Identifying why the Bayesian calibration of process-based models with unbalanced quantities of calibration data can be challenging: The significance of model structural deficiencies and data biases.

D. R. Cameron, F. Hartig, F. Minnuno, J. Oberpriller, B. Reineking, M. Van Oijen and M. Dietze

2021-06-25

Abstract

Calibrating ecological process-based models using multiple data streams often improves the identifiability of model parameters, provides model predictions after calibration that are more consistent with underlying processes and helps to avoid compensating errors. However, using multiple constraints can sometimes also lead to problems, for example causing predictions for some variables to get worse, and 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. Here we use a series of simulated data experiments to illustrate the challenges that can occur when combining data, with a particular focus on unbalanced data and the role of systematic errors (i.e. biases) in models and data. We find 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 even in the presence of (random) observation errors. Despite the popularity of ad-hoc approaches that focus on "weighting" data sets differently, this illustrates that the issue at hand is not one of sample size or information content per se. However, when there are systematic errors in the model or the data, we cannot recover the correct parameters, and as a consequance the modeled dynamics of the low data volume variables can often depart significantly from the true values. To address these challenges we provide a diagnostic tool to help identify whether the issue of model/data error is the root cause of poor outcomes when calibrating with unbalanced data. The tool also helps to identify what extent of the imbalance before the calibration starts to ignore the more sparce data. Finally, we demonstate that representing uncertainty due to model structural errors and data biases in the calibration can greatly improve the model fit to low-volume data and thus make good predictions with a quantification of uncertaiinty that includes the true system. This emphasises the importance of considering model structural deficiencies and data systematic biases in the Bayesian calibration of ecological process-based models and demonstrates the lack of utility of ad-hoc measures that change the reliability of calibration data to achieve a false balance and hence a less useful arbitrary quantification of uncertainty in model predictions.

1 Introduction

Modellers in ecology and earth system sciences increasingly rely on complex computer simulations (Fisher and Koven 2020, Fisher et al 2018), coupled with methods to combine models and data to generate precise forecasts and improve system understanding.

In principle, the process of constraining model uncertainties via calibration (aka inverse modelling or model-data fusion, e.g. Hartig et al., 2012, Dietze et al. 2018) is relatively straightforward. The idea is to infer models (and their parameters) that are in agreement with the observed ecological and environmental data. This can be achieved via informal calibration or optimization procedures (citations), but as increasingly more data has become available in the recent years (citations), the field has moved towards formal statistical calibration methods based on likelihood or Bayesian statistics (CITATIONS). The methods allow formal parameter estimations for process-based models, and to project these uncertainties to model predictions, for example under climate change (citation).

In practice however, the Bayesian Calibration (BC) of ecological process-based models (PBMs) can be challenging. For example there can be difficulties in calibrating PBMs with observational data representing multiple parts of the system being predicted by the model. In principle, the combination of heterogeneous data types in BC is straightforward, and simply amounts to multiplying the likelihoods for the individual data streams. In practice, however, researchers often observe problems when attempting a naive calibration with multiple constraints. Issues are particularly common when the data streams differ greatly in the quantity of data available. Such an imbalanced calibration dataset is now common since low-volumes of manually-collected field data are frequently combined with high-volumes of automatically-collected data either from in situ sensors (e.g. eddy-covariance) or via remote sensing. A common observation in this situation is that the outputs from BC (e.g. calibrated model parameters) are virtually identical to the results achieved by fitting the model to the high-volume data by itself, and in obvious disagreement with the lower-volume data type (e.g. Cameron et al in review 2018).

Since each data point is usually modelled as an independent piece of information in a Likelihood, the influence of the sparse observations can often be overwhelmed by the higher frequency data (Cameron et al in review 2018). In essence, the BC ignores the low volume data, which gets swamped by the much larger sample size of the high volume data. Moreover, if the model cannot achieve a good fit to both data streams at the same time, it will logically favour fitting the common data stream, at the expense of worse predictions for model outputs with fewer data (Oberpriller et al., 2021).

This is highly undesirable as the lower-volume data often represents a part of the system that is crucial for the future projections (eg soil carbon and nitrogen), has relatively high uncertainty and has required higher labour costs to collect. As increasingly more data becomes available this issue of extremely imbalanced 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 ~5PB/yr in 2018-2020 to a staggering ~50PB/yr, as soon as 2022 (https://earthdata.nasa.gov/eosdis/cloud-evolution).

Since the apparent issue is the large imbalance in the data it would then seem naively logical to try and correct that balance in some way. Unfortunately it is difficult to follow such an approach without resorting to ad-hoc solutions. For example, it is common to thin-out, aggregate or reweigh the calibration datasets so that they have a more balanced influence on the BC. The main purpose is to down-weigh the high-volume data so that the different data models are more balanced. For example, Medvigy et al 2009 constrained the ED2 model to nine data constraints, including eddy covariance at the annual, monthly, and hourly scale and forest growth and mortality data, and weighted each part of the likelihood equally. (Keenan et al. 2013) similarly weighted each dataset equally when calibrating the FöBAAR model to 16 distinct data constraints. [Cailleret et al. 2020] also equally weighted basal area increment and stem number distribution in the calibration of the forest model ForClim. (Thum et al. 2017) constrained ORCIDEE with multiple constraints, weighting each by sample size. (Richardson et al. 2010) calibrated the DALEC model by optimizing the product of the log-Likelihoods across six data constraints, which similarly weights all data sets as equally important. Unfortunately, this approach has no basis in probability theory largely since it makes no logical sense that the weight or significance 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 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. Indeed, the purpose of BC is to update our subjective prior knowledge with more objective evidence taken from observations. By reweighing the data we reintroduce our subjective control over the calibration by some measure of how close we want the model to fit the observations after calibration. While pragmatic, reweighing the observation does not seem to be the right approach unless there are no other possible solutions. Much better to form our solutions based on the roots causes that lead to poor outcomes in the BC of PBMs with unbalanced data.

The underlying reasons for challenges with unbalanced data in the BC of process-based models, namely model and data error, are known (Oberpriller et al., 2021) but are not immediately apparent and hence the practise of resorting to ad-hoc and hence fundamentally less useful approaches is still common. This motivates one aim of this paper which is identify as simply as possible using clear-cut virtual data experiments the underlying reasons for the issues that are commonly found when using unbalanced datasets in BC. Secondly we offer a diagnostic tool to help researchers identify whether issues that they are facing in BC could be attributed to

the interaction of imbalanced calibration data with model/data error rather than some other cause. Finally we demonstrate, as simply as possible that including uncertainty due to model structural error and data systematic bias in the BC improves model predictions and provides a demonstrably useful quantification of uncertainty that we argue has greater utility than can be found using more ad-hoc arbitrary methods.

2 Methods

2.1 VSEM model

Here we present the Very Simple Ecosystem Model (VSEM). The model was created to help illustrate the main ideas that we present here. The model was designed to be very simple rather then realistic, but yet resemble many typical, but more complicated, process-based ecosystem models (PBMs) that are commonly used in carbon growth type ecosystem modelling.

In essence, the model determines the 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. The timestep of the VSEM is daily.

2.1.1 VSEM input data: Photosythetically active radiation (PAR)

The VSEM requires only one input dataset to drive the model namely daily PAR.

Since we are interested in virtual experiments here we generate the PAR input data using an sinusoidal function.

$$PAR = (|\sin(Days/365 \times \pi) + \epsilon|) \times 10 \tag{1}$$

$$\tag{2}$$

- PAR Photosynthetically active radiation
- ϵ Gaussian noise added
- Days number of days

2.1.2 Photosynthesis equation

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 operating on Leaf Area Index (LAI). A parameter (GAMMA) determines the fraction of GPP that is autotrophic respiration, giving the Net Primary Productivity (NPP).

$$GPP = PAR \times LUE \times (1 - \exp^{(-KEXT \times LAR \times C_v)}) \tag{3}$$

$$NPP = (1 - GAMMA) * GPP \tag{4}$$

- PAR Photosynthetically active radiation (MJ m^{-2} day^{-1})
- LUE Light use efficiency of NPP (Ra implicit)
- KEXT Beer's law light extinction coeff
- C_v Vegetation carbon
- LAR is the leaf area ratio
- GAMMA is the ratio of autotrophic respiration to GPP

2.1.3 Carbon pool state equations

There are three state equations representing the change in time of vegetation (C_v), root (C_r) and soil (C_s) carbon pools. The Net Primary Productivity (NPP) is allocated to above (vegetation) and below(root)

ground carbon pools via a fixed allocation fraction. Carbon is lost from the plant pools to a single soil pool via fixed vegetation and root turnover rates. Heterotropic respiration in the soil is determined via a soil turnover rate.

$$\frac{dC_v}{dt} = A_v \times NPP - \frac{C_v}{\tau_v} \tag{5}$$

$$\frac{dC_r}{dt} = (1.0 - A_v) \times NPP - \frac{C_r}{\tau_r} \tag{6}$$

$$\frac{dC_s}{dt} = \frac{C_r}{\tau_r} + \frac{C_v}{\tau_v} - \frac{C_s}{\tau_s} \tag{7}$$

2.1.4 VSEM model parameters

parameter name	variable name
Light extinction coeff	KEXT
Leaf area ratio	LAR
Light use efficiency	LUE
Ratio of autotrophic resp to GPP	GAMMA
Vegetation turnover rate	tauV
Soil decomposition rate	tauS
Root turnover rate	tauR
Allocation frac to vegetation	Av
Initial vegetation pool size	Cv
Initial soil pool size	Cs
Initial root pool size	Cr

2.2 Bayesian Calibration

In Bayesian Calibration, our aim is to quantify the probability of the model parameters (θ) being correct given the calibration data (D) (P(θ | D). Since this is not straightforward to calculate we make use of Bayes equation.

$$P(\theta|D) \propto P(\theta)L(D|\theta) \tag{8}$$

Where P(θ | D), P(θ) and L(D|θ) are known as the posterior, prior and likelihood respectively. Since it is not possible to calculate the likelihood for a numerical model such as VSEM analytically we sample from it and the prior using a Monte Carlo approach to sample from the posterior. As a way of making this sampling more efficient we use the DREAMzs algorithm in a Markov Chain Monte Carlo (MCMC) sampling.

- TODO: brief summary of DREAMzs algorithm

The DREAMzs algorithm and MCMC functions that we use here are from the BayesianTools package (https://cran.r-project.org/web/packages/BayesianTools/index.html).

2.2.1 Prior

Here we adopt a very simple uniform prior since our aim here is to identify the issue using a simple and therefore easy to interpret modelling approach.

We need to be able to control the values for two parameters, allocation to vegetation (Av) and initial root pool for the virtual experiments described below. Since the root pool is not part of the model with the error we also exclude tauR from the calibration

Of the remaining parameters LAR and GAMMA were removed from the calibration to avoid nonidentifiability issues.

The remaining parameters are listed below along with the uniform prior ranges used.

parameter	min	max
KEXT	0.2	1.0
LUE	0.0002	0.004
tauV	200	3000
tauS	4000	50000
Cv	0.0	400
Cs	0.0	1000

2.3 Idealised experiments with virtual data from VSEM

2.3.1 Likelihood

Given that we added Gaussian noise to the model output to produce the virtual data, a univariate Gaussian likelihood is the obvious choice. In section (2.4) we discuss modifications to this simple likelihood to represent model structural error and data systematic bias.

2.3.2 Perfect model

A central theme that we consider here is the significance of a perfect model structure that is to say where all the processes are represented perfectly. 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. Gaussian noise is added to the model output to represent system varability that is not captured by the model but crucially can be represented perfectly by the likelihood function that we use in the BC. The observations are for the full 2048 day length of the VSEM for Net Ecosystem Exchange (NEE), vegetative carbon and soil carbon.

For the vegetative carbon we create a sparse dataset 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.

2.3.3 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 model. Here we remove the root pool completely from the VSEM to simulate a major structural error. This is done by initialising 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 of course shuts off any senescence from the root pool to the soil. This gives the model a major structural error as we might have in a real situation whilst being sufficiently simple that we can still interpret the influence of the error.

2.3.4 Observational data with known bias

In addition to considering model structural error, we also wish to investigate the influence of observations with biases since all observational data will to a greater or lesser extent contain biases. Here we simulate data biases by multiplying the soil data by 0.8 to represent a considerable multiplicative bias in the observations of soil carbon.

2.4 Modified Likelihood to represent structural errors in the model and systematic biases in the data.

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, by representing model structural error

and data systematic bias in the likelihood function by very simple multiplicative and additive constants to the model outputs. We add terms for each of the three outputs for which we have calibration data namely (NEE, Cs and Cv). Therefore we have six extra parameters to represent addition and multiplicative error 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 the table below.

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

3 Identifying the issue

In this section we investigate the underlying issue that can cause problems when we try to calibrate a model with a data set that has very unbalanced numbers of observations from different parts of the system. We do this by breaking the problem into parts to investigate the individual influence of model structural error and data bias when calibrating with balanced and unbalanced datasets. We start with the idealised situation of a perfect model and a perfect calibration dataset with a balanced number of observations for each part of the system.

In the following sections will refer to posterior marginal parameter plots in Fig (1) and timeseries plots in Fig (3).

3.1 Perfect model and balanced data Pb

Looking first at the parameters we find that the 'true' parameters are largely recaptured by the calibration. The marginal posterior distributions are centred around the 'truth' line and the uncertainty versus the prior has reduced significantly. The model outputs for NEE, Cv and Cs are also centred around the truth line (Fig 2) with the 50% quantile line matching the truth line closely. The posterior uncertainty is small and the predictive interval matches the uncertainty in the data as would be expected. This first calibration can be considered as a control against which all subsequent calibrations can be compared.

3.2 Perfect model and unbalanced data Pu

We now consider what happens when we have a large imbalance in the calibration data. We do this by thinning out the number of observations for Cv from 2048 to just six observations (as described in section 2.3.2) whilst retaining the original 2048 observations for NEE and Cs; thus creating an O(3) imbalance. After calibration the parameters are still largely centred on the 'truth' line. For KEXT and especially 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. In Fig({fig:timeseries}) we see timeseries outputs for Cv and Cs for Pu. 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. For Cs also, there is little change from before (Pb) when the data was balanced. The Cv plot shows the six observations that were retained in the calibration. The posterior is still centred on the truth line with a larger posterior uncertainty as might be expected since far fewer data have been included. These results show that creating an imbalanced does not cause an issue in the calibration other than to increase the uncertainty.

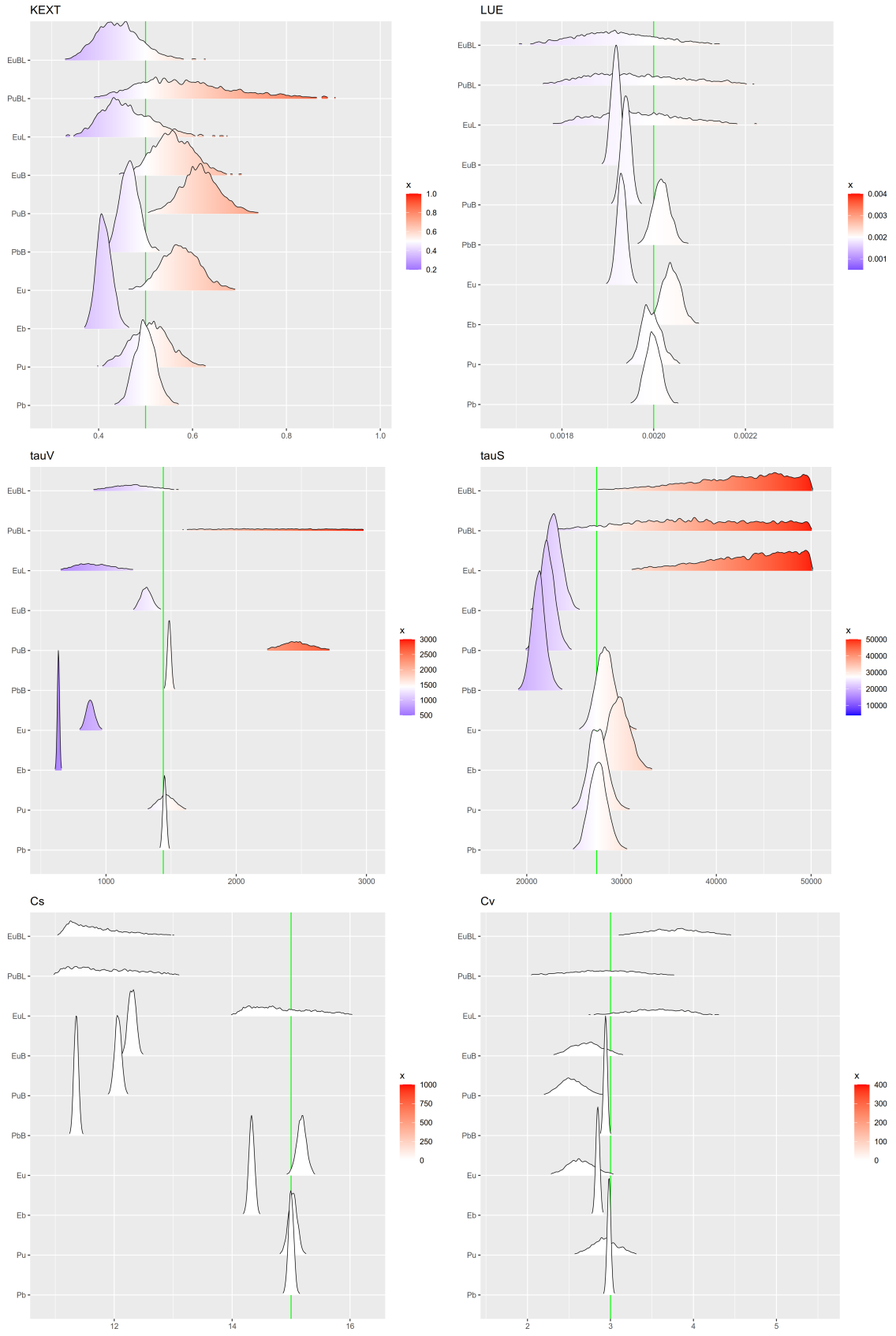


Figure 1: Parameters plot.

