). Heterotrophic respiration in the soil is determined via a soil turnover rate (τ_s).

$$\begin{aligned} \frac{dC_v}{dt} &= A_v \times \mathrm{NPP} &&- \frac{C_v}{\tau_v} \\ \frac{dC_r}{dt} &= (1.0 - A_v) \times \mathrm{NPP} &&- \frac{C_r}{\tau_r} \\ \frac{dC_s}{dt} &= \frac{C_r}{\tau_r} + \frac{C_v}{\tau_v} &&- \frac{C_s}{\tau_s} \end{aligned}$$

2.2 | Bayesian calibration

We use a Bayesian approach to model calibration, though we note that the issues we raise, and their solutions, are not limited to Bayesian approaches but extend equally to other forms of statistical model calibration (e.g. Maximum Likelihood). In Bayesian Calibration (BC), our aim is to quantify the posterior probability of the model parameters (θ). The posterior probability $P(\theta \mid D)$ is calculated using Bayes' equation,

$$P(\theta \mid D) \propto P(\theta) L(D \mid \theta),$$

where $P(\theta)$ and $L(D|\theta)$ are the prior and likelihood, respectively.

Since it is not possible to calculate the posterior distribution for VSEM analytically, we estimate it with Markov Chain Monte Carlo sampling, using the DREAMzs algorithm (Vrugt et al., 2009) implemented in the R package BayesianTools (Hartig et al., 2019).

2.2.1 | Prior

We used simple uniform priors (Table 2) since our aim is to identify the issues associated with multiple constraints using a simple and easy to interpret modelling approach. We focus the calibration on a subset of the parameters in Table 1 because we manipulated the values for two parameters, allocation to vegetation (Av) and initial root pool (Cr), as part of the simulated data experiments described below. Since the root pool is not part of the model with the error in these experiments, we also exclude tauR from the calibration. The parameters LAR and GAMMA were removed from the calibration to avoid non-identifiability issues. During calibration, tauR, LAR and GAMMA were fixed to the 'true' values used when generating simulated data.

2.2.2 | Likelihood

For the likelihood, we use a univariate Gaussian distribution. This is a typical choice and as we simulated the calibration data under the same assumptions (see Section 2.3.1), we know this form of the likelihood to be appropriate. In Section (2.4), we discuss modifications to this simple likelihood to represent model structural error and data systematic bias. Because heteroskedasticity is a common feature of carbon cycle data, each σ^2 was modelled as proportional to the variable in question (net ecosystem exchange [NEE], soil carbon, vegetation carbon) via a single coefficient of variation parameter, included in the calibration.

TABLE 1 Very simple ecosystem model (VSEM) model parameters

Parameter	Variable name	Default	Units
Light extinction coeff.	KEXT	0.5	m^2 ground area/m^2 leaf area
Leaf area ratio	LAR	1.5	m^2 leaf area/kg aboveground vegetation
Light use efficiency	LUE	0.002	kg C MJ^{-1} PAR
Ratio of autotrophic resp. to GPP	GAMMA	0.4	—
Vegetation turnover rate	tauV	1440	Days
Soil decomposition rate	tauS	27,370	Days
Root turnover rate	tauR	1440	Days
Allocation fraction to vegetation	Av	0.5	—
Initial vegetation pool size	Cv	3	kg C m^{-2}
Initial soil pool size	Cs	15	kg C m^{-2}
Initial root pool size	Cr	3	kg C m^{-2}

Abbreviation: GPP, gross primary productivity.

TABLE 2 Uniform priors ranges used for model calibration experiments

Parameter	Min	Max
KEXT	0.2	1.0
LUE	0.0002	0.004
tauV	200	3000
tauS	4000	50,000
Cv	0.0	400
Cs	0.0	1000

2.3 | Experiments with virtual data from VSEM

To illustrate the impacts of relative data volume (balanced vs. unbalanced) and different sources of model and data errors on the outcome of calibrating models to multiple data constraints, we designed a series of calibration experiments. Specifically, we simulated data from VSEM assuming a Gaussian observation error and then calibrated the model to these pseudo-observations for one flux, NEE, and two pools, vegetative carbon and soil carbon, that represent likely real-world data constraints. Model assessment focused on both the quantitative ability to recover the 'true' model parameters and the ability of the calibrated models to reconstruct the observed time series. The experiments described below are summarized in Table 3.

2.3.1 | Perfect model

A central theme that we consider here is the significance of a 'perfect' model structure where all the processes are represented

TABLE 3 Summary of computational experiments. Model structure indicates whether the model used for calibration was identical to that used to simulate the data (perfect) or contained a structural error. Data volume indicates whether all three data constraints had the same sample size (balanced) or whether vegetative carbon data were sparse (unbalanced). Data error indicates whether the observation errors were uncorrelated random Normal noise or whether soil carbon observations included a multiplicative bias. Likelihood indicates whether the statistical likelihood was a Normal or a Normal (N) with an additional linear bias correction. Not all model experiment permutations were needed to identify the patterns of model calibration error

Experiment	Model structure	Data volume	Data errors	Likelihood
Pb	Perfect	balanced	random	Normal
Pu	Perfect	unbalanced	random	Normal
Eb	Error	balanced	random	Normal
Eu	Error	unbalanced	random	Normal
PbB	Perfect	balanced	Bias + random	Normal
PuB	Perfect	unbalanced	Bias + random	Normal
EuB	Error	unbalanced	Bias + random	Normal
EuL	Error	unbalanced	random	N + Linear
PuBL	Perfect	unbalanced	Bias + random	N + Linear
EuBL	Error	unbalanced	Bias + random	N + Linear

correctly. The only way to ensure such a perfect model is to take the output from the VSEM and consider this as virtual data in the BC. In the first experiment (Pb), the observations are available for the full 2048 day length of the VSEM simulation. For the second experiment (Pu), which isolates the impact of relative data volume, we create a sparse dataset for vegetative carbon to simulate having an imbalance between observations available for vegetative carbon, soil carbon and NEE. The sparse dataset has six observations for days 2, 404, 780, 1100, 1500 and 1840.

2.3.2 | Model with known structural error

To simulate a model with a known structural error, we consider a situation where a major model process/structure is unknown and therefore missing in the calibrated model (but not the pseudodata). Here we remove the root pool completely from the VSEM to simulate a major structural error. This was done by initializing the root pool to zero and setting the root allocation fraction to zero so that all the NPP is now allocated to the vegetation pool. This also shuts off any senescence from the root pool to the soil. This gave the model a structural error as we might have in a real situation while being sufficiently simple that we can still interpret the influence of the error. This experiment was run both with balanced data (Eb) and unbalanced data (Eu).

2.3.3 | Observational data with known bias

In addition to considering model structural error, we also investigated the influence of observations with systematic biases since all observational data will to a greater or lesser extent contain biases. Here we multiplied the soil data by 0.8 to represent a considerable multiplicative bias in the observations of soil carbon. The observation bias experiment was repeated for the perfect model/balanced data case (PbB), for the case where data were unbalanced (PuB), and when there is both unbalanced data and a model structural error (EuB).

