TG15 manuscript

David Cameron, Mike Dietze and Marcel van Oijen

2020-12-09

Contents

1 Introduction

We live in an era of rapid environmental change, with multiple drivers of ecosystem processes shifting simultaneously (climate, CO2, nutrient deposition, pollution, species introductions, habitat destruction) and continuously. Because of this the natural systems that we seek to understand, manage, and conserve are in a period dominated by transient conditions and will continue to be so for the foreseeable future. In the face of this change there is an urgent need for ecologists and other environmental scientists to be able to better understand and predict both these transient dynamics and their long-term implications (Dietze et al. 2018). Fortunately, we are aided in this endeavor by an increasing volume, variety, and velocity of environmental data (LaDeau et al. 2017, Farley et al. 2017, Shiklomanov et al. 2019) as well as increasingly sophisticated models (Fisher and Koven 2020, Fisher et al 2018). In other words, while we face unprecedented challenges, we also have unprecedented capacity to address those challenges.

As the variety of data available increases it is often the case that there are multiple types of data available to constrain our understanding of ecological processes, state variables, and/or model parameters. Sometimes these are alternative ways of measuring the same thing (e.g. field versus remotely sensed estimates of leaf area). In other instances they are observations of different parts of a coupled problem. Most ecological processes are complex, and thus ecological models in general, and mechanistic models in particular, often need to predict multiple interacting variables, such as different species, life-history stages, demographic processes, or biogeochemical pools and fluxes. In both of these cases, there is information to be gained through *data fusion* – the use of multiple different data types to constrain a single model. Using multiple constraints is often essential; very often no single data type provides us with a complete understanding of a process. We need to confront different parts of a model with different observations to make sure we are getting the right answer for the right reason, and have not just calibrated a model to produce the "right" answer for one variable through a series of compensating errors (Medlyn et al. 2015)

In principle data fusion is conceptually straightforward. When writing down a statistical Likelihood for either a frequentist (maximum likelihood) or Bayesian model, one writes down multiple data models that are are connected to the same underlying process model. In practice, there are a number of places where things can go awry (Zipkin and Saunders 2018, Zipkin et al. 2021). One particularly common challenge is the need to combine unbalanced datasets, where one or more data types is available in a much larger volume than the other data constraints. This frequently occurs when combining low-volumes of manually-collected field data with high-volumes automatically-collected data collected either from in situ sensors or via remote sensing. For example, in studies of the terrestrial carbon cycle there is often two orders of magnitude difference or more imbalance between the number of measurements available from automated measurements such as eddy covariance data and manual measurements such as soil and plant carbon stocks.

A common observation when combining unbalanced data is that the outputs from data fusion (e.g. calibrated model parameters) are virtually identical to the results achieved by fitting the model to the high-volume data by itself. Since each data point is usually modeled 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 data fusion ignores the low volume data, which gets swamped out by the much larger sample size of the high volume data. Needless to say, this can be disappointing as the lower-volume data often represents considerable labor and, as noted earlier, is often collected to ensure that the model is getting the right answer for the wrong reason. As more and 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).

To add insult to injury, it is also often the case that the addition of high-volume data can cause the predictions for other, low-volume output variables to perform worse. **For example.. <>.**

In response to data fusion gone awry, a number of *ad hoc* solutions have been employed. For example, one might simply thin the high volume data until the data constraints are more balanced (**CITE + example**). Another common solution is to average the data over time or space, for example by aggregating high-frequency sensor data up to a daily, monthly, or annual number.

Richardson, Moore, Riciutto « add more examples »

Williams et al 2009 “Temporal error correlations (systematic biases) are likely to be severe between sub-daily fluxes, for models and observations. To handle this data redundancy, one can include error correlations in R (a rather difficult task), sub-sample the whole data set, or use a mean diurnal cycle over a relatively long period (Santaren et al., 2007).”

(MacBean et al. 2016) has a review of existing multiple constraint studies ***

While neither of these approaches is technically wrong, they are definitely disappointing, as they involve throwing out data which results in a loss of information. Another option that has been employed is to apply weights to the different data models within the overall Likelihood, with the most common option being to down-weight 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 but doesn’t have any basis in probability theory.

Weighting likelihoods has the intuitive appeal of retaining every observation, but it has the more subtle issue of not being grounded in probability theory, which is the basis for both Likelihood and Bayesian statistics. Mathematically there’s also nothing stopping one from ‘upweighting’ datasets, which amounts to pretending that you have more data than you actually do, leading to falsely overconfident parameter estimates and predictions. As a point of fact, in all of these options (thinning, averaging, weighting) the choices made shift not only the mean, but also have large impacts on the parameter uncertainties and model confidence intervals.

The big problem with these approaches, whether thinning, averaging, or weighting data, is that they are all very subjective. There is a full continuum of options available, from working with data at its raw frequency all the way up to averaging all the data to a single point, and the outcome of the analysis will change, often dramatically, depending on the averaging/thinning/weighting one chooses. Indeed, practitioners are resorting to thinning/averaging/weighting precisely because it changes the outcome, but there’s currently no objective advice on how to do so.

One possible path forward is to look for a more objective way to weight data. In this regard, some have pointed to the “independence” assumption and suggested that greater attention needs to be paid to the information content of a dataset (Dietze 2017). Spectral analyses of high-frequency environmental data, such as eddy covariance, often show distinct peaks at the annual and daily scale, reflecting strong seasonal and diurnal cycles (Dietze et al. 2011, STOY CITE). In such cases aggregating data that’s at a high frequency (e.g. 1 observation per minute) to 5 minute data will retain almost the same information content but the sample size will be cut by five. While aggregating is ad hoc, formally accounting for non-independence of observations, for example by accounting for autocorrelation in a Likelihood, is not (**CITE Examples**). Because modeling spatial and temporal autocorrelation can be computationally demanding, others have sought to approximate this by weighting data models based on calculations of effective sample size (Fer et al. 2018).

While autocorrelation is probably part of the challenge of working with high-volume data, there’s reason to believe that its not the only issue at hand. For example, autocorrelation doesn’t explain why calibration to a high-volume dataset would cause a model to perform worse at predicting a different output variable. Also, many high-volume data sources have little to no replication. For example, is an eddy covariance tower that produces 17520 half-hourly observations per year count as n=17520 data points or n=1 tower? If there were a whole population of sensors we could calculate a sample mean and variance. There is every reason to believe that any one sensor will be different from the mean, but with only one sensor there is no way of knowing what the mean is or how different the data we have is from that mean.

The goal of this paper is to use simulated data experiments to identify more clearly the different issues associated with fusing unbalanced data in parameter calibration. We aim to develop a general methodology for identifying whether and where the issue becomes a problem, and then to start to explore simple modifications to the Likelihood (i.e. including additive and multiplicative bias terms for variables constrained by high-volume observations) to see to what extent they can help ameliorate the identified issue. This approach of modifying the Likelihood is complementary to the computationally more demanding approach of bias correction during calibration with Gaussian process (Oberpriller et al. XXX).

For these simulated data experiments we developed a very simple process-based ecosystem model (VSEM) as a testbed. Our aim is to present a model that is simple enough that results are easily understood, but sufficiently complex that we can be confident that the model/data issues identified here would also be seen in more complex/realistic process-based ecosystem models. By calibrating the model (and variants with structural model error) to variations of it's own output (with added observation error), we are able to experimentally separate the influence of unbalanced data in the calibration from the influence of artificially-introduced errors in the data or the model. Overall we find that systematic model and data errors appear to be at the root of many of the issues attributed to unbalanced data. So data imbalance *per se* would not necessarily be a problem if model and data were unbiased. It is the presence of multiple constraints, and the power that comes from high-volume data, that together shine a light on errors that otherwise often go unnoticed. Because all models are wrong (but some are useful), simply "fixing the model" is not always sufficient given the statistical power big data provides. We show that simple fixes that account for model and data bias can lead to improvements in prediction, but point to the need for research.

INTRO CITATIONS (to be moved)

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. Ecol. Appl., 30, e02021.

Dietze et. al. 2011. Identifying the time scales that dominate model error: A North American synthesis of the spectral properties of ecosystem models. JGR-Biogeosciences 116, G04029, doi:10.1029/2011JG00166

Dietze M. 2017. Ecological Forecasting. Princeton University Press. ISBN: 9780691160573

Dietze MC, A Fox, L Beck-Johnson, JL Betancourt, MB Hooten, CS Jarnevich, TH Keitt, MA Kenney, CM Laney, LG Larsen, HW Loescher, CK Lunch, B Pijanowski, JT Randerson, EK Read, AT Tredennick, R Vargas, KC Weathers, EP White. 2018. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. Proceedings of the National Academy of Sciences 115 (7) 1424-1432 https://doi.org/10.1073/pnas.1710231115.

Farley, S. S., A. Dawson, S. J. Goring, and J. W. Williams. 2018. Situating Ecology as a Big-Data Science : Current Advances , Challenges , and Solutions. Bioscience 68:563–576.

Fer I, R Kelly, P Moorcroft, AD Richardson, E Cowdery, MC Dietze. 2018. Linking big models to big data: efficient ecosystem model calibration through Bayesian model emulation. Biogeosciences 15, 5801–5830, 2018 https://doi.org/10.5194/bg-15-5801-2018

Fisher RA, CD Koven, WRL Anderegg, BO Christoffersen, MC Dietze, C Farrior, JA Holm, G Hurtt, RG Knox, PJ Lawrence, M Longo, AM Matheny, D Medvigy, HC Muller-Landau, TL Powell, SP Serbin, H Sato, J Shuman, B Smith, AT Trugman, T Viskari, H Verbeeck, E Weng, C Xu, X Xu, T Zhang, P Moorcroft. 2018. Vegetation Demographics in Earth System Models: a review of progress and priorities. Global Change Biology 24(1):35-54 DOI: 10.1111/gcb.13910

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, e2018MS001453. https://doi.org/ 10.1029/2018MS001453

LaDeau, S. L., B. A. Han, E. J. Rosi-Marshall, and K. C. Weathers. 2017. The Next Decade of Big Data in Ecosystem Science. Ecosystems 20:274–283.

Medlyn, B., Zaehle, S., De Kauwe, M. et al. Using ecosystem experiments to improve vegetation models. Nature Clim Change 5, 528–534 (2015). https://doi.org/10.1038/nclimate2621

Medvigy, D. M., S. C. Wofsy, J. W. Munger, D. Y. Hollinger, and P. R. Moorcroft. 2009. Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem Demography model version 2. Journal of Geophysical Research 114:1–21.

Shiklomanov, A. N., B. A. Bradley, K. M. Dahlin, A. M Fox, C. M. Gough, F. M. Hoffman, E. M Middleton, S. P. Serbin, L. Smallman, and W. K. Smith. 2019. Enhancing global change experiments through integration of remote-sensing techniques. Frontiers in Ecology and the Environment 17:215–224.

Zipkin EF and Saunders SP. 2018. Synthesizing multiple data types for biological conservation using integrated population models. Biol Conserv 217: 240–50.

Zipkin EF, ER Zylstra, AD Wright, SP Saunders, AO Finley, MC Dietze, MS Itter, MW Tingley. 2021. "Linking ecological processes across scales with data integration" Frontiers in Ecology and the Environment *in press*

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.

$$\begin{align} PAR = (|\sin(Days/365 \times \pi) + \epsilon|) \times 10 \tag{1} \\ \tag{2} \end{align}$$

- 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).

$$\begin{align} GPP &= PAR \times LUE \times (1 - \exp^{(-KEXT \times LAR \times C_v)}) \tag{3} \\ NPP &= (1 - GAMMA) * GPP \tag{4} \end{align}$$

- 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 (Gill 1980) 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.

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

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

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

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 (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

- brief summary of DREAMzs algorithm

The DREAMzs algorithm and MCMC functions that we use here are from the BayesianTools package.

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 set 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 (@ref(modLike)) 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 with all the processes modelled perfectly. The only way to ensure 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 (as is away the case) 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 NEE, vegetative carbon and soil carbon.

For the vegetaive 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 two 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. {modLike}

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. Whilst we could include these terms for all of the parts of the system where we have calibration data available we opted here to only include terms on the parts of the system where we had plentiful data (NEE and soil carbon). Therefore we have four extra parameters to represent addition and multiplicative error for each of NEE and soil carbon (modaddNEE, modmultNEE, modaddCs and modmultCs). The priors for each of these are as follows.

parameter name	min	max
modmultNEE	0.1	2.0
modmultCs	0.1	2.0
modaddNEE	-0.01	0.01
modaddCs	-1.0	1.0

3 Identifying the issue

3.1 Perfect model and balanced data

- refer to Fig. (1) and (2)
- ‘true’ parameters largely found in posterior
- posterior parameter controlling data error close to 0.1 coefficient of variance originally imposed to create observations from ‘true’ model output.
- 50% quantile red line very close to green ‘truth’ line
- posterior very narrow marked by 2.5% and 97.5% quantiles
- most data within predictive interval

3.2 Perfect model and unbalanced data

- refer to Fig. (3) and (4)
- despite unbalanced data
  - parameter values still close to true values in posterior
  - model output still close to truth even for vegetative carbon
  - although now greater uncertainty for Cv then for balanced calibration

3.3 Model with error and balanced data

- refer to Fig. (5) and (6)
- Root pool is effectively removed from the model by initialising pool to zero and setting allocation to roots to zero. The loss of the root pool has introduced a significant structural error to the model.
- Parameter posteriors now quite far away from ‘true’ values.
  - Especially parameter which controls turnover of vegetation so that rate of turnover to soil is now more than doubled.
- Parameters calibration seems to have somewhat ‘absorbed’ the model structural error so that
  - outputs where there was data included in the BC are still not too far away from the ‘truth’ line.
  - Cv now has too much variability but average increase not too bad.
  - most data still within predictive interval.