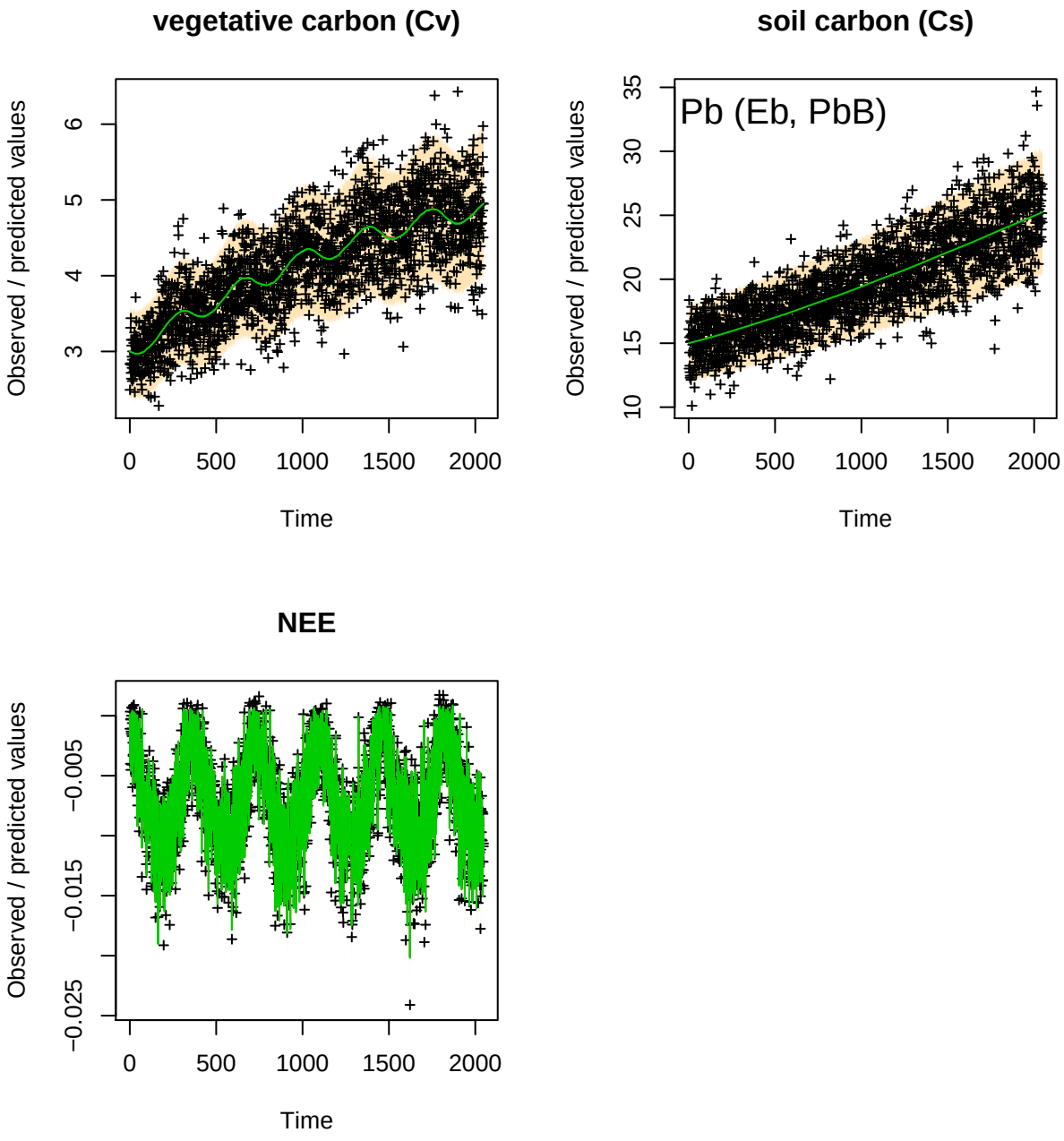


Figure 2: Perfect model, balanced data (NEE, Cv, Cs: 2048 obs). Observations included in the calibration marked with a '+'. Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

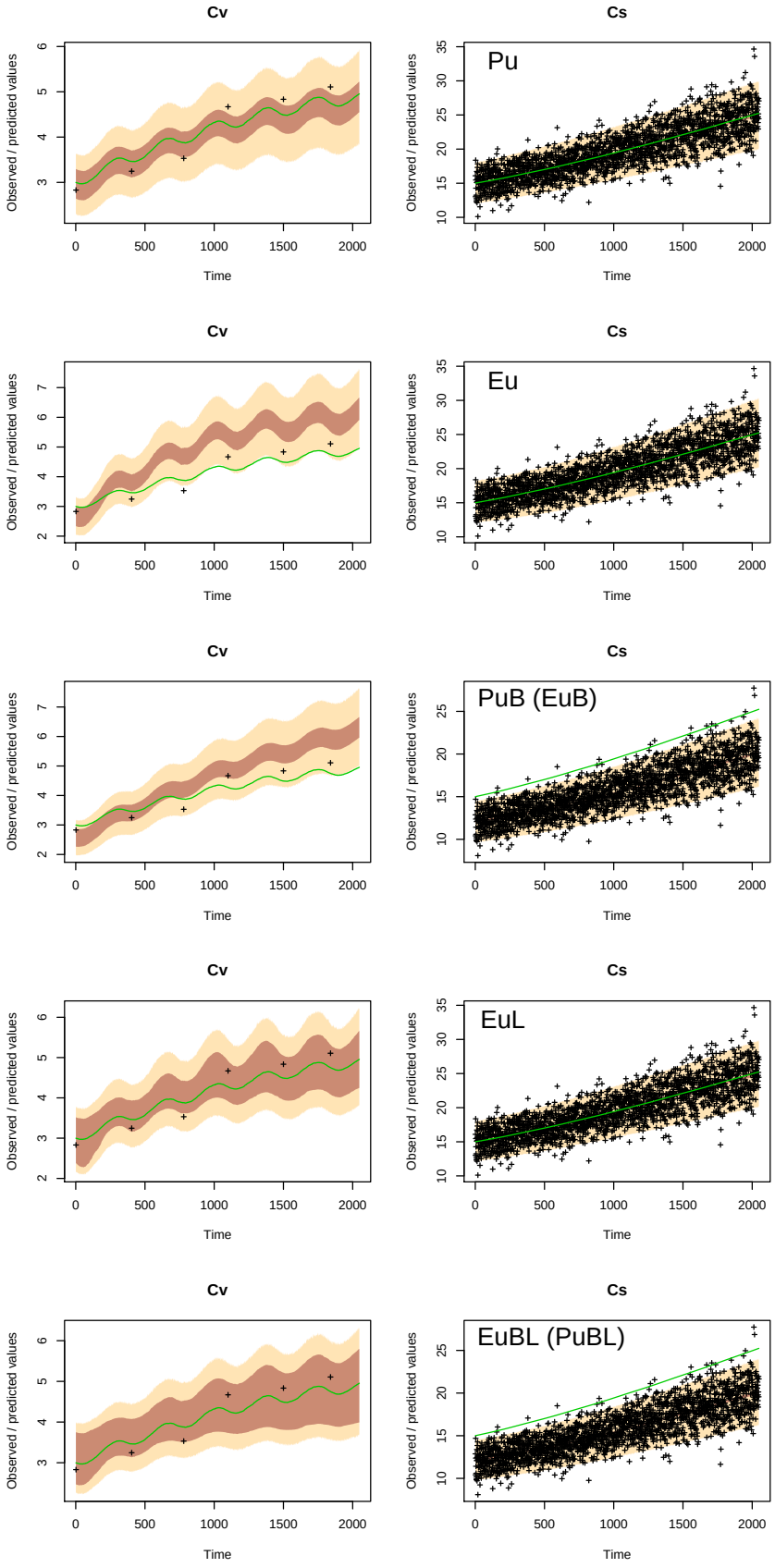


Figure 3: Timeseries plot.

3.3 Model with error and balanced data Eb

Here we create a known significant structural error in the model by effectively removing the root pool from the model see section (2.3.3). After calibration a number of parameters are now quite far away from their 'true' values. This is especially dramatic for tauV, which controls the turnover of vegetation, and is now lower, so that the rate of turnover of the vegetation pool has now more than doubled. The model error causes all the new carbon to be allocated to the vegetation pool only. The increased turnover rate tries to compensate for this error in the model. Hence, the large departure of the parameters from their 'true' values has the effect of 'absorbing' some of the influence of the model structural error. The result is that the model outputs have not changed significantly (see supplementary material) from the perfect model run. These results illustrate that model performance against calibration data can still be acceptable even when very signifcant model errors are present so long as changed parameters settings somewhat 'absorb' the influence of the error.

3.4 Model with error and unbalanced data Eu

We now combine the influences of unbalanced data and model error, investigated in the previous two sections, into the calibration. Looking first at the marginal parameter distributions after calibration there are changes versus the Eb calibration which are significant but not huge. In production, KEXT has increased and LUE has decreased slightly compensating for each other. Belowground, parameters Cs and tauS are now closer to their 'true' value than in Eb. Similarly aboveground, tauV is now closer to its 'true' value then in the Eb calibration. In general, the change in parameters to compensate for the model structural error is less than for Eb. Looking at the timeseries outputs for Cv and Cs, the model prediction is fine for Cs (and also NEE not shown) but drifts away significantly from the six vegetation measurements. This is the typical behaviour for calibrations with a large data imbalance, the sparsely measured parts of the system are ignored at the expense of the parts of the system with many observations. This calibration, along with the result from the previous two (Pu and Eb), make it clear that the model structural error is key in creating an issue when calibrating a model with a large imbalance in data.

3.5 Perfect model and balanced data with a multiplicative bias PbB

We now investigate the influence of data bias on the calibration. As presented in section (2.3.4), we create a multiplicative data bias by multiplying the soil carbon pool by 0.8. Similarly to Eb, parameters in the calibration do not all recover their 'true' values and hence 'absorb' the influence of data error. As might be expected this is most dramatic for the belowground parameters. The initial Cs parameter decreases significantly and tauS also decreases, increasing the turnover. This has the effect of decreasing the soil carbon pool to match the erroneous data. As before, these departures of the parameters from their 'true' value allows there to be a reasonably close match between the model outputs after calibration and the data (supplementary material).

3.6 Perfect model and unbalanced data with a multiplicative bias PuB

We now add the effect of unbalanced data to the calibration with the significant data bias. We look first at the parameter marginal distributions after calibration. In carbon production, KEXT is now larger than its true value increasing the carbon inputted to the system. This is counteracted by a lower LUE. Aboveground Cv is smaller and tauV significantly larger decreasing the turnover to the soil. This has the combined effect of passing on less carbon to the soil. Belowground, tauS is slightly closer to its true value than PbB, Cs has increased versus the PbB calibration pushing it back towards its true value. As the timeseries output plots show, the calibration is now somewhat ignoring the six vegetation observations in a similar way to what was found for the calibration with a model error. The main the 'effort' in the calibration is on matching the many erroneous soil carbon observations. These results show there can be issues calibrating with unbalanced datasets whether there is a model structural error or a significant data bias.

3.7 Model with error and unbalanced data with a multiplicative bias EuB

Now we combine the model structural error with the data bias and run the calibration with the unbalanced dataset. The two errors reinforce each other since the erroneous increase in the vegetation pool due to the missing root pool model error add to the issue of trying to match the erroneously low soil carbon observations. The additive effect of the two errors increase further the poor results that are found with unbalanced data with the model outputs after calibration even further away from the six Cv observations (supplementary material).

4 Diagnosing the issue

4.1 Comparing model output with virtual data as truth.

Moving on from identifying the issue in the previous section, here we develop a tool for helping to diagnose at what point and to what extent having unbalanced data in Bayesian calibration (BC) becomes an issue when models and data are imperfect.

This is done by running a number of calibrations with perfect and imperfect models where the quantity and imbalance of data used increases with each calibration. Here we chose an increasing power series of two (2^3,2^4 ... 2^{11}) for the increase in the quantity of calibration data; eight calibrations in all. In the balanced data BC case, quantities of NEE, vegetative carbon and soil carbon data included in the BC 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 for each of the eight calibrations. After running the calibrations the VSEM was rerun with the maximum a posteriori (MAP) vector and the RMS difference with the 'true' data was calculated and plotted (Fig. 4).

The figure shows broad similarity in results except for vegetative carbon when the model has an error and where there is an imbalanced in calibration data. In general, the RMS difference has a tendency to decrease as the quantity of data included in calibration increases. There is also a marked grouping of results with the perfect model getting closer to the data than the model with the error, as might be expected. 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. This is in marked contrast to vegetative carbon where RMS differences increase significantly as quantity of calibration data increases when the model has an error and when there is an imbalanced in calibration data. This increase in RMS difference for vegetative carbon occurs in tandem with the decreases noted already from NEE and soil carbon. This signature of increasing RMS difference for the low quantity data output versus the decreasing RMS difference for the high quantity can be used to diagnose when large imbalances in calibrations data with imperfect models and data start to become an issue. In this case, it appears after the quantity of data included in the calibration exceeds 32 but this will be different for each model, likelihood function and for each dataset used in calibrations.

4.2 Comparing model output against "obervations"

The diagnosis made in the previous section had the benefit of access to the 'true' data and a perfect model. Unfortunately this is never the case for real world ecological model calibrations. Therefore, here we have repeated the previous graph Fig.(4) with just the imperfect model and the imbalanced calibration, but with RMS differences now calculated against observations (NEE: 2048 points, vegetative carbon: 6 points, soil carbon: 2048 points) (Fig. 5). While there are clear differences in the RMS values versus the previous graph, as might be expected, the broad-scale signature of increasing RMS difference for vegetative carbon and decreasing RMS difference for NEE and soil carbon is retained. As before, this graph can be used to diagnose when the imbalanced in data is starting to interact with the erroneous model. In this case, as before, this occurs for a data quantity greater than 32.