2.4 | Modified likelihood to represent structural errors in the model and systematic biases in the data

Here we address the question of whether modifications to the likelihood function can help compensate for the errors introduced above (model structural errors, biased observational errors). A general principle in modelling is to begin with the simplest approach and only move on to more complicated solutions if the simple approach fails. We adopt that approach here, representing model structural error and data systematic bias via very simple multiplicative (α_1) and additive (α_0) corrections to the model outputs (i.e. a linear bias correction),

$$L(D|\,\theta) = N\left(\alpha_0 + \alpha_1 \mathrm{VSEM}(\theta), \sigma^2\right).$$

We add terms for each of the three outputs for which we have calibration data, and therefore have six additional parameters to represent additive and multiplicative errors for each of NEE, soil carbon and vegetative carbon (modaddNEE, modmultNEE, modaddCs, modmultCs, modaddCv and modmultCv). The priors for each of these are given in Table 4.

3 | RESULTS

3.1 | Identifying the underlying issue

Here, we investigate the underlying issues when calibrating a model with unbalanced data (experiments Pb-EuB, Table 3). The following sections refer to posterior marginal parameter plots in Figure 1 and time series plots in Figure 3. Plots of the coefficient of variance (Section 2.2.2) are shown in the Supplementary Material.

3.1.1 | Perfect model and balanced data (Pb)

For the initial case where the data are balanced and the model is perfect, the 'true' parameters are recaptured by the calibration and the uncertainty versus the prior has reduced significantly. The model outputs for NEE, Cv and Cs are also centred around the truth (Figure 2), the posterior uncertainty is small, and the predictive interval matches the uncertainty in the data. This first calibration can be considered a control for all subsequent calibrations.

3.1.2 | Perfect model and unbalanced data (Pu)

When we have a large imbalance in the calibration data (Cv reduced from 2048 to six observations, Section 2.3.1), the parameters are still largely centred on the 'truth' line. For KEXT, tauV and Cv, there has been an increase in marginal uncertainty but this would be expected since we have included less information in the calibration (Figure 1). For Cs and Cv, there is little change from experiment Pb (Figure 3, top row). For the remaining calibrations, we do not include further plots of NEE as the plot does not show much change from that shown previously for Pb. Overall, these results show that unbalanced data, by itself, does not cause an issue in the calibration other than to increase the uncertainty.

TABLE 4 Uniform priors ranges used for the systematic bias parameters

Parameter name	Min	Max
modmultNEE	0.1	2.0
modmultCs	0.1	2.0
modmultCv	0.1	2.0
modaddNEE	−0.01	0.01
modaddCs	−1.0	1.0
modaddCv	−1.0	1.0

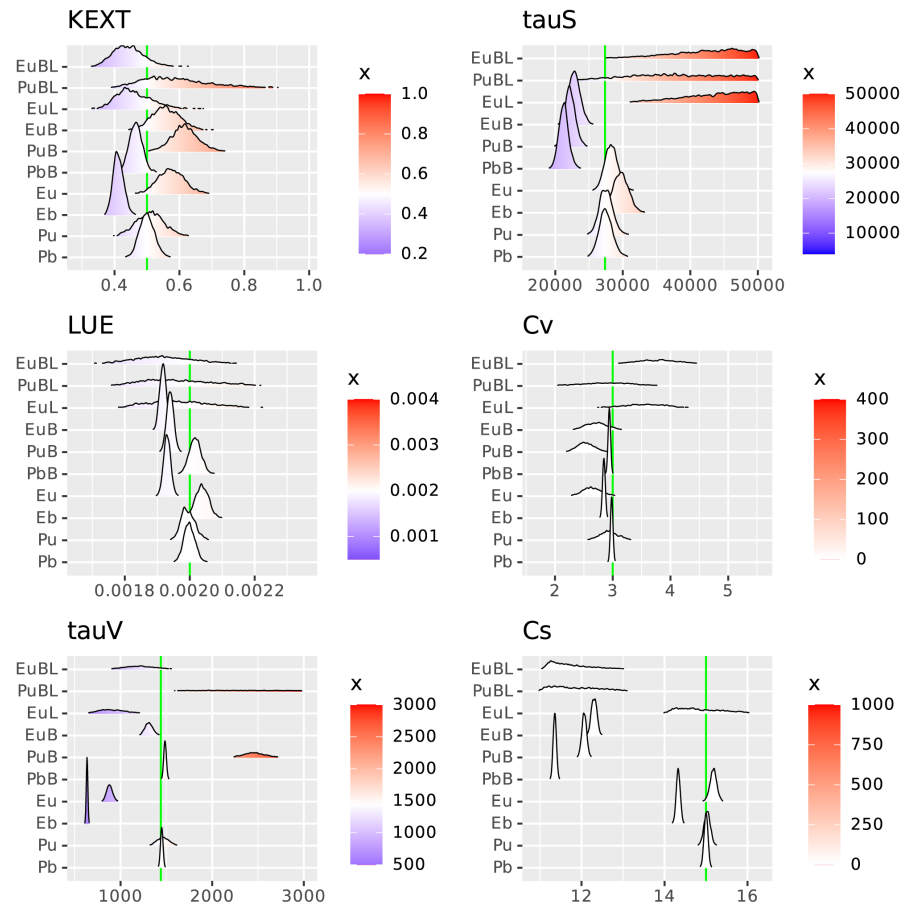

FIGURE 1 Ridge plots of model posterior parameter distributions for each of the experiments in Table 3. The 'true' value for each parameter (Table 1) is indicated by a green vertical line. Positive and negative biases in parameter estimates are coloured red and blue, respectively, and indicate the prior range. Biases in the model or data frequently result in parameter estimates that are confidently wrong (do not overlap with the true value), while the inclusion of a linear bias correction often result in an increase in parameter uncertainty.

