Bayesian Inference and Predictive Performance of Soil Respiration Models in the Presence of Model Discrepancy

Ahmed S. Elshall[1,2], Ming Ye[3,*], Guo-Yue Niu[4,5] and Greg A. Barron-Gafford[4,6]

[1] Department of Earth Sciences, University of Hawai'i Manoa, Honolulu, Hawaii, USA
[2] Water Resources Research Center, University of Hawai'i Manoa, Honolulu, Hawaii, USA
[3] Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, Florida, USA
[4] Biosphere 2, University of Arizona, Tucson, Arizona, USA
[5] Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA
[6] School of Geography and Development, University of Arizona, Tucson, Arizona, USA

*Corresponding Author: Ming Ye, Telephone: (850) 644-4587, Email: mye@fsu.edu

Submitted for publication in Geoscientific Model Development

April, 2019

48 **Key Points**

50 (1) Bayesian inference and prediction are useful to evaluate multiple soil respiration models

51 with different levels of model complexity.

52 (2) Data models used in Bayesian inference have substantial impacts on model parameter

53 distributions and subsequently model predictions.

54 (3) Using exponential power distribution and considering heteroscedasticity in data models

55 improve Bayesian inference and prediction.

65

66

67

68 Keywords: Soil respiration, modeling, Bayesian, likelihood function, data model, autocorrelation,

69 heteroscedasticity, skew exponential power distribution, cross-validation, scoring rule

70

71

Abstract

Bayesian inference of microbial soil respiration models is often based on the assumptions that the residuals are independent (i.e. no temporal or spatial correlation), identically distributed (i.e. Gaussian noise) and with constant variance (i.e. homoscedastic). In the presence of model discrepancy, since no model is perfect, this study shows that these assumptions are generally invalid in soil respiration modeling such that residuals have high temporal correlation, an increasing variance with increasing magnitude of CO_2 efflux, and non-Gaussian distribution. Relaxing these three assumptions stepwise results in eight data models. Data models are the basis of formulating likelihood functions of Bayesian inference. This study presents a systematic and comprehensive investigation of the impacts of data model selection on Bayesian inference and predictive performance. We use three mechanistic soil respiration models with different levels of model fidelity (i.e. model discrepancy) with respect to number of carbon pools and explicit representations of soil moisture controls on carbon degradation, and accordingly have different levels of model complexity with respect to the number of model parameters. The study shows data models have substantial impacts on Bayesian inference and predictive performance of the soil respiration models such that: (i) the level of complexity of the best model is generally justified by the cross-validation results for different data models; (ii) not accounting for heteroscedasticity and autocorrelation might not necessarily result in biased parameter estimates or predictions, but will definitely underestimate uncertainty; (iii) using a non-Gaussian data model improves the parameter estimates and the predictive performance; and (iv) separate accounting for autocorrelation or joint inversion of correlation and heteroscedasticity can be problematic and requires special treatment. Although the conclusions of this study are empirical, the analysis may provide insights for selecting appropriate data models for soil respiration modeling.

1   Introduction

Developing accurate soil respiration models is important for realistic projection of global carbon [C] cycle, as global soils store 2,300Pg carbon, an amount more than 3 times that of the atmosphere (Schmidt et al., 2011) and release 60–75 Pg C/yr, about 7 times more CO_2 to the atmosphere than all human-caused emissions (Le Quéré et al., 2014). The major work on soil respiration modeling has been focused on advancing knowledge about model inputs and calibration data (e.g. Janssens et al., 2003; Peters et al., 2007; Scott et al., 2009; Barron-Gafford et al., 2011; Hilton et al., 2014) and on developing more advanced models for better representing soil microbial processes (e.g. Schimel and Weintraub, 2003; Allison et al., 2010; Davidson et al., 2011; Wieder et al., 2013, 2015; Xu et al., 2014; Zhang et al., 2014) . Integration of data and models is indispensable for improving predictability of the terrestrial carbon cycle, and statistical modeling is a vital tool for the model-data integration (Luo et al., 2011, 2014; Wieder et al., 2015). In addition, use of state-of-the-art statistical methods is necessary to accurately quantify uncertainty in parameters and structures of soil respiration models for improvement and practical uses of the models (Katz et al., 2013). A data model that is also known as a residuals model or an error model is used to characterize residuals (i.e., the difference between data and corresponding model simulations). While a large number of data models have been used (e.g. Elshall et al., 2018; Scholz et al., 2018) to our knowledge comprehensive and systematic evaluation of data models for soil respiration modeling has not been reported in literature.

The objectives of this study are to evaluate the impacts of data models on Bayesian inference and predictive performance of three mechanistic soil respiration models, and to use the evaluation results to make broader recommendations. The three models were developed by Zhang et al. (2014) to simulate the Birch effect (the peak soil microbial respiration pulses in response to episodic

118   rainfall pulses) at a site scale and a short temporal scale; understanding the Birch effect is important

119   for gaining mechanistic understanding of CO_2 efflux production (Högberg and Read, 2006; Vargas

120   et al., 2011). The models of Zhang et al. (2014) are based on an existing four-carbon pool model,

121   but have additional carbon pools and/or explicit representations of soil moisture controls on carbon

122   degradation and microbial uptake rates. The models were calibrated, and Bayesian model selection

123   was used to select the best model (Zhang et al., 2014). However, this effort was based on a single

124   data model. It is unknown whether the best model still remains the best (in terms of reproducing

125   the both calibration data and the cross-validation data) if a different data model is used. In addition,

126   since predictive performance of the models was not evaluated in Zhang et al. (2014), it is unknown

127   whether the best model will give the best predictions. These two questions are addressed in this

128   study by considering eight data models and by evaluating predictive performance in a manner of

129   cross-validation. The top two models (also the two most high fidelity models) ranked by Zhang et

130   al. (2014) are considered in this study, and the worst model (also the low fidelity model) is also

131   considered in this study for comparison. We use the terms model fidelity and model discrepancy

132   interchangeably. Model fidelity refers to the degree of realism of representing our scientific

133   knowledge with respect to the real world system. That is a high fidelity model has less discrepancy.

134   Evaluating predictive performance for the three models with different degrees of fidelity provides

135   more insights than a single model.

136   Bayesian inference in general uses the Bayes' theorem to update the prior distributions of

137   model parameters to posterior parameter distributions given a likelihood function of data. The

138   mathematical formulation of the (formal and informal) likelihood function requires a probabilistic

139   data model that however is intrinsically unknown due to unknown errors in all model components

140   such as model structures, parameters, and driving forces. Bayesian inference of soil respiration

models often adopts the assumption of independent, normally distributed and homoscedastic residuals (e.g. Ahrens et al., 2014; Bagnara et al., 2015, 2018; Barr et al., 2013; Barron-gafford et al., 2014; Braakhekke et al., 2014; Braswell et al., 2015; Correia et al., 2012; Du et al., 2015, 2017; Hararuk et al., 2014; Hashimoto et al., 2011; He et al., 2018; Klemedtsson et al., 2008; Menichetti et al., 2016; Raich et al., 2002; Ren et al., 2013; Richardson and Hollinger, 2005; Steinacher and Joos, 2016; Tucker et al., 2014; Tuomi et al., 2008; Xu et al., 2006; Yeluripati et al., 2009; Yuan et al., 2012, 2016; Zhang et al., 2014; Zhou et al., 2010). These assumptions are conveniently adopted to satisfy the requirement of using an unknown probability model in Bayesian statistics, which is called "a basic dilemma" by (Box and Tiao, 1992).

Postulating the data models is always based on assumptions about residual statistics, and the most widely used assumptions are paired as follows: (i) independent vs. correlated residuals, (ii) homoscedastic vs. heteroscedastic residuals, and (iii) Gaussian vs. non-Gaussian residuals. For soil respiration modeling few studies have relaxed the non-correlation assumption(e.g. Cable et al., 2008, 2011; Li et al., 2016b), the homoscedasticity assumption (e.g. Berryman et al., 2018; Elshall et al., 2018; Ogle et al., 2016; Tucker et al., 2013), and the non-Gaussian and homoscedasticity assumptions (e.g. Elshall et al., 2018; Ishikura et al., 2017; Kim et al., 2014). The recent study of Scholz et al. (2018) relaxed these three assumptions using the generalized likelihood function developed by Schoups and Vrugt (2010). However, few studies have focused on investigating appropriateness and impact of these assumptions for soil respiration modeling, by relaxing the independent residuals assumption (Ricciuto et al., 2011) and the Gaussian residuals assumption (Ricciuto et al., 2011; van Wijk et al., 2008). By relaxing these three assumptions stepwise resulting in eight data models, to our knowledge this is the first study that systematically evaluates the impact of data model selection on Bayesian inference and predictive performance of soil

164 respiration modeling. In addition, to our knowledge this is the first soil respiration modeling study

165 that investigates the impact of data models in relation to model fidelity.

166 Relaxing these three assumption results in eight data models, which are shown in details in

167 Section 2. For example, combining the assumptions of independent, homoscedastic, and Gaussian

168 residuals leads to the standard least squares data model. This model is the simplest one among the

169 eight data models, since it requires only one parameter, i.e., the constant variance of the Gaussian

170 distribution. Note that there is a difference between the soil respiration model parameters and the

171 data model parameters. They technically can be jointly estimated, but one arises from assumptions

172 about soil respiration processes, and the other from assumptions about the residuals. Relaxing the

173 homoscedastic assumption to heteroscedastic gives the weighted least squares data model. It is

174 more complex because it has extra parameters to account for multiple variances for multiple data.

175 Whenever one or combinations of the three assumptions (independence, homoscedasticity, and

176 normality) are relaxed, the resulting data models become more complex and require more

177 parameters. Such systematic evaluation of data models (McInerney et al., 2017; Smith et al. 2010b,

178 2015) is necessary to evaluate appropriateness of residuals assumptions and their impacts on

179 Bayesian inference.

180 The assumptions of heteroscedastic, correlated, and non-Gaussian residuals are accounted for

181 by using the method of Schoups and Vrugt (2010) in the following procedure: (i) the correlation

182 is removed from the residuals by using an autoregressive model; (ii) the resulting residuals are

183 normalized by a linear model of variance; and (iii) the normalized residuals are characterized by

184 using the skew exponential power distribution. The data model parameters (i.e., coefficients of the

185 autoregressive model, the linear variance model, and the skew exponential power distribution) are

186 not specified by users, but estimated together with soil respiration model parameters during the

187   Bayesian inference. The skew exponential power distribution is general in that by adjusting the

188   values of its kurtosis and skewness parameters the distribution can produce other distributions such

189   as the Laplace distribution (van Wijk et al., 2008; Ricciuto et al., 2011) and other distributions

190   through using an exponential model with different kurtosis parameters (Tang and Zhuang, 2009).