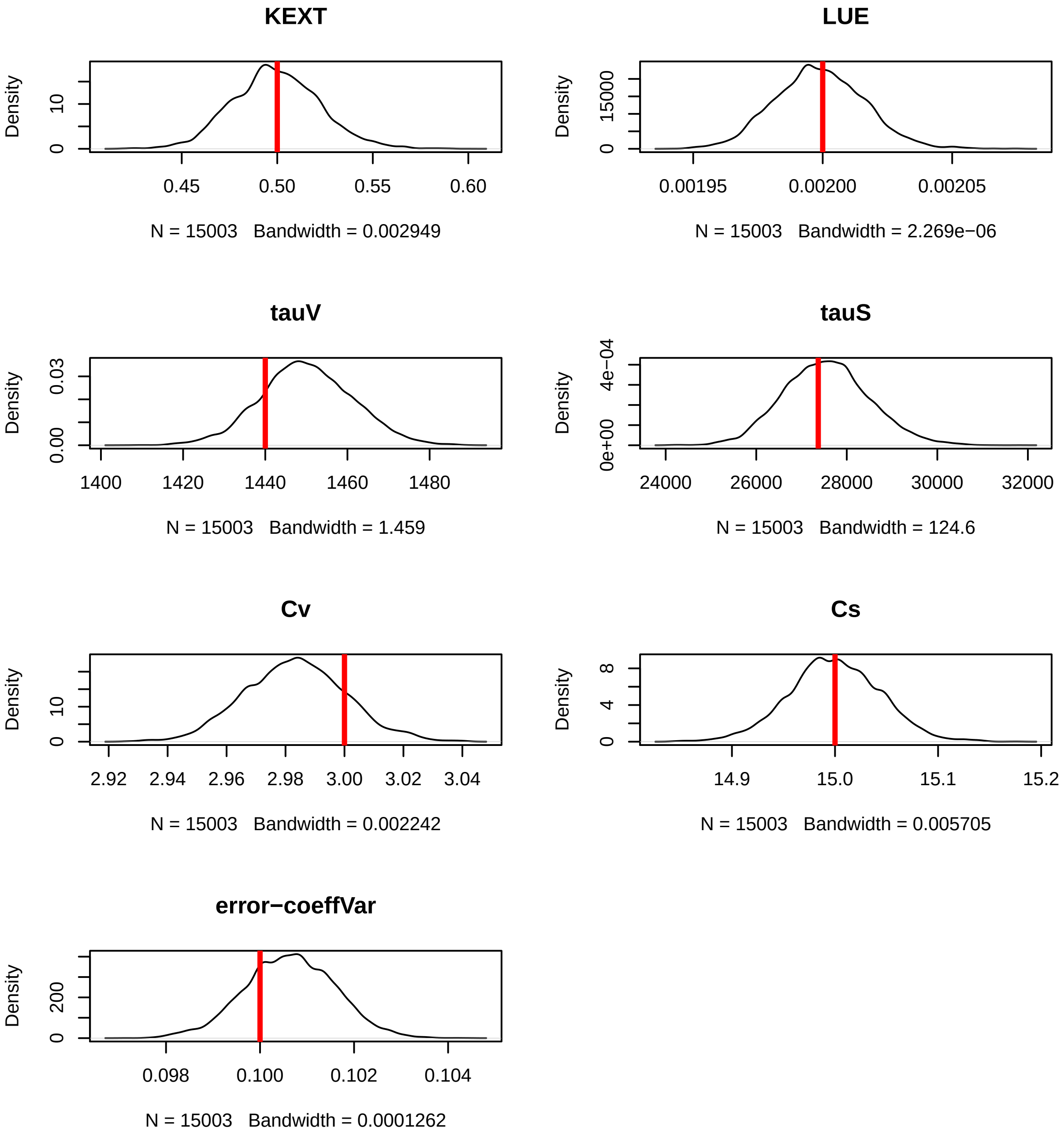

Figure 1: Perfect model, balanced data (NEE, Cv, Cs: 2048 obs). Marginal posterior distribution of model parameters and intital states. The red line marks the 'true' parameter values.

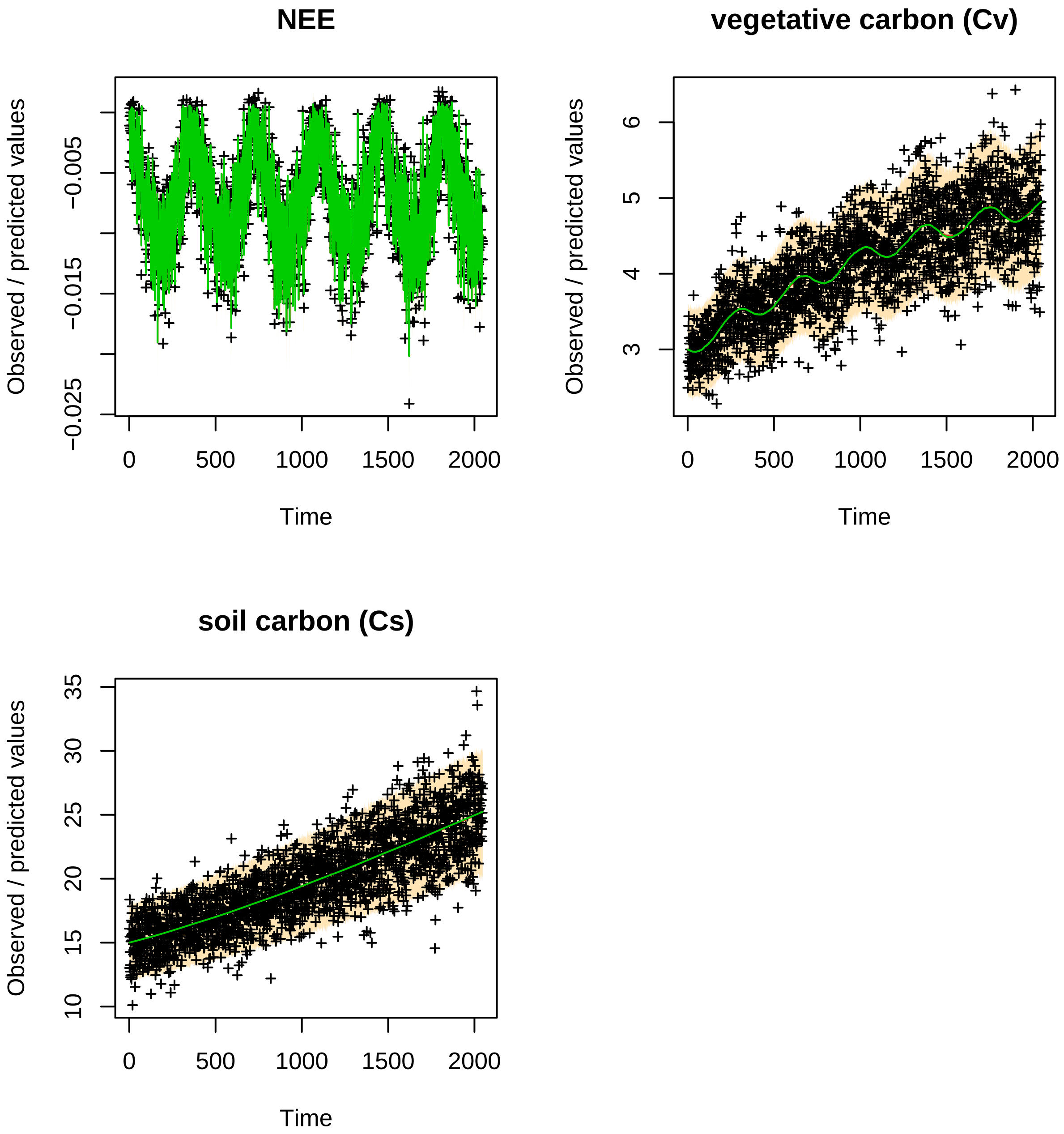

Figure 2: Perfect model, balanced data (NEE, Cv, Cs: 2048 obs). Observations included in the calibration marked with a '+.' Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5%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

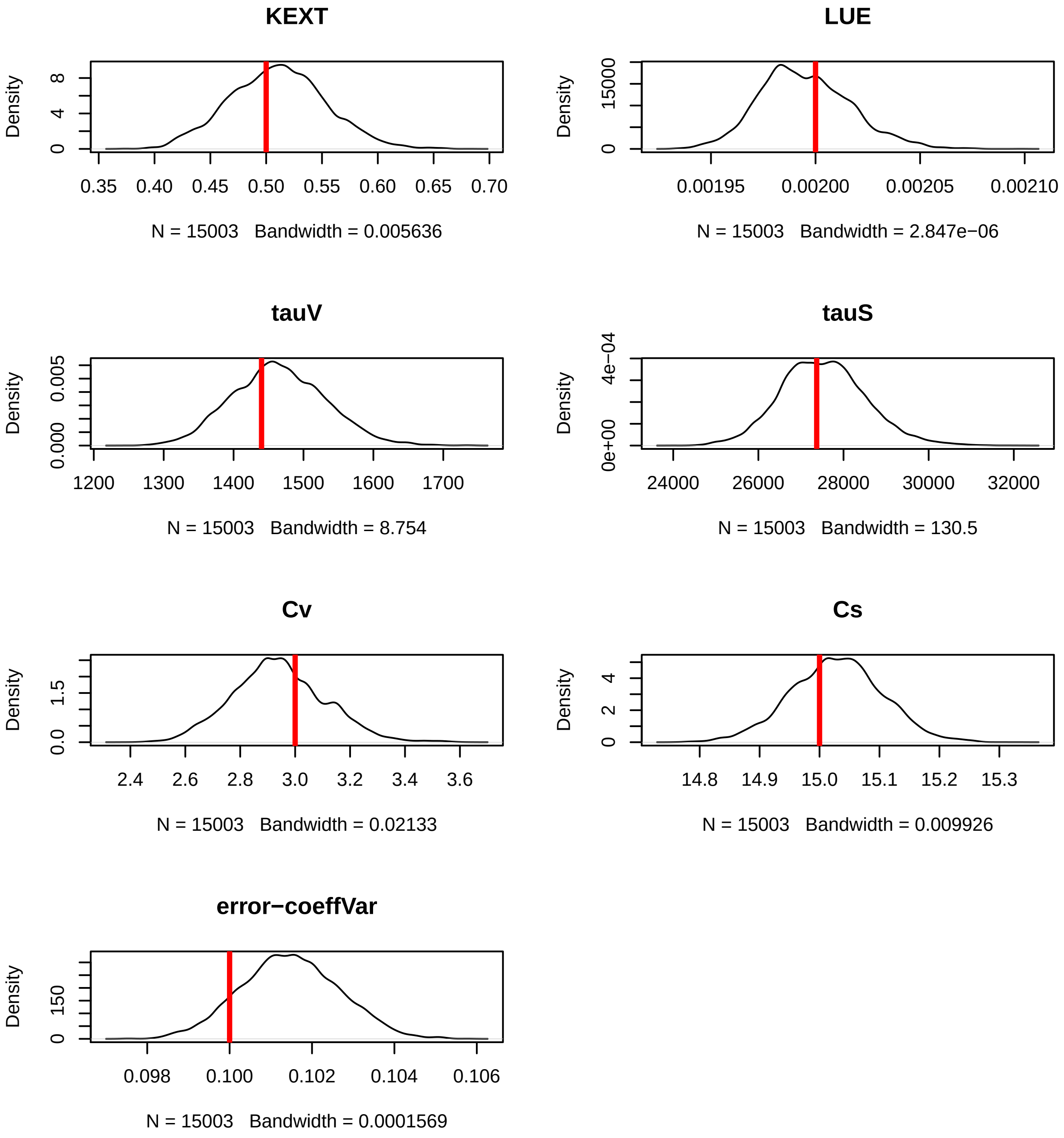

Figure 3: Perfect model, unbalanced data (NEE, Cs: 2048 obs, Cv: 6 obs). Marginal posterior distribution of model parameters and intital states. The red line marks the 'true' parameter values.