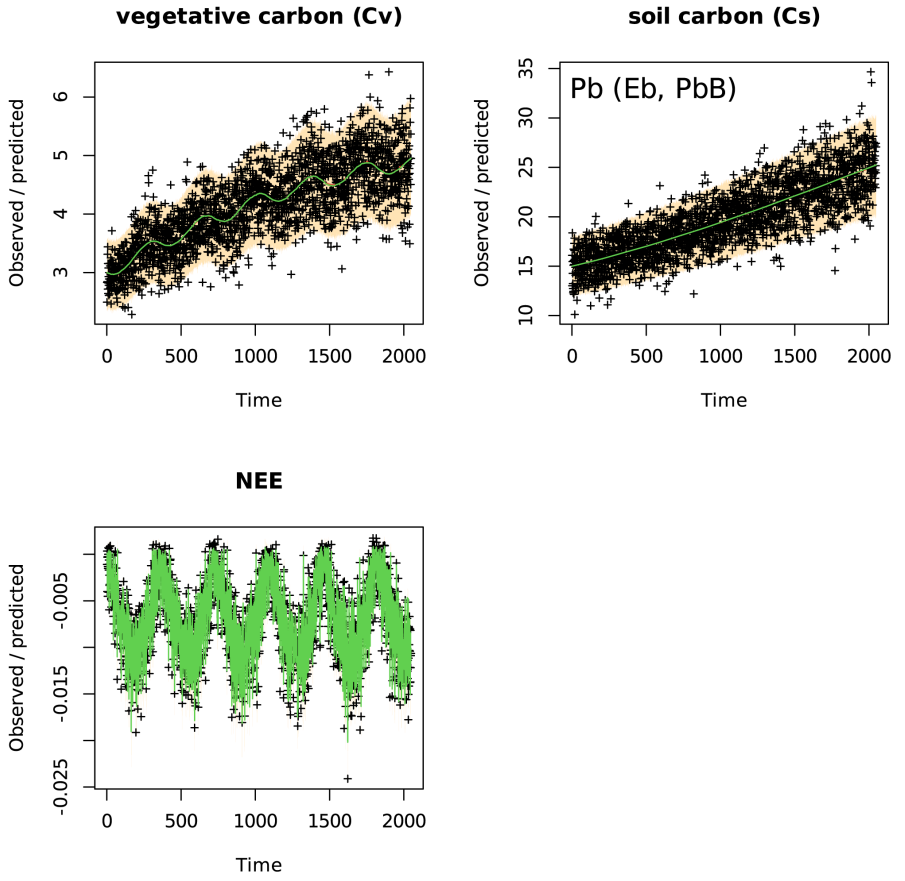

FIGURE 2 Experiment Pb: Perfect model, balanced data (NEE, Cv, Cs: 2048 obs). Observations included in the calibration marked with a '+'. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval. Experiment names in parentheses are not shown but are qualitatively equivalent to Pb. Complete set of figures available as Supplementary Material.

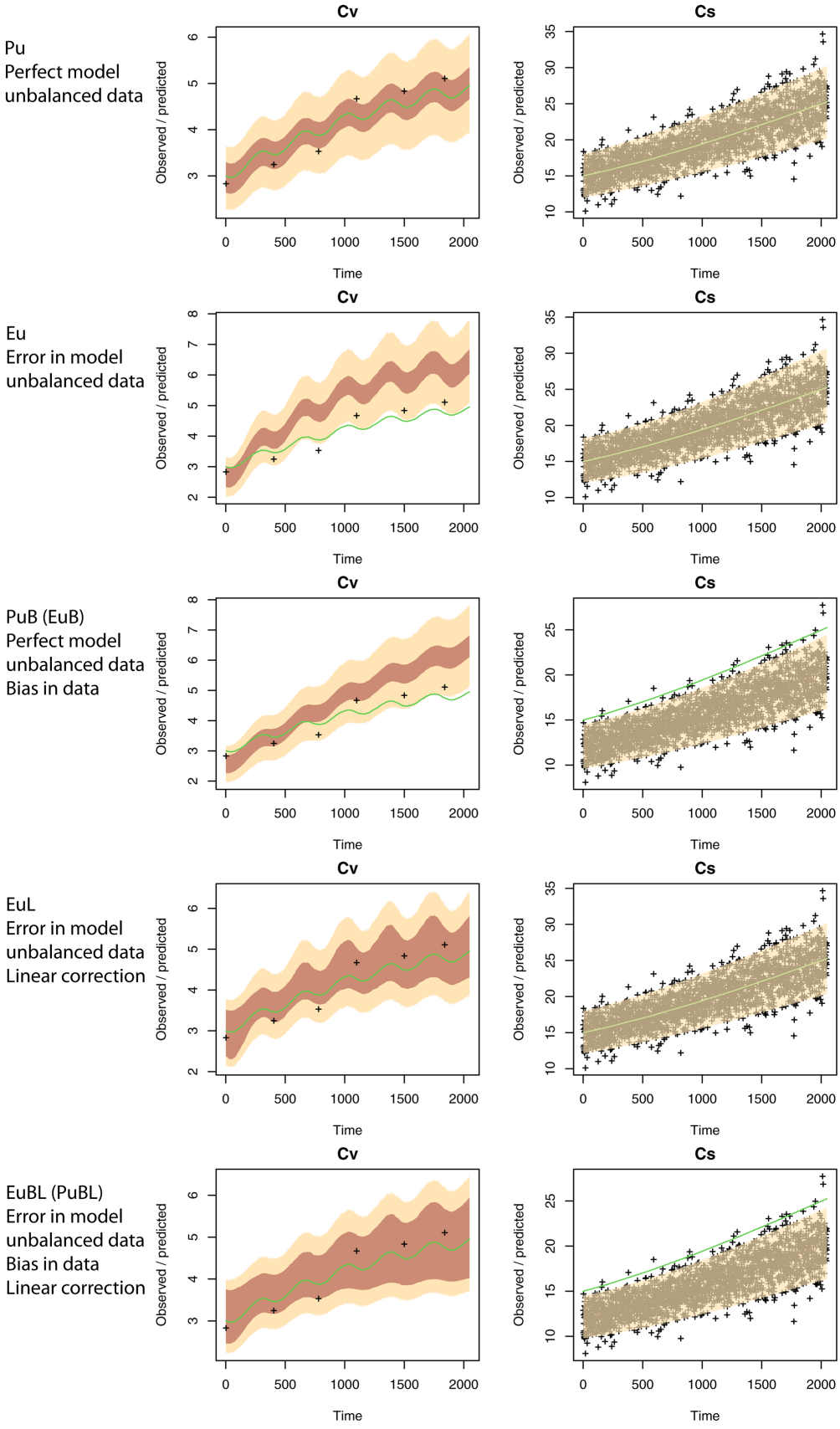

FIGURE 3 Time-series plots for vegetation and soil pools under the different experiments (symbols as in Figure 2). Experiment names in parentheses are not shown but are qualitatively equivalent to the row of figures indicated. Net ecosystem exchange (NEE) not shown because all cases were qualitatively equivalent to Figure 2. Complete set of figures available as Supplementary Material.