191   It is worth pointing out that there exist other methods to account for the three assumptions. Evin

192   et al. (2013) suggested accounting for residual heteroscedasticity before accounting for residual

193   autocorrelation. Lu et al. (2013) developed an iterative two-stage procedure to separately estimate

194   physical model parameters and data model parameters. Evin et al. (2014) developed a similar

195   procedure to first estimate model parameters and then estimate heteroscedasticity and

196   autocorrelation parameters. While this study uses the method of Schoups and Vrugt (2010),

197   exploring other methods is warranted in future studies.

198   After investigating the impacts of the data models on Bayesian inference, this study evaluates

199   the impacts of the data models on predictive performance of the three soil respiration models.

200   Using random samples generated during the Bayesian inference, a prediction ensemble is produced

201   for each soil respiration model. The ensemble is used to evaluate predictive performance of the

202   models in a stochastic sense by estimating to what extent the models can predict future events. The

203   evaluation in this study is done in a cross-validation manner by splitting the dataset of CO_2 efflux

204   into two parts for Bayesian inference and cross-validation, respectively. The evaluation of

205   predictive performance is important because different data models may give different parameter

206   distributions and accordingly different predictive performance. For example, the study of van Wijk

207   et al. (2008) concluded that the choice of the residual function is crucial to achieve accurate model

208   prediction and parameter estimation. Shi et al. (2014) showed that the posterior parameter

209   distributions and predictive performance given by two data models (weighted least square and

210 skew exponential power distribution after removing heteroscedasticity and autocorrelation) are

211 dramatically different, and a definitive conclusion was drawn that one data model is better than

212 the other. The evaluation of predictive analysis is conducted for the following two cases: (1) the

213 prediction ensemble is generated by random samples of the soil respiration models only (i.e.

214 credible interval), and (2) the prediction ensemble is generated by random samples of not only the

215 soil respiration models but also the data models (i.e. predictive interval). The two cases lead to

216 different conclusions about the predictive performance. It is expected that the evaluation of

217 predictive performance conducted in this study can help select the most appropriate data model to

218 achieve optimal model predictions.

219   The remainder of the paper is organized as follows. Section 2 starts with a description of the

220 evolving data models and their corresponding likelihood functions used in Bayesian inference,

221 followed by a brief summary of the three soil respiration models. The results of Bayesian inference

222 are discussed in Section 3 and Section 4, addressing the data model implications on parameter

223 estimation and predictive performance, respectively. Section 5 summarizes the key findings and

224 limitations of this study, and provides recommendations for approaching data model selection.

225 **2   Methodology**

226   This section starts with a description of the eight data models that account for the three pairs

227 of assumptions about residuals in a stepwise manner in Section 2.1. The data models are used to

228 build the likelihood functions used in Section 2.2 for Bayesian inference. The three soil respiration

229 models and observations of CO_2 efflux are described in Sections 2.3 and 2.4, respectively. Metrics

230 for evaluating predictive performance are presented in Section 2.5.

231

2.1   Data models

This study considers eight evolving data models starting from a data model that assumes independent, homoscedastic, and Gaussian residuals to a data model that relaxes all the three assumptions. The eight data models are based on the generic normalized residual,

$$a_t = \frac{\varepsilon_t}{\sigma_t} \qquad a_t \sim X, \tag{1}$$

where $\varepsilon_t = d_t - Y_t$ is the residual (the difference between data d_t and its corresponding model simulation Y_t) at time or location t; σ_t is the standard deviation of the residual; and X is the probability density function (PDF) of a_t. The eight data models are formulated with different forms of ε_t, σ_t, and X. The standard least square (SLS) data model is

$$a_t = \frac{\varepsilon_t}{\sigma_0} \qquad a_t \sim N(0,1), \tag{2}$$

where $\sigma_t = \sigma_0$ is a constant for all the data (i.e., homoscedasticity), and X is the standard normal distribution, $N(0,1)$. The unknown parameter σ_0 is estimated jointly with unknown physical model parameters. If σ_t is not a constant (i.e., heteroscedastic), SLS becomes the weighted least squared (WLS) data model. While heteroscedasticity can be accounted for through residuals transformation (e.g. Thiemann et al., 200; Smith et al., 2010b) or other similar approaches (Gragne et al., 2015) , a linear heteroscedastic model $\sigma_t = \sigma_0 + \sigma_1 Y_t$ is assumed here by following the studies of Thyer et al. (2009), Schoups and Vrugt (2010), and Evin et al. (2013, 2014). With the linear model, there is no need to estimate σ_t for each data. Instead, σ_t is calculated by estimating only two parameters, σ_0 and σ_1. The WSL data model is written as

$$a_t = \frac{\varepsilon_t}{\sigma_0 + \sigma_1 Y_t} \qquad a_t \sim N(0,1). \tag{3}$$

252   The two unknown parameters σ_0 and σ_1 are estimated jointly with unknown physical model

253   parameters. The linear model assigns smaller weight to the data with larger simulation, Y_t. If the

254   simulation is small and $\sigma_0 \gg \sigma_1 Y_t$, the weight becomes constant for all data. Both SLS and WLS

255   assume that a_t is independently and identically distributed.

256   It is not uncommon that residuals are correlated in space and time, due to propagation of

257   measurement errors (Tiedeman and Green, 2013) and model structure errors (Evin et al., 2014;

258   Kavetski et al., 2013; Lu et al., 2013). The temporal correlation that occurs in the numerical

259   example of this study can be accounted for by using a p-order autoregressive model. This leads to

260   the data model of standard least square with autocorrelation (SLS-AC),

261
$$a_t = \frac{\varepsilon_t - \sum_{i=1}^{p} \phi_i \varepsilon_{t-i}}{\sigma_0} \qquad a_t \sim N(0,1) \tag{4}$$

262   where p is the order of autocorrelation, and ϕ_i is an autocorrelation coefficient. The unknown ϕ_i

263   and σ_0 are estimated together with unknown model parameters. By extending the concept of

264   correlated residuals to WLS leads to the weight least square with autocorrelation (WLS-AC),

265
$$a_t = \frac{\varepsilon_t - \sum_{i=1}^{p} \phi_i \varepsilon_{t-1}}{\sigma_0 + \sigma_1 Y_t} \qquad a_t \sim N(0,1) \tag{5}$$

266   The unknown parameters of σ_0, σ_1, and ϕ_i are estimated jointly with physical model

267   parameters. Equations (2) – (5) assume that the residuals are Gaussian.

268   The next four data models are similar to the previous four models except that the standard

269   normal distribution of a_t is replaced by the skew exponential power distribution, $SEP(0,1,\xi,\beta)$,

270   with zero mean and unit standard deviation (Schoups and Vrugt, 2010)

271 $$p(a_t \mid \xi, \beta) = \frac{2\sigma_\xi}{\xi + \xi^{-1}} \omega_\beta \exp\left[-c_\beta \left|a_{\xi,t}\right|^{2/(1+\beta)}\right],$$ (6)

272 where ξ is skewness, β is kurtosis, $a_{\xi,t} = (\mu_\xi + \sigma_\xi a_t)\big/\xi^{sign(\mu_\xi + \sigma_\xi a_t)}$, $\mu_\xi = M(\xi - \xi^{-1})$,

273 $$\omega_\beta = \frac{\Gamma^{1/2}[3(1+\beta)/2]}{(1+\beta)\Gamma^{3/2}[(1+\beta)/2]}, \qquad \sigma_\xi = \sqrt{(1 - M^2)(\zeta^2 + \zeta^{-2}) + 2M^2 - 1},$$

274 $$M = \frac{\Gamma[1+\beta]}{\Gamma^{1/2}[3(1+\beta)/2]\Gamma^{1/2}[(1+\beta)/2]}, \text{ and } c_\beta = \left(\frac{\Gamma[3(1+\beta)/2]}{\Gamma[(1+\beta)/2]}\right)^{1/(1+\beta)}$$ are derived variables of β and

275 ξ, and $\Gamma[.]$ is the gamma function. The kurtosis parameter $\{\beta \in \mathbb{R} : -1 \le \beta \le 1\}$ determines the

276 peakness of the pdf such that the β values of -1, 0, and 1 give uniform, Gaussian and Laplace

277 distributions, respectively. The skewness parameter $\{\xi \in \mathbb{R} : 0.1 \le \xi \le 10\}$ determines the

278 skewness of the pdf such that the ξ values of 0.1, 1, and 10 give positively skewed, symmetric,

279 and negatively skewed distributions, respectively. Setting $\beta = 0$ and $\xi = 1$ leads to $\mu_\xi = 0$, $\sigma_\xi = 1$

280 , $\omega_\beta = 1/\sqrt{2\pi}$, $c_\beta = 1/2$ and $a_{\xi,t} = a_t$, and the skew exponential power distribution

281 $SEP(0,1,\xi = 1, \beta = 0)$ becomes the standard normal distribution,

282 $$p(a_t \mid \xi = 1, \beta = 0) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(a_t)^2\right].$$ (7)

283 which is the data model of SLS in equation (2).

284   Replacing $a_t \sim N(0,1)$ with $a_t \sim SEP(0,1,\xi,\beta)$ in equations (2) – (5) leads to the data models

285 SEP, WSEP, SEP-AC, and WSEP-AC as follows,

286 $$a_t = \frac{\varepsilon_t}{\sigma_0} \qquad a_t \sim SEP(0,1,\xi,\beta)$$ (8)

287 $$a_t = \frac{\varepsilon_t}{\sigma_0 + \sigma_1 Y_t} \qquad a_t \sim SEP(0,1,\xi,\beta).$$ (9)

288 $$a_t = \frac{\varepsilon_t - \sum_{i=1}^{p} \phi_i \varepsilon_{t-1}}{\sigma_0} \qquad a_t \sim SEP(0,1,\xi,\beta) \tag{10}$$

289 $$a_t = \frac{\varepsilon_t - \sum_{i=1}^{p} \phi_i \varepsilon_{t-1}}{\sigma_0 + \sigma_1 Y_t} \qquad a_t \sim SEP(0,1,\xi,\beta) \tag{11}$$

290 In comparison with the Gaussian data models, the SEP-based data models have two more

291 parameters (ξ and β) to be estimated jointly with physical model parameters. Data model WSEP-

292 AC, which is known as the generalized likelihood function, is the most commonly used SEP-based

293 data model (e.g. Vrugt and Ter Braak, 2011; Hublart et al., 2016; Scholz et al., 2018). A summary

294 table of the eight data models with corresponding parameters is provided in the supplementary

295 materials.