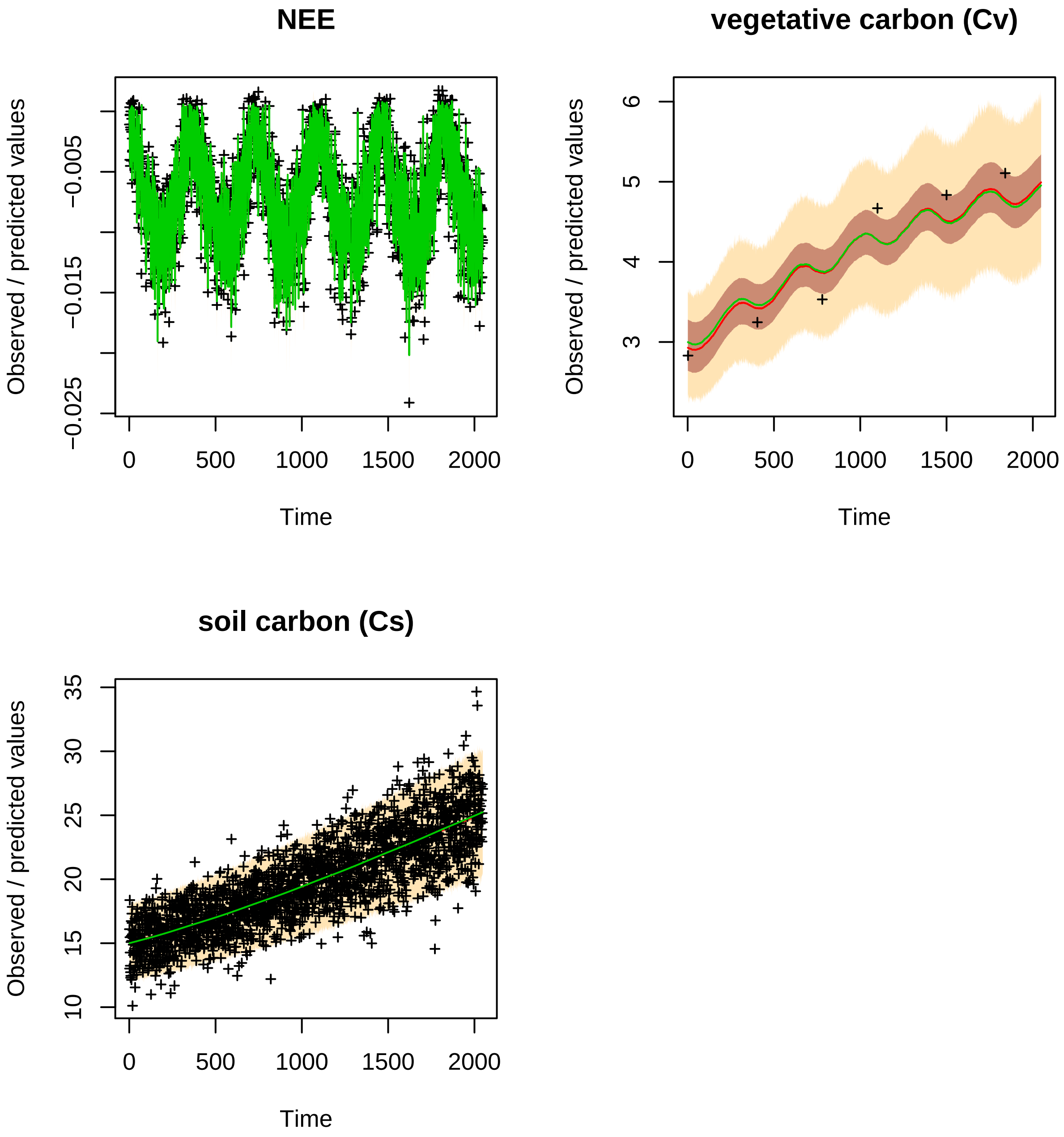

Figure 4: Perfect model, unbalanced data (NEE, Cs: 2048 obs, Cv: 6 obs). Observations included in the calibration marked with a '+.' Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

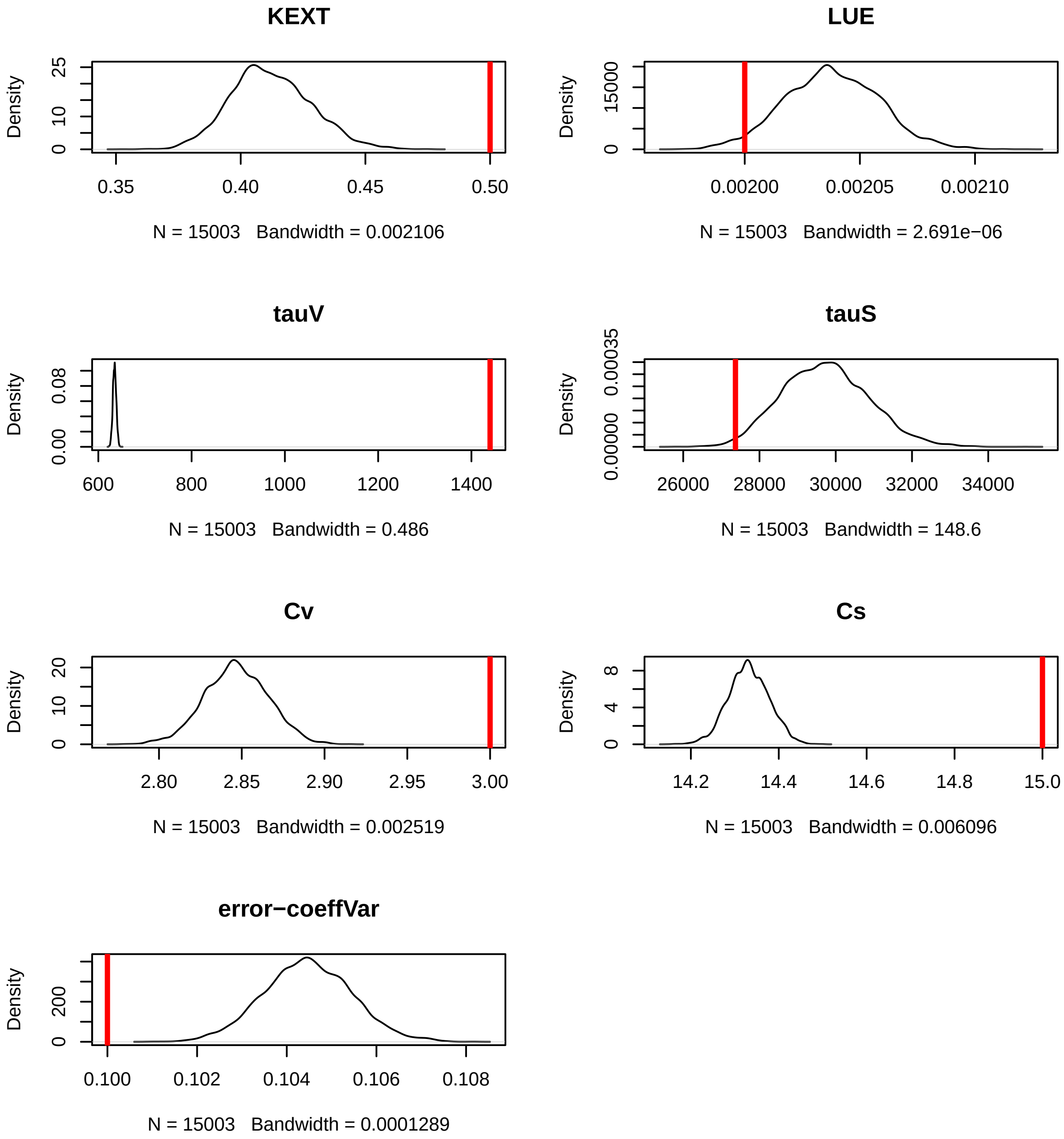

Figure 5: Model with error, balanced data. Marginal posterior distribution of model parameters and intital states. The red line marks the ‘true’ parameter values.

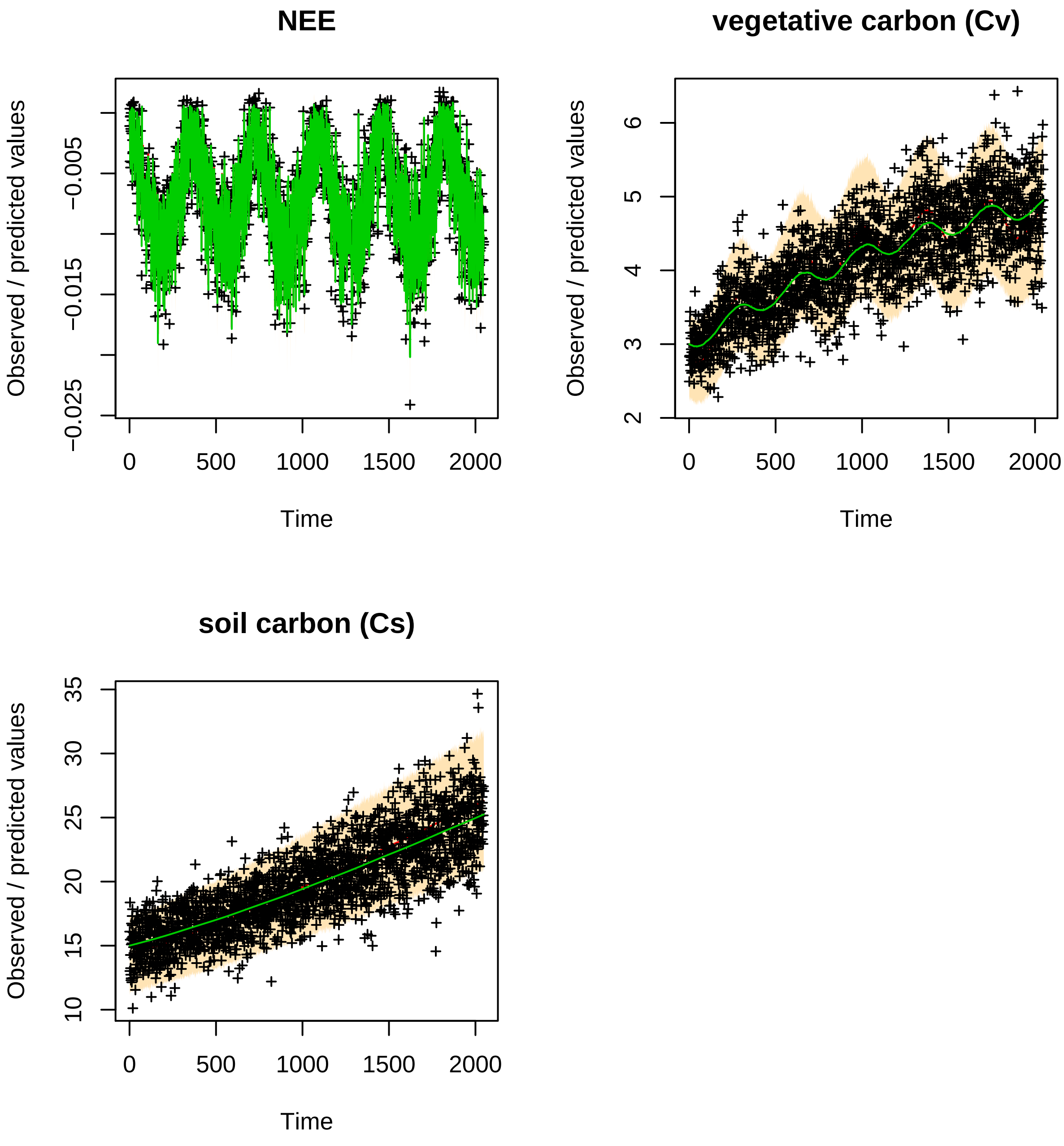

Figure 6: Model with error, balanced data. Observations included in the calibration marked with a '+.' Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

3.4 Model with error and unbalanced data

- refer to Fig. (7) and (8)
- KEXT smaller, LUE larger, Cv larger, tauS smaller, tauV much larger, Cs much larger
  - Generally parameters closer to their 'true' value. Less 'absorbing' of the model structural error.
- NEE and soil carbon pools look largely unchanged to balanced run and close to data and predictive interval.
- significant change to soil vegetation pool
  - data outside of posterior and close to one edge of predictive interval
  - departure from 'truth' line growing in time
- General sense is that six vegetation data points are being somewhat ignored by the calibration in favour of the more plentiful NEE and soil carbon data.

3.5 Perfect model and balanced data with a multiplicative bias

- refer to Fig. (9) and (10)
- soil carbon pool data multiplied by two to represent data that has a systematic bias
- Parameters KEXT larger, tauV smaller, tauS much smaller, Cs much larger
- as for model error parameter calibration 'absorb' the error
  - initial value of soil pool parameter more than double 'true' value
  - parameter controlling turnover time of soil approximately doubled to keep soil carbon pool high.
- calibrated outputs again reasonably close to observations for NEE and Cv
  - slope of soil carbon pool with time too shallow (so some data outside predictive interval) but compares well with 'true' line (influence of other more accurate NEE and vegetative carbon data?)
- data and calibrated output of soil carbon greater than 'true' line by a factor of two as might be expected to match calibration data.

3.6 Perfect model and unbalanced data with a multiplicative bias

- refer to Fig. (11) and (12)
- most parameters are far away from their 'true' values
- soil carbon improved fit to data versus previous balanced data calibration
- six vegetative carbon data points effectively ignored and overpowered by NEE and soil carbon data in the calibration.

3.7 Model with error and unbalanced data with a multiplicative bias

- refer to Fig. (13) and (14)
- similar to calibration with bias in data above.
- vegetative carbon perhaps very slightly closer to the data than when only data bias was present. This may indicate compensating errors in the model and data.