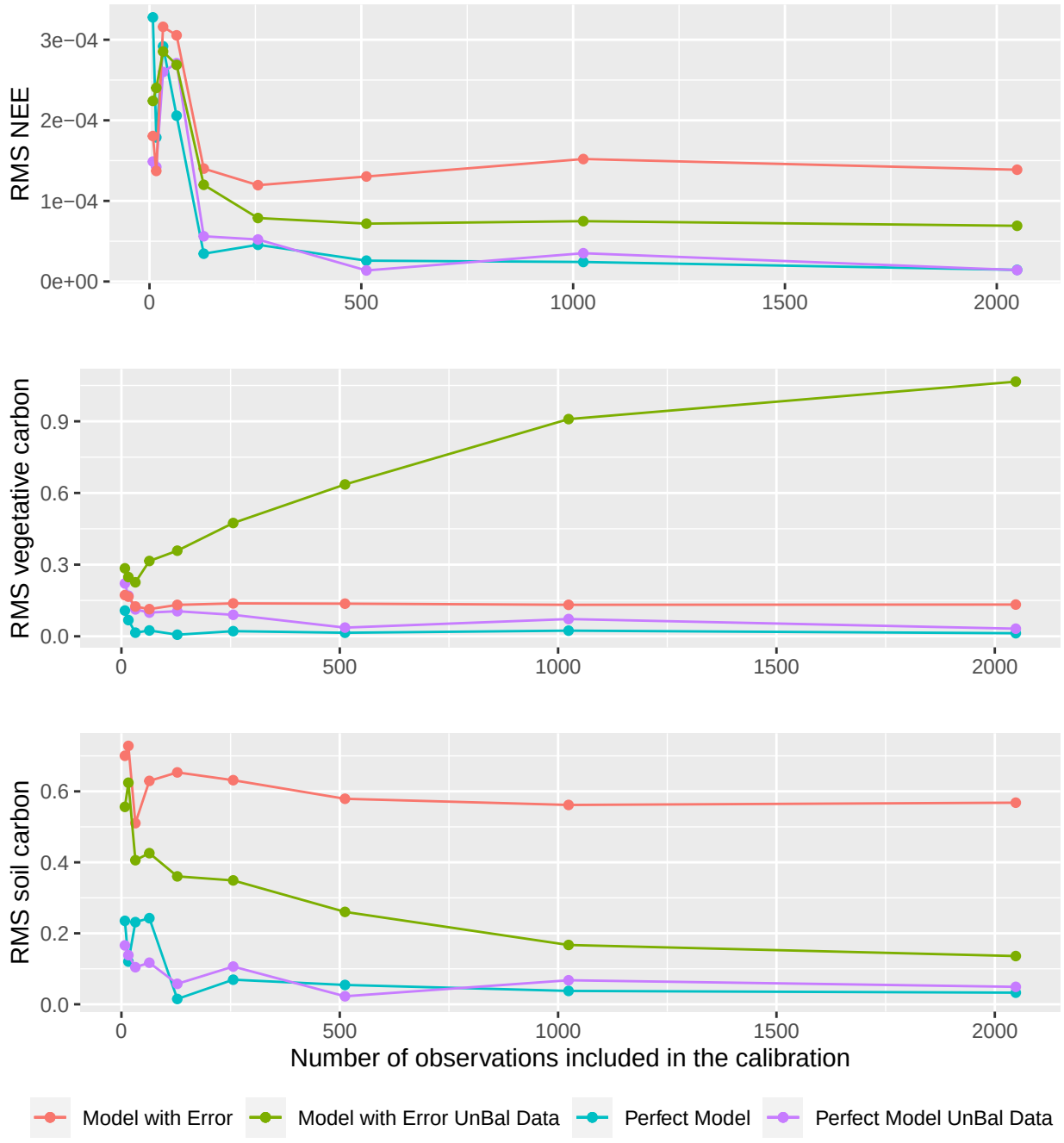


Figure 4: Each point in the three graphs (NEE, vegetative carbon, and soil carbon) represents the RMS difference between the VSEM model and the 'truth' run with different maximum a posteriori (MAP) vectors. The MAP vector at each point is obtained from a Bayesian calibration (BC) where the quantity of data included in the BC increases in a sequence along the x-axis following the exponentiation of base two. For the balanced calibration case (red and cyan) vegetative carbon data increases 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.

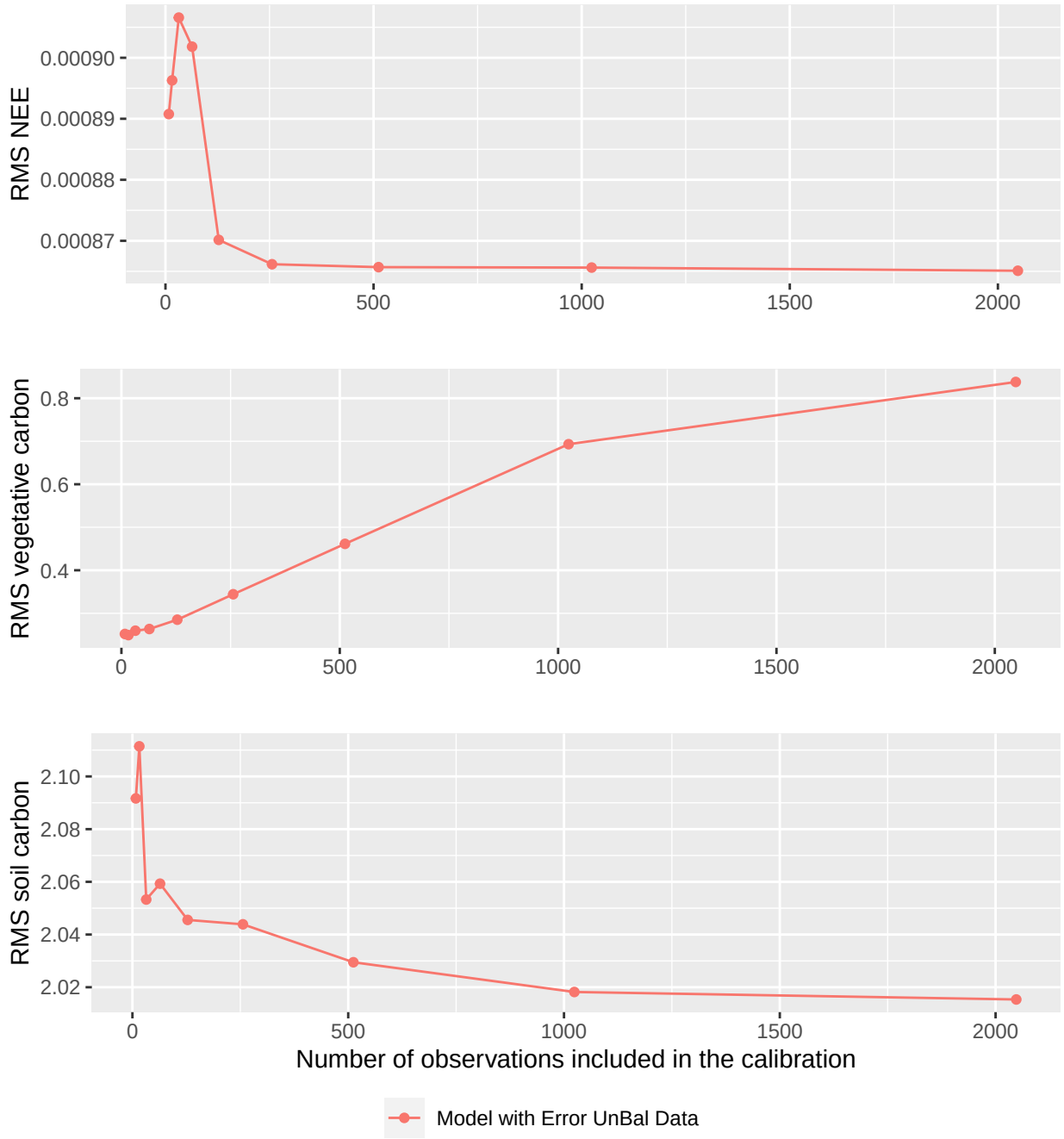


Figure 5: Each point in the three graphs (NEE, vegetative carbon, and soil carbon) represents the RMS difference between the VSEM model and virtual observations run with different maximum a posteriori (MAP) vectors. The MAP vector at each point is obtained from a Bayesian calibration (BC) where the quantity of data included in the BC for NEE and soil carbon increases in a sequence along the x-axis following the exponentiation of base two. The quantity of vegetaive carbon data is held fixed at six for all points in the graphs. The VSEM model used has a known error relative to the virtual observations that was derived from it.

5 Changes to the Likelihood to represent model and data errors

The results from section (3) demonstrate that the underlying issue with including unbalanced data in the calibration is not the imbalance itself but that there are significant model structural errors or data systematic biases or both effecting the calibration. Therefore, here we aim to introduce terms in the likelihood which represent our uncertainty about what these errors could be. This uncertainty exists and hence it needs to be represented. Otherwise, as we saw in section (3), the posterior uncertainty will be too small and parameters will find their greatest posterior probability far from their 'true' values, to try and compensate for the unrepresented errors in the calibration.

As presented in section (2.4) here we represent these model and data errors as simply as possible using additive and multiplicative terms in the calibration data for NEE, Cv and Cs. This introduces six new parameters to the calibration.

Before presenting the detailed results a general theme in Fig (1) is that introducing the new terms has significantly increased the posterior uncertainty in the parameters which makes it more likely that the marginal posterior distribution for the parameters includes the true value. In tandem with this, Fig (3) shows that including these new terms has increased the uncertainty and improved the fit to data considerably (EuL, EuBL).

5.1 Model with error and unbalanced perfect data with additive and multiplicative parameters to represent model error. EuL

Our first test using the new Likelihood is in the calibration with a significant model error and unbalanced data. To see the influence of the new terms we compare first marginal posterior parameter distribution with those for Eu Fig (1). Comparing with Eu the posterior marginal distribution of LUE is now larger and includes the true value. The uncertainty in KEXT, tauV, Cs and Cv has also increased so that these distributions are now closer to the true value. An outlier is tauS where the uncertainty has decreased but the distribution is now clearly further away from the true value. As presented already the main influence of the model error is to allocate too much new carbon to the aboveground vegetation pool, since there is no allocation now to the root pool. In this calibration KEXT is lower than in Eu this reduces the carbon produced and helps to counteract the effect of the error. For NEE to stay close to the data there must be a compensating drop in respiration and hence the soil turnover needs to be slower. A larger tauS helps with this and is counteracted by the distribution of modmultCs being centred at 0.9 (supplementary material) so that a higher tauS still allows for a good fit to the Cs data.

Looking at the output timeseries (Fig 3) the influence of the error hasn't been removed, as already noted for posterior parameter distributions, but there has been a significant improvement in the predictions, versus Eu, with the centre of the posterior now much closer to the 'truth' line. In addition, the uncertainty has increased so that 5 of the 6 data points are now inside the posterior confidence interval. The very simple multiplicative and additive terms introduced have not removed the influence of the error suggesting that more complex terms may be beneficial. There is however, a much greater sense that the sparse Cv data are influencing the calibration.

5.2 Perfect model and and unbalanced data with a multiplicative bias and additive and multiplicative parameters to represent the bias. PuBL

This calibration introduces the Likelikehood error terms to the previous calibration (PuB) with a large data bias and unbalanced data. Looking at changes in the posterior parameter distributions versus PuB (Fig 1), as noted previously for EuL, a key influence is that uncertainty is increased.

Some parameters (KEXT, LUE, tauV, Cs and Cv) are significantly closer to their true values. The centre of the tauS distribution is now considerably lower due to the lower modmultCs parameter slightly over compensating for the data error (supplementary material) however, the significantly larger uncertainty means that the true value is now included in the distribution.

Similarly to EuL, the vegetative carbon timeseries is also much improved (supplementary material). This shows that the extra terms are having a similar beneficial influence on this calibration as was found in EuL. For the Cs timeseries the model prediction without the Likelihood terms (not shown) is close to the truth line as might be expected.

5.3 Model with error and unbalanced data with a multiplicative bias and additive and multiplicative parameters to represent model error and the data bias. EuBL

In this final calibration, we combine the influence of the model error, the data bias and the unbalanced data now with the new terms in the Likelihood representing additive and multiplicative errors.

We compare against the posterior parameter distributions for the calibration EuB (Fig 1). Similarly to the previous calibrations (EuL and PuBL) the uncertainty has increased significantly for a number of parameters (KEXT, LUE tauV, Cs and Cv) parameter distributions so that in general they are now closer to parameter's 'true' value. The tauS distribution has changed in a similar way to PuBL as discussed above.

The timeseries predictions including the new Likelihood terms are a clear improvement over EuB (Fig 3 and supplementary material) the uncertainty has increased and the posterior predictions have a much better fit to the data especially for Cv. As for PuBL the Cs timeseries the model prediction without the Likelihood terms (not shown) is close to the truth line as might be expected.

6 Discussion

6.1 Unbalanced data in Bayesian calibration: Identifying the issue

Our aim was to identify as clearly as possible the root cause behind why including unbalanced data in Bayesian Calibration (BC) can cause challenges. Firstly, we demonstrated that there was no issue with including very unbalanced data in the BC if the data had unbiased errors and the model was perfect. So there is no intrinsic issue in using very unbalanced data in BC. The issues that we often find must be caused by something else. Next, we introduced a very significant model error (removing an important model process) but we initially kept the quantities of data balanced. The model predictions after BC were close to the data, which is typically what we find when using balanced data in BC. In a 'real world' calibrations, with data as measurements taken from the environment, we do not know the extent of the model error present nor the best or true settings of the parameters, so this BC with balanced data would be considered a successful calibration. In our artificial BC we had access to information that we would not typically have because we know the true parameter settings. Given this, we found that after BC 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 or model structural error and has no means to change the structure of the model so model-data difference is reduced by solely by the parameters departing significantly from their true values. In this way, the calibration 'absorbs' the model error into wrong settings of the parameters so that model differences against the data used in the calibration are reduced. Other outputs from the model may be very poor but we have no data available to assess this. Therefore, in this 'balanced data' calibration the model error gave problems in the BC but these issues would normally be hidden from us and we would deem the calibration a success. It is normally only when we include unbalanced data in the calibration that the model predictions against the more sparse calibration data are poor that we identify an issue in the calibration and potentially wrongly perceive that it is an issue with the unbalanced data. Indeed, while the model predictions were poorer after calibration with the unbalanced data, the parameters were if anything closer to their true values and less confidently wrong. This was because the erroneous model was fitted to fewer data points and from the perspective of the calibration it was not so important that the model-data difference was larger for the output with fewer observations present. Therefore, some parameters need not depart so greatly from their true values to reduce model-data difference for outputs with plentiful data. These idealised calibrations demonstrate cleanly that the underlying issue with calibrating models with unbalanced data is not the unbalance in the data, but that

models have structural deficiencies that often remain hidden when we calibrate with balanced datasets but whose influence is only seen in poor predictions after calibration with unbalanced datasets. The root issue identified here is the presence of the model error not any imbalance in the calibration data.

Further, we repeated the above analysis but now included a large bias in the calibration data rather than an error in the model. We found very similar results to those with the model error because 'from the perspective of the calibration' what we have introduced is just a slightly different model-data difference that as before is reduced in the calibration by setting the parameter values away from their true values so that the data bias is effectively absorbed by erroneous parameter values. Here we have demonstrated that very similar issues occur with unbalanced data whether we have a model error or a data bias because in the likelihood and hence the calibration, the issue is effectively the same. The likelihood cannot distinguish whether the model-data difference is due to model systematic error, data bias or poorly set parameters and the only means to reduce the difference is to change the parameter settings. As for the previous case, with model error, the underlying data bias can remain hidden and it is possible to erroneously perceive that the issue in the calibration is due to unbalanced data rather than a systematic bias in the data.

6.2 Diagnostic tool (Fig. 5)

In the virtual experiments that we have presented here it is relatively easy to perceive what is happening because we have access to the true parameter settings and also to the 'true' system timeseries. In a 'real-world' calibrations it can be much harder to identify why a calibration is going wrong. Our aim in developing a diagnostic tool was firstly to identify the characteristic behaviour or signature that model or data errors or both are causing issues when calibrating with unbalanced datasets. We first used the privileged access that we had to the true system and then we went on to develop a tool that could be used in a 'real-world' calibration. We first illustrated with the perfect model (Fig. 4) that starting with a balanced calibration dataset that RMS output error goes down for all the model outputs when the quantity of data in the calibration increases, increasing the imbalance. In contradistinction when the model error is present (Fig. 5), the RMS output error increases as the quantity of data in the calibration increases (increasing the imbalance) for the sparse data output in tandem with the RMS output error decreasing for the plentiful data outputs. This is the signature behaviour that demonstrates and hence diagnoses the influence of the model discrepancy (or data bias) on the calibration. Indeed, when we repeat these set of eight calibrations in a more 'real-world' setting where the RMS is calculated against the observational data rather than the 'true' data we see the same signature. That is to say starting with a balanced calibration and adding more observations, thus increasing the imbalance, the predictions of the plentifully observed outputs improve whilst the poorly observed outputs get worse. In addition, to the signature curve showing the presence of the issue the curves also can 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.

6.3 Reweighting the calibration data to restore balance

We argued in the introduction that using ad-hoc methods such as thinning the calibration data to give a more balanced dataset was the wrong approach. The idealised experiments that we have conducted in this study provide another reason to avoid ad-hoc data reliability changing methods. In general, it is much preferable to 'treat' the underlying cause of a problem rather than try and mitigate the symptoms. Therefore, we should deal directly with model and data errors rather than address 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 unfortunately, it is only possible to achieve this to some extent and we are left with the need to continue to make useful predictions with our imperfect models informed by imperfect data. Models will never be a perfect representation of the ecological system. Also, there are known issues with observational data for example eddy covariance data doesn't close the energy budget (ref?) so it is not possible to fully match such data with a model that conserves energy. Fortunately, making useful predictions with imperfect models is not so problematic as long as there is a good estimation of the model reliability. That is to say that we can accurately quantify the uncertainty in the model predictions. In BC we infer the posterior uncertainty in the model parameters but as we illustrate here it is also very important to quantify the uncertainty due to model structural errors and

data systematic biases. We also hope to avoid the overfitting of the model parameters in 'absorbing' the errors present in the model and data which leads to the erroneous calibrations that we have illustrated herein.