296 **2.2   Bayesian inference and likelihood functions**

297 Consider a Bayesian inference problem for a nonlinear model, f, used to simulate state

298 variables (e.g., CO_2 efflux), $d = Y(\theta) + \varepsilon$, where d is a vector of data, θ is a vector of model

299 parameters, and ε is a vector of residuals that may include errors in data, model parameters, and

300 model structures. The goal of Bayesian inference is to estimate the posterior distributions, $p(\theta|d)$,

301 of model parameters, θ, given data, d, using Bayes' theorem (Box and Tiao, 1992)

302 $$p(\theta \mid d) = \frac{p(d \mid \theta) p(\theta)}{\int p(d \mid \theta) p(\theta) d\theta} \tag{12}$$

303 where $p(\theta)$ is the prior distribution, and $p(d|\theta)$ is the likelihood function to measure goodness-of-

304 fit between model simulations, $Y(\theta)$, and data, d. The prior distribution can be obtained from data

305 of previous studies (e.g. Elshall and Tsai, 2014) or expert judgment. When prior information is

306 lacking, a common practice is to assume uniform distributions with relatively large parameter

307 ranges so that the prior distributions do not affect the estimation of posterior distributions.

308   The data models above can be used to construct the likelihood functions. For the Gaussian data

309   models given in equations (2) – (5), the corresponding Gaussian likelihood functions are

310   straightforward, and an example is equation (7). For the SEP data models, the corresponding

311   likelihood that is called generalized likelihood function is (Schoups and Vrugt, 2010)

312
$$p(\boldsymbol{d}\,|\,\boldsymbol{\theta}) = p(\boldsymbol{\varepsilon}_t\,|\,\boldsymbol{\theta}) = \prod_{t=1}^{n} \sigma_t^{-1} \frac{2\sigma_\xi}{\xi + \xi^{-1}} \omega_\beta \exp\left(-c_\beta \left|a_{\xi,t}\right|^{2/(1+\beta)}\right).$$
(13)

313   where n is the dimension of $\boldsymbol{d}$. The Gaussian likelihood functions are special case of the generalized

314   likelihood functions. For example, by setting $\beta=0$, $\xi=1$, $\phi_i=0$, $\sigma_t=\sigma_0$, $\sigma_\xi=1$, $\mu_\xi=0$,

315   $\omega_\beta = 1/\sqrt{2\pi}$, $c_\beta = 1/2$, and $a_{\xi,t}=a_t$, equation (13) becomes the likelihood function corresponding

316   to the SLS data model. Replacing $\sigma_t=\sigma_0$ by $\sigma_t=\sigma_0+\sigma_1 E_t$, equation (13) becomes the likelihood

317   function of the WLS data model.

318   In this study, the posterior distributions of the data model parameters are jointly estimated with

319   the soil respiration model parameters using the MT-DREAM(ZS) code (Laloy and Vrugt, 2012).

320   MT-DREAM(ZS) implements a Markov chain Monte Carlo (MCMC) algorithm by running

321   multiple Markov chains in parallel with adaptive proposal distribution, multiple-try sampling, and

322   sampling from an archive of past states. These state-of-the-art features assist in overcoming

323   common challenges in the sampling space such as multimodality, ill-conditioning, and high

324   dimensionality, and thus allow for accurate exploration of the targeted distributions.

325   **2.3   Soil respiration models**

326   Zhang et al. (2014) studied the Birch effect (the peak soil microbial respiration pulses in

327   response to episodic rainfall pulses), and developed five models, evolving from an existing four-

328   carbon pool model to models with additional carbon pools and/or explicit representations of soil

329   moisture controls on carbon degradation and microbial uptake rates. Three of the five models are

330   used in this study, and they are dented as 4C, 5C, and 6C. Note that model 4C is model 4C_NOSM

331   of Zhang et al. (2014), not their model 4C. Figure 1 is the diagram of model 6C, the most complex

332   one among the five models. The simplest one, model 4C, has four carbon pools, i.e., soil organic

333   carbon (SOC), dissolved organic carbon (DOC), microbial biomass (MIC), and enzymes (ENZ),

334   and does not consider the soil moisture control on carbon degradation and microbial uptake rates.

335   Models 5C and 6C have an explicit representation of soil moisture controls on the rates. Based on

336   the dual Arrhenius and Michaelis–Menten kinetics model, the original SOC degradation rate,

337   V_{decom}, is (Davidson et al., 2011; Davidson and Janssens, 2006)

338
$$V_{decom} = V_{\max} C_{ENZ} \frac{C_{SOC}}{K_m + C_{SOC}} \qquad (14)$$

339   where $V_{\max}$ [s^{-1}] is the maximum SOC degradation rate per unit enzyme when the substrates is not

340   limiting, C_{ENZ} [gCm^{-3}] is enzyme pool size, C_{SOC} [gCm^{-3}] is SOC pool size, and K_m is the half-

341   saturation for SOC. The original microbial uptake rate, V_{uptake}, is (Davidson et al., 2011; Davidson

342   and Janssens, 2006)

343
$$V_{uptake} = V_{\max\_up} C_{MIC} \frac{C_{DOC}}{K_{m\_up} + C_{DOC}} \frac{C_{O2}}{K_{m\_upO2} + C_{O2}}, \qquad (15)$$

344   where $V_{\max\_up}$ [s^{-1}] is the maximum DOC uptake rate when the substrates is not limiting, C_{MIC}

345   [gCm^{-3}] is the MIC pool size, C_{DOC} [gCm^{-3}] is the DOC pool size, C_{O2} [m^{3}m^{-3}] is the gas

346   concentration of O$_2$ in the soil pore, and K_{m_up} [gCm^{-3}] and K_{m_upO2} [m^{3}m^{-3}] are the corresponding

347   half-saturation constants for DOC and O$_2$, respectively. With the explicit representation of soil

348   moisture control, the two rates become (Zhang et al., 2014)

349
$$V_{decom} = V_{\max} C_{ENZ} \frac{C_{SOC}}{K_m + C_{SOC}} \left(\frac{\theta}{\theta_s} \right) \qquad (16)$$

350 $$V_{uptake} = V_{max\_up} C_{MIC} \frac{C_{DOC}}{K_{m\_up} + C_{DOC}} \frac{C_{O2}}{K_{m\_upO2} + C_{O2}} \left(\frac{\theta}{\theta_s} \right) \qquad (17)$$

351 where θ [-] is the volumetric soil moisture, and θ_s [-] is the porosity.

352   In addition to using the new rate equations, models 5C and 6C have more carbon pools. In

353 model 5C, DOC is split into two sub-pools for wet zone and dry zone of soil pores, and only the

354 wet DOC is used by MIC, as shown in Figure 1. The moisture-controlled microbial uptake rate

355 becomes

356 $$V_{uptake} = V_{max\_up} C_{MIC} \frac{C_{DOC\_w}}{K_{m\_up} + C_{DOC\_w}} \frac{C_{O2}}{K_{m\_upO2} + C_{O2}} \left(\frac{\theta}{\theta_s} \right). \qquad (18)$$

357 where C_{DOC_w} [gCm^{-3}] is the DOC pool size in the wet soil pores. Model 6C is more complex in

358 that ENZ is further split into two sub-pools for wet and dry pores, and both the wet and dry ENZ

359 are subject to degradation, as shown in Figure 1. The moisture-controlled SOC degradation rate

360 becomes

361 $$V_{decom} = V_{max} C_{ENZ\_W} \frac{C_{SOC}}{K_m + C_{SOC}} \left(\frac{\theta}{\theta_s} \right) \qquad (19)$$

362 for the wet ENZ and

363 $$V_{decom} = V_{max} C_{ENZ\_D} \frac{C_{SOC}}{K_m + C_{SOC}} \left(1 - \frac{\theta}{\theta_s} \right) \varepsilon_D \qquad (20)$$

364 for the dry ENZ, where C_{ENZ_W} [gCm^{-3}] is the wet soil pores enzyme pool size, C_{ENZ_D} [gCm^{-3}] is

365 the enzyme pool size in the dry soil pores, and ε_D is the catalysis efficiency of the dry zone enzyme.

366   Due to considering the moisture control and adding more soil pools, model 5C is expected to

367 be significantly better than model 4C for simulating the Birch effect. Since the accumulated ENZ

368 in dry soil is secondary, model 6C is expected to be slightly better than model 5C. In terms of

369   model structural error, model 4C has the largest model structure error, model 5C has significantly

370   less model structure error, and model 6C has the smallest model structural error. In other words,

371   model 6C has the highest model fidelity (i.e. lowest model discrepancy) among the three models.

372   As shown below, the degree of model structural error is reflected in the process of Bayesian

373   inference and verified by the cross-validation.

374   **2.4   Observations and parameter estimation**

375   Figure 2 plots the time series of 17,016 observations of soil moisture and CO_2 efflux used in

376   this study. The observations were obtained during the entire year of 2007, covering a long period

377   of dry season prior to monsoon and episodic rainfall events during monsoon. The first two third of

378   this dataset is used for the Bayesian inference, and the last one third is used for cross-validation.

379   The inference and cross-validation periods have both dry and wet periods, as shown in Figure 2.

380   The observation site is located within the Santa Rita Experimental Range (SRER, 31.8214°N,

381   110.8661°W, elevation 1,116 m) outside of Tucson, Arizona (Barron-Gafford et al., 2011; Scott

382   et al., 2009). This savanna site was covered by 22% of perennial grass, forbs and subshrubs and

383   35% of mesquite. The soils are uniformly Comoro loamy sand (77.6% sand, 11.0% clay, and

384   11.4% silt). The half-hourly atmospheric forcing data were collected from measurements through

385   an eddy covariance tower (Scott et al., 2009). This includes downward shortwave, longwave,

386   precipitation, wind, air temperature, humidity, and pressure. Volumetric CO_2 concentration was

387   measured at a half-hourly interval through compact probes. The CO_2 efflux was estimated from

388   the gradient of CO_2 concentration measured at two depths of 2 cm and 10 cm through Fick's first

389   law of diffusion, and the estimates were validated against measurements from a portable CO_2 gas

390   analyzer.

391   The parameters estimated in this study include the parameters of the soil respiration models

392   (4C – 6C) and the parameters of the data models described in Section 2.1. The estimated

393   parameters of models 4C and 5C include the microbial carbon use efficiency (CUE) [g/g], enzyme

394   production rate, k_e [g/m^{3}s], microbial turnover rate, τ_m [1/s], and enzyme turnover rate τ_e [1/s].