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

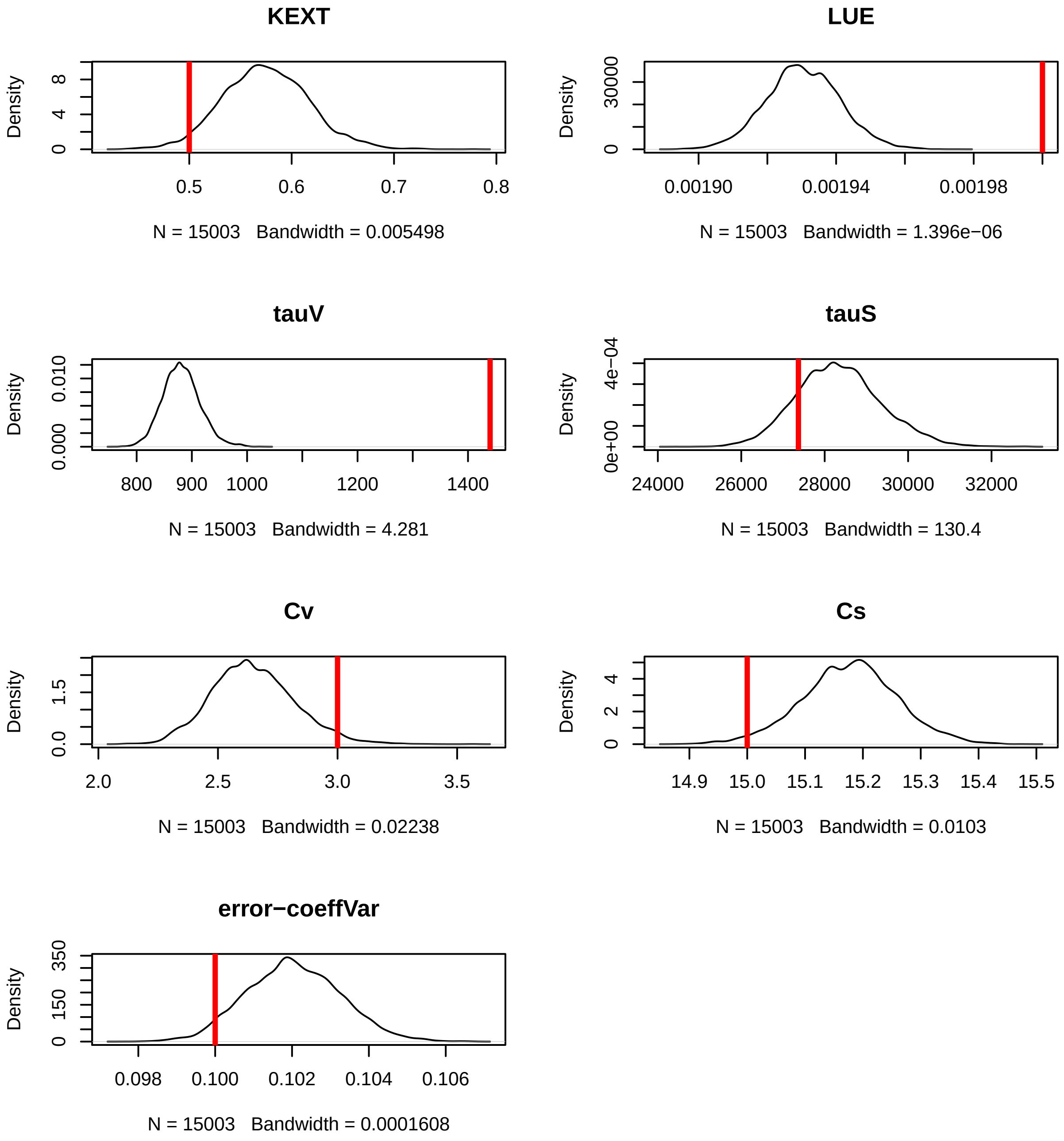

Figure 7: Model with error, unbalanced data (NEE, Cs: 2048 obs, Cv: 6 obs). Marginal posterior distribution of model parameters and intital states. The red line marks the 'true' parameter values.

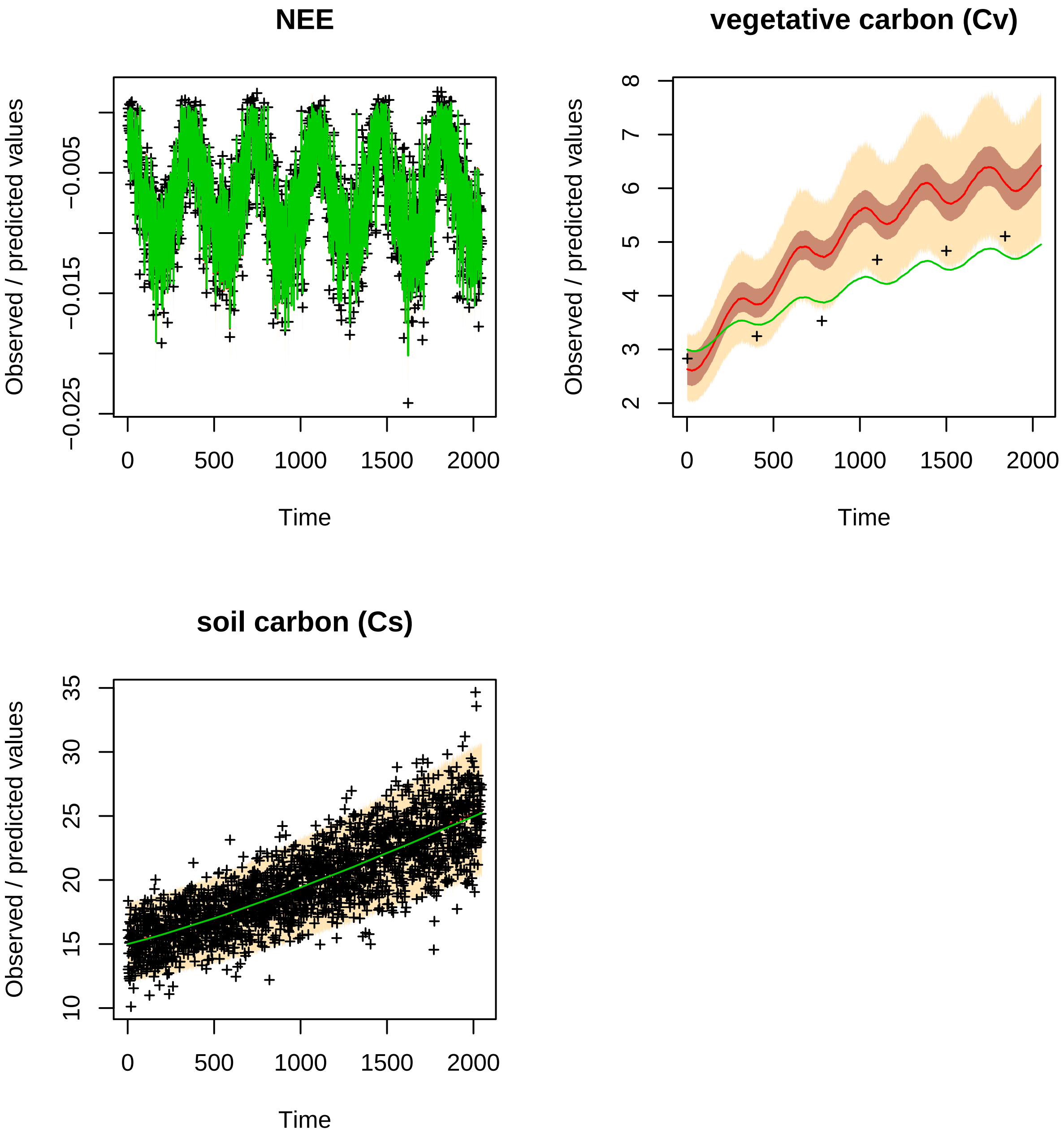

Figure 8: Model with error, unbalanced data (NEE, Cs: 2048 obs, Cv: 6 obs). Observations included in the calibration marked with a '+.' Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

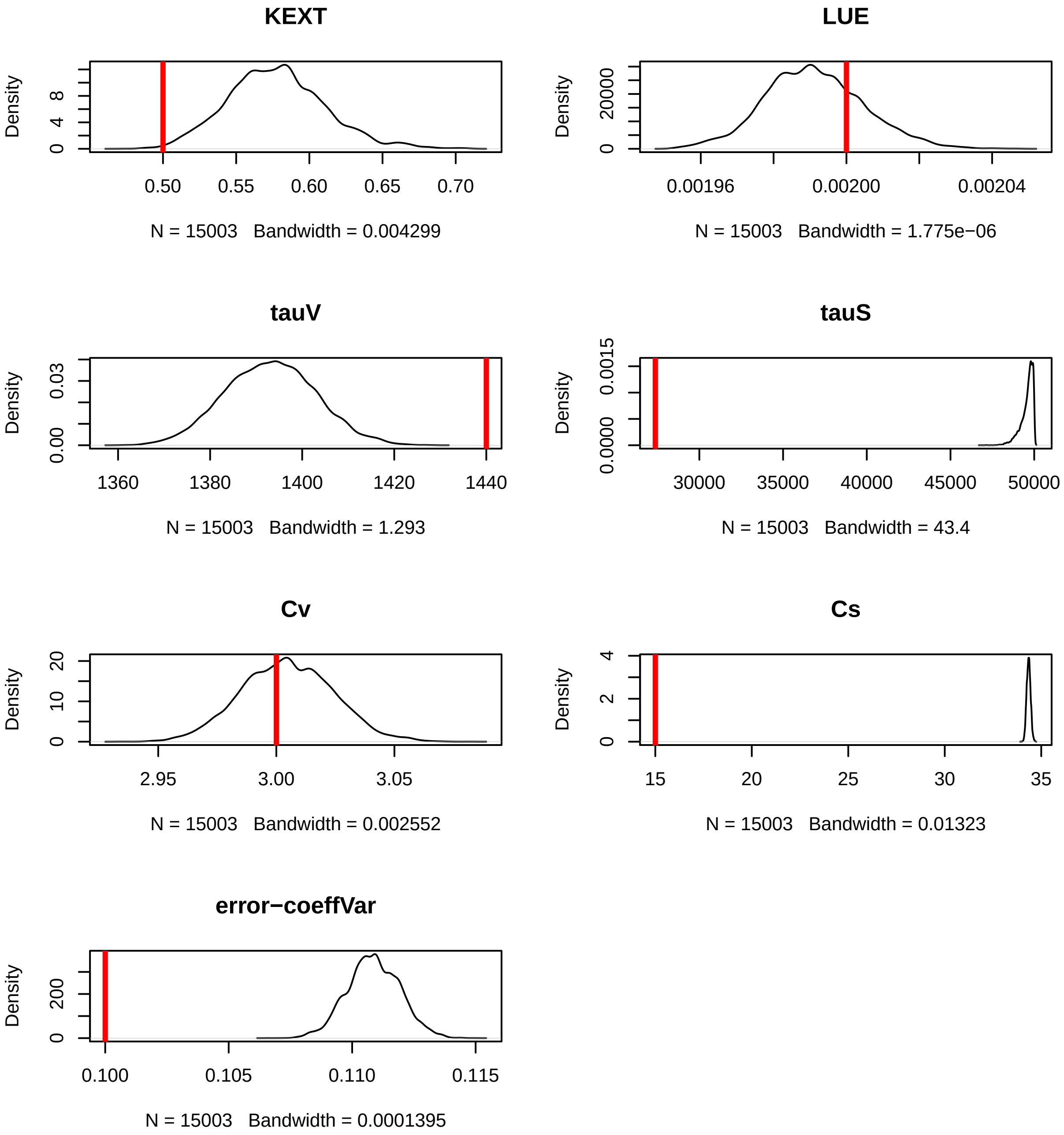

Figure 9: Perfect model and balanced data with a multiplicative bias. Marginal posterior distribution of model parameters and intital states. The red line marks the 'true' parameter values.

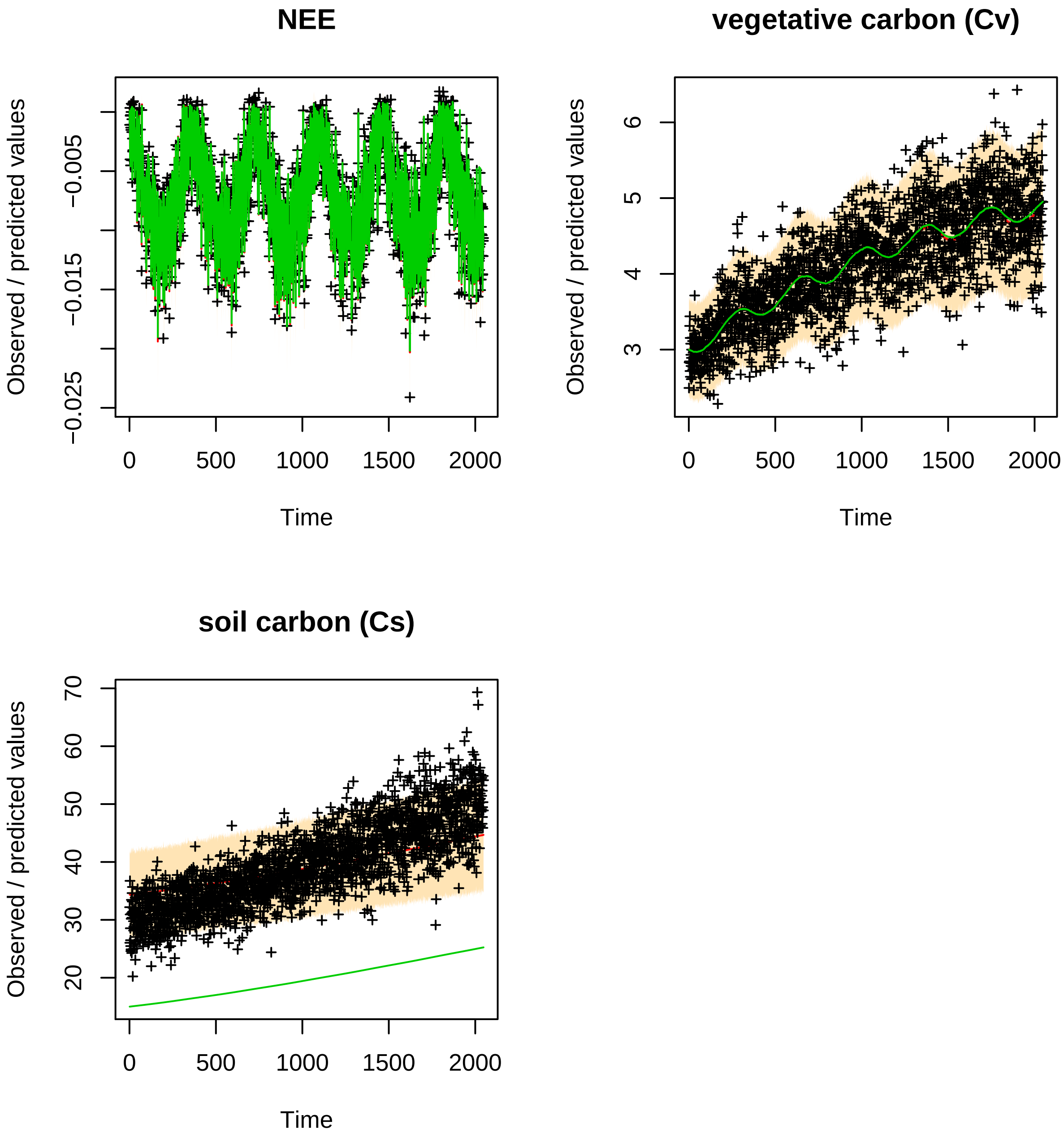

Figure 10: Perfect model and balanced data with a multiplicative bias. Observations included in the calibration marked with a '+.' Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