6.4 Illustrating the benefits of including terms into he calibration that represent model and data error

For the reasons given above, in the final section of this study we included very simple linear terms in the BC to represent our uncertainty about model structural errors and data systematic biases. As identified, the underlying issue in BC with model and data error present is that the only means to increase the fit of the model predictions and the data in the likelihood is by pushing the model parameters away from their true or best value and if there is sufficient data then the posterior joint distribution will be confidently wrong. When we have access to balanced calibration datasets then the model predictions may fit well with the data so it is often only when the calibration includes unbalanced data that we notice issues because we see a poor fit to the data. Our main purpose in introducing the new terms in the likelihood is to recognise our uncertainty about model and data error to avoid a over confident fit to the wrong parametrisation. Therefore, what we would hope to see as a result of adding such terms would be that we are less confidently wrong and hence that our uncertainties are large enough to include the true system. Our results with very simple terms largely bear this out with the main effect being to increase the uncertainty in the joint posterior parameter distribution. This make it much 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. The simple approach used here in these idealised calibrations are not offered as solutions that will necessarily work well with real models and data in BC. Nevertheless, as in all modelling we advocate beginning with simple approaches as we have followed here and only complicating the modelling should that be necessary. The simple terms used here did make a significant improvement to the model predictions but were only partially successful in recovering the true parameter setting. In particular, the soil turnover parameter was pushed further away from the true value suggesting that further work is required to improve the discrepancy model even in this very idealised situation. Our purpose here is illustrative and to motivate greater efforts to be made in model discrepancy modelling rather than to provide guidance on specific solutions which in any case will be very model and context specific. Therefore, it doesn't make sense to pursue more complex modelling further within this study. Indeed, the topic of identifying and creating good statistical models of model discrepancy (and data bias) is not straightforward, is being actively researched in its own right (refs) and is outwith the remit of this study. Our purpose here is merely to illustrate the kinds of benefits that can be expected when such terms are introduced. One possible avenue would be to pursue...

(Include refs/discussion on more complex model discrepancy approaches eg GP, Oberpriller paper?)

7 Supplementary material

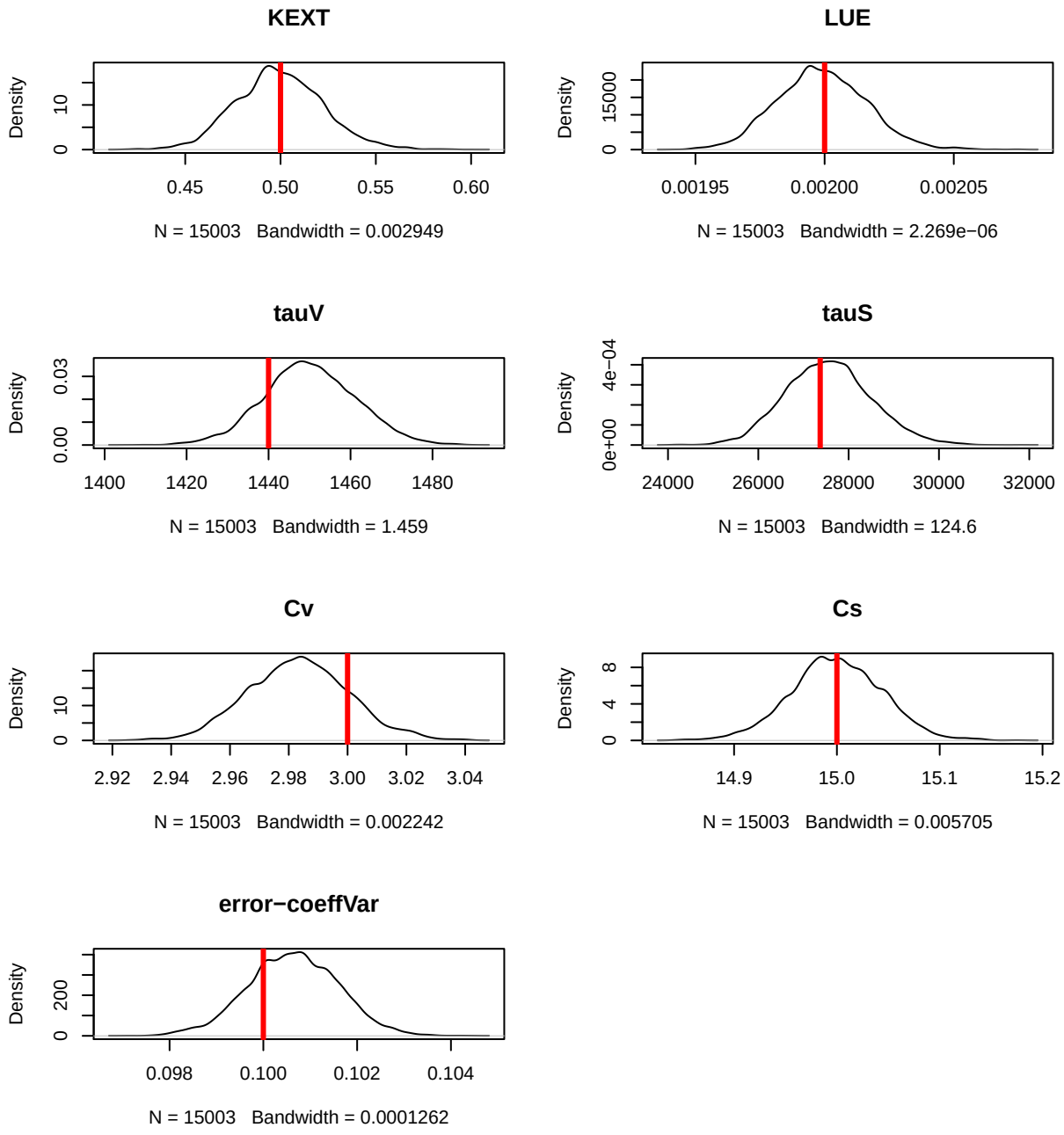


Figure 6: Perfect model, balanced data (NEE, Cv, Cs: 2048 obs). Marginal posteriors distribution of model parameters and intital states. The red line marks the 'true' parameter values.

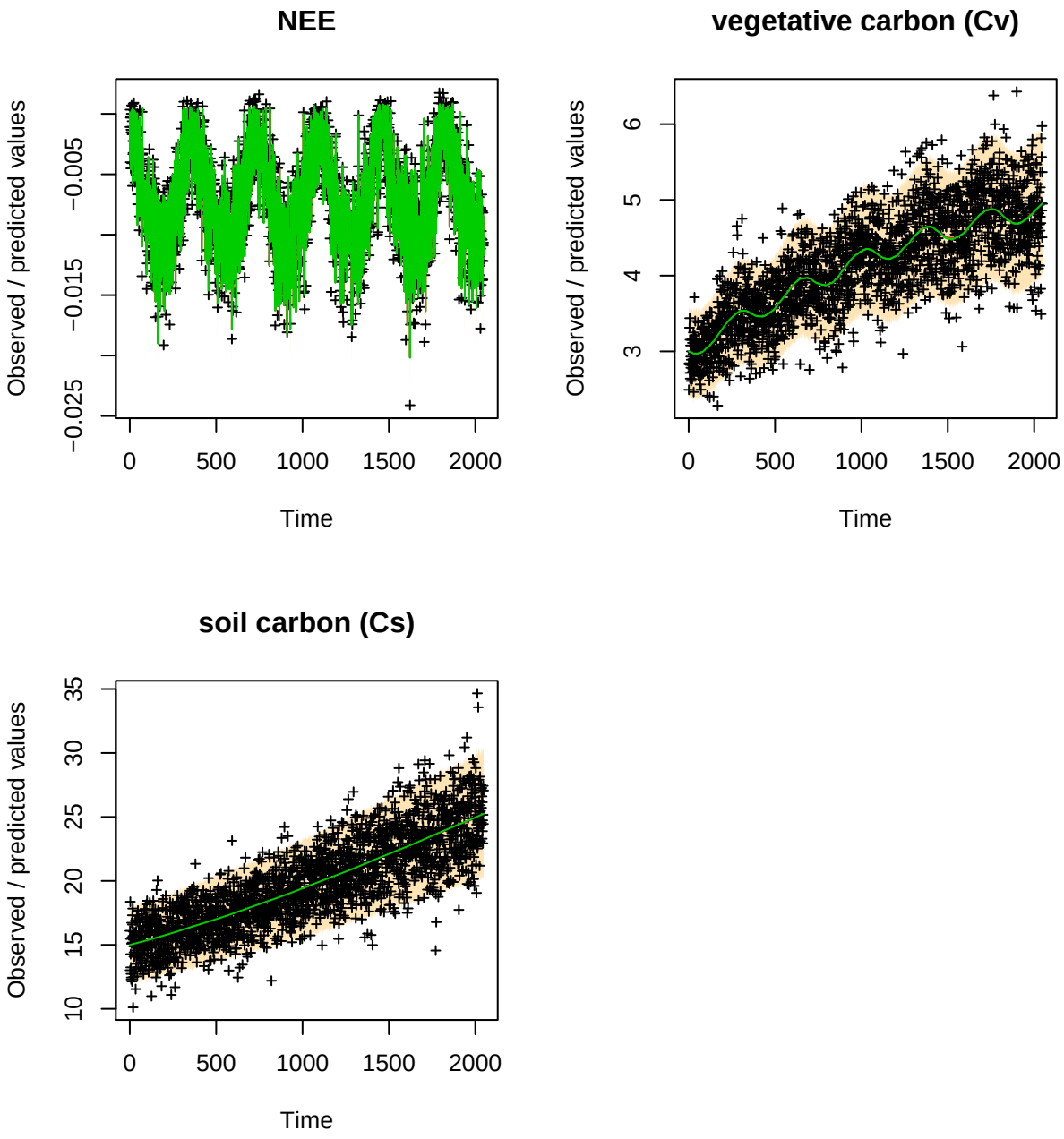


Figure 7: Perfect model, balanced data (NEE, Cv, Cs: 2048 obs). Observations included in the calibration marked with a '+'. Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

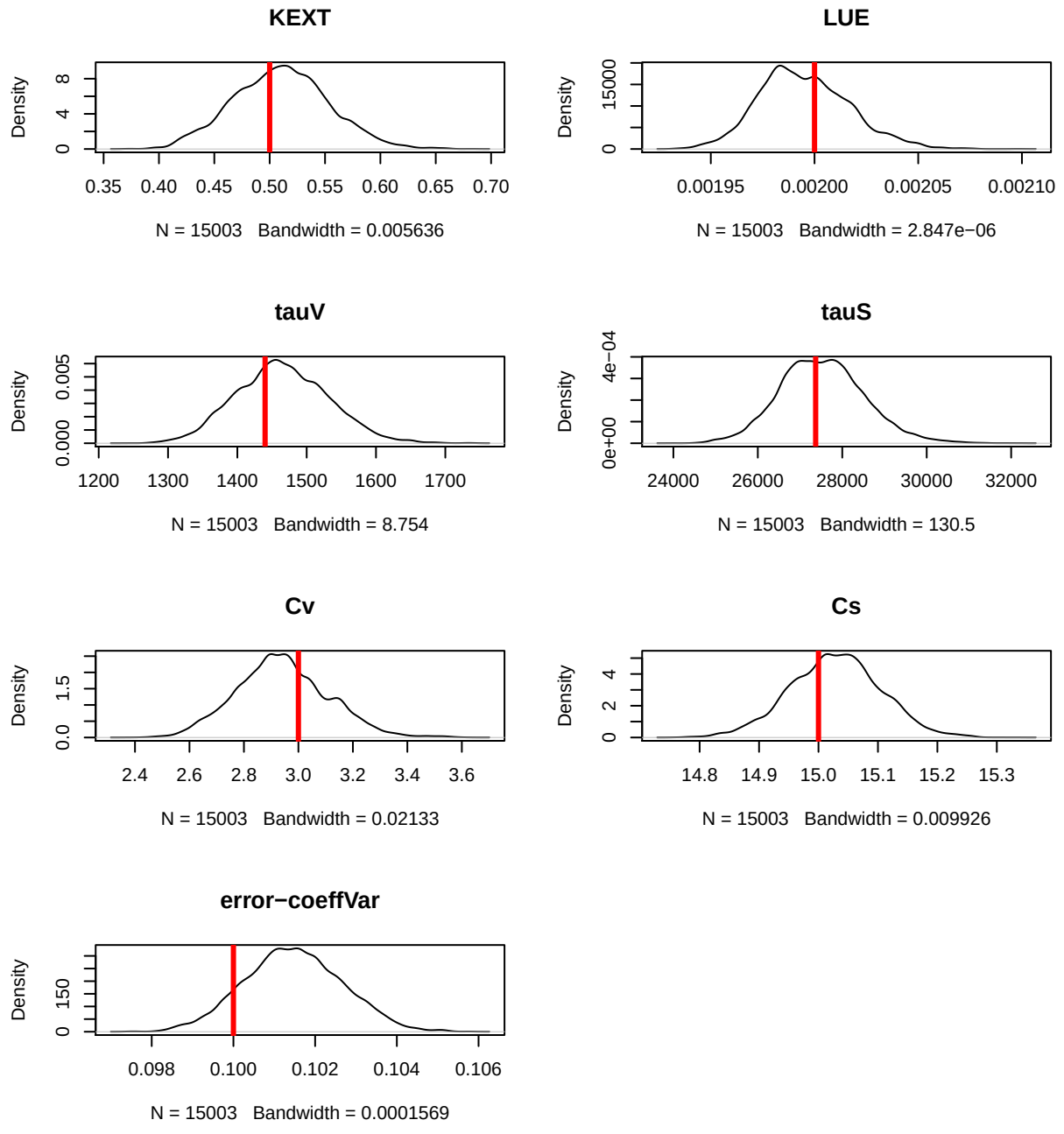


Figure 8: Perfect model, unbalanced data (NEE, Cs: 2048 obs, Cv: 6 obs). Marginal posterior distribution of model parameters and intital states. The red line marks the 'true' parameter values.

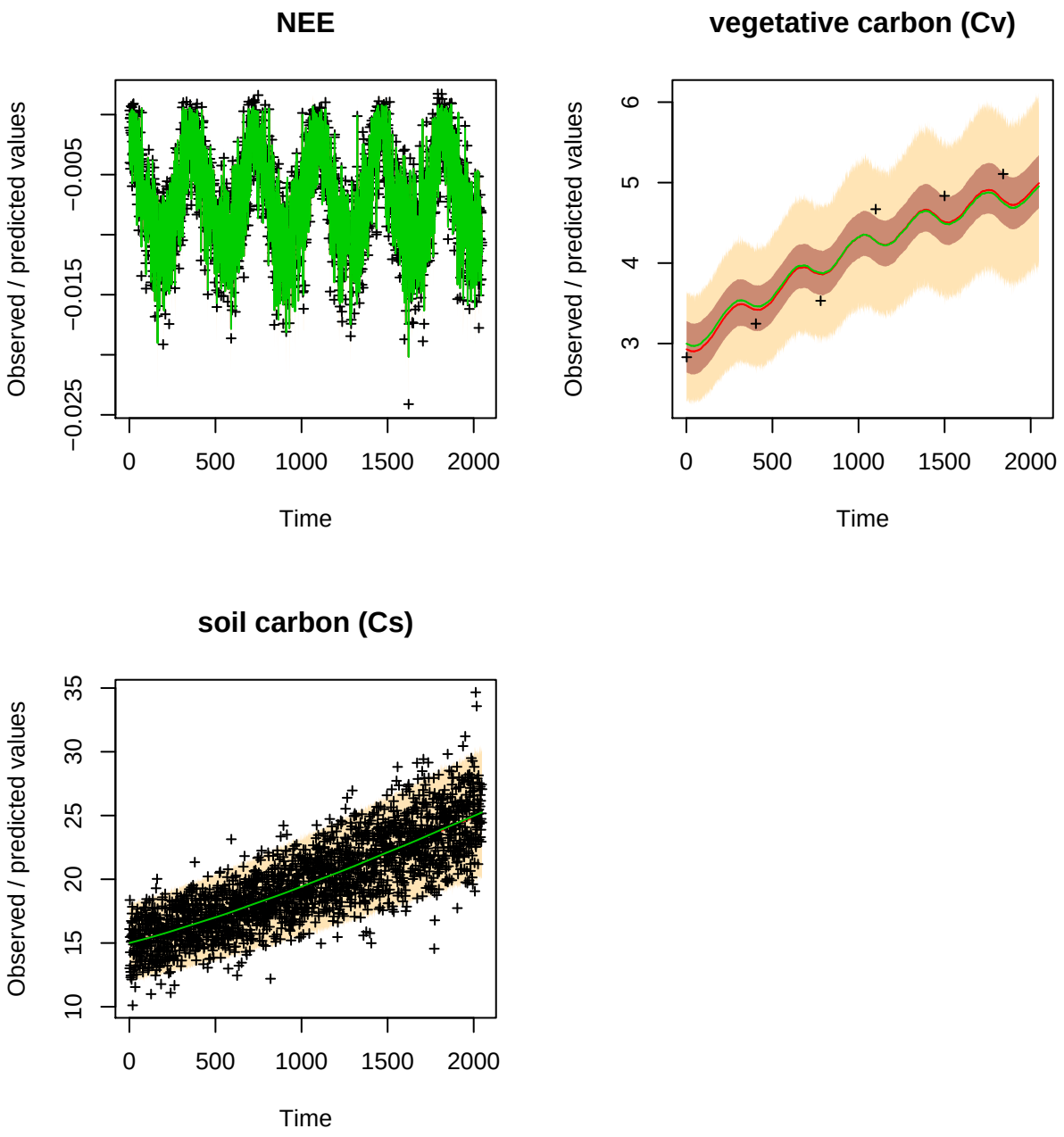


Figure 9: Perfect model, unbalanced data (NEE, Cs: 2048 obs, Cv: 6 obs). Observations included in the calibration marked with a '+'. Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