395   Uniform distributions are used as the prior in the Bayesian inference, and the ranges of the four

396   parameters are 0.2 – 1.00, 1×10^{-12} – 1×10^{-7}, 1×10^{-12} – 1×10^{-5} and 1×10^{-11} – 1×10^{-6}, respectively.

397   The values of other parameters are fixed at the values used in Allison et al. (2010). Model 6C has

398   two more parameters, and they are the catalysis efficiency ε_D [-] and the turnover rate of the dry-

399   zone enzymes τ_{en} [1/s]. The prior of the two parameters are uniform distributions with the ranges

400   of 0.2 – 0.8 and 1×10^{-12} – 1×10^{-8}, respectively.

401   The DREAM-based MCMC simulation is conducted for a total of 24 cases, the combinations

402   of eight data models and three soil respiration models. For each case, the parameter distributions

403   are obtained after drawing a total of 5×10^5 samples using five Markov chains. The Gelman and

404   Rubin (1992) R-statistic is used for convergence diagnostic, and it approaches one in less than

405   40,000 samples. The initial 50% of the samples are discarded during the burn-in period.

406   **2.5   Metrics for evaluating predictive performance**

407   Three criteria are used to evaluate the predictive performance of the soil respiration models

408   and data models, and they are central mean tendency, dispersion, and reliability. Each criterion is

409   measured by a single metric. In addition, a newly defined metric by (Elshall et al., 2018) is also

410   used for simultaneously measuring the three criteria.

411   The central mean tendency is measured in this study using the Nash-Sutcliffe model efficiency

412   (NSME) coefficient (Nash and Sutcliffe, 1970),

413    $$NSME = 1 - \sum_{i=1}^{n}(d_i - \overline{\mathbf{Y}}_i)^2 \left/ \sum_{i=1}^{n}(d_i - \overline{\mathbf{d}})^2 \right. , \qquad (21)$$

414   where n is the number of cross-validation data, d_i is the i-th data, $\bar{\mathbf{d}}$ is the mean of the data, and

415     $\overline{\mathbf{Y}_i}$ is the mean of the prediction ensemble, $\mathbf{Y}_i$, for d_i. NSME ranges from $-\infty$ to 1, with $NSME = 1$

416   corresponding to a perfect match between data and mean prediction, i.e., the ensemble is centered

417   on the data. $NSME = 0$ indicates that the model predictions are as only accurate as the mean of the

418   data, while an efficiency $NSME < 1$ indicates that the mean of data is a better prediction than the

419   mean prediction.

420   In addition to the central mean tendency, it is also desirable that the ensemble is precise with

421   small dispersion and reliable to cover all the data. This study uses a nonparametric metric for

422   dispersion, and it is the sharpness of a prediction interval (e.g. Smith et al., 2010a)

423     $$Sharpness = 1/n \sum\nolimits_{i=1}^{n} \left[Max(\mathbf{Y}_i) - Min(\mathbf{Y}_i) \right] \tag{22}$$

424   where $\mathbf{Y}_i$ is the prediction ensemble within the 95% prediction interval, the Bayesian credible

425   interval, not the confidence interval used in nonlinear regression (Lu et al., 2013). Smaller values

426   of sharpness indicate better prediction precision. Reliability is measured using predictive coverage.

427   (e.g. Hoeting et al., 1999), which is the percentages of data contained in the prediction interval.

428   Larger predictive coverage values are preferred.

429   To account for the trade-off between the three metrics, Elshall et al. (2018) defined relative

430   model score (RMS) that simultaneously measure all the three criteria. Scoring rules are commonly

431   used in hydrology to assess predictive performance (e.g., Weijs et al., 2010; Westerberg et al.,

432   2011). RMS is used in this study to measure the relative predictive performance of the

433   combinations of soil respiration models and data models. For combination M_j, RMS is defined as

434 $$RMS(M_j) = \sum_{i=1}^{n} \frac{p(d_i \mid \mathbf{Y}_{ij}, M_j)}{\sum_{j=1}^{m} p(d_i \mid \mathbf{Y}_{ij}, M_j)} \times 100 \qquad (23)$$

435 where m is the number of combinations; the ensemble prediction $\mathbf{Y}_{ij}$ is similar to $\mathbf{Y}_i$ above with

436 index i over time and index j specific to the j-th combination. The density function $p(d_i|\mathbf{Y}_{ij})$ can be

437 evaluated by first obtaining the density function $p(\mathbf{Y}_{ij})$ of the ensemble prediction $\mathbf{Y}_{ij}$ (e.g., by using

438 the kernel density function) and then evaluating $p(d_i|\mathbf{Y}_{ij})$ using interpolation methods based on the

439 intersection of $\mathbf{Y}_{ij}$ and d_i. More details about evaluating RMS can be found in Elshall et al. (2018).

440 This evaluation is based purely on the model predictions, and does not involve any assumptions

441 on the models, their parameters, and likelihood functions. Larger RMS values indicate better

442 overall predictive performance. A figure of our workflow scheme is presented in the

443 supplementary materials.

444 **3   Results of Bayesian Inverse Modeling**

445 This section analyzes the residuals of the best realization (with the highest likelihood value) of

446 the MCMC simulation to understand whether the assumptions of the eight data models hold. The

447 impacts of the data models on the posterior parameter distributions are also analyzed.

448 **3.1   Residual characterization**

449 Figure 3 shows residual plots for model 6C based on data models SLS and WSEP-AC. SLS is

450 the simplest data model with the assumptions of homoscedastic, independent, and Gaussian

451 residuals, and the WSEP-AC is the most complex one without the assumptions. Model 6C is the

452 most complex model and also the best one as ranked by Zhang et al. (2014) using Bayesian model

453 selection. The variable a_t plotted in Figures 3a-3c and Figures 3d-3f is defined in equations (2) and

454 (11), respectively. Figures 3a – 3c show that all the three residual assumptions are violated when

455 SLS is used, because (i) the residual variance is not constant, but increases as a function of the

456   simulated CO_2 efflux (Figure 3a); (ii) the autocorrelation function at most lags is beyond the 95%

457   confidence interval (Figure 3b); (iii) the standard normal density function cannot adequately

458   characterize the residuals (Figure 3c). Figures 3d-f show that, after relaxing the three assumptions,

459   the processed residuals, a_t, can be well characterized by WSEP-AC. Figure 3d shows that, after

460   normalizing ε_t with the linear variance ($\sigma_t = 0.034 + 0.099 E_t$), the variation of the variance of

461   a_t becomes significantly smaller, although the variance is still not constant. Figure 3e shows that,

462   after removing a first-order autoregressive model from ε_t, a_t becomes less correlated, although the

463   correlation is not fully removed. The two coefficients of the autoregressive model are $\phi_1 = 0.989$

464   and $\phi_2 = 4.5 \times 10^{-6}$; the small value of ϕ_2 indicates that there is no need to attempt an autoregressive

465   model of higher order. Figure 3f shows that a_t follows the SEP distribution with the estimated

466   skewness coefficient of $\xi = 0.933$ and kurtosis coefficient of $\beta = 0.998$. As a summary, Figure

467   3 shows that it is important to examine the residuals and to determine whether the selected data

468   model is adequate for charactering the residuals. Although WSEP-AC still cannot perfectly

469   characterize ε_t, it is significantly better than SLS.

470   Although the Gaussian assumption used in SLS is violated for model 6C (Figure 3c), this is

471   not generally the case for other data models and soil respiration models. This is shown in Figure

472   4, which presents the quantile-quantile (Q-Q) plot for the eight data models and the three soil

473   respiration models. For SLS, WLS, SLS-AC, and WLS-AC, the theoretical quantiles are based on

474   the standard normal distribution, $N(0,1)$; for SEP, WSEP, SEP-AC, and WSEP-AC, the theoretical

475   quantiles are based on the standard skew exponential power distribution, $SEP(0,1,1,0)$. If the

476   residuals follow the assumed standard distributions, the Q-Q plots fall on the 1:1 lines, marked as

477   the theoretical lines in Figure 4. If the residuals are Gaussian or SEP but not standard, the Q-Q

478   plots fall on a straight line but not the 1:1 line. Figures 4a and 4e show that, for all the soil

479   respiration models, the Q-Q plots of SLS and SEP deviate significantly from the theoretical lines

480   and exhibit fat-tail behaviors, which is an indication of outliers (Thyer et al., 2009). The deviation

481   is reduced after accounting for autocorrelation in SLS-AC and SEP-AC, as shown in Figures 4c

482   and 4g. It is interesting to observe from the two figures that the Q-Q plots of the three models are

483   almost visually identical. The deviation is almost fully removed after accounting for

484   heteroscedasticity in WLS and WSEP in that their corresponding Q-Q plots fall on the 1:1 lines,

485   especially for models 5C and 6C, as shown in Figures 4b and 4f. However, the Q-Q plots start

486   deviating from the 1:1 lines as shown in Figures 4d and 4h, after accounting for both

487   heteroscedasticity and autocorrelation in WLS-AC and WSEP-AC. As a summary, Figure 4 shows

488   that, for the numerical example of this study, either the Gaussian or the SEP distribution is valid if

489   heteroscedasticity is accounted for in the data models. However, accounting for autocorrelation in

490   the data models does not help improve the characterization of the residual distributions.

491   **3.2   Posterior parameter distributions**

492   While Figures 3 and 4 help understand validity of the three assumptions used in the data

493   models, the impacts of the data models on estimating model parameter distributions must be

494   evaluated separately. This section discusses the impact of the data model selection on parameter

495   estimation with the objective of understanding whether incorrect specification of the data model

496   necessarily leads to biased parameter estimates. Such assessment is not a trivial task for two main

497   reasons. First, microbial soil respiration models aggregate complex natural processes and spatial

498   details into simpler conceptual representations. As a results several model parameters are effective

499   values of several complex natural processes that cannot be actually measured in the field as

500   discussed by Vrugt et al. (2013). In addition, even for model parameter that can be measured in

501   the field, since the model structure is imperfect, calibrated parameter values are sometimes beyond

502 their physically reasonable range, as discussed by Pappenberger and Beven (2006). This is often

503 undesirable, if we seek to make the models more mechanistically descriptive.