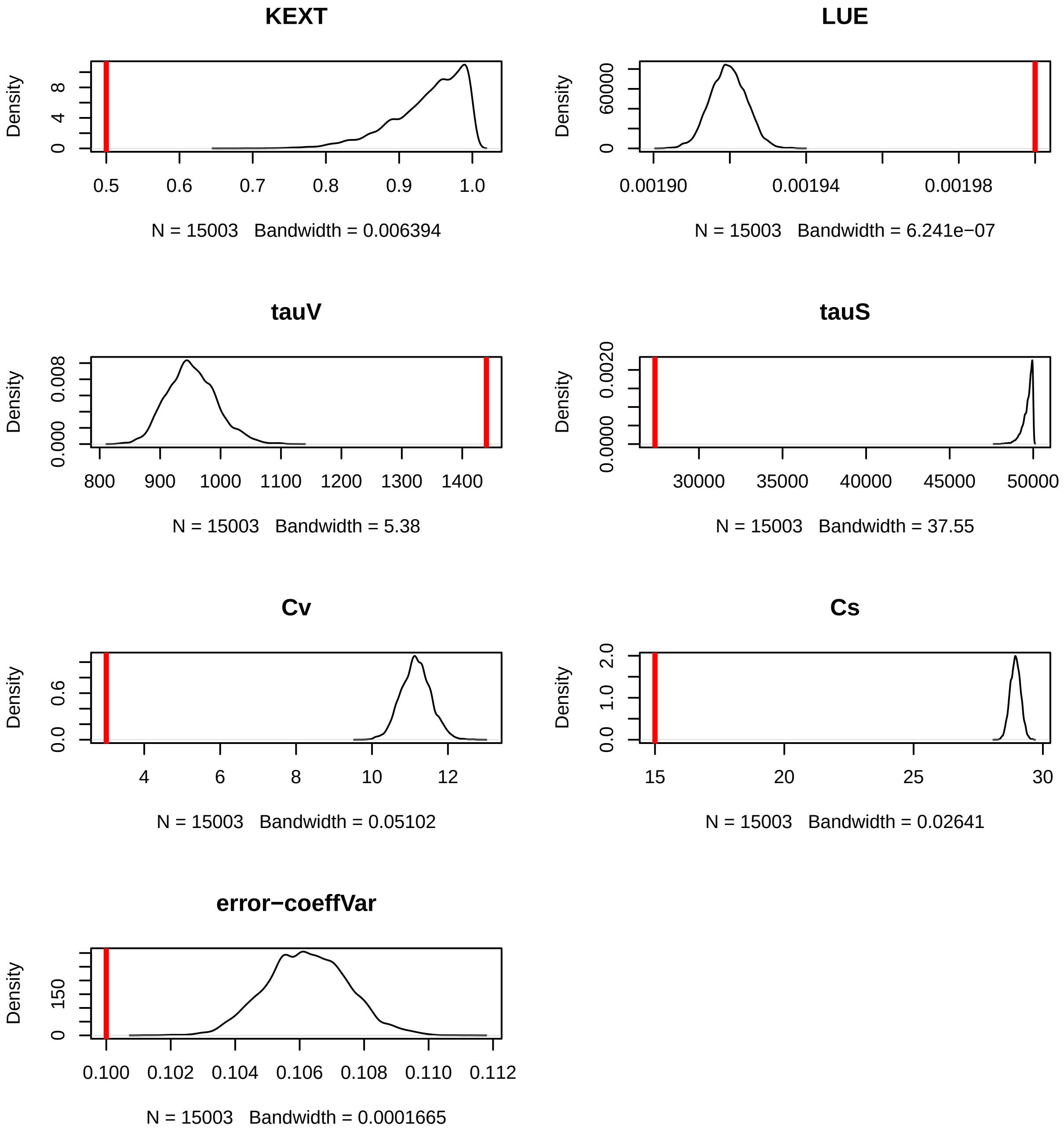

Figure 11: Perfect model and unbalanced data with a multiplicative bias. Marginal posterior distribution of model parameters and intital states. The red line marks the ‘true’ parameter values.

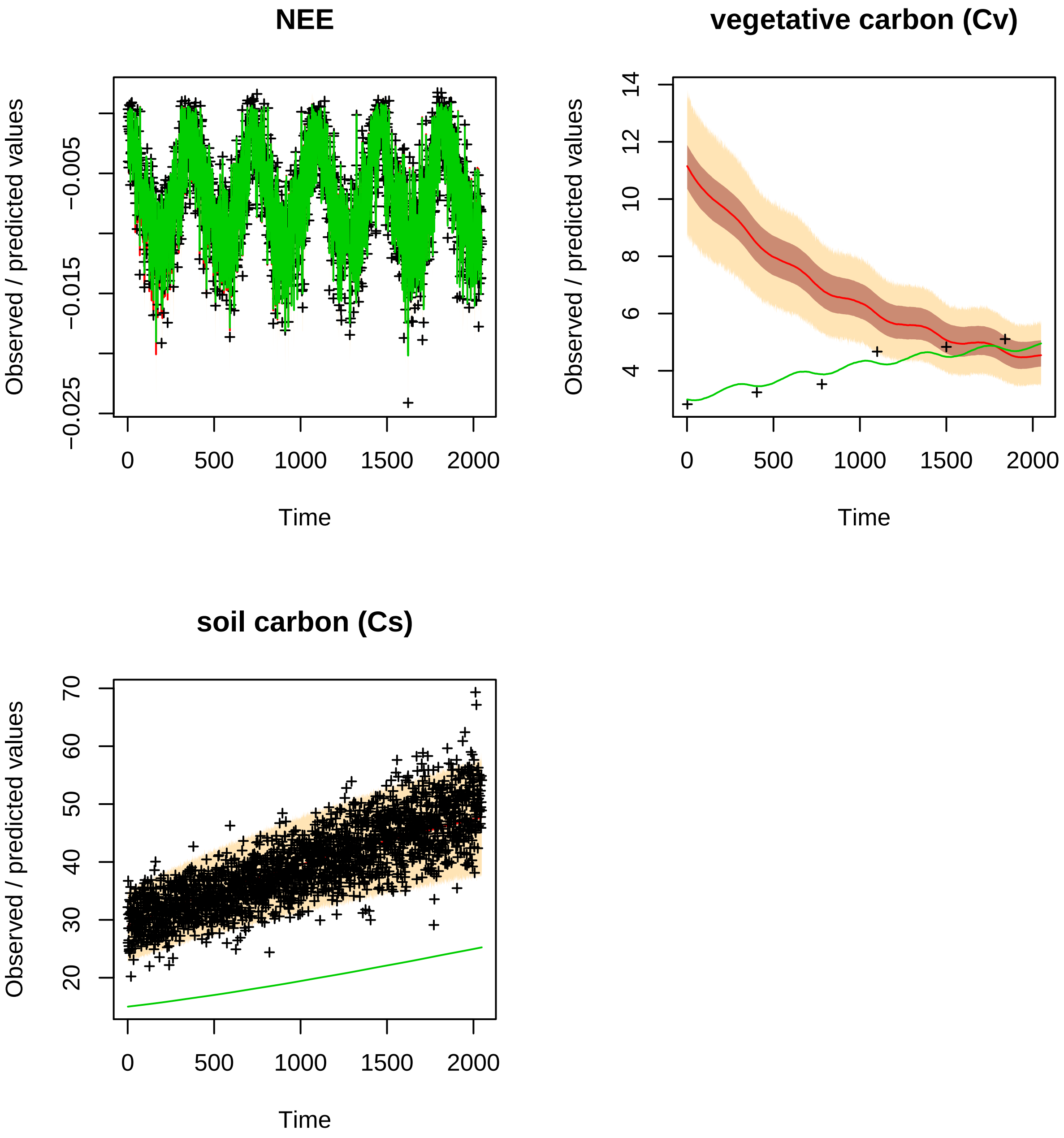

Figure 12: Perfect model and unbalanced data with a multiplicative bias. Observations included in the calibration marked with a '+.' Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

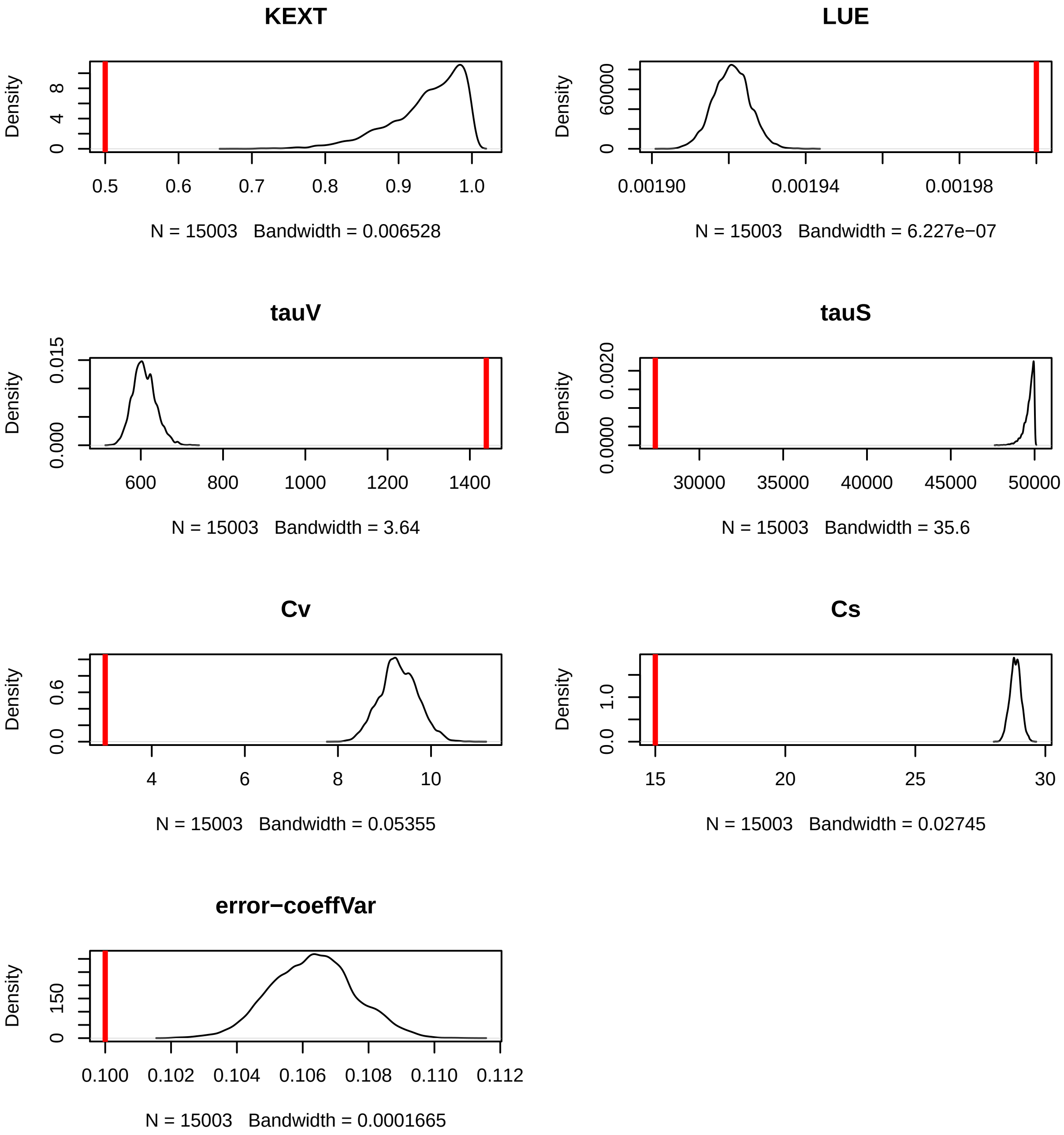

Figure 13: Model with error and unbalanced data with a multiplicative bias. Marginal posterior distribution of model parameters and intital states. The red line marks the ‘true’ parameter values.