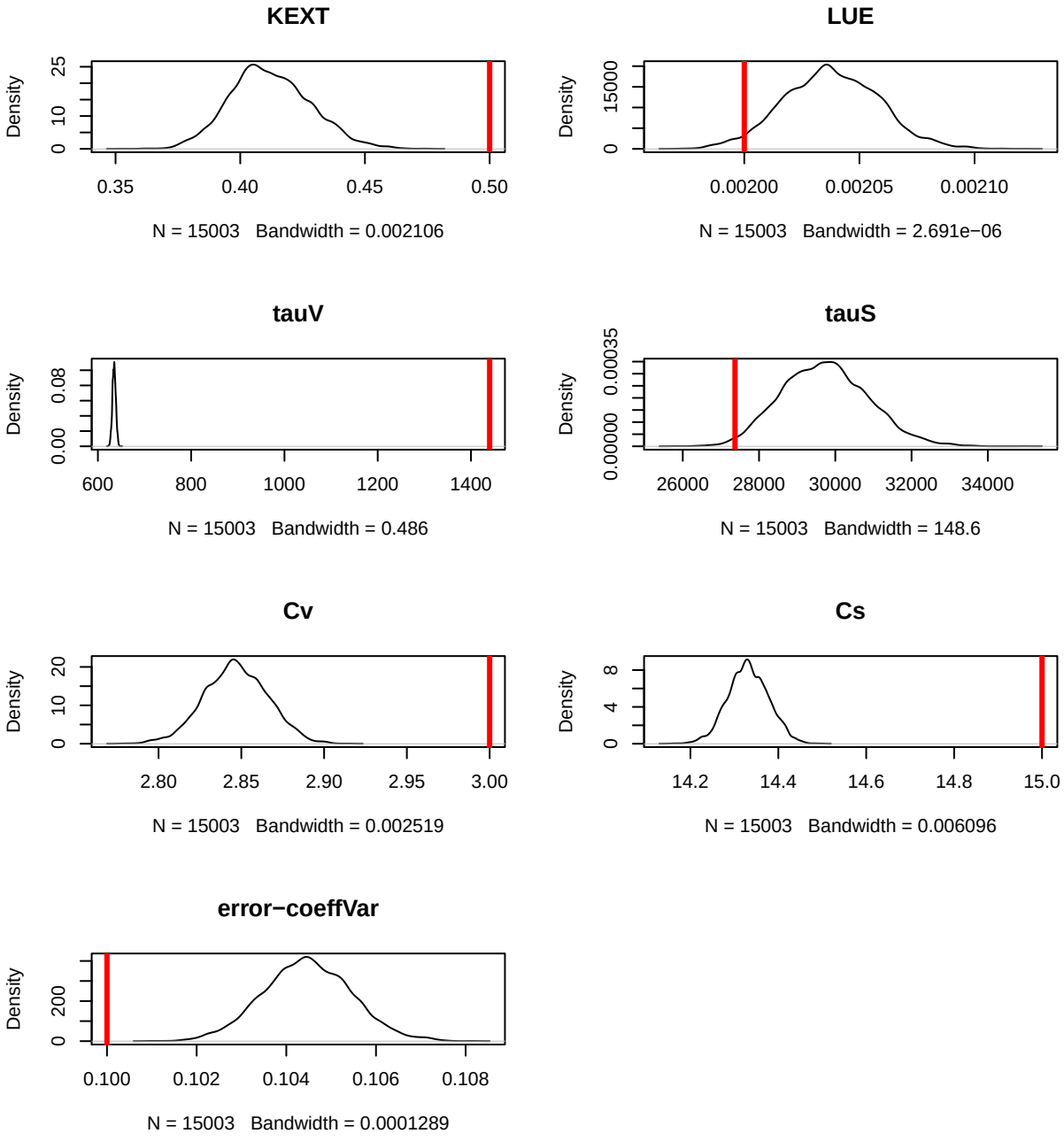


Figure 10: Model with error, balanced data. Marginal posterior distribution of model parameters and intital states. The red line marks the 'true' parameter values.

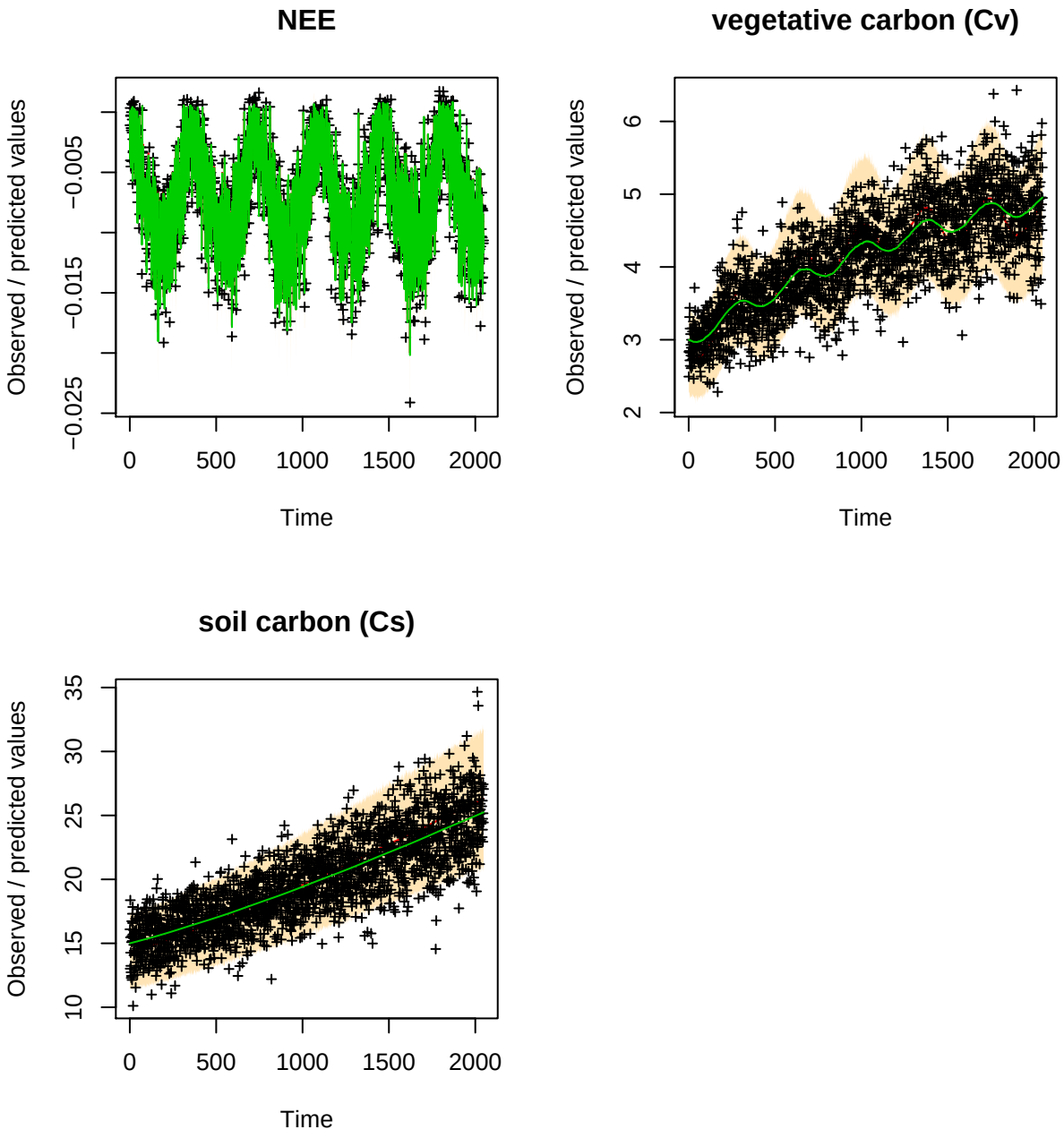


Figure 11: Model with error, balanced data. Observations included in the calibration marked with a '+'. Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

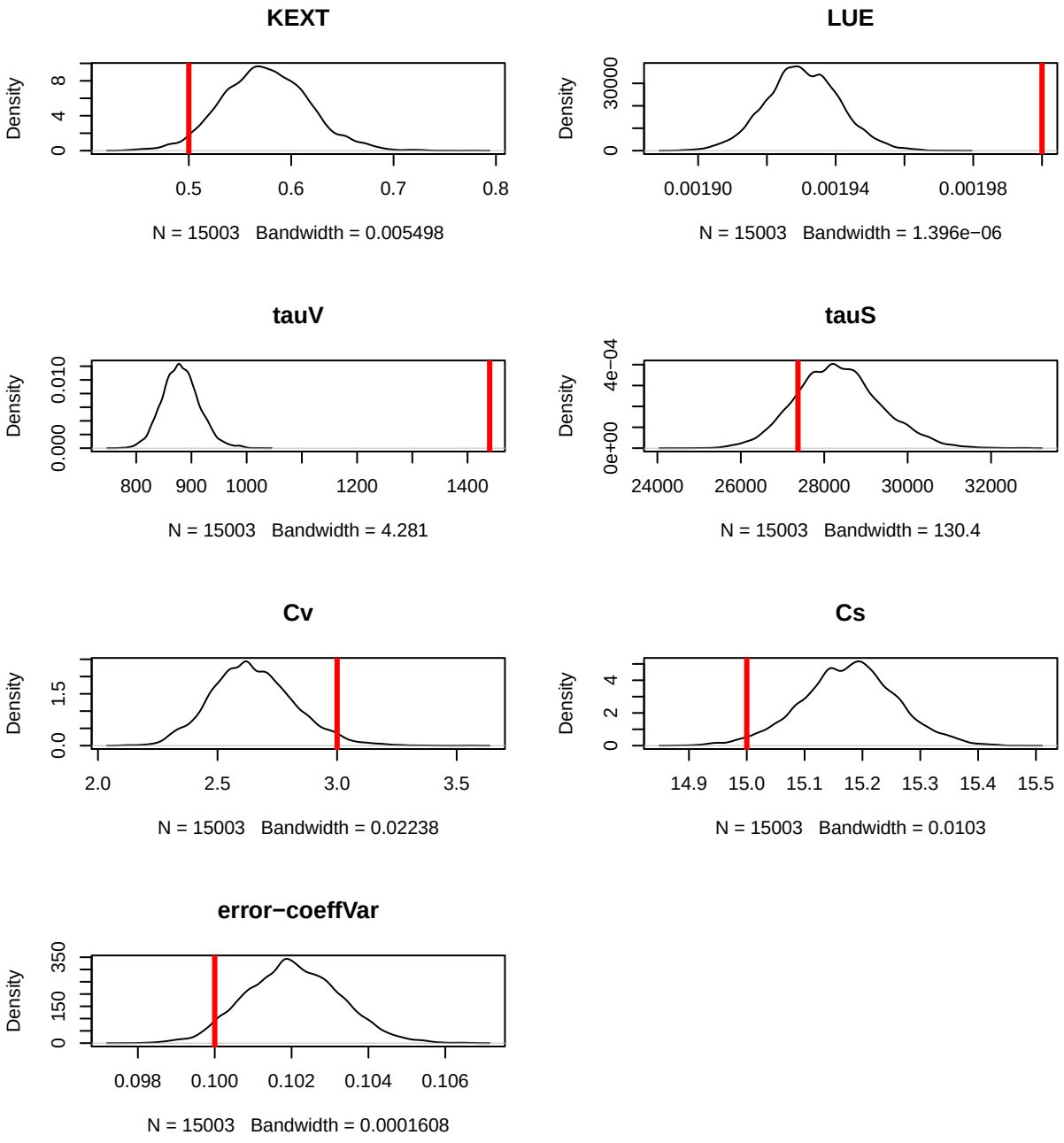


Figure 12: Model with error, unbalanced data (NEE, Cs: 2048 obs, Cv: 6 obs). Marginal posterior distribution of model parameters and intital states. The red line marks the 'true' parameter values.

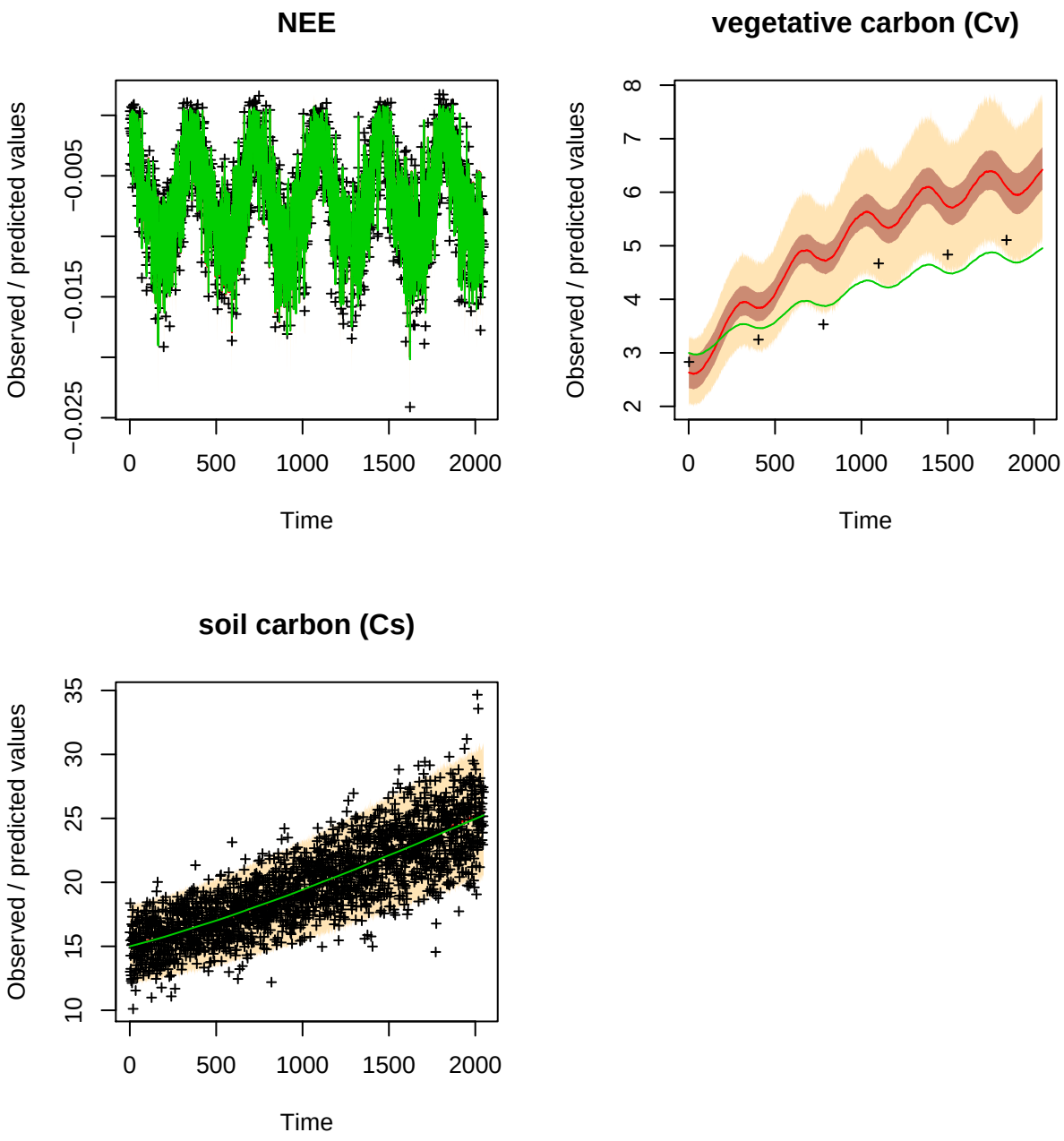


Figure 13: Model with error, unbalanced data (NEE, Cs: 2048 obs, Cv: 6 obs). Observations included in the calibration marked with a '+'. Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