504   We focus our discussion on carbon use efficiency (CUE) for microbial growth due to two

505 reasons: (1) CUE is a fundamental parameter in microbial soil respiration models, and (2) a

506 physically reasonable range for CUE can estimated. The concept of microbial CUE(Allison et al.,

507 2010; Bradford et al., 2008; Manzoni et al., 2012; Wieder et al., 2013) has been used to present

508 fundamental microbial processes in recent microbial enzyme models (Allison et al., 2010; German

509 et al., 2011; Schimel and Weintraub, 2003; Wang et al., 2013). The microbial CUE, which is

510 marked between MIC and CO2 in Figure 1, controls microbial growth, enzyme production and

511 microbial respiration. A physically reasonable range of CUE can be estimated from the physical

512 viewpoint (Tang and Riley, 2014). Sinsabaugh et al. (2013) showed that the thermodynamic

513 calculations support a maximum CUE of 0.60 and that previous studies that estimate CUE in

514 terrestrial systems report a mean value of 0.55. Theoretically, there is no lower limit for CUE as it

515 can approach zero, and CUE< 0.1 has been reported for terrestrial ecosystems (e.g., Fernández-

516 Martínez et al., 2014) and used in modeling studies (Li et al., 2014). Note that, for inverse modeling

517 with MCMC sampling, we did not assume CUE maximum value of 0.6. In other words, for

518 parameter estimation and predictive performance we did not impose the constraint that CUE is less

519 than 0.6. We merely use this CUE maximum value of 0.6 to evaluate whether the posterior CUE

520 parameter samples obtained using different data models and different soil respiration models are

521 within the physically reasonable range of 0 ~ 0.6.

522   Figure 5 plots the CUE posterior marginal density of the three soil respiration models obtained

523 using the eight data models. The physical range between zero and 0.6 is marked in yellow. Figure

524 5 shows that the CUE posterior parameter distribution of Model 6C obtained using the data models

525   that does not account for autocorrelation are within the physically reasonable range. For models

526   4C and 5C, the posterior parameter samples are outside the range for six data models. For model

527   4C, the posterior parameters are within the physical range only for data models SEP and WSEP;

528   for model 5C, the two data models are WLS and WSEP. It is not surprising to find the posterior

529   parameter distribution of models 4C and 5C, which have a certain degree of model structure error,

530   to be out of the physically plausible range. This can be attributed to two reasons. First, the model

531   solution can be biased toward the missing processes in the model structure such as the additional

532   carbon pool in both 4C and 5C or missing the explicit accounting for soil moisture in 4C. Second,

533   biased parameter estimation can compensate for model structure inadequacy and other sources of

534   discrepancy in both the physical models and the data models.

535   In addition, it is important to understand how accounting for autocorrelation, heteroscedasticity

536   and non-Gaussian residuals can affect the parameter estimation. First, it is observed in Figure 5e-

537   h that biased parameter estimates are outside the physically reasonable range when autocorrelation

538   is explicitly accounted for. This may suggest again that accounting for heteroscedasticity is

539   desirable but accounting for autocorrelation is not. A possible reason is that filtering

540   autocorrelation may reduce the residual space such that the transformed residual space cannot

541   correspond to the parameter space of the models. In other words, parameter information may be

542   lost due to filtering out autocorrelation. However, it is not fully understood why this does not occur

543   for the model 6C under data model SLS-AC (Figure 5e), and more research is warranted. Second,

544   unlike accounting for auto-correlation, accounting only for heteroscedasticity (i.e., WLS and

545   WSEP) only amplifies or reduces the variance without affecting the structure of the residual space.

546   Figures 5c-d show that account for heteroscedasticity (i.e. WLS and WSEP) tends to improve the

547   parameter estimation in comparison with homoscedastic data models (i.e., SLS and SEP) shown

548   in Figure 5a-b. Finally, with respect to non-Gaussian residuals, Schoups and Vrugt (2010)

549   suggested that, compared to Gaussian pdf, the peaked pdf of the SEP with a longer tail is useful

550   for making parameter inference robust against outliers. To a certain degree, this can be

551   substantiated by the results in Figure 5a-d, in that SEP and WSEP provide more favorable

552   parameter estimates than SLS and WLS.

553   Finally, Figure 5a shows that the posterior parameter distributions of SLS are very narrow for

554   the three soil respiration models. The narrow distributions can be attributed to several reasons.

555   Since SEP distribution can have longer tails than Gaussian distribution, this can further increase

556   the samples acceptance ratio from tails resulting in wider distribution (Figure 5b). In addition,

557   accounting for heteroscedasticity will result in wider posterior parameter distribution (Figure 5c)

558   due to accepting higher variances at peak effluxes. Moreover, filtering correlation (Figure 5e-h)

559   increases the entropy, and leads to wider distributions.

560   **4.   Results of Predictive Performance**

561   Based on the last one third of the CO_2 efflux observations, a cross-validation test was

562   conducted for the combinations of three soil respiration models and eight data models. For the

563   cross-validation period, the predictive performance is examined using the four statistical metrics

564   that are defined in Section 2.5. The metrics are also calculated for the calibration period. This is

565   not to perform Bayesian model selection given the calibration data, but to better understand the

566   impact of data models on predictive performance of the three soil respiration models. For each

567   calibration and each cross-validation data, a prediction ensemble is generated from the two

568   perspectives of parametric uncertainty only and total uncertainty, as presented in Section 4.1 and

569   4.2, respectively.

570

4.1   Predictive performance with parametric uncertainty of soil respiration model

In this section the ensemble is generated by running the soil respiration models with the posterior samples (obtained from the Bayesian inference) of the physical model parameters. In other words, the ensemble addresses parametric uncertainty of the soil respiration models only. Considering the relative contribution of parametric uncertainty only will provide insights for modeling approaches that attempt to segregate various sources of uncertainty (e.g., Thyer et al., 2009 ; Tsai and Elshall, 2013). The four statistics above (i.e. NSME, sharpness, coverage, and RMS) are calculated for the three soil respiration models and the eight data models. Taking data models SLS and WSEP-AC as an example, Figure 6 plots the data (for the calibration and cross-validation periods separately) along with the mean and 95% credible intervals of the prediction ensemble for the three models.

Figure 6 shows that the data models affect model simulations for all the models. The statistics, especially RMS, indicate that WSEP-AC has better predictive performance than SLS. This is most visually obvious for model 6C during the cross-validation period after 330 days, as the prediction ensemble of SLS (Figure 6k) cannot cover the observations, whereas the prediction ensemble of WSEP-AC can (Figure 6l). This conclusion that WSEP-AC outperforms SLS agrees with that drawn from Figures 3 and 4.

Figure 7 plots the four statistics for all the soil respiration models and data models. Figures 7a and 7b show the predictive performance with respect to the central mean tendency measured by NSME for both the calibration and cross-validation periods respectively. The results indicate that, under all data models, the low fidelity model 4C over-fits the data and results in biased predictions, in that the NSME values become significantly worse (e.g., from 0.6 to -0.6) from the calibration to the cross-validation period. This is confirmed by the visual inspection of Figures 6a and 6g for

594 data model SLS and of Figures 6b and 6h for data model WSEP-AC. For models 5C and 6C, their

595 NSME values vary with the data models; and the central mean accuracy is the worst for SLS-AC

596 that considers only autocorrelation (Figure 6b).

597 With respect to parametric uncertainty estimation, Figures 7c and 7d show that sharpness

598 generally increases when the three assumptions in the data models are gradually relaxed from SLS

599 to WSEP-AC. This is even more obvious during the validation period. Given that the prediction

600 ensemble does not center on the data, the increasing sharpness is desirable as it improves

601 reliability. This is confirmed by the reliability plots in Figures 7e and 7f. The exceptions are again

602 for SLS-AC and SEP-AC that generally have the lowest coverage.

603 With respect to the overall predictive performance measured by RMS, the same variation

604 pattern and exception are also observed in the RMS plots in Figures 7g and 7h. This is not

605 surprising because RMS is the metric that can be used to measure all the three criteria (central

606 mean tendency, sharpness, and reliability). Since the prediction ensemble is not centered on the

607 data, the sharpness and reliability are the decisive factors for evaluating the predictive

608 performance.

609 As a summary, while it is necessary to account for heteroscedasticity in a data model, caution

610 is needed when accounting for autocorrelation in the manner described in Section 2.1. In addition,

611 after comparing the RMS values of the residuals using the Gaussian and SEP distributions, the

612 conclusion is that the SEP distribution outperforms the Gaussian distribution with respect to

613 predictive performance. Finally, uncertainty underestimation is evidenced by the very small

614 predictive coverage. The underestimation of uncertainty for all the physical models with all the

615 data model is not unexpected because only parametric uncertainty is considered in this study.

616 Considering the overall predictive uncertainty is the subject of the next section.

4.2 Predictive performance with total uncertainty

The simulated output $\mathbf{Y}(\boldsymbol{\theta}_p)$ is generally not equal to the observed output $\mathbf{d}$, and we have a residual term $\boldsymbol{\varepsilon}$ due to measurement, input and model structure errors such that $\mathbf{d} = \mathbf{Y}(\boldsymbol{\theta}_p) + \boldsymbol{\varepsilon}$. Accounting for the error term $\boldsymbol{\varepsilon}$ can be through separating various error terms. For example, in section 4.1 we obtained uncertainty due to the physical model parameters. Accounting for other sources of uncertainty can be done using a single model approach (e.g. Thyer et al., 2009) or a multi-model approach (e.g. Tsai and Elshall, 2013). Alternatively, we can quantify the uncertainty based on total residuals that separates out parametric uncertainty, so the residual error includes errors in measurements, model inputs, and model structures (e.g. Thyer et al., 2009; Schoups and Vrugt, 2010). This lumped approach is based on sampling the residuals model $\boldsymbol{\varepsilon}(\boldsymbol{\theta}_\varepsilon)$ with parameters $\boldsymbol{\theta}_\varepsilon$. SLS has one fixed parameter that is the constant variance, and other data models have two to six parameters. Thus in this section the prediction ensemble addresses parametric uncertainty of not only the soil respiration models but also the data models. When generating the prediction ensemble in the procedure described by Schoups and Vrugt (2010), an ensemble of residuals is first generated by running the data models with posterior samples of the data model parameters for the positive carbon efflux domain; the residual ensemble is then added to the prediction ensemble generated in Section 4.1.

We start by a visual assessment of the predictive performance. Figure 8 is similar to Figure 6 with the exception that Figure 8 considers the overall predictive uncertainty (i.e. parametric and output uncertainty), while Figure 6 considers the parametric uncertainty only. Figure 8 reveals a practical observation about accounting for the overall uncertainty through the lumped approach of sampling the data models. For example, Figure 8b shows that, despite the wide prediction interval of model 4C, the model with significant model structure error cannot capture the birch pulse around

640   day 180. It indicates that proper using a data model for model residuals cannot compensate

641   significant model structure error.