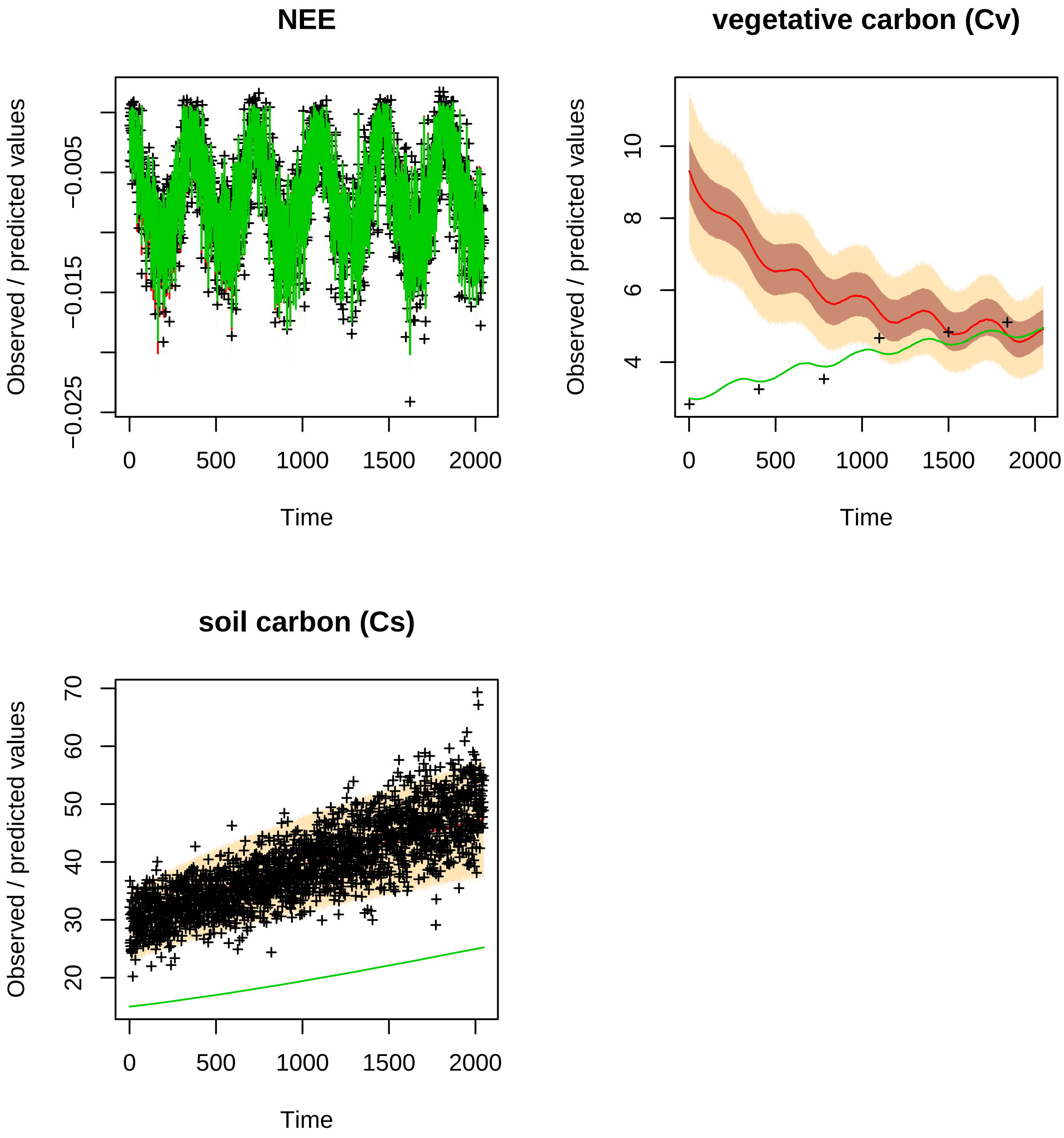

Figure 14: Model with error and unbalanced data with a multiplicative bias. Observations included in the calibration marked with a '+.' Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

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. 15).

The figure shows broad similarity in results except for vegetative carbon case when the model has an error and where there is an imbalanced in calibration data. In general, the RMS difference has a tendency to go down 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.(15) 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. 16). 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.

5 Changes to the Likelihood to represent model and data errors

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

- refer to Fig. (17) and (18)
- KEXT, LUE, Cv, Cs and error-coeffVar are now significantly closer to the 'true' values. tauS is not but is much more uncertain.
- vegetative carbon much improved
  - 5 out of 6 data points are now inside the posterior confidence interval
  - 50% quantile line now much closer to the 'true' line.

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

- refer to Fig. (19) and (20)
- many model parameters (KEXT, LUE, tauV, tauS, initial Cv) much closer to 'true' values
- multiplicative bias multiplication parameter modmultCs centred around 2.25 which is close to the multiplication factor applied to the data.
- NEE and soil carbon close to the data and within the predictive interval.
- vegetative carbon pool much improved with all data points covered by the posterior credible interval.

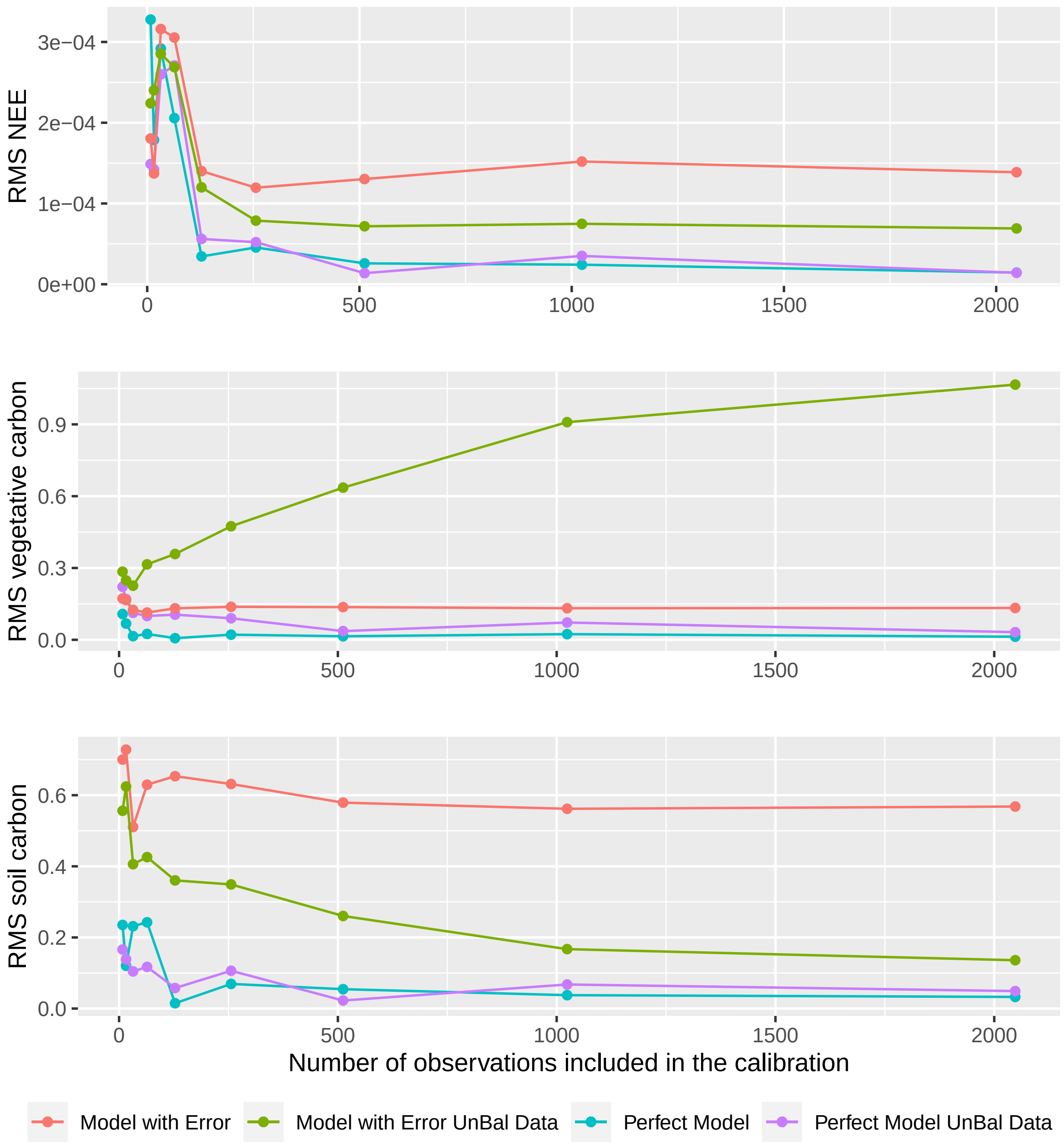

Figure 15: Each point in the three graphs (NEE, vegetative carbon, and soil carbon) represents the RMS difference between the VSEM model and the ‘truth’ run with different maximum a posteriori (MAP) vectors. The MAP vector at each point is obtained from a Bayesian calibration (BC) where the quantity 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.

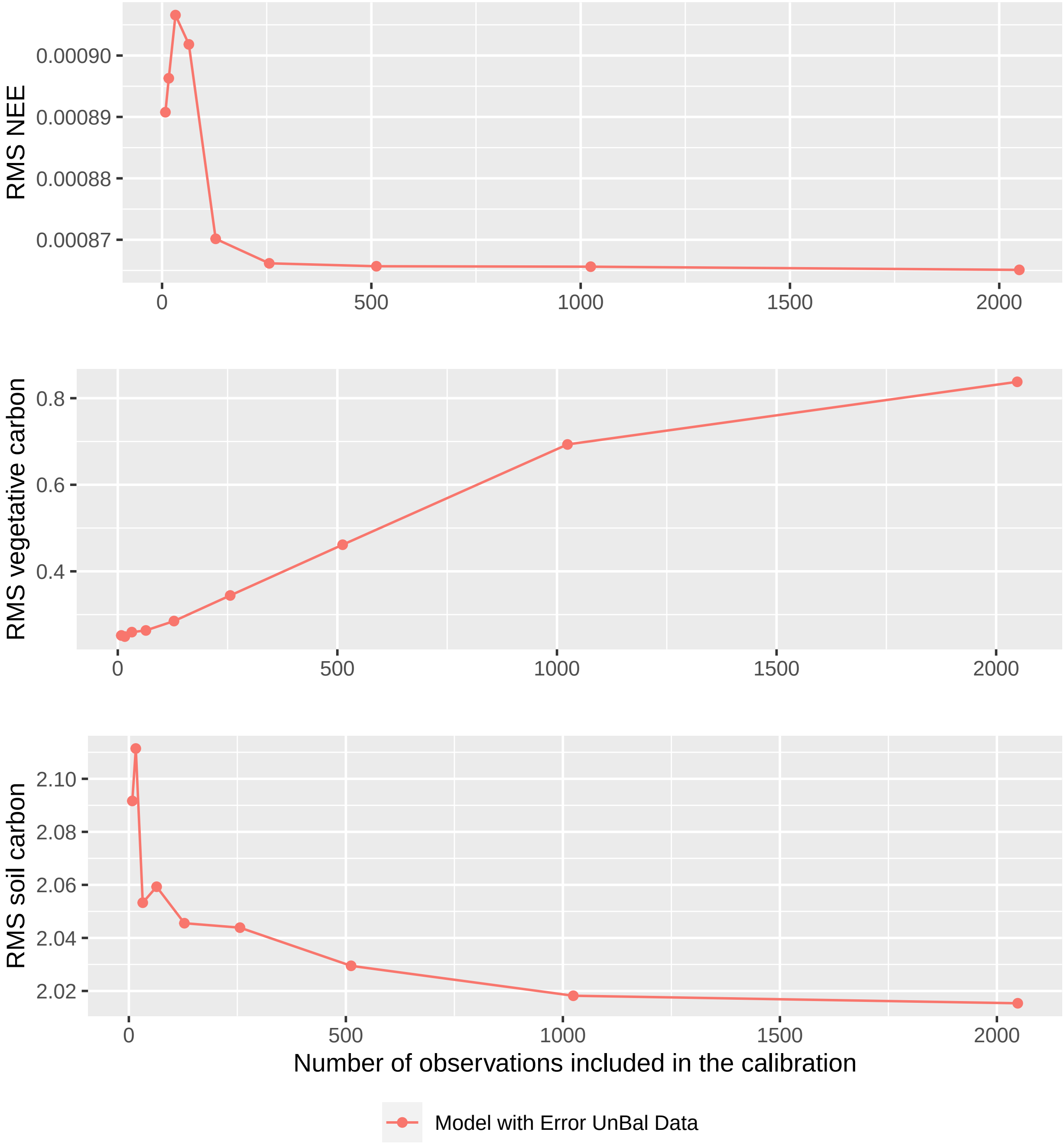

Figure 16: Each point in the three graphs (NEE, vegetative carbon, and soil carbon) represents the RMS difference between the VSEM model and virtual observations run with different maximum a posteriori (MAP) vectors. The MAP vector at each point is obtained from a Bayesian calibration (BC) where the 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.