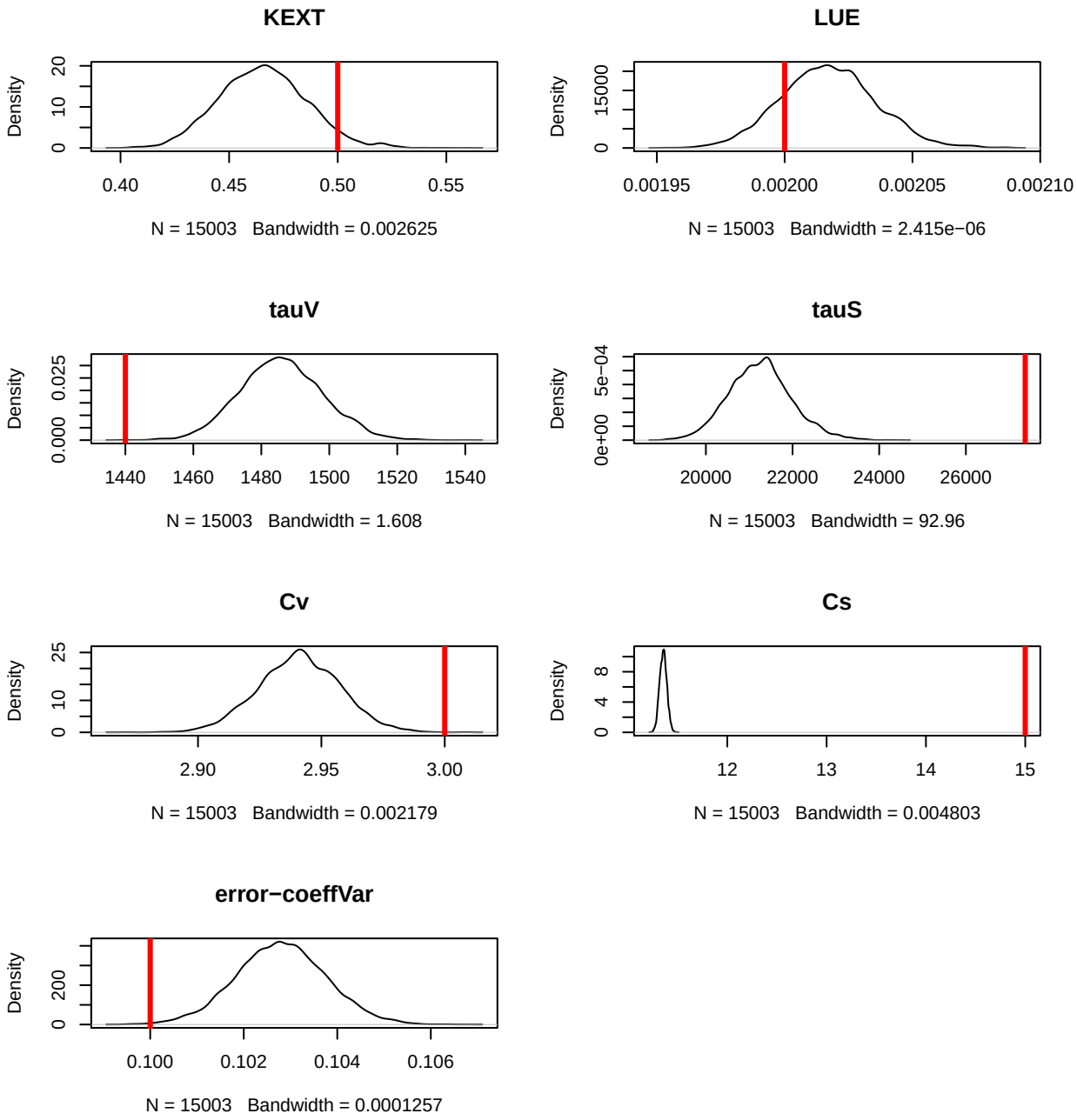


Figure 14: Perfect model and balanced data with a multiplicative bias. Marginal posterior distribution of model parameters and intital states. The red line marks the 'true' parameter values.

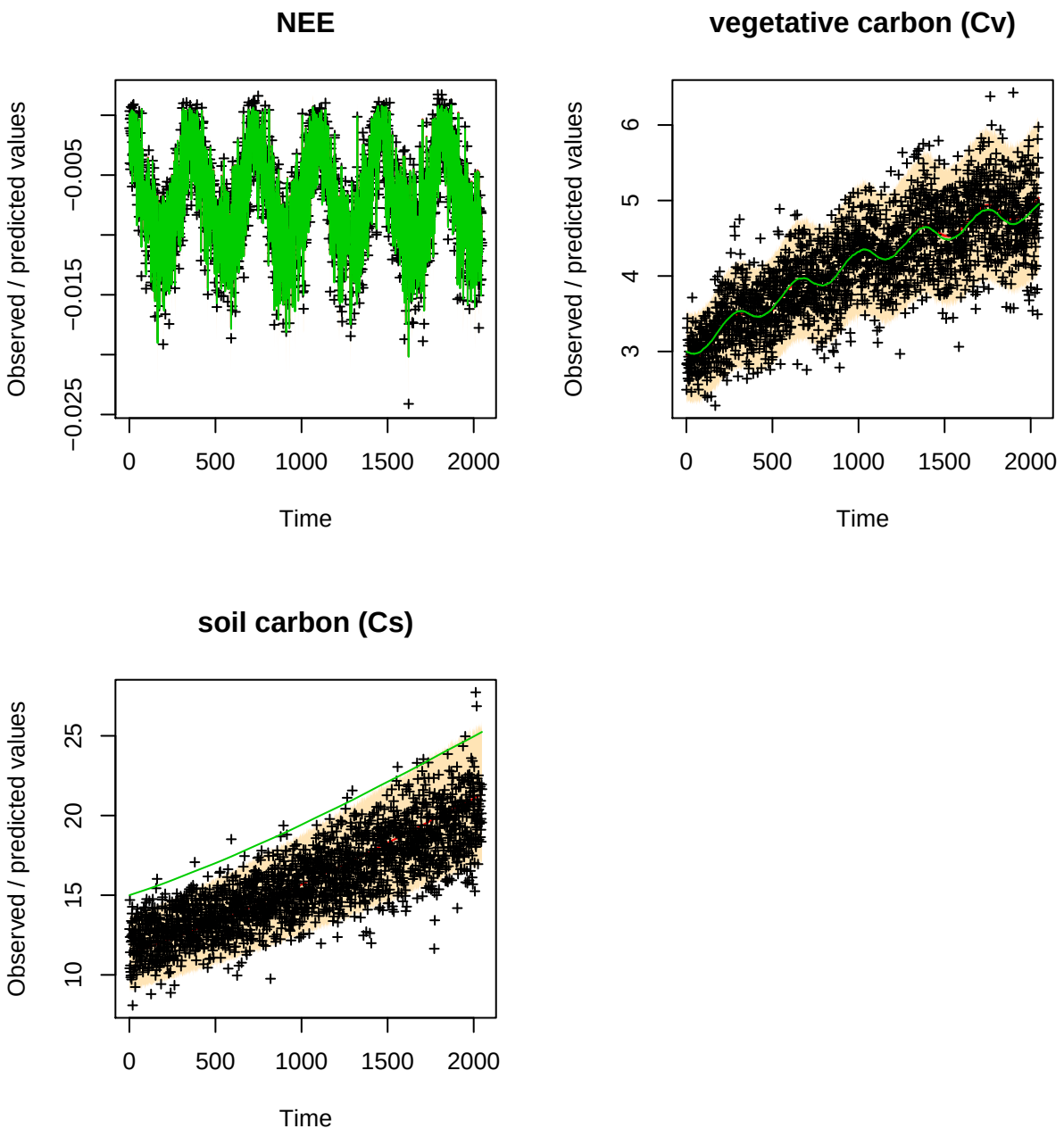


Figure 15: Perfect model and balanced data with a multiplicative bias. Observations included in the calibration marked with a '+'. Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

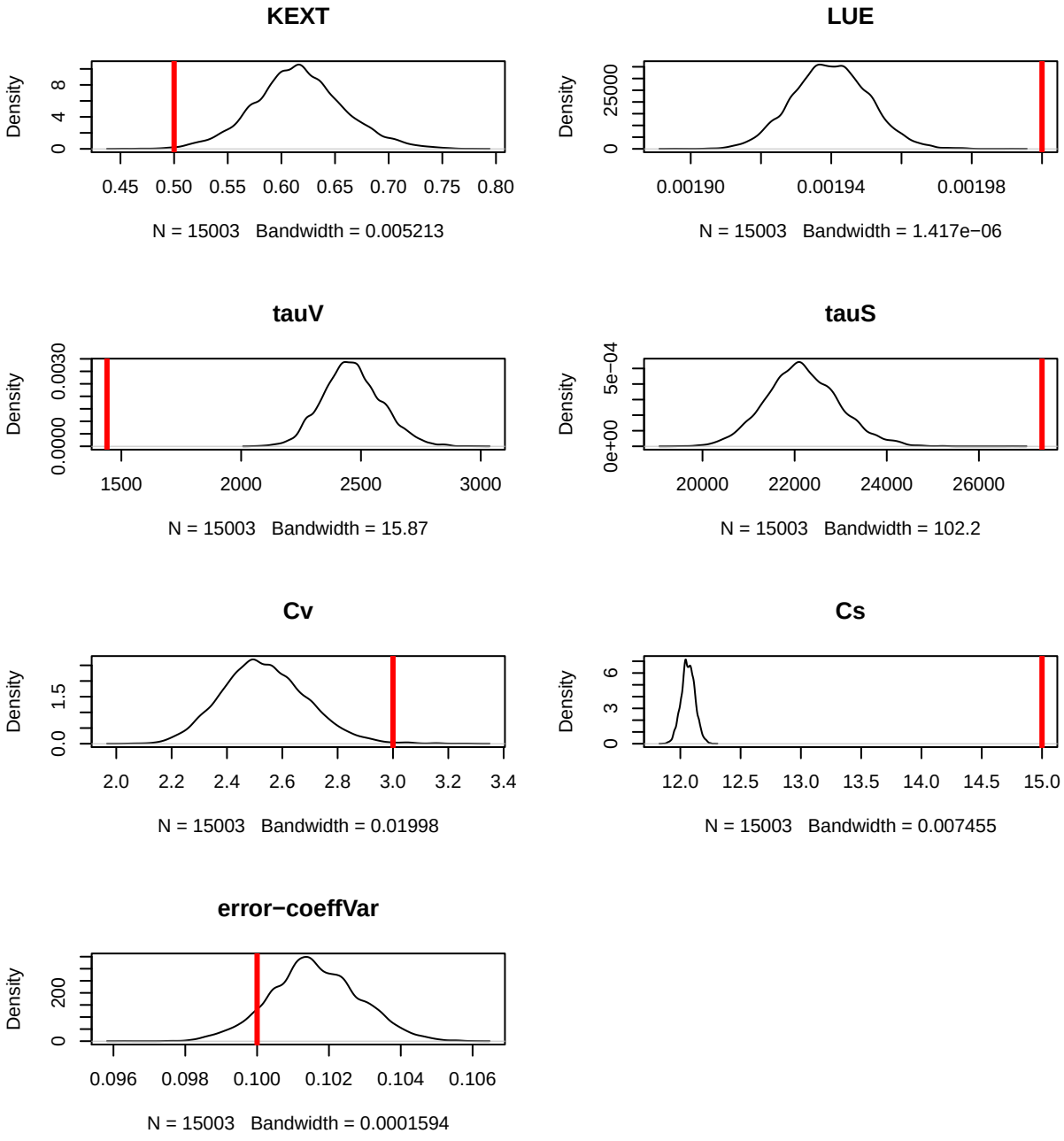


Figure 16: Perfect model and unbalanced data with a multiplicative bias. Marginal posterior distribution of model parameters and intital states. The red line marks the 'true' parameter values.

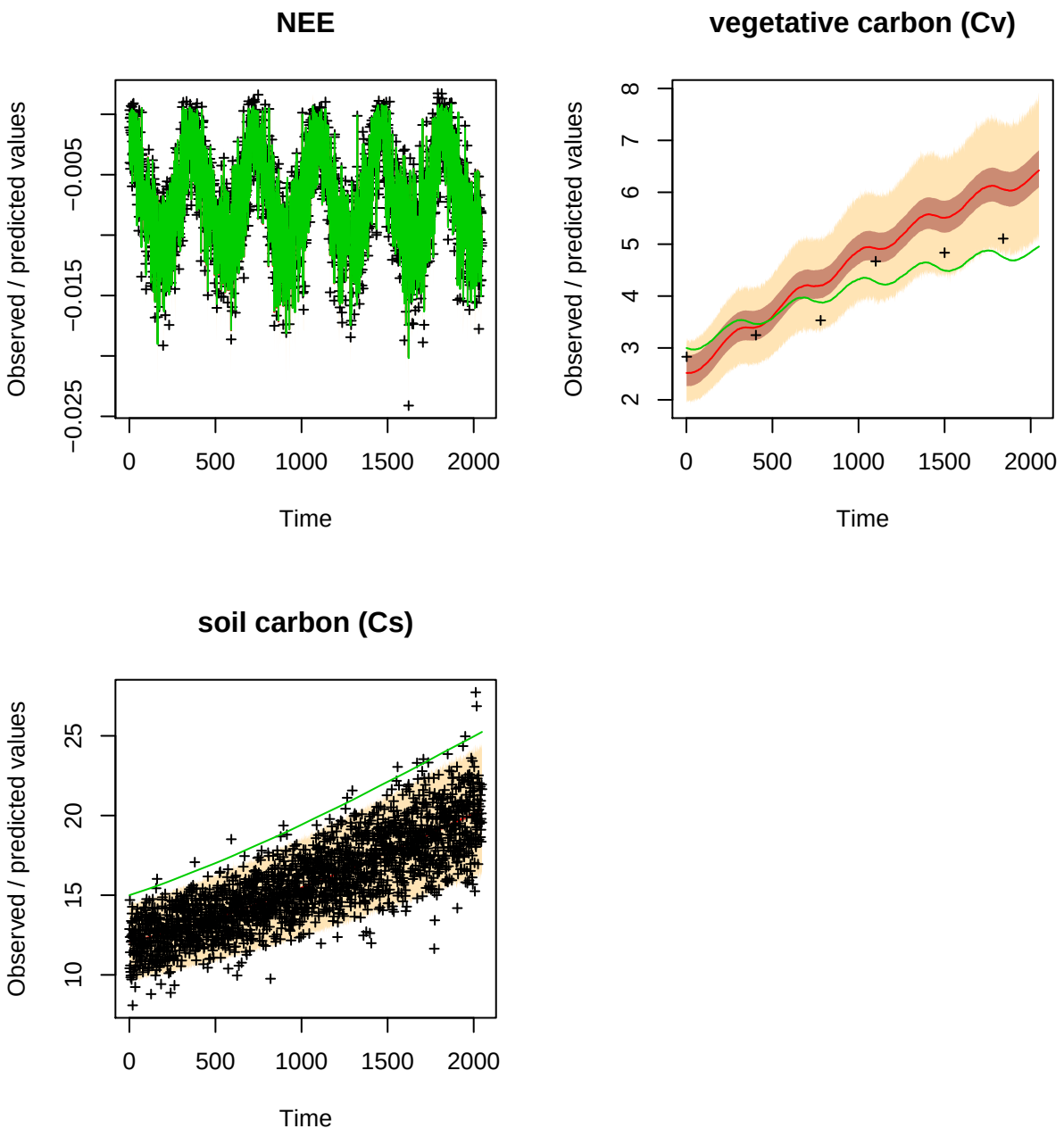


Figure 17: Perfect model and unbalanced data with a multiplicative bias. Observations included in the calibration marked with a '+'. Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