642   Figure 9 plots the four statistics (NSME, sharpness, predictive coverage, and RMS) of the three

643   soil respiration models under the eight data models to assess the predictive performance. With

644   respect to central mean tendency, the NSME values in Figures 9a-9b are visually the same as those

645   in Figures 7a-7b, indicating that the central mean accuracy under parametric uncertainty is the

646   same as that under predictive uncertainty.

647   With respect to uncertainty, the values of sharpness and predictive coverage increase

648   substantially (Figures 9c – 9f). In particular, Figures 9e and 9f show that, except for SLS and SEP,

649   the predictive coverage of the rest of the six data models are close to 100% for all the three soil

650   respiration models, indicating that the prediction intervals cover almost all the data. This is

651   demonstrated in Figures 6 for WSEP-AC. Similar to Figures 7c and 7d, Figures 9c and 9d also

652   show a general pattern that the sharpness increases when the three assumptions in the data models

653   are gradually relaxed from SLS to WSEP-AC. The data models that account for autocorrelation

654   are still the exceptions.

655   With respect to the overall predictive performance, the RMS values are largely determined by

656   the mean accuracy and sharpness as the predictive coverage is similar for different data models.

657   Figures 9g and 9h of RMS show that the predictive performance of the four data models that

658   account for autocorrelation is worse than that of the other four data models. This suggests again

659   that one needs to be cautious when building autocorrelation into a data model. This is consistent

660   with the finding of Evin et al. (2013, 2014) that accounting for autocorrelation before accounting

661   for heteroscedasticity or jointly accounting for autocorrelation and heteroscedasticity can result in

662   poor predictive performance. In summary, Figures 9g and 9h show for both the calibration and

663 prediction periods that accounting for heteroscedasticity in WLS and WSEP gives the best overall

664 predictive performance, and accounting for autocorrelation without heteroscedasticity in SLS-AC

665 and SEP-AC gives the worst overall predictive performance. Finally, for the three soil respiration

666 models, RMS shows that model 4C has the worst predictive performance for both the calibration

667 and cross-validation data. Generally speaking, the high fidelity model 6C outperforms model 5C

668 for both the calibration and cross-validation data, which justifies the complexity of model 6C.

669   To demonstrate the impacts of the data models on predictive performance of the soil respiration

670 models, Figure 10 plots the model simulations and predictions given by model 6C during the

671 calibration and cross-validation periods using all the eight data models. Figure 10 is used to

672 investigate predictive performance characteristics of the different data models. By examining the

673 predictive performance of model 6C, specific predictive performance patterns can be identified.

674 Figures 10a – 10d show that SLS and SEP have similar predictive performance with SEP generally

675 having better predictive performance especially during the validation period. Not accounting for

676 heteroscedasticity will underestimate the predication uncertainty (Figure 10b and Figure 10d). This

677 is mainly because the variance of the efflux residuals increases with the magnitude of the carbon

678 effluxes (Figure 3a), and thus assuming constant variance is not representative. Accordingly,

679 accounting for heteroscedasticity using WLS (Figure 10e) or WSEP (Figure 10h) will make the

680 predictions more sensitive to peck carbon effluxes. This will generally improve the predictive

681 coverage on the expense of sharpness and the central mean tendency. While WLS and WSEP have

682 similar predictive performance, WSEP has better central mean tendency and overall predictive

683 performance than WLS. Figures 10i – 10l show that accounting for autocorrelation using SLS-AC

684 and SEP-AC results in wider uncertainty bands and insensitivity to peak carbon effluxes as

685 compared to SLS and SEP (Figures 10a-d), which may be due to reduction of information content

686   of the residuals. This results in deteriorating the sharpness, the central mean tendency and the

687   capturing of peak carbon fluxes, especially during the validation period. Figures 10m – 10p show

688   that accounting for both heteroscedasticity and autocorrelation using WLS-AC and WSEP-AC

689   makes the inference robust against peak carbon effluxes. However, due to the loss of information

690   content, the uncertainty bands are still wider, and uncertainty becomes overestimated especially

691   during validation period as compared to WLS and WSEP (Figures 10e – 10h). The results of

692   Models 4C and 5C, which are not shown here, also show the same prediction patterns with respect

693   to non-Gaussian residuals, heteroscedasticity, and autocorrelation.

694   Finally, we observe in Figure 10 that the data models that have good overall predictive

695   performance as measured by RMS during the calibration period will maintain this good predictive

696   performance during the validation period. For model 6C, RMS values for the calibration and

697   validation periods are very well correlated with a correlation coefficient of 0.92. However, we note

698   that for models 4C and 5C the overall predictive performances during the calibration and validation

699   periods are not that well correlated as 6C, with correlation coefficients of 0.52 for model 4C and

700   0.61 for model 5C. This suggests that model 6C is more robust than 4C and 5C for forecasting and

701   hindcasting.

702   **4.3   Discussion on handling residual correlation**

703   Accounting for autocorrelation can lead to biased parameter estimation (Figure 5) and poor

704   predictive performance (Figure 10). Auto-correlated residuals may be attributed to model

705   discrepancy, as shown in Lu et al. (2013). The most obvious solution to handle the autocorrelation

706   is to reduce the autocorrelation by improving the soil respiration model. If model improvement is

707   difficult for practical reasons, we can improve the data model to better characterize the

708 autocorrelation. Addressing autocorrelation in a data model is challenging since it involves several

709 interlinked factors as follows:

710 (1) Non-stationarity due to wet-dry periods could be a reason for this problem. By drawing on

711 similarity from surface hydrology, the study of Ammann et al. (2018) suggests that auto-

712 correlated residuals might be attributed to non-stationarity due to wet-dry periods with half-

713 hourly data. Accounting for non-stationarity could better address the problem of auto-

714 correlated residuals (Ammann et al., 2018; Smith et al., 2010b).

715 (2) The way of implementing autocorrelation could have an impact. Autocorrelation could be

716 applied to raw residuals directly (e.g., Li et al., 2015), to transformed residuals based on

717 covariance matrix of residuals L($\mathbf{e}$) (e.g., Lu et al., 2013), or to normalized residuals L($\mathbf{a}$) (e.g.,

718 Schoups and Vrugt, 2010; Evin et al., 2013). Note that $\mathbf{e}$ is a vector of transformed residuals,

719 while $\mathbf{a}$ denotes a vector of independent and identically distributed random errors with zero

720 mean and unit standard deviation. The L($\mathbf{e}$) approach based on covariance matrix of residuals

721 is generally limited to Gaussian data models (e.g. Lu et al., 2013), while the L($\mathbf{a}$) approach for

722 normalized residuals can be readily adopted for non-Gaussian data models.

723 (3) The autocorrelation model could have an impact. Using an autoregressive model is a popular

724 technique to account for auto-correlated residuals. However, using an autoregressive model

725 with either joint inversion approach (e.g., this study and Schoups and Vrugt, 2010) or

726 sequential approaches (e.g., Evin et al., 2013, 2014; Lu et al., 2013) removes correlation errors

727 through a filter approach, which can lead to a loss of information content. As this may cause

728 overcorrection of prediction especially at surge events, Li et al. (2015) developed a restricted

729 autoregressive model to overcome this adverse effect. Other autocorrelation models include

730 moving average model and mixed autoregressive-moving averaging model (Chatfield, 2004).

731   (4) Joint versus sequential inversion for autocorrelation could have an impact. Sequential inversion

732   approaches include two-step procedures (e.g. Evin et al., 2013, 2014; Lu et al., 2013) or the

733   multi-step procedure (Li et al., 2016a). These sequential approach estimates the autoregressive

734   parameters sequentially in a later step after estimating the physical model parameters and other

735   data model parameters. Evin et al. (2013, 2014) used a sequential approach to avoid the

736   interaction between the parameters of the heteroscedasticity model and the autocorrelation

737   model. In addition, the autoregressive model parameters can be deterministically calculated as

738   an internal variables of the data model similar to Lu et al. (2013), and not as calibration

739   parameters (e.g. Schoups and Vrugt; Evin et al. 2013; 2014). While the first step in the

740   sequential approach would avoid the biased parameter estimation (Figure 10a-d), the second

741   step can still lead a poor predicative performance since we are essentially using a filter

742   approach to remove residual correlation. To address this problem, Li et al. (2016) multi-step

743   procedure that is based on Gaussian data model uses restricted autoregressive model.

744   Generally, Ammann et al. (2018) states that the joint inversion is still preferred, and

745   understanding the conditions where accounting for auto-correlation can be achieved remain

746   poorly understood.

747   **5.   Conclusions**

748   In parameter estimation and prediction of soil carbon fluxes to the atmosphere, one often

749   assumes that residuals, which include errors in observations, model inputs, parameter estimates,

750   and model structures, are normally distributed, homoscedastic and uncorrelated. We study these

751   assumptions by calibrating three soil inspiration models, which have varying degrees of model

752   structure errors. We further explore eight data models that characterize the residuals statistically

753   by starting with the standard least squares (SLS) and skew exponential power (SEP) data models

754   that assume homoscedastic and non-correlated residuals. For these two distributions, we evaluate

755   six other data models that account for heteroscedasticity (WLS and WSEP), autocorrelation (SLS-

756   AC and SEP-AC), and joint inversion of heteroscedasticity and autocorrelation (WLS-AC and

757   WSEP-AC). To our knowledge this is the first study that provides such detailed analysis for soil

758   reparation inverse modeling. We also use three soil respiration models with different degrees of

759   model fidelity (i.e., model discrepancy) and model complexity (i.e. number of model parameters)

760   to understand the impact of model discrepancy on the calibration results under different data

761   models. We analyze the results with respect to (1) residual characterization, (2) parameter

762   estimation, (3) predictive performance, and (4) impacts of model discrepancy. The main findings

763   of this study are summarized as follows:

764   (1) With respect to residual characterization, residual analysis results suggest that the common

765   assumption of not accounting for heteroscedasticity and residual autocorrelation in the data

766   models SLS and SEP results in poor characterization of residuals. Explicit accounting for

767   heteroscedasticity in WLS and WSEP results in significantly improved characterization of the

768   residuals, and the improvement is larger than that obtained by accounting for both

769   heteroscedasticity and autocorrelation in WSL-AC and WSEP-AC. Accounting for

770   autocorrelation only in SLS-AC and SEP-AC does not significantly improve the

771   characterization of the residuals.

772   (2) With respect to parameter estimation, the impacts of the data models are evaluated by focusing

773   on carbon use efficiency (CUE), which is a central parameter in soil respiration modeling.