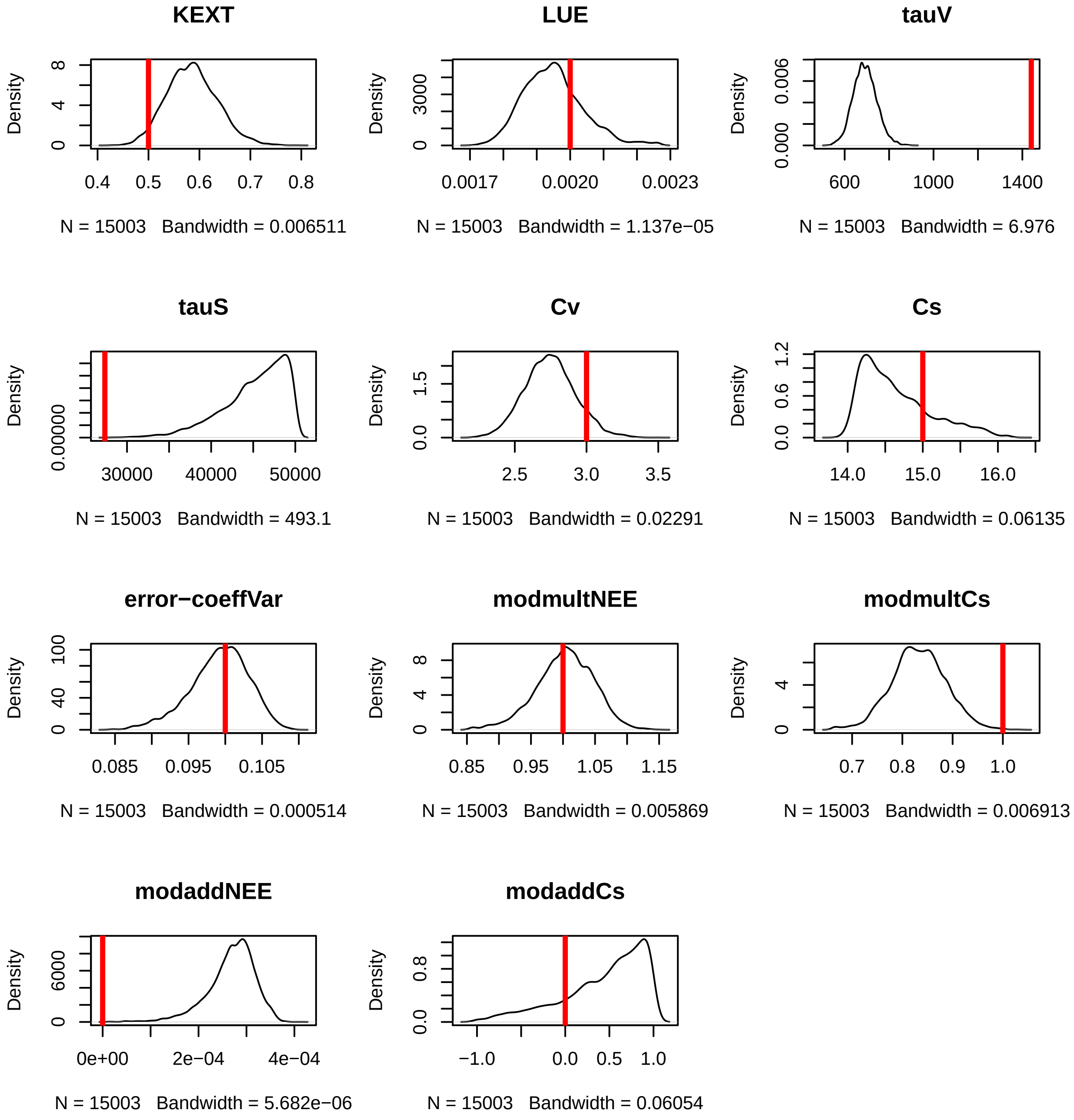

Figure 17: Model with error and unbalanced data with additive and multiplicative parameters to represent model error. Marginal posterior distribution of model parameters and intital states. The red line marks the ‘true’ parameter values.

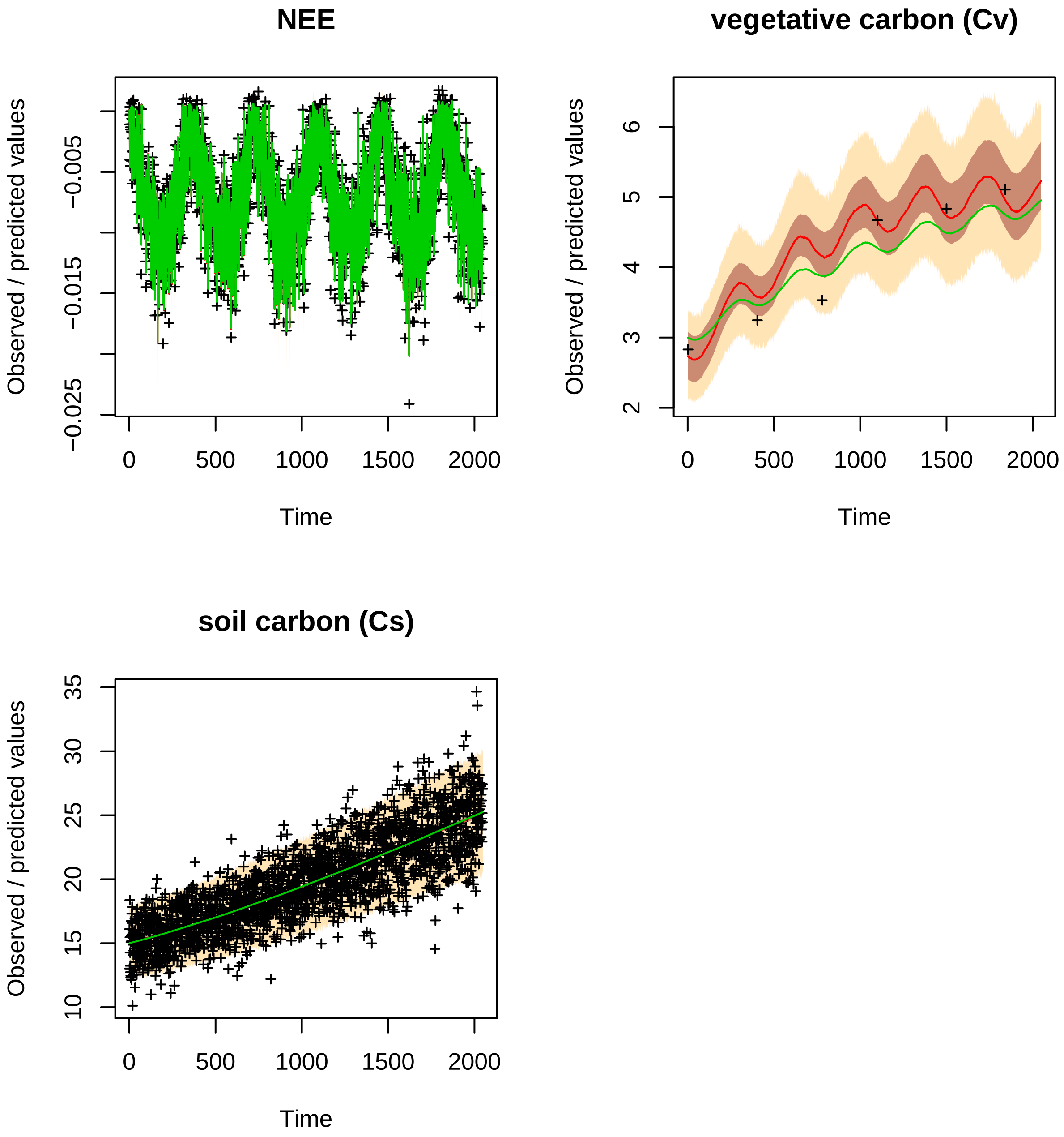

Figure 18: Model with error and unbalanced data with additive and multiplicative parameters to represent model error. Observations included in the calibration marked with a '+.' Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading 2.5% 97.5% quantile posterior distribution. Light brown shading 2.5% 97.5% predictive interval.

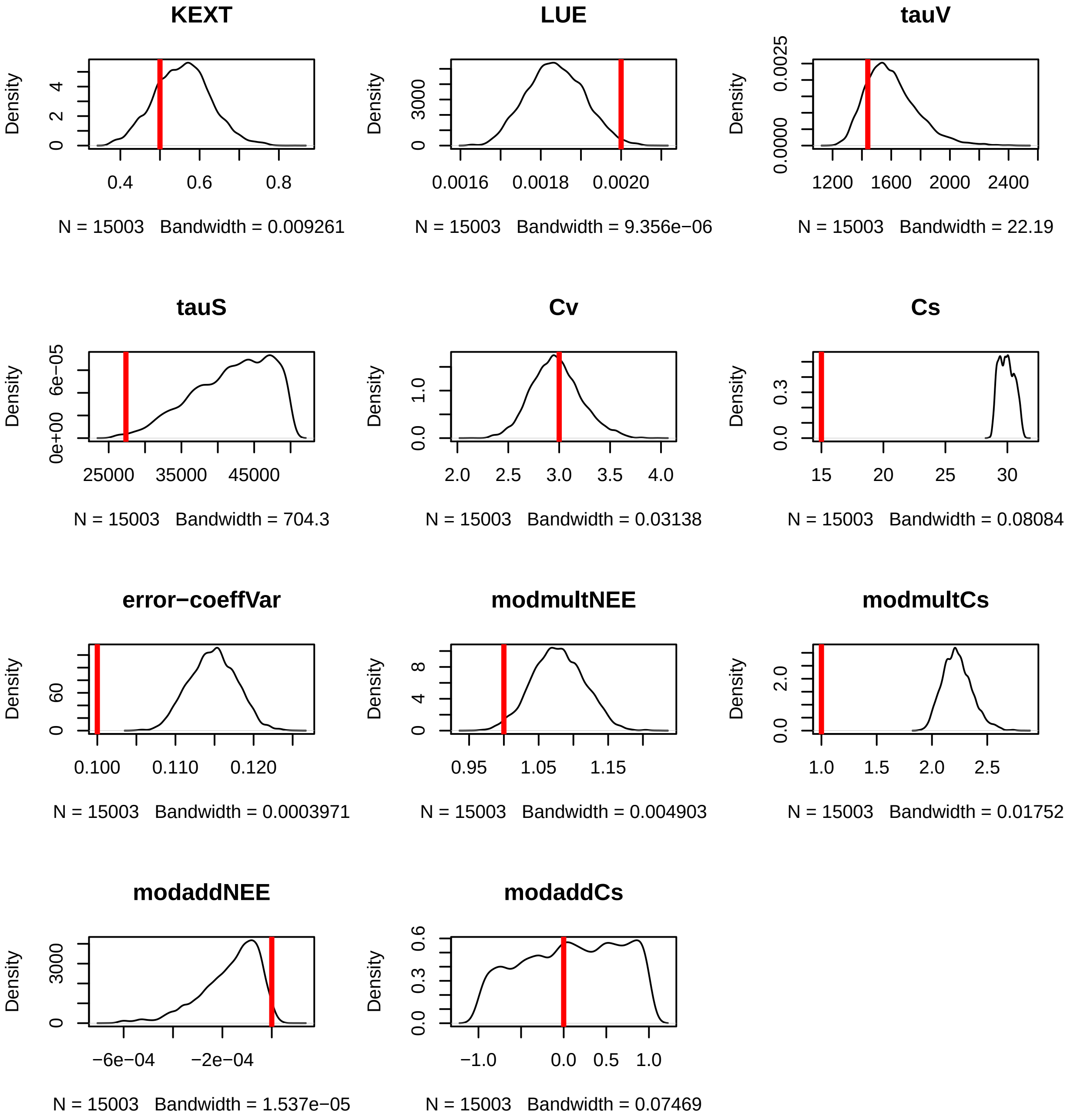

Figure 19: Perfect model and and unbalanced data with a multiplicative bias and additive and multiplicative parameters to represent the bias. Marginal posterior distribution of model parameters and intital states. The red line marks the 'true' parameter values.

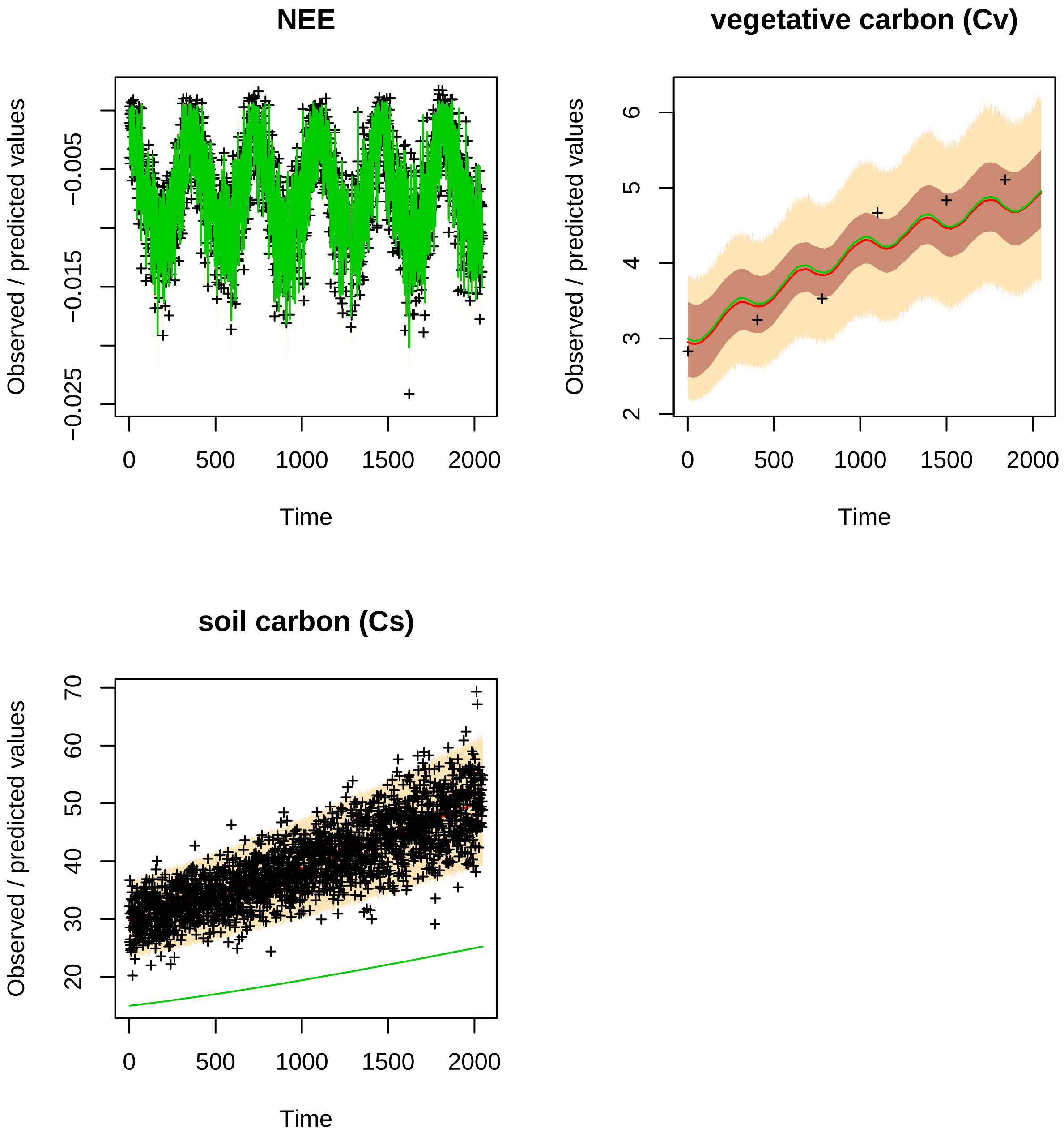

Figure 20: Perfect model and and unbalanced data with a multiplicative bias and additive and multiplicative parameters to represent the bias. Observations included in the calibration marked with a '+.' Red line 50% quantile posterior distribution. Green line is the 'true' model output. Dark brown shading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