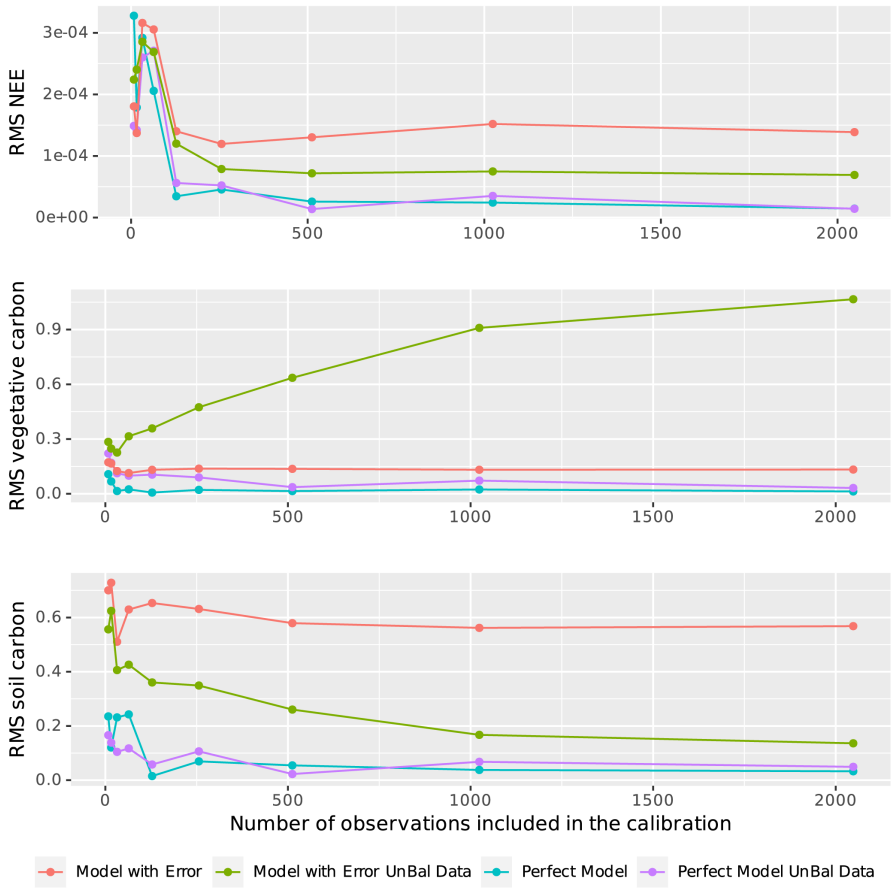

FIGURE 4 Each point in the three graphs (net ecosystem exchange [NEE], vegetative carbon and soil carbon) represents the RMS difference between the very simple ecosystem model (VSEM) model and the 'truth' run with different maximum a posteriori (MAP) parameters. The MAP parameters at each point are obtained by increasing the quantity of data included in the calibration along the x-axis. For the balanced calibration case (red and cyan), vegetative carbon data increase in tandem with NEE and soil carbon. For the unbalanced calibration case (green and purple), the quantity of vegetative carbon data is held fixed at six data values for each point along the *x*-axis. The VSEM model is either 'perfect' (cyan and purple) or has a known error (red and green) relative to the 'true' data that was derived from it.

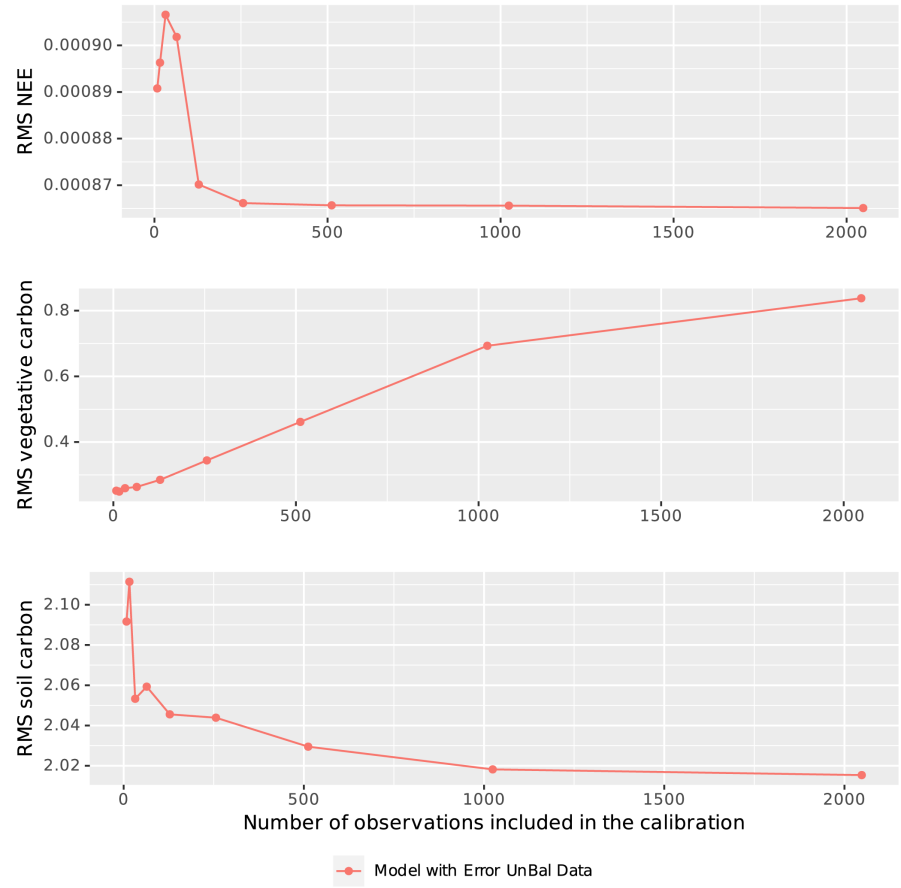

FIGURE 5 Each point in the three graphs (net ecosystem exchange [NEE], vegetative carbon and soil carbon) represents the root mean square (RMS) difference between the very simple ecosystem model (VSEM) model and virtual observations run with different maximum a posteriori (MAP) vectors obtained from a calibration where the quantity of data included for NEE and soil carbon increases along the *x*-axis and the quantity of vegetative carbon data is held fixed at six points. The VSEM model used has a known error relative to the virtual observations that were derived from it (Eu).

3.1.3 | Model with error and balanced data (Eb)

Here we created a known significant structural error in the model by effectively removing the root pool from the model (Section 2.3.2). After calibration, a number of parameters are now quite far away from their 'true' values. This is especially dramatic for tauV, where the rate of turnover of the vegetation pool has now more than doubled to compensate for the lack of root allocation and turnover. The large departures of the parameters from their 'true' values 'absorb' some of the model structural error, resulting in model outputs that have not changed significantly from the perfect model run (Figures S5 and S6). These results illustrate that model performance can still be acceptable, even when significant model errors are present, so long as parameter trade-offs can absorb the influence of the error.

3.1.4 | Model with error and unbalanced data (Eu)

When combining the influences of unbalanced data and model error, parameter changes versus the Eb calibration are significant but not huge. KEXT increased and LUE decreased slightly, compensating for each other, while parameters Cs, tauS and tauV are now closer to their 'true' value than in Eb. In general, the change in parameters to compensate for the model structural error is less than for Eb. Looking at the time series (Figure 3), the model does well for Cs (and NEE, not shown) but drifts away significantly from the six vegetation measurements. Consistent with the issues raised in the Introduction, this example illustrates a common behaviour for calibrations with a large data imbalance, where the sparsely measured parts of the system are ignored at the expense of the parts of the system with many observations.