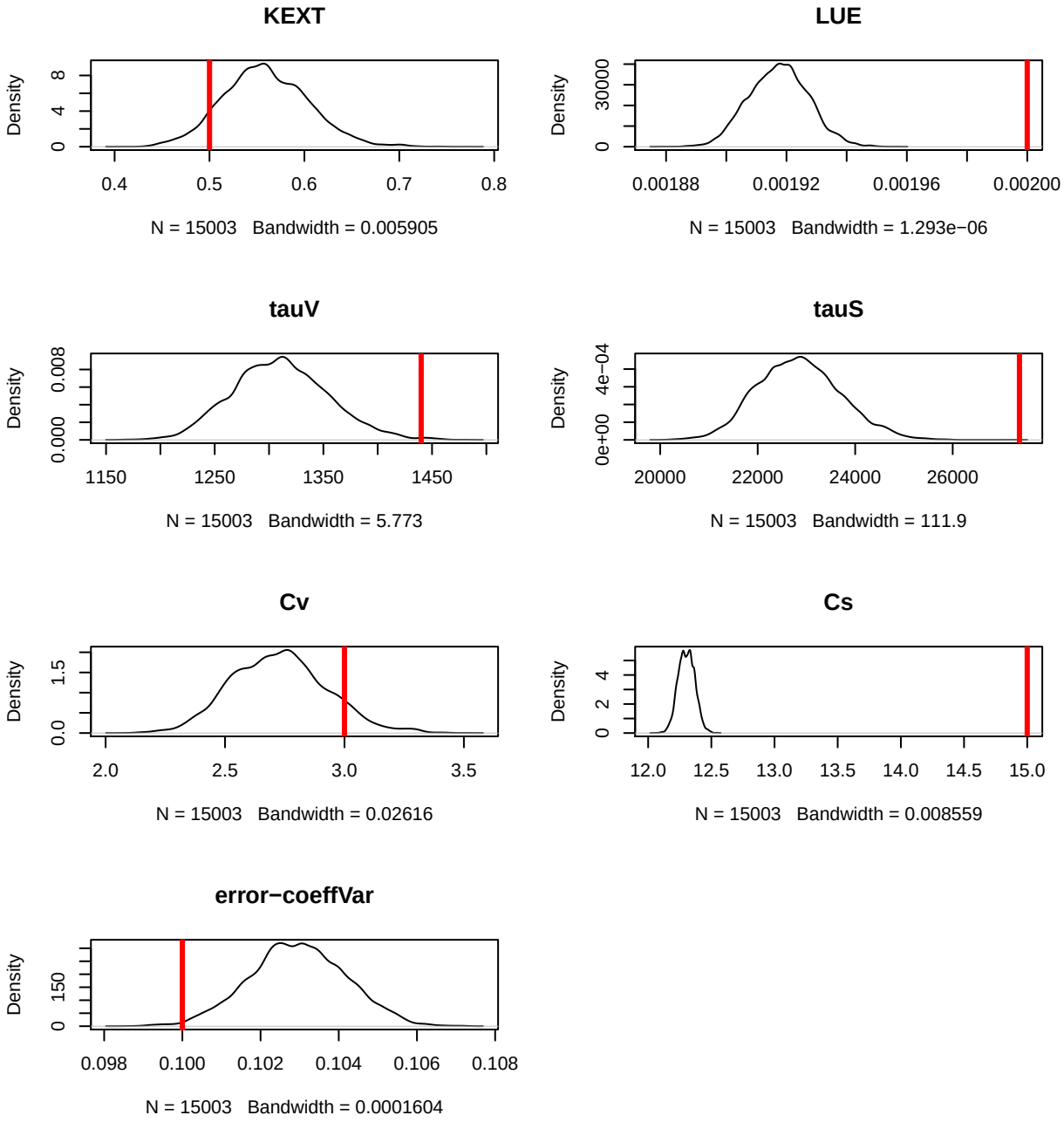


Figure 18: Model with error and unbalanced data with a multiplicative bias. Marginal posterior distribution of model parameters and intital states. The red line marks the 'true' parameter values.

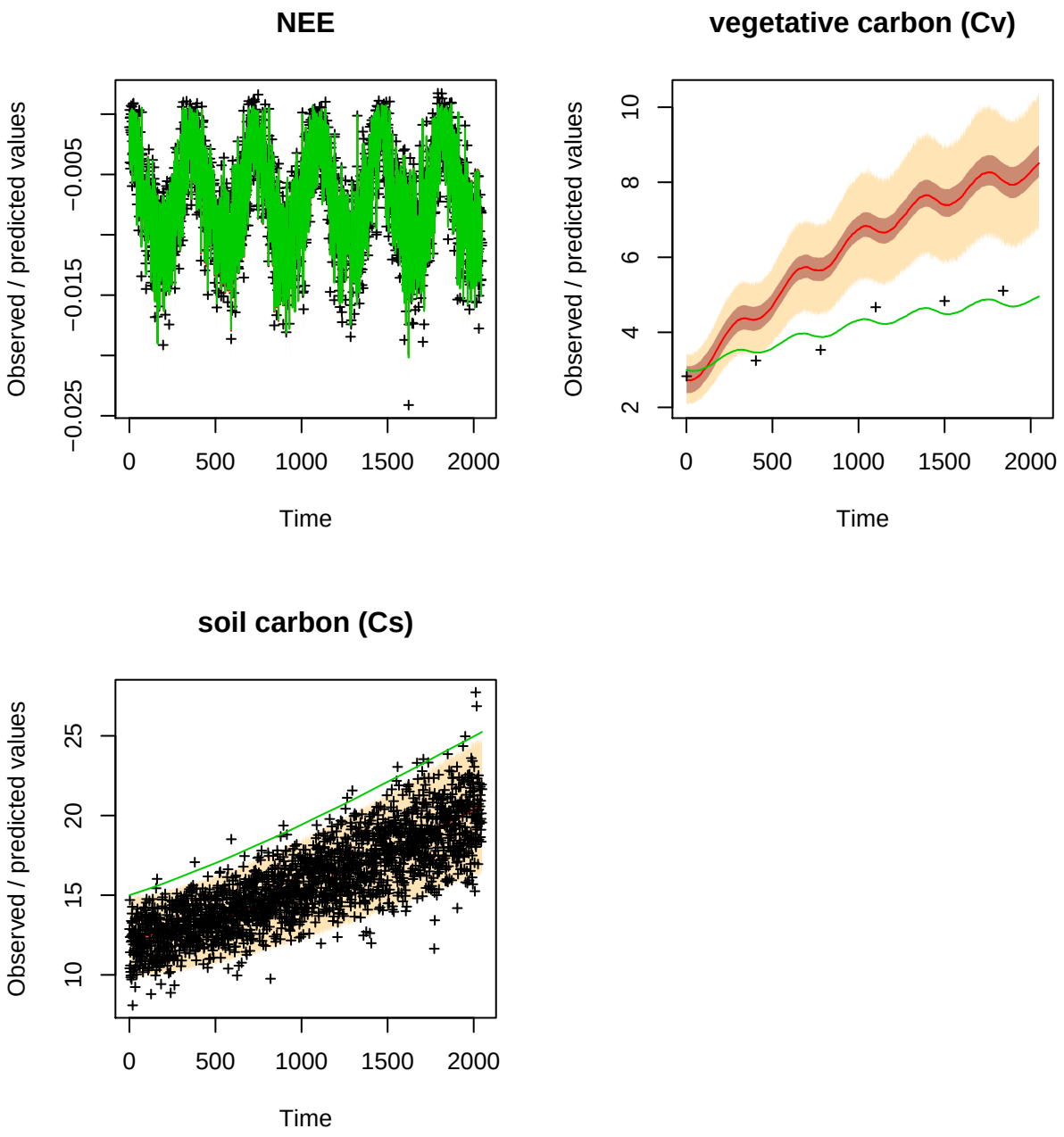
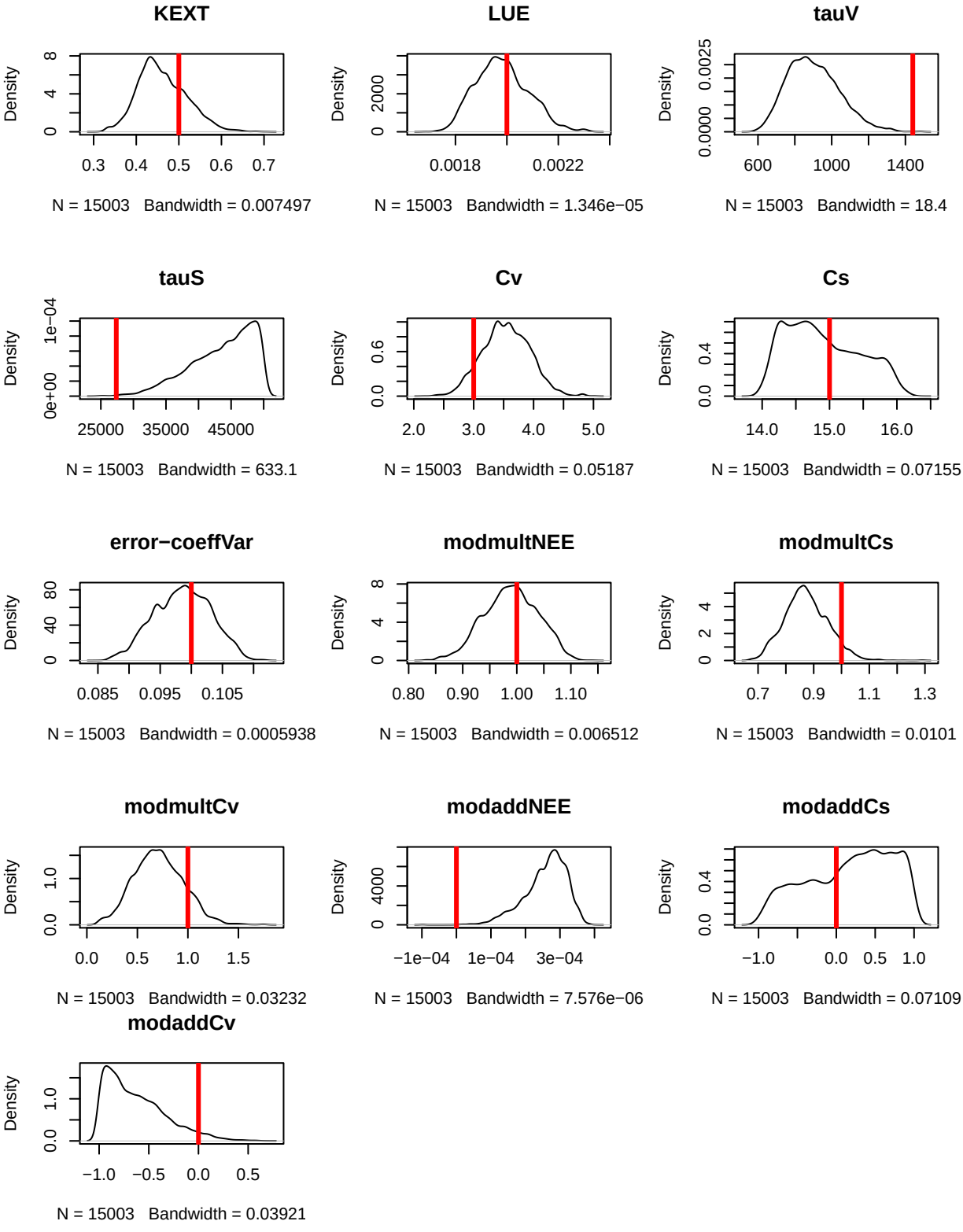


Figure 19: Model with error and unbalanced data with a multiplicative bias. Observations included in the calibration marked with a '+'. Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

7.1 Model with error and unbalanced data with a multiplicative bias EuB



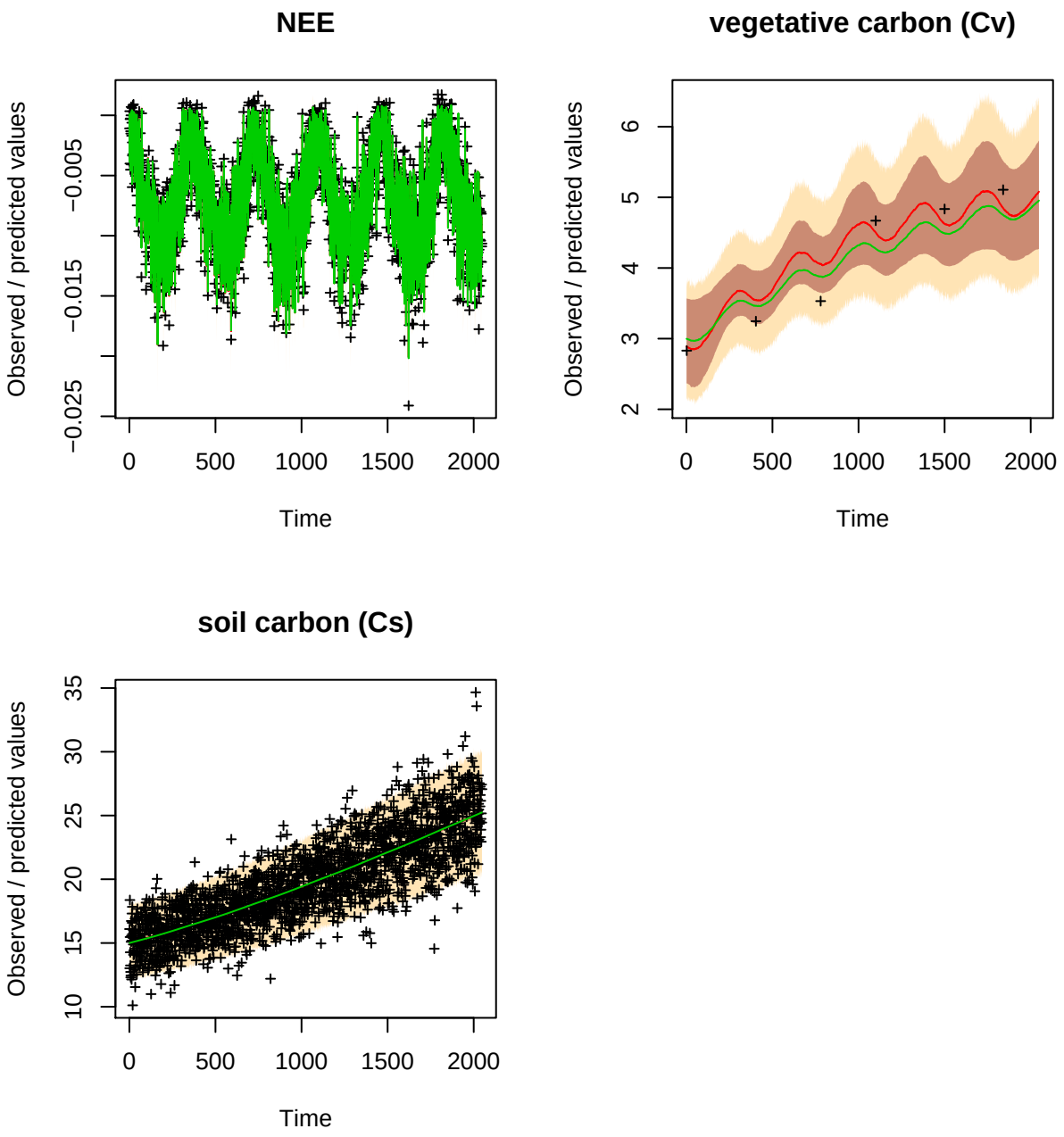
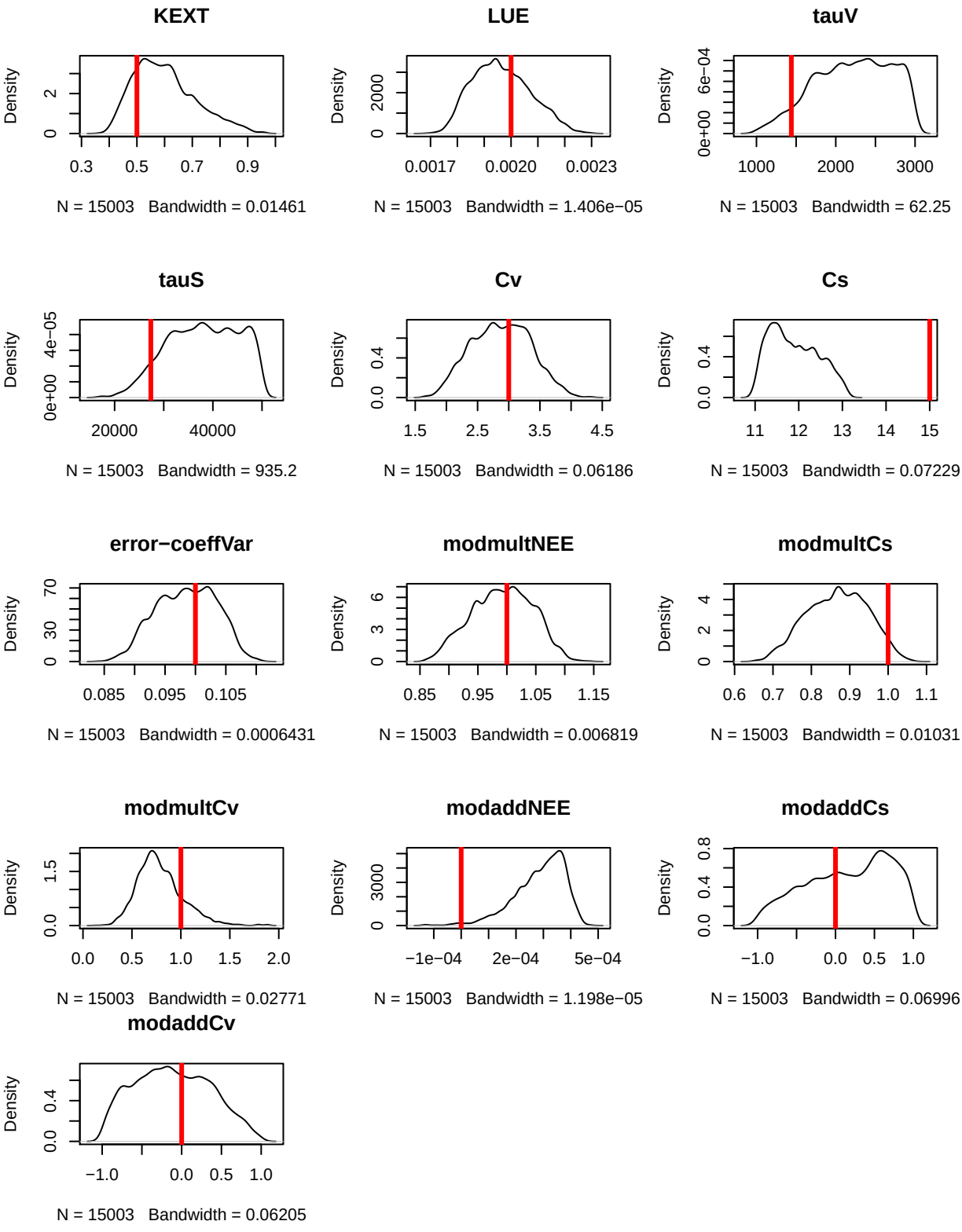


Figure 20: Model with error and unbalanced data with additive and multiplicative parameters to represent model error. Observations included in the calibration marked with a '+'. Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.