774   Using SLS yields relatively reasonable posterior parameter distributions for CUE , yet very

775   narrow posterior. The data models SLS-AC, SEP-AC, WLS-AC and WSEP-AC that consider

776   autocorrelation tend to yield CUE estimates that are physically unreasonable. We speculate

777   that filtering residual correlation can affect the mapping of the model physics (as implicitly

778   included in the residuals) into the parameter space, which might result in biased parameter

779   estimates that are physically unreasonable.

780   (3) With respect to predictive performance, it is measured by four statistical criteria: central mean

781   tendency, sharpness, coverage, and relative model score for both the calibration and the cross-

782   validation periods. Results show that accounting for autocorrelation in SLS-AC, SEP-AC,

783   WLS-AC, and WSEP-AC deteriorates the predicative performance, such that the predictive

784   performance is inferior to that of SLS in terms of the central mean tendency and overall

785   predictive performance (measured by the relative model score), especially during the cross-

786   validation period. Results also indicates that using the SEP distribution can potentially improve

787   the predictive performance. The same is true for accounting for heteroscedasticity. Using SEP

788   distribution and accounting for heteroscedasticity (i.e. WSEP) can potentially improve the

789   predictive performance.

790   (4) With respect to the impact of model discrepancy, the high fidelity model (6C) gives the best

791   results with respect to parameter estimation and predictive performance. Model 6C generally

792   maintains its superior performance under different data models. This justifies the complexity

793   of model 6C relative to model 5C that has one less carbon pool. Model 4C with the lowest

794   fidelity maintains its poor performance for different data models, because the model has only

795   four carbon pools and lacks the explicit representation of soil moisture control.

796   Based on the empirical findings above, we conclude the following:

797   (1) Not accounting for heteroscedasticity and autocorrelation using a Gaussian or non-Gaussian

798   data model might not necessarily result in biased parameter estimates or biased predictions

799   with respect to central mean tendency, but will definitely underestimate uncertainty resulting

800   in lower overall predictive performance.

801   (2) Using a non-Gaussian data model can improve parameter estimation and predictive

802   performance with respect to central mean tendency and uncertainty quantification.

803   (3) Accounting for heteroscedasticity improves the uncertainty estimation with respect to

804   reliability at the cost of having a wider predictive interval.

805   (4) This study confirms other empirical findings and theoretical analyses (Evin et al., 2013; 2014;

806   Li et al., 2015, Ammann et al. 2018) that separately accounting for autocorrelation or jointly

807   accounting for autocorrelation and heteroscedasticity can be problematic. While the reasons

808   remain poorly understood (Ammann et al., 2018), it might be attributed to non-stationarity due

809   to wet-dry periods with half-hourly data (Ammann et al., 2018) or to the method of handling

810   autocorrelation (e.g., Schoups and Vrugt, 2010, Evin et al., 2013; 2014; Lu et al., 2013; Li et

811   al., 2015, 2016a; Ammann et al. 2018). Further investigation to address autocorrelation in soil

812   respiration modeling is warranted in a future study.

813   The above conclusions are subject to several limitations. First, the conclusions are specific to

814   the soil respiration models developed and validated for semi-arid savannah. Performance

815   variations across different soil respiration models with different levels of complexities is possible.

816   Second, the conclusions are conditioned on the data that were obtained at the half-hour interval

817   over a one-year period. Different conclusions are possible if the data are thinned to daily or weekly

818   scales or data of longer observation periods are used. Third, our study investigates effects of the

819   residual assumptions of formal likelihood functions through direct conditioning of the residuals

820   model parameters, yet this can also be done through other approaches such as residuals

821   transformation (Thiemann et al., 2001), autorgressive bias model (Del Giudice et al., 2013),

822   approximate Bayesian computation (Sadegh and Vrugt, 2013), and data assimliation (Spaaks and

823   Bouten, 2013). Comparing different methods for accounting the residual assumptions are beyond

824   the scope of this work. Fourth, this study focuses on formal Bayesian computation using formal

825   likelihood functions, and comparison with other inference functions such as informal likelihood

826   functions or approximate Bayesian computation is warranted in a future study.

827   Based on the aforesaid conclusions and limitations, we recommend to start calibrating soil

828   respiration models with simple SLS or SEP likelihood function. If the residuals characterization is

829   adequate (e.g., Scharnagl et al., 2011), then the underlying assumptions are met. Otherwise,

830   increase complexity of the data model until satisfactory results are obtained in terms of residuals

831   characterization, posterior parameter estimation, and predictive performance. This is similar to the

832   procedure given in Smith et al. (2015). Although the empirical findings of this study provide

833   general guidelines for data model selection for soil respiration modeling, more comparative studies

834   are needed to validate and refute the findings of this study.

835   **Acronyms**

836	4C	Four carbon pool model
837	5C	Five carbon pool model
838	6C	Six carbon pool model
839	CUE	Microbial carbon use efficiency
840	DOC	Dissolved organic carbon
841	ENZ	Enzymes
842	MCMC	Markov chain Monte Carlo
843	MIC	Microbial biomass
844	NSME	Nash-Sutcliffe model efficiency
845	PDF	Probability density function
846	RMS	Relative model score
847	SEP	Skew exponential power distribution
848	SEP-AC	Skew exponential power distribution with autocorrelation
849	SLS	Standard least square
850	SLS-AC	Standard least square with autocorrelation
851	SOC	Soil organic carbon
852	WLS	Weighted least squared
853	WLS-AC	Weight least square with autocorrelation

854   WSEP   Weighted skew exponential power distribution
855   WSEP-AC   Weighted skew exponential power distribution with autocorrelation
856

857 **Code and data availability**

858 The data and codes and models used to produce this paper are available on contact of the

859 corresponding author at mye@fsu.edu. We cannot publicly share the workflow because MT-

860 DREAM$_{(ZS)}$ code (Laloy and Vrugt, 2012) , which is a main component in the workflow, is in the

861 process of becoming a commercial code.

862 **Author contributions**

863 ASE developed and implemented the code for the eight data models for soil respiration modeling,

864 and prepared the manuscript with contribution of all co-authors. MY developed the research idea

865 and outline, and supervised the research implementation when ASE was a post-doc at Florida State

866 University. GN developed the soil respiration models. GAB collected and processed the eddy-

867 covariance data used for model calibration.

868 **Competing interests**

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

870 **Acknowledgement**

871 The first two authors were supported by the U.S. Department of Energy grant DE-SC0008272.

872 The first author was also partly supported by the U.S. National Science Foundation Award# OIA-

873 1557349. The second author was also partly supported by U.S. Department of Energy grant DE-

874 SC0019438 and U.S. National Science Foundation grant EAR-1552329. We thank two anonymous

875 reviewers for providing comments that helped to improve the manuscript.

876 **References**

877 Ahrens, B., Reichstein, M., Borken, W., Muhr, J., Trumbore, S. E. and Wutzler, T.: Bayesian

878   calibration of a soil organic carbon model using ΔC measurements of soil organic carbon

and heterotrophic respiration as joint constraints, Biogeosciences, 11(8), 2147–2168, doi:10.5194/bg-11-2147-2014, 2014.

Allison, S. D., Wallenstein, M. D. and Bradford, M. A.: Soil-carbon response to warming dependent on microbial physiology, Nat. Geosci., 3, 336 [online] Available from: http://dx.doi.org/10.1038/ngeo846, 2010.

Ammann, L., Reichert, P. and Fenicia, F.: A framework for likelihood functions of deterministic hydrological models, Hydrol. Earth Syst. Sci., (August), 2018.

Bagnara, M., Sottocornola, M., Cescatti, A., Minerbi, S., Montagnani, L., Gianelle, D. and Magnani, F.: Bayesian optimization of a light use ef fi ciency model for the estimation of daily gross primary productivity in a range of Italian forest ecosystems, Ecol. Modell., 306, 57–66, doi:10.1016/j.ecolmodel.2014.09.021, 2015.

Bagnara, M., Oijen, M. Van, Cameron, D., Gianelle, D., Magnani, F. and Sottocornola, M.: Bayesian calibration of simple forest models with multiplicative mathematical structure : A case study with two Light Use E ffi ciency models in an alpine forest, Ecol. Modell., 371(January), 90–100, doi:10.1016/j.ecolmodel.2018.01.014, 2018.

Barr, J. G., Engel, V., Fuentes, J. D., Fuller, D. O. and Kwon, H.: Modeling light use efficiency in a subtropical mangrove forest equipped with CO2eddy covariance, Biogeosciences, 10(3), 2145–2158, doi:10.5194/bg-10-2145-2013, 2013.

Barron-gafford, G. A., Cable, J. M., Bentley, L. P., Scott, R. L., Huxman, T. E., Jenerette, G. D. and Ogle, K.: Quantifying the timescales over which exogenous and endogenous conditions affect soil respiration, New Phytol., 2014.

Barron-Gafford, G. A., Scott, R. L., Jenerette, G. D. and Huxman, T. E.: The relative controls of temperature, soil moisture, and plant functional group on soil CO_2 efflux at

902   diel, seasonal, and annual scales, J. Geophys. Res. Biogeosciences, 116(1), 1–16,

903   doi:10.1029/2010JG001442, 2011.

904   Berryman, E. M., Frank, J. M., Massman, W. J. and Ryan, M. G.: Agricultural and Forest

905   Meteorology Using a Bayesian framework to account for advection in seven years of

906   snowpack CO 2 fl uxes in a mortality-impacted subalpine forest, Agric. For. Meteorol.,

907   249(April 2017), 420–433, doi:10.1016/j.agrformet.2017.11.004, 2018.

908   Box, G. E. P. and Tiao, G. C.: Bayesian inference in statistical analysis, Wiley., 1992.

909   Braakhekke, M. C., Beer, C., Schrumpf, M., Ekici, A., Ahrens, B., Hoosbeek, M. R., Kruijt, B.,

910   Kabat, P. and Reichstein, M.: The use of radiocarbon to constrain current and future soil

911   organic matter turnover and transport in a temperate forest, J. Geophys. Res.

912   Biogeosciences, 372–391, doi:10.1002/2013JG002420.Received, 2014.

913   Bradford, M. A., Davies, C. A., Frey, S. D., Maddox, T. R., Melillo, J. M., Mohan, J. E., Reynolds,

914   J. F., Treseder, K. K. and Wallenstein, M. D.: Thermal adaptation of soil microbial

915   respiration to elevated temperature, Ecol. Lett., 11(12), 1316–1327, doi:10.1111/j.1461-

916   0248.2008.01251.x, 2008.