3.1.5 | Perfect model and balanced data with a multiplicative bias (PbB)

We investigate the influence of data bias on the calibration by multiplying the soil carbon pool by 0.8 (Section 2.3.3). Similarly to when there is a model structural error (Eb), parameters in the calibration do not all recover their 'true' values and hence 'absorb' the influence of data error, particularly for the below-ground parameters. The initial Cs and tauS both decrease, increasing the turnover and decreasing the soil carbon pool to match the erroneous data. As before, these departures of the parameters from their 'true' value allow there to be a reasonably close match between the model outputs after calibration and the data (Figure S9 and S10).

3.1.6 | Perfect model and unbalanced data with a multiplicative bias (PuB)

Adding the effect of unbalanced data to the calibration with data bias, KEXT is now larger than its true value, increasing the carbon input into the system, but this is counteracted by a lower LUE. Cv is smaller and tauV larger, which has the combined effect of passing on less carbon to the soil. Belowground, tauS is slightly closer to its true value than PbB and Cs has increased versus the PbB calibration, pushing it back towards its true value. Similar to Eu, the model diverges from the vegetation observations, while similar to PbB the model is 'steered' towards fitting the many erroneous soil carbon observations. These results show there can also be issues calibrating with unbalanced datasets, even if the model is correct, if there are systematic biases in the observations.

3.1.7 | Model with error and unbalanced data with a multiplicative bias (EuB)

Combining the model structural error with the data bias for the unbalanced calibration causes the two errors to reinforce each other. The erroneous increase in the vegetation pool, due to the missing root pool model error, adds to the issue of trying to match the erroneously low soil carbon observations. The combined effect of the two errors pushes the model prediction even further away from the six Cv observations (S13 and S14).

3.2 | How to diagnose the issue in real applications

3.2.1 | Comparing model output with virtual data as truth

Modellers typically neither know the true model parameters nor the model structural error. Therefore, they cannot be sure if calibrations have the issues demonstrated in Section 3.1.

Here we develop a tool to help diagnose the presence of such issues. Calibrations are made with perfect and imperfect models where the quantity and imbalance of data increase with each calibration. Here we chose an increasing power series (2^3,2^4 … 2^{11}) for the quantity of calibration data, running eight calibrations in all. In the balanced data case, quantities of NEE, vegetative carbon and soil carbon data included in the calibration all increased in tandem in each subsequent calibration. For the unbalanced BC case, NEE and soil carbon data increased as before, but the quantity of vegetative carbon data included in the BC was held fixed at six data points. After running the calibrations, the VSEM was rerun with the maximum a posteriori (MAP) vector and the root mean square (RMS) difference with the 'true' data was calculated and plotted (Figure 4).

For most variables and experiments, we observe the expected pattern, whereby the RMS difference decreases as the quantity of data increases, with the perfect model getting closer to the data than the model with the error. Furthermore, for NEE and soil carbon with an imperfect model, the unbalanced calibration gets closer to the data than the balanced calibration, especially as the quantity

of calibration data increases. However, when the model has an error and there is unbalanced data, the vegetative carbon RMS difference increases as the quantity of calibration data increases. This signature of increasing RMS difference diagnoses when unbalanced data starts to become an issue. In this case, it is after the calibration exceeds 32 data points, but this will be different for each model, likelihood and dataset used in calibrations.

3.2.2 | Comparing model output against 'observations'

The diagnosis made in Section 3.2.1 had access to the 'true' data and a perfect model, which is never the case for real-world ecological model calibrations. Here we repeat the previous analysis, but focus just on the imperfect model and the unbalanced calibration, but with RMS differences now calculated against the 'pseudo-observed' data (Figure 5). While there are differences in the RMS values, the broad-scale signature of increasing RMS difference for vegetative carbon and decreasing RMS difference for NEE and soil carbon is retained.

3.3 | Changes to the Likelihood to represent model and data errors

The results from Section 3.1 demonstrate that the underlying issue with including unbalanced data in the calibration is not the imbalance itself but systematic errors in the model structure, data or both affecting the calibration. As presented in Section (2.4), we aim to introduce linear bias-correction terms in the likelihood which represent our uncertainty about what these systematic errors could be. Since we would not normally know the error present in the model or the data, these terms are not designed to address the specific errors present in the model and data here but rather as a simple linear correction to the model outputs.

We now repeat calibrations Eu, PuB and EuB but with the new likelihood. For all three experiments (EuL, PuBL and EuBL), the uncertainty has increases significantly for a number of parameters (KEXT, LUE, tauV, Cs and Cv; Figure 2 also Figures S15–S17) so that in general they are now closer to the parameter's 'true' value. An outlier is tauS where the uncertainty has increased but centre of the distribution is further away from the true value. Looking at the output time series (Figure 3), the influence of the error has not been removed, but there has been a significant improvement in the predictions, versus Eu, and EuB (PuB), with the centre of the posterior now much closer to the 'truth' line, especially for Cv. In addition, the uncertainty has increased so that, in general, data points are now inside the posterior predictive interval. The linear terms introduced have not completely removed the influence of the error, but there is a much greater sense that the sparse Cv data are influencing the calibration.

4 | DISCUSSION

4.1 | Unbalanced data in model calibration: Identifying the underlying issue

Our aim was to identify as cleanly as possible the underlying causes behind issues for model calibration caused by unbalanced data. First, we demonstrated that unbalanced data by itself is not the problem—there was no issue with including very unbalanced data so long as the observation error in the data was unbiased and the model was perfect. This finding runs counter to the hypothesis implicit in weighting data streams, which is that poor fits reflect an imbalance in information content, and thus that this imbalance needs to be corrected by reweighting.

Second, if we introduce a very significant model error or data bias, but keep the data streams balanced, the model predictions after calibration remain close to the data. In real-world calibrations, where we do not know the extent of the systematic model and data errors present, nor the 'true' settings of the parameters, these calibrations with balanced data would be considered a successful calibration. Given that we had access to the true parameter settings, however, we found that after calibration the parameters were far from their true values with high confidence. 'From the perspective of the calibration', the goal is to diminish the model-data difference. The likelihood cannot distinguish between model-data difference due to parameter error, model structural error or observation error, and has no means to change the structure of the model, so model-data difference is reduced solely by the parameters departing significantly from their true values. In this way, the calibration 'absorbs' the model and data errors into wrong settings of the parameters such that the model delivers fair performance on all data streams it is calibrated to. Other outputs from the model may still be very poor but there is no data available to assess this.

Third, it is only when we combine unbalanced data with a systematic error in either the model or data that the model predictions against the more sparse calibration data become poor and we identify an issue in the calibration. Because most real-world calibrations against multiple data constraints involve unbalanced data, it is easy to wrongly attribute the issue to unbalanced data. Indeed, while the model predictions were poorer after calibration with the unbalanced data (Eu, PuB), the parameters were if anything closer to their true values and less confidently wrong.