KEXT
Density
0
2
0.3
0.5
0.7
0.9
N = 15003 Bandwidth = 0.01461
LUE
Density
0
2000
0.0017
0.0020
0.0023
N = 15003 Bandwidth = 1.406e−05
tauV
Density
0e+00
6e−04
1000
2000
3000
N = 15003 Bandwidth = 62.25
tauS
Density
0e+00
4e−05
20000
40000
N = 15003 Bandwidth = 935.2
Cv
Density
0.0
0.4
1.5
2.5
3.5
4.5
N = 15003 Bandwidth = 0.06186
Cs
Density
0.0
0.4
11
12
13
14
15
N = 15003 Bandwidth = 0.07229
error−coeffVar
Density
0
30
70
0.085
0.095
0.105
N = 15003 Bandwidth = 0.0006431
modmultNEE
Density
0
3
6
0.85
0.95
1.05
1.15
N = 15003 Bandwidth = 0.006819
modmultCs
Density
0
2
4
0.6
0.7
0.8
0.9
1.0
1.1
N = 15003 Bandwidth = 0.01031
modmultCv
Density
0.0
1.5
0.0
0.5
1.0
1.5
2.0
N = 15003 Bandwidth = 0.02771
modaddNEE
Density
0
3000
−1e−04
2e−04
5e−04
N = 15003 Bandwidth = 1.198e−05
modaddCs
Density
0.0
0.4
0.8
−1.0
0.0
0.5
1.0
N = 15003 Bandwidth = 0.06996
modaddCv
Density
0.0
0.4
−1.0
0.0
0.5
1.0
N = 15003 Bandwidth = 0.06205

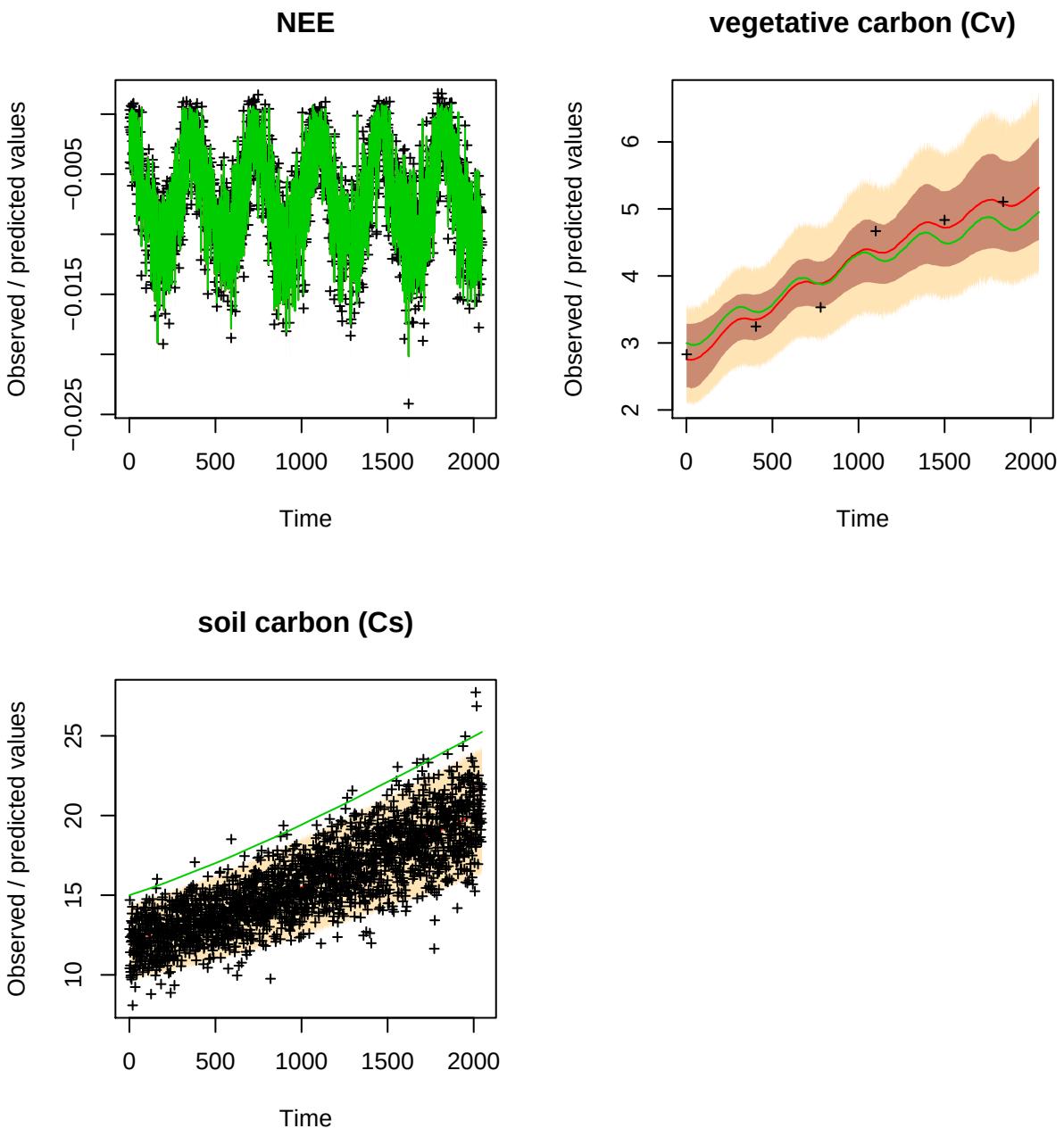
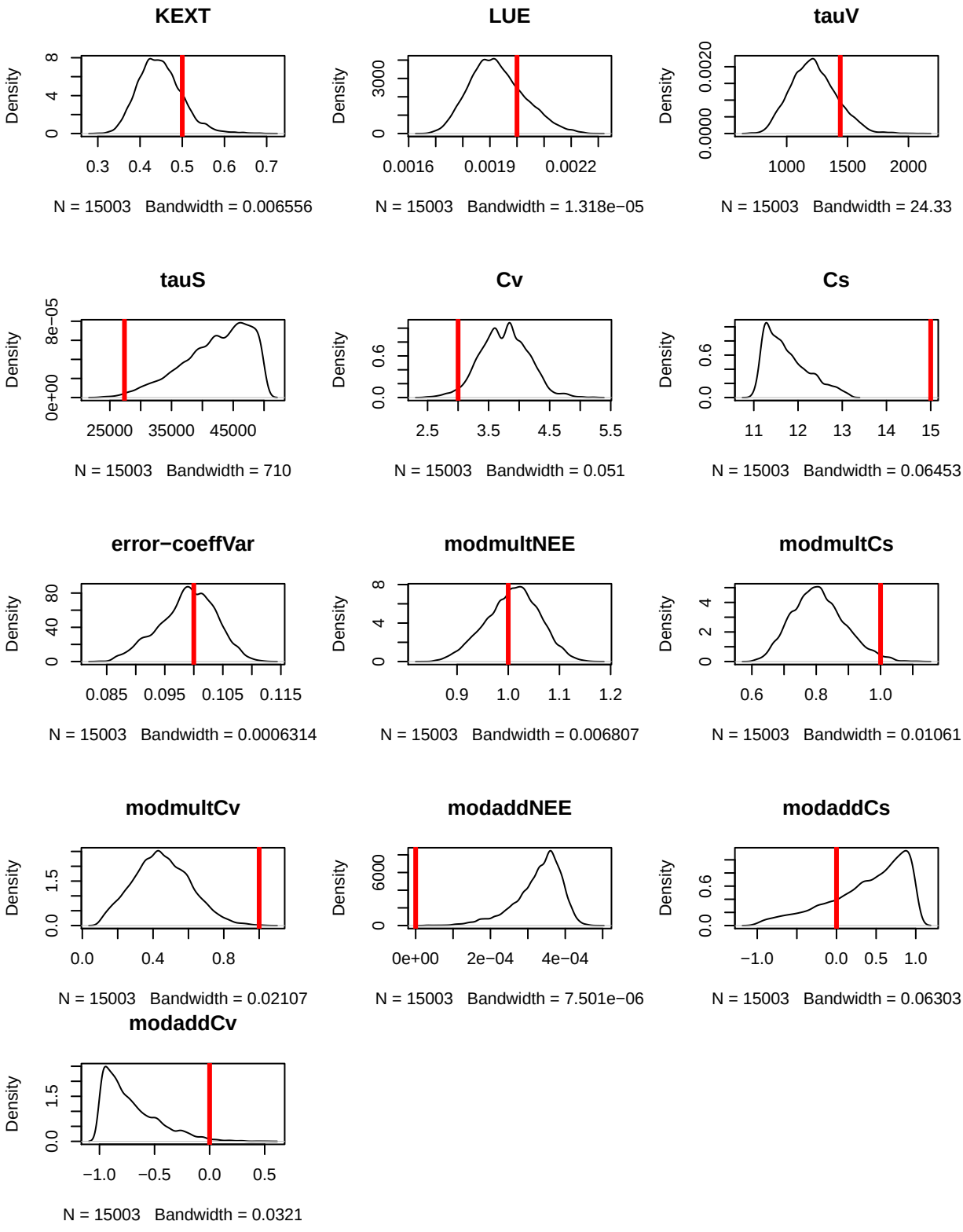


Figure 21: Perfect model and and unbalanced data with a multiplicative bias and additive and multiplicative parameters to represent the bias. Observations included in the calibration marked with a '+'. Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.


KEXT
Density
0
4
8
0.3
0.4
0.5
0.6
0.7
N = 15003 Bandwidth = 0.006556
LUE
Density
0
3000
0.0016
0.0019
0.0022
N = 15003 Bandwidth = 1.318e−05
tauV
Density
0.0000
0.0020
1000
1500
2000
N = 15003 Bandwidth = 24.33
tauS
Density
0e+00
8e−05
25000
35000
45000
N = 15003 Bandwidth = 710
Cv
Density
0.0
0.6
2.5
3.5
4.5
5.5
N = 15003 Bandwidth = 0.051
Cs
Density
0.0
0.6
11
12
13
14
15
N = 15003 Bandwidth = 0.06453
error−coeffVar
Density
0
40
80
0.085
0.095
0.105
0.115
N = 15003 Bandwidth = 0.0006314
modmultNEE
Density
0
4
8
0.9
1.0
1.1
1.2
N = 15003 Bandwidth = 0.006807
modmultCs
Density
0
2
4
0.6
0.8
1.0
N = 15003 Bandwidth = 0.01061
modmultCv
Density
0.0
1.5
0.0
0.4
0.8
N = 15003 Bandwidth = 0.02107
modaddNEE
Density
0
6000
0e+00
2e−04
4e−04
N = 15003 Bandwidth = 7.501e−06
modaddCs
Density
0.0
0.6
−1.0
0.0
0.5
1.0
N = 15003 Bandwidth = 0.06303
modaddCv
Density
0.0
1.5
−1.0
−0.5
0.0
0.5
N = 15003 Bandwidth = 0.0321

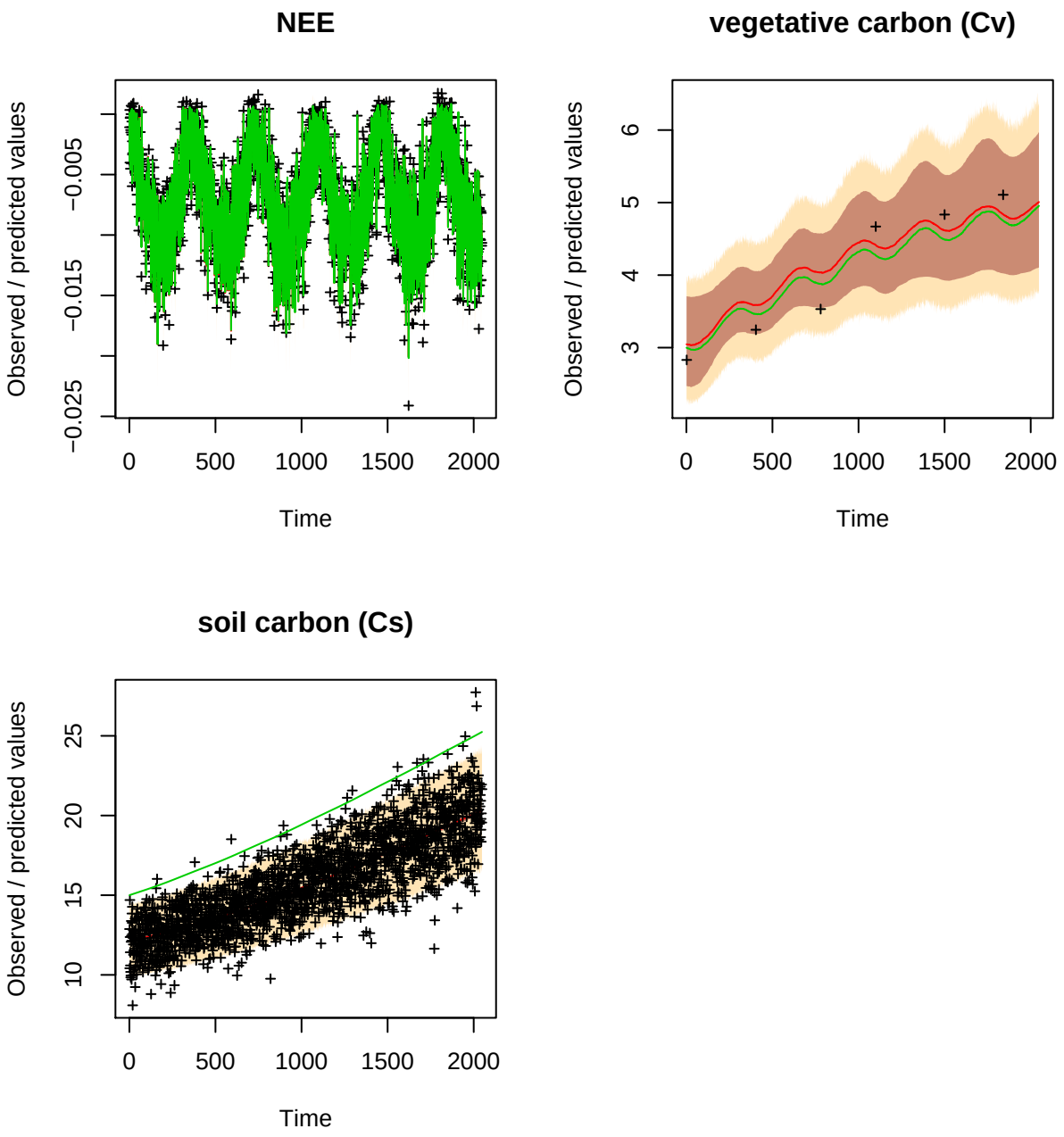


Figure 22: Model with error and and unbalanced data with a multiplicative bias and additive and multiplicative parameters to represent model error and the data bias. Observations included in the calibration marked with a '+'. Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.