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

- refer to Fig. (21) and (22)
- as above many parameters improved (KEXT, LUE, tauS, initial Cv)
- modmultCs value centered ~1.8 compromise between value in Fig. (19) for bias only and Fig. (17) model error only.
- vegetative pool much improved with
  - 5 out of 6 data points within posterior credible interval
  - 50% quantile line now much closer to the 'true' line.
  - similar to Fig. (18)

6 Discussion

6.1 Identifying the issue with unbalanced dataset BC

- Unbalanced data are not an issue though uncertainty is larger.
- For a model with a significant structural error or systematic bias in the calibration data parameters 'absorb' the error so that model output is not too far away from the data.
- With a model structural error or data with a systematic bias general sense is that sparse data are somewhat ignored in the BC in favour of the plentiful data.
  - This is what we often observe with unbalanced datasets in BC. For example, Cameron et al (2018) but results here make it apparent that the issue is the presence of the structural error in the model rather than that the calibration data are unbalanced.

6.2 Diagnostic tool introduced

As the Box aphorism says "All models are wrong but some are useful." A key consideration then is to what extent errors in the model (and biases in the data) give rise to the issues in calibration with imbalanced data that we have identified here. Or to put it another way, can we identify a signature for the issue that can be diagnosed so that modellers will know whether and how sever the issue may be in their calibration? To help with this we have developed a methodology that can be used as a diagnostic tool. We perform the diagnosis by starting with equal numbers of observations in the calibration and incrementally increasing the imbalance by allowing in more of the plentiful observations. If after calibration, model RMS errors versus observations increase for the sparse observations whilst RMS errors decrease for the plentiful observations, as shown in Fig. (22), as the imbalance increases, then we know that our calibration has this issue. If we create the diagnostic graph then it is possible also to assess how sever the problem is and at what point the increasing imbalance in the data starts to have a significant impact on the calibration. To overcome the issue there are ad-hoc measures that can be taken but as we have identified here, the underlying issue that needs to be addressed to make progress is to take account of model structural and data bias errors in the calibration.

6.3 Representing model and data error in BC helps to alleviate the issue

- In this very simple example we were able to demonstrate a significant improvement by including terms in the likelihood to represent model and data error.
  - In more real-world applications, representing model and data error in BC will be much more challenging but the analysis demonstrated here shows how to deal with the issue of calibrating with unbalanced datasets without resorting to ad hoc methods.

6.4 Is observational error too simple?

- Likelihood term is the same as the observational error term.

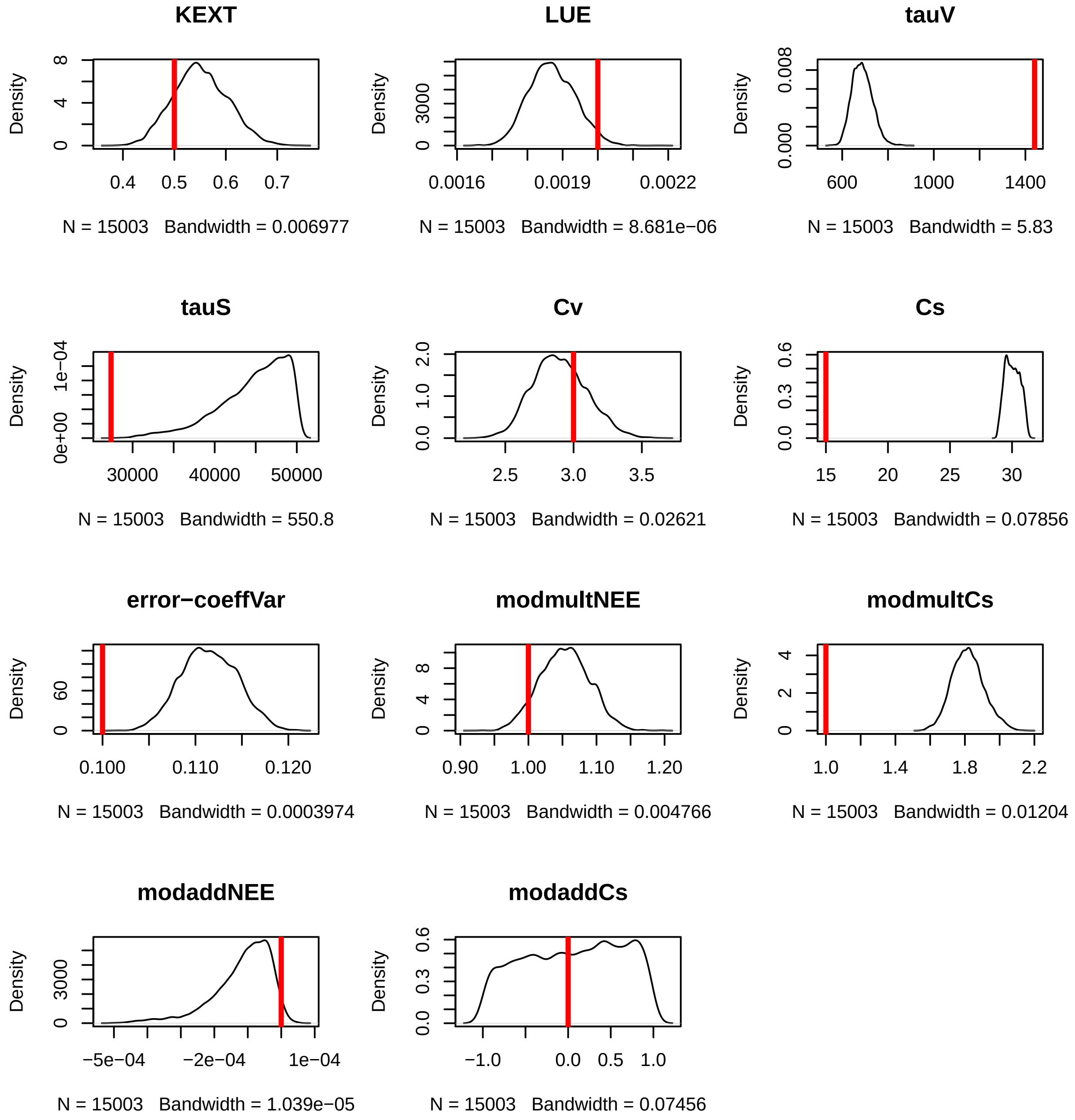

Figure 21: Model with error and and unbalanced data with a multiplicative bias and additive and multiplicative parameters to represent model error and the data bias. Marginal posterior distribution of model parameters and intital states. The red line marks the ‘true’ parameter values.

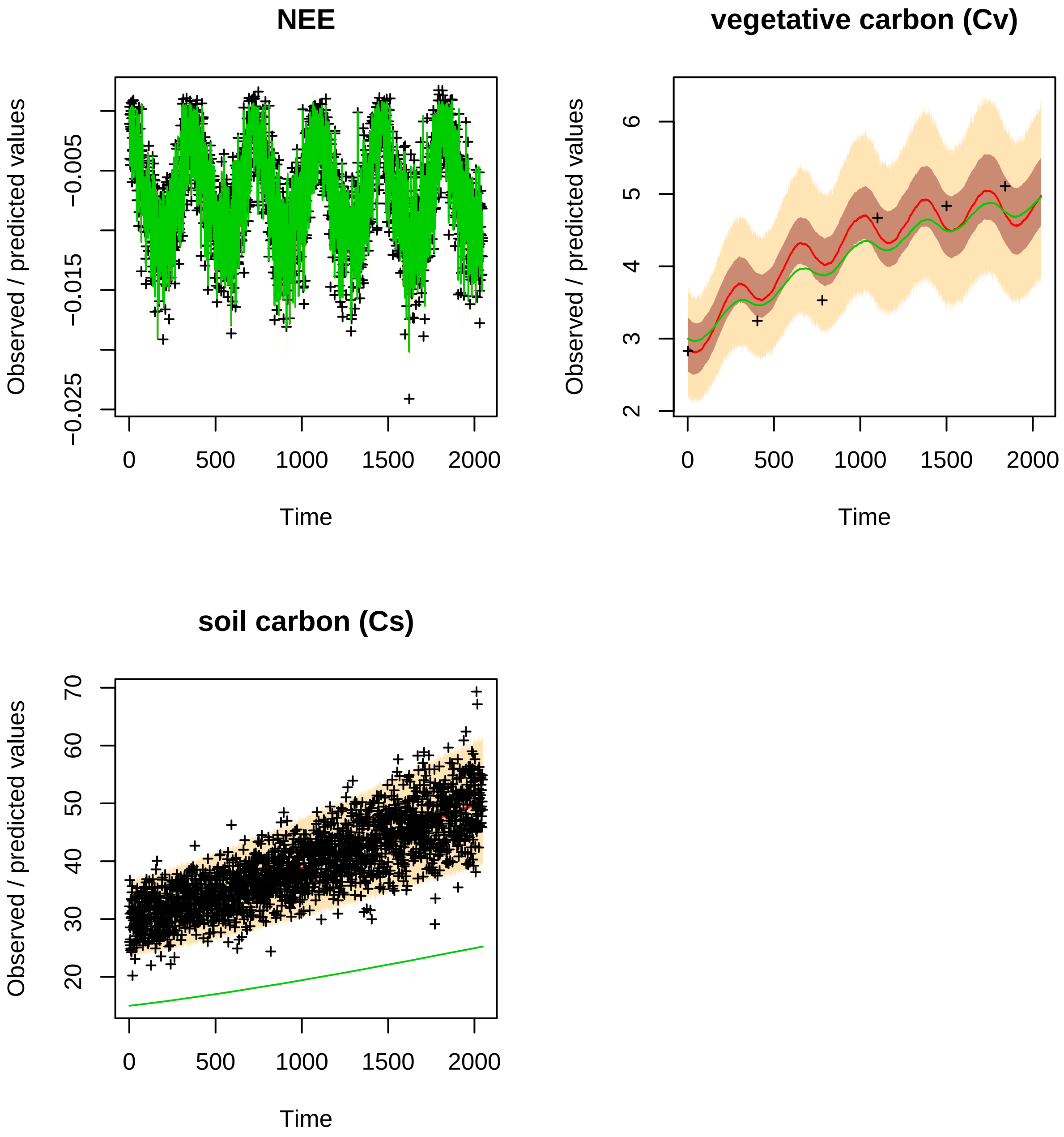

Figure 22: Model with error and and unbalanced data with a multiplicative bias and additive and multiplicative parameters to represent model error and the data bias. Observations included in the calibration marked with a '+.' Red line 50% quantile posterior distribution. Green line is the 'true' model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

6.5 Eddy covariance data doesn't close the budget

- Would never be able to match the data with a model that conserves energy.

6.6 Model with error and balanced data doesn't show that the model has an error.

- Model is able to match the data quite well so doesn't 'uncover' the serious model structural error.

7 References

Gill, A. E. 1980. "Some Simple Solutions for Heat-Induced Tropical Circulation." *Quart. J. R. Met. Soc.* 106: 447–62.

Keenan, Trevor F., Eric A Davidson, J William Munger, and Andrew D Richardson. 2013. "Rate my data: quantifying the value of ecological data for the development of models of the terrestrial carbon cycle." *Ecological Applications* 23 (1): 273–86. http://www.ncbi.nlm.nih.gov/pubmed/23495651.

MacBean, Natasha, Philippe Peylin, Frédéric Chevallier, Marko Scholze, and Gregor Schürmann. 2016. "Consistent assimilation of multiple data streams in a carbon cycle data assimilation system." *Geoscientific Model Development Discussions* 9: 3569–88. https://doi.org/10.5194/gmd-2016-25.

Richardson, Andrew D., Mathew Williams, David Y. Hollinger, David J P Moore, D Bryan Dail, Eric a Davidson, Neal a Scott, et al. 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.

Thum, T., N. MacBean, P. Peylin, C. Bacour, D. Santaren, B. Longdoz, D. Loustau, and P. Ciais. 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.