4.2 | Diagnostic tool

Our (Figures 4 and 5) aim in developing a diagnostic tool was first to identify the characteristic behaviour, or signature, that model or data errors are causing issues when calibrating with unbalanced datasets. We illustrated with a perfect model (Figure 4) that the RMS difference goes down for all model outputs when the quantity of

data in the calibration increases, regardless of whether the data are balanced. However, when a model error is present (Figure 4), the RMS difference increases for the sparse data output as the data imbalance increases. This is the signature behaviour that diagnoses the influence of the model discrepancy (or data bias) on the calibration. We showed that this diagnostic plot could also be created where the true model and data are unknown (Figure 5); hence, this tool can be used in real-world calibrations. In addition, the diagnostic figure can also be used to identify what size of imbalance in the data leads to a significant problem. This could be used to estimate how severely model and data errors are detrimentally influencing a calibration with unbalanced data and which variables are most effected, which may give clues about the underlying model and data systematic errors at issue.

4.3 | The role of data autocorrelation when calibrating with unbalanced datasets

Another class of potentially important errors in observations is due to autocorrelation. Accounting for autocorrelation in observations is an important way to discriminate between raw sample size and information content of the data to generate a more appropriate weighting among data constraints. Sample size and contribution of the data stream in the likelihood are not the same thing as contribution can be lowered, for example, by higher variance, or by data with a high degree of autocorrelation present. Nevertheless, we have shown here that the fundamental problem with data imbalance in model calibrations is not one of differences in information content/weight, but one of systematic errors in the model and/or data. So while it would be 'best practice' to incorporate autocorrelation into calibration we have shown that it will not solve the problem that we have identified here.

4.4 | Addressing underlying causes rather than symptoms

We argued in the Introduction that using ad-hoc methods, such as reweighting the calibration data to give a more balanced dataset, was logically the wrong approach. The virtual data experiments that we have conducted in this study provide another reason to avoid ad-hoc methods. In general, it is much preferable to 'treat' the underlying cause of a problem rather than try and mitigate the symptoms. Therefore, it is better to address model and data errors directly rather than trying to mitigate the symptoms by reweighing the data to arbitrarily adjust its reliability. Ideally, the best approach would be to make changes to the model and the data collection to eradicate the damaging systematic and structural errors. In reality, all models are approximations and data are also imperfect so it is only possible to achieve this to some extent. For example, in our terrestrial carbon flux example, there are known issues with eddy covariance data due to a lack of closure of the energy budget (Wilson et al., 2002); it is not possible to fully match such data with models that conserve mass and energy. As a solution, Maunder et al. (2017) state that ideally model misspecification would be eliminated but that this is often difficult to diagnose (Carvalho et al., 2017; Maunder & Piner, 2017). Hence, this study provides further evidence that calibration without any explicit recognition of model discrepancy (systematic error) is potentially 'dangerous' (Brynjarsdóttir & O'Hagan, 2014). It can lead to model parameters 'absorbing' the errors present in the model and data, as we have illustrated herein, causing poor posterior inference of model parameters and hence poor predictions (Brynjarsdóttir & O'Hagan, 2014). In agreement with Brynjarsdóttir and O'Hagan (2014), we suggest that model discrepancy and biases in the data should be accounted for. More widely the same conclusions have been found in other research areas. In climate modelling, Sexton et al. (2012) showed that calibrating models without recognizing discrepancy increased the risk of making predictions that were overconfident. In fisheries modelling, Maunder and Piner (2017) and Carvalho et al. (2017) argue that down-weighing or eliminating conflicting data may not be appropriate as it may not resolve model misspecification. Stewart and Monnahan (2017) state that 'analysts should be aware that they cannot weigh their way out of a misspecified model'. They further suggest that inclusion of 'process variation', and not excessive down-weighing of data, is more likely to provide robust estimation. In Bayesian inference of soil respiration models, Elshall et al. (2019) suggest that there is often an assumption of independent, normally distributed and homoscedastic residuals. Furthermore, they suggest not accounting for these may not result in biased predictions and parameter estimates however, it will lead to underestimated posterior uncertainties and poorer predictions.

4.5 | Model and data discrepancy modelling recommendations

Given the complexities of many mechanistic models and the processes that we are aiming to model, it will often be very challenging to find a good discrepancy model. In many cases, the discrepancy may be highly nonlinear. Indeed, given the very large variation in models and processes and hence in model discrepancies it is not possible to offer a general approach that will work in most circumstances. Brynjarsdóttir and O'Hagan (2014) and others (Oberpriller et al., 2021; Van Oijen, 2020) have advocated the use of a Gaussian Process (GP) as a flexible and powerful approach to discrepancy modelling and indeed this may be a good approach for many but it can have significant downsides. Brynjarsdóttir and O'Hagan (2014) show that such an approach can only avoid possible identifiability issues between model parameters and model error, finding the true parameter values and hence be useful for extrapolation predictions if good prior information is known on the GP parameters which they acknowledge will in many cases be very challenging. In addition, using GPs ignores physical mechanisms and can often be very expensive computationally because it involves the inversion of a potentially large covariance matrix. In general, in modelling, it is

good practise to try simple solutions first and only to progress to more complex solutions such as GP when needed. This motivated our choice of simple linear bias correction term in the calibration to represent our uncertainty about model structural errors and data systematic biases. Similar approaches have been used to correct for systematic biases in greenhouse gas emission measurement and modelling by Van Oijen et al. (2011) and biases in soil respiration data by Fer et al. (2018). With this simple discrepancy model, we were able to illustrate that the linear bias correction increased the uncertainty in the joint posterior parameter distribution, making it more likely that the true parameter value was somewhere in the joint posterior distribution and that the model included the 'true' system in the posterior predictions. This facilitated a significant improvement of the fit of model predictions to the data even with very unbalanced datasets. Although even in this very simple case the linear discrepancy model did not fully recapture all the true model parameter settings. Indeed, in many real-world calibrations, a simple linear modelling approach may be found to be too simplistic; nevertheless, it has been usefully employed here to illustrate the importance of addressing model discrepancy and data bias in model calibration; especially where large calibration data imbalances are present. The topic of identifying and creating good statistical models of model discrepancy (and data bias) is not straightforward, and is an important area for future research and tool development (Chandler, 2013; Van Oijen, 2020). Nevertheless, as in all modelling, we advocate beginning with simple approaches, as we have followed here, and adding complexity incrementally.

5 | CONCLUSIONS

The virtual data calibrations presented here demonstrate cleanly that the underlying issue calibrating models with multiple constraint unbalanced data is not the unbalance in the data, but that models and data have systematic errors that remain hidden when we calibrate with balanced datasets, but whose influence is only seen in poor predictions after calibration with unbalanced datasets. This issue is likely even more rampant in the common case of calibrating models against a single constraint, but it only becomes apparent when such models are tested against additional types of observations. By addressing the underlying cause and including terms in the calibration for systematic error (discrepancy), we demonstrated that the model fit to low-volume data can be greatly improved with a quantification of uncertainty that has sufficient coverage to include the true system.

AUTHOR CONTRIBUTIONS

The original ideas for this study came from David Cameron and Marcel Van Oijen. Michael Dietze, Florian Hartig, Björn Reineking and Francesco Minunno contributed ideas and further refinement of the virtual experiments in the COST Action FP104 PROFOUND. Johannes Oberpriller helped in defining the final focus of the study. Michael Dietze and Francesco Minunno contributed, in particular, to the production of the figures. David Cameron and Michael Dietze led the writing with all other authors making significant contributions.

ACKNOWLEDGEMENTS

This work greatly benefited from discussions within COST Action FP1304 PROFOUND. FM was supported by the European Space Agency (4000135015/21/I-NB) and the Strategic Research Council under the Academy of Finland (IBC-Carbon project number 312559).

CONFLICT OF INTEREST

No conflicts of interest are known.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.14002.

DATA AVAILABILITY STATEMENT

All the data generated in this manuscript were created using code written in R. The R scripts that generated this data (Cameron et al., 2022) are available at the url https://doi.org/10.5281/zenodo.7115671.

ORCID

David Cameron https://orcid.org/0000-0001-8938-0908
Florian Hartig https://orcid.org/0000-0002-6255-9059
Francesco Minnuno https://orcid.org/0000-0002-7658-6402
Johannes Oberpriller https://orcid.org/0000-0001-6007-0041
Björn Reineking https://orcid.org/0000-0001-5277-9181
Marcel Van Oijen https://orcid.org/0000-0003-4028-3626
Michael Dietze https://orcid.org/0000-0002-2324-2518

REFERENCES

Aber, J. D., & Federer, C. A. (1992). A generalized, lumped-parameter model of photosynthesis, evapotranspiration and net primary production in temperate and boreal forest ecosystems. *Oecologia*, *92*(4), 463–474. https://doi.org/10.1007/BF00317837

Brynjarsdóttir, J., & O'Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, *30*, https://doi.org/10.1088/0266-5611/30/11/114007

Cailleret, M., Bircher, N., Hartig, F., Hülsmann, L., & Bugmann, H. (2020). Bayesian calibration of a growth-dependent tree mortality model to simulate the dynamics of European temperate forests. *Ecological Applications*, *30*(1), e02021. https://doi.org/10.1002/eap.2021

Cameron, D. R., Hartig, F., Minunno, F., Oberpriller, J., Van Oijen, M., & Dietze, M. (2022). COST-FP1304 PROFOUND/TG15: Code for "Issues in calibrating models with multiple unbalanced constraints: Significance of systematic model and data errors" (v1.0.0). *Zenodo*. https://doi.org/10.5281/zenodo.7115671

Carvalho, F., Punt, A. E., Chang, Y.-J., Maunder, M. N., & Piner, K. R. (2017). Can diagnostic tests help identify model misspecification in integrated stock assessments? *Fisheries Research*, *192*, 28–40, ISSN 0165-7836. https://doi.org/10.1016/j.fishres.2016.09.018

Chandler, R. E. (2013). Exploiting strength, discounting weakness: Combining information from multiple climate simulators. *Philosophical Transactions of the Royal Society A: Mathematical*,

Physical and Engineering Sciences, *371*(1991), 20120388. https://doi.org/10.1098/rsta.2012.0388

Dietze, M. C., Fox, A., Beck-Johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., Keitt, T. H., Kenney, M. A., Laney, C. M., Larsen, L. G., Loescher, H. W., Lunch, C. K., Pijanowski, B. C., Randerson, J. T., Read, E. K., Tredennick, A. T., Vargas, R., Weathers, K. C., & White, E. P. (2018). Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(7), 1424–1432.

Elshall, A. S., Ye, M., Niu, G.-Y., & Barron-Gafford, G. A. (2019). Bayesian inference and predictive performance of soil respiration models in the presence of model discrepancy. *Geoscientific Model Development*, *12*(5), 2009–2032. https://doi.org/10.5194/gmd-12-2009-2019

Farley, S. S., Dawson, A., Goring, S. J., & Williams, J. W. (2018). Situating ecology as a big-data science: Current advances, challenges, and solutions. *Bioscience*, *68*(8), 563–576. https://doi.org/10.1093/biosci/biy068

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*(19), 5801–5830. https://doi.org/10.5194/bg-15-5801-2018

Fisher, R. A., & Koven, C. D. (2020). Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *Journal of Advances in Modeling Earth Systems*, *12*(4), e2018MS001453. https://doi.org/10.1029/2018MS001453

Fisher, R. A., Koven, C. D., Anderegg, W. R. L., Christoffersen, B. O., Dietze, M. C., Farrior, C. E., Holm, J. A., Hurtt, G. C., Knox, R. G., Lawrence, P. J., Lichstein, J. W., Longo, M., Matheny, A. M., Medvigy, D., Muller-Landau, H. C., Powell, T. L., Serbin, S. P., Sato, H., Shuman, J. K., … Moorcroft, P. R. (2018). Vegetation demographics in Earth System Models: A review of progress and priorities. *Global Change Biology*, *24*(1), 35–54. https://doi.org/10.1111/gcb.13910

Grimm, V., & Railsback, S. F. (2012). Pattern-oriented modelling: A 'multi-scope' for predictive systems ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1586), 298–310. https://doi.org/10.1098/rstb.2011.0180

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S., & Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156–162. https://doi.org/10.1890/120103

Hartig, F., Dyke, J., Hickler, T., Higgins, S. I., O'Hara, R. B., Scheiter, S., & Huth, A. (2012). Connecting dynamic vegetation models to data – an inverse perspective. *Journal of Biogeography*, *39*(12), 2240–2252. https://doi.org/10.1111/j.1365-2699.2012.02745.x

Hartig, F., Minunno, F., & Paul, S. (2019). BayesianTools: General-purpose MCMC and SMC samplers and tools for bayesian statistics. https://CRAN.R-project.org/package=BayesianTools

Hilborn, R., & Mangel, M. (1997). *The ecological detective confronting models with data: Confronting models with data (MPB-28) (Monographs in Population Biology)*. Princeton University Press.

Keenan, T. F., Davidson, E. A., Munger, J. W., & Richardson, A. D. (2013). Rate my data: Quantifying the value of ecological data for the development of models of the terrestrial carbon cycle. *Ecological Applications*, *23*(1), 273–286. https://doi.org/10.1890/12-0747.1

Kell, L. T., De Bruyn, P., Maunder, M. N., Piner, K. R., & Taylor, I. G. (2014). Likelihood component profiling as a data exploratory tool for north Atlantic albacore. *Collective Volumes of Scientific Papers ICCAT*, *70*(3), 1288–1293.

LaDeau, S. L., Han, B. A., Rosi-Marshall, E. J., & Weathers, K. C. (2017). The next decade of big data in ecosystem science. *Ecosystems*, *20*(2), 274–283. https://doi.org/10.1007/s10021-016-0075-y

Marquet, P. A., Allen, A. P., Brown, J. H., Dunne, J. A., Enquist, B. J., Gillooly, J. F., Gowaty, P. A., Green, J. L., Harte, J., Hubbell, S. P., O'Dwyer, J., Okie, J. G., Ostling, A., Ritchie, M., Storch, D., & West, G. B. (2014). On theory in ecology. *Bioscience*, *64*(8), 701–710. https://doi.org/10.1093/biosci/biu098

Maunder, M. N., Crone, P. R., Punt, A. E., Valero, J. L., & Semmens, B. X. (2017). Data conflict and weighting, likelihood functions and process error. *Fisheries Research*, *192*, 1–4, ISSN 0165–7836. https://doi.org/10.1016/j.fishres.2017.03.006

Maunder, M. N., & Piner, K. R. (2017). Dealing with data conflicts in statistical inference of population assessment models that integrate information from multiple diverse data sets. *Fisheries Research*, *192*, 16–27, ISSN 0165–7836. https://doi.org/10.1016/j.fishres.2016.04.022

Medlyn, B. E., Zaehle, S., De Kauwe, M. G., Walker, A. P., Dietze, M. C., Hanson, P. J., Hickler, T., Jain, A. K., Luo, Y., Parton, W., Prentice, I. C., Thornton, P. E., Wang, S., Wang, Y.-P., Weng, E., Iversen, C. M., McCarthy, H. R., Warren, J. M., Oren, R., & Norby, R. J. (2015). Using ecosystem experiments to improve vegetation models. *Nature Climate Change*, *5*(6), 528–534. https://doi.org/10.1038/nclimate2621

Medvigy, D., Wofsy, S. C., Munger, J. W., Hollinger, D. Y., & Moorcroft, P. R. (2009). Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem Demography model version 2. *Journal of Geophysical Research: Biogeosciences*, *114*(G1). https://doi.org/10.1029/2008JG000812

Oberpriller, J., Cameron, D. R., Dietze, M. C., & Hartig, F. (2021). Towards robust statistical inference for complex computer models. *Ecology Letters*, *24*(6), 1251–1261. https://doi.org/10.1111/ele.13728

Parton, W. J., Scurlock, J. M. O., Ojima, D. S., Gilmanov, T. G., Scholes, R. J., Schimel, D. S., Kirchner, T., Menaut, J.-C., Seastedt, T., Garcia Moya, E., Kamnalrut, A., & Kinyamario, J. I. (1993). Observations and modeling of biomass and soil organic matter dynamics for the grassland biome worldwide. *Global Biogeochemical Cycles*, *7*(4), 785–809. https://doi.org/10.1029/93GB02042

Ricciuto, D. M., King, A. W., Dragoni, D., & Post, W. M. (2011). Parameter and prediction uncertainty in an optimized terrestrial carbon cycle model: Effects of constraining variables and data record length. *Journal of Geophysical Research: Biogeosciences*, *116*(G1). https://doi.org/10.1029/2010JG001400

Richardson, A. D., Williams, M., Hollinger, D. Y., Moore, D. J. P., Dail, D. B., Davidson, E. A., Scott, N. A., Evans, R. S., Hughes, H., Lee, J. T., Rodrigues, C., & Savage, K. (2010). Estimating parameters of a forest ecosystem C model with measurements of stocks and fluxes as joint constraints. *Oecologia*, *164*(1), 25–40. https://doi.org/10.1007/s00442-010-1628-y

Sexton, D. M. H., Murphy, J. M., Collins, M., & Webb, M. J. (2012). Multivariate probabilistic projections using imperfect climate models part I: Outline of methodology. *Climate Dynamics*, *38*, 2513–2542. https://doi.org/10.1007/s00382-011-1208-9

Siddeek, M. S. M., Zheng, J., Punt, A. E., & Pengilly, D. (2017). Effect of data weighting on the mature male biomass estimate for Alaskan golden king crab. *Fisheries Research*, *192*, 103–113, ISSN 0165–7836. https://doi.org/10.1016/j.fishres.2017.02.001

Stewart, I. J., & Monnahan, C. C. (2017). Implications of process error in selectivity for approaches to weighting compositional data in fisheries stock assessments. *Fisheries Research*, *192*, 126–134, ISSN 0165–7836. https://doi.org/10.1016/j.fishres.2016.06.018

Thum, T., MacBean, N., Peylin, P., Bacour, C., Santaren, D., Longdoz, B., Loustau, D., & Ciais, P. (2017). The potential benefit of using forest biomass data in addition to carbon and water flux measurements to constrain ecosystem model parameters: Case studies at two temperate forest sites. *Agricultural and Forest Meteorology*, *234–235*, 48–65. https://doi.org/10.1016/j.agrformet.2016.12.004

Van Oijen, M. (2020). *Bayesian compendium*. Springer, xiv+204 pp. https://doi.org/10.1007/978-3-030-55897-0

Van Oijen, M., Cameron, D. R., Butterbach-Bahl, K., Farahbakhshazad, N., Jansson, P.-E., Kiese, R., Rahn, K.-H., Werner, C., & Yeluripati, J. B. (2011). A Bayesian framework for model calibration, comparison and analysis: Application to four models for the biogeochemistry of

a Norway spruce forest. *Agricultural and Forest Meteorology*, *151*(12), 1609–1621. https://doi.org/10.1016/j.agrformet.2011.06.017

Van Oijen, M., Rougier, J., & Smith, R. (2005). Bayesian calibration of process-based forest models: Bridging the gap between models and data. *Tree Physiology*, *25*(7), 915–927. https://doi.org/10.1093/treephys/25.7.915

Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., Hyman, J. M., & Higdon, D. (2009). Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation*, *10*(3), 273–290. https://doi.org/10.1515/IJNSNS.2009.10.3.273

Wang, S. P., Maunder, M. N., Nishida, T., & Chen, Y. R. (2015). Influence of model misspecification, temporal changes, and data weighting in stock assessment models: Application to swordfish (*Xiphias gladius*) in the Indian Ocean. *Fisheries Research*, *166*, 119–128.

Williams, M., Richardson, A. D., Reichstein, M., Stoy, P. C., Peylin, P., Verbeeck, H., Carvalhais, N., Jung, M., Hollinger, D. Y., Kattge, J., Leuning, R., Luo, Y., Tomelleri, E., & Trudinger, C. M. (2009). Improving land surface models with FLUXNET data. *Biogeosciences*, *6*(7), 1341–1359.

Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., Bernhofer, C., Ceulemans, R., Dolman, H., Field, C., Grelle, A., Ibrom, A., Law, B. E., Kowalski, A., Meyers, T., Moncrieff, J., Monson, R., Oechel, W., Tenhunen, J., … Verma, S. (2002). Energy balance closure at FLUXNET sites. *Agricultural and Forest Meteorology*, *113*(1), 223–243. https://doi.org/10.1016/S0168-1923(02)00109-0

SUPPORTING INFORMATION

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

How to cite this article: Cameron, D., Hartig, F., Minnuno, F., Oberpriller, J., Reineking, B., Van Oijen, M., & Dietze, M. (2022). Issues in calibrating models with multiple unbalanced constraints: the significance of systematic model and data errors. *Methods in Ecology and Evolution*, *13*, 2757–2770. https://doi.org/10.1111/2041-210X.14002