917   Braswell, B. H., Sacks, W. J., Linder, E. and Schimel, D. S.: Estimating diurnal to annual

918   ecosystem parameters by synthesis of a carbon flux model with eddy covariance net

919   ecosystem exchange observations, Glob. Chang. Biol., 335–355, doi:10.1111/j.1365-

920   2486.2005.00897.x, 2015.

921   Cable, J. M., Ogle, K., Williams, D. G., Weltzin, J. F. and Huxman, T. E.: Soil Texture Drives

922   Responses of Soil Respiration to Precipitation Pulses in the Sonoran Desert : Implications

923   for Climate Change, Ecosystems, 961–979, doi:10.1007/s10021-008-9172-x, 2008.

924   Cable, J. M., Ogle, K., Lucas, R. W., Huxman, T. E., Loik, M. E., Smith, S. D., Tissue, D. T.,

925   Ewers, B. E., Pendall, E., Welker, J. M., Charlet, T. N., Cleary, M., Griffith, A., Nowak,

926   R. S., Rogers, M., Steltzer, H., Sullivan, P. F. and Gestel, N. C. Van: The temperature

927   responses of soil respiration in deserts : a seven desert synthesis, Biogeochemistry, 71–90,

928   doi:10.1007/s10533-010-9448-z, 2011.

929 Chatfield, C.: The analysis of time series : an introduction, Chapman & Hall/CRC. [online]

930   Available   from: https://www.crcpress.com/The-Analysis-of-Time-Series-An-

931   Introduction-Sixth-Edition/Chatfield/p/book/9781584883173 (Accessed 9 April 2019),

932   2004.

933 Chevallier, F. and O'Dell, C. W.: Error statistics of Bayesian CO2flux inversion schemes as seen

934   from GOSAT, Geophys. Res. Lett., 40(6), 1252–1256, doi:10.1002/grl.50228, 2013.

935 Correia, A. C., Minunno, F., Caldeira, M. C., Banza, J., Mateus, J., Carneiro, M., Wingate, L.,

936   Shvaleva, A., Ramos, A., Jongen, M., Bugalho, M. N., Nogueira, C., Lecomte, X. and

937   Pereira, J. S.: Agriculture , Ecosystems and Environment Soil water availability strongly

938   modulates soil CO 2 efflux in different Mediterranean ecosystems : Model calibration

939   using   the   Bayesian   approach,   Agric.   Ecosyst.   Environ.,   161,   88–100,

940   doi:10.1016/j.agee.2012.07.025, 2012.

941 Davidson, E. A. and Janssens, I. A.: Temperature sensitivity of soil carbon decomposition and

942   feedbacks   to   climate   change,   Nature,   440,   165   [online]   Available   from:

943   http://dx.doi.org/10.1038/nature04514, 2006.

944 Davidson, E. A., Samanta, S., Caramori, S. S. and Savage, K.: The Dual Arrhenius and Michaelis–

945   Menten kinetics model for decomposition of soil organic matter at hourly to seasonal time

946   scales, Glob. Chang. Biol., 18(1), 371–384, doi:10.1111/j.1365-2486.2011.02546.x, 2011.

947 Du, Z., Nie, Y., He, Y., Yu, G. and Wang, H.: Tellus B : Chemical and Physical Meteorology

948   Complementarity of flux- and biometric-based data to constrain parameters in a terrestrial

949   carbon model Complementarity of flux- and biometric-based data to constrain parameters

950   in a terrestrial carbon model, Tellus B Chem. Phys. Meteorol., 0889,

951   doi:10.3402/tellusb.v67.24102, 2015.

952 Du, Z., Zhou, X., Shao, J., Yu, G., Wang, H., Zhai, D., Xai, J. and Luo, Y.: Journal of Advances

953   in Modeling Earth Systems, J. Adv. Model. Earth Syst., 548–565,

954   doi:10.1002/2016MS000687.Received, 2017.

955 Elshall, A. S. and Tsai, F. T.-C.: Constructive epistemic modeling of groundwater flow with

956   geological structure and boundary condition uncertainty under the Bayesian paradigm, J.

957   Hydrol., 517, doi:10.1016/j.jhydrol.2014.05.027, 2014.

958 Elshall, A. S., Ye, M., Pei, Y., Zhang, F., Niu, G.-Y. and Barron-Gafford, G. A.: Relative model

959   score: a scoring rule for evaluating ensemble simulations with application to microbial soil

960   respiration modeling, Stoch. Environ. Res. Risk Assess., doi:10.1007/s00477-018-1592-3,

961   2018a.

962 Elshall, A. S., Ye, M., Pei, Y., Zhang, F., Niu, G.-Y. and Barron-Gafford, G. A.: Relative model

963   score: a scoring rule for evaluating ensemble simulations with application to microbial soil

964   respiration modeling, Stoch. Environ. Res. Risk Assess., 32(10), 2809–2819,

965   doi:10.1007/s00477-018-1592-3, 2018b.

966 Evin, G., Kavetski, D., Thyer, M. and Kuczera, G.: Pitfalls and improvements in the joint inference

967   of heteroscedasticity and autocorrelation in hydrological model calibration, Water Resour.

968   Res., 49(7), 4518–4524, doi:10.1002/wrcr.20284, 2013.

969 Evin, G., Thyer, M., Kavetski, D., McInerney, D. and Kuczera, G.: Comparison of joint versus

970   postprocessor approaches for hydrological uncertainty estimation accounting for error

971   autocorrelation and heteroscedasticity, Water Resour. Res., 50(3), 2350–2375,

972   doi:10.1002/2013WR014185, 2014.

973 Fernández-Martínez, M., Vicca, S., Janssens, I. A., Sardans, J., Luyssaert, S., Campioli, M.,

974   Chapin III, F. S., Ciais, P., Malhi, Y., Obersteiner, M., Papale, D., Piao, S. L., Reichstein,

975   M., Rodà, F. and Peñuelas, J.: Nutrient availability as the key regulator of global forest

976   carbon balance, Nat. Clim. Chang., 4, 471 [online] Available from:

977   http://dx.doi.org/10.1038/nclimate2177, 2014.

978 Gelman, A. and Rubin, D. B.: Inference from Iterative Simulation Using Multiple Sequences, Stat.

979   Sci., 7(4), 457–472, doi:10.1214/ss/1177011136, 1992.

980 German, D. P., Marcelo, K. R. B., Stone, M. M. and Allison, S. D.: The Michaelis–Menten kinetics

981   of soil extracellular enzymes in response to temperature: a cross-latitudinal study, Glob.

982   Chang. Biol., 18(4), 1468–1479, doi:10.1111/j.1365-2486.2011.02615.x, 2011.

983 Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P. and Rieckermann, J.:

984   Improving uncertainty estimation in urban hydrological modeling by statistically

985   describing bias, Hydrol. Earth Syst. Sci., 17(10), 4209–4225, doi:10.5194/hess-17-4209-

986   2013, 2013.

987 Gragne, A. S., Sharma, A., Mehrotra, R. and Alfredsen, K.: Improving real-time inflow forecasting

988   into hydropower reservoirs through a complementary modelling framework, Hydrol. Earth

989   Syst. Sci., 19(8), 3695–3714, doi:10.5194/hess-19-3695-2015, 2015.

990 Hararuk, O., Xia, J. and Luo, Y.: Evaluation and improvement of a global land model against soil

991   carbon data using a Bayesian Markov chain Monte Carlo method, J. Geophys. Res.

992   Biogeosciences, 119(3), 403–417, doi:10.1002/2013JG002535, 2014.

993 Hashimoto, S., Morishita, T., Sakata, T., Ishizuka, S., Kaneko, S. and Takahashi, M.: Simple

994   models for soil CO2, CH4, and N2O fluxes calibrated using a Bayesian approach and

995   multi-site data, Ecol. Modell., 222(7), 1283–1292, doi:10.1016/j.ecolmodel.2011.01.013,

996   2011.

997 He, H., Meyer, A., Jansson, P., Svensson, M., Rütting, T. and Klemedtsson, L.: Simulating

998   ectomycorrhiza in boreal forests : implementing ectomycorrhizal fungi model MYCOFON

999   in CoupModel (v5), Geosci. Model Dev., 725–751, 2018.

1000 Hilton, T. W., Davis, K. J. and Keller, K.: Evaluating terrestrial CO2flux diagnoses and

1001   uncertainties from a simple land surface model and its residuals, Biogeosciences, 11(2),

1002   217–235, doi:10.5194/bg-11-217-2014, 2014.

1003 Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T.: Bayesian model averaging: a

1004   tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by

1005   the authors, Stat. Sci., 14(4), 382–417, doi:10.1214/ss/1009212519, 1999.

1006 Högberg, P. and Read, D. J.: Towards a more plant physiological perspective on soil ecology,

1007   Trends Ecol. Evol., 21(10), 548–554, doi:10.1016/j.tree.2006.06.004, 2006.

1008 Hublart, P., Ruelland, D., De Cortázar-Atauri, I. G., Gascoin, S., Lhermitte, S. and Ibacache, A.:

1009   Reliability of lumped hydrological modeling in a semi-arid mountainous catchment facing

1010   water-use changes, Hydrol. Earth Syst. Sci., 20(9), 3691–3717, doi:10.5194/hess-20-3691-

1011   2016, 2016.

1012 Ishikura, K., Yamada, H., Toma, Y., Takakai, F., Darung, U., Limin, A. and Limin, S. H.: Soil

1013   Science and Plant Nutrition Effect of groundwater level fluctuation on soil respiration rate

1014   of tropical peatland in Central Kalimantan , Indonesia, Soil Sci. Plant Nutr., 63(1), 1–13,

1015   doi:10.1080/00380768.2016.1244652, 2017.

1016 Janssens, I. A., Freibauer, A., Ciais, P., Smith, P., Nabuurs, G.-J., Folberth, G., Schlamadinger, B.,

1017   Hutjes, R. W. A., Ceulemans, R., Schulze, E.-D., Valentini, R. and Dolman, A. J.: Europe's

1018   terrestrial biosphere absorbs 7 to 12% of European anthropogenic CO2 emissions.,

1019   Science, 300(5625), 1538–42, doi:10.1126/science.1083592, 2003.

1020 Katz, R. W., Craigmile, P. F., Guttorp, P., Haran, M., Sansó, B. and Stein, M. L.: Uncertainty

1021   analysis in climate change assessments, Nat. Clim. Chang., 3, 769 [online] Available from:

1022   http://dx.doi.org/10.1038/nclimate1980, 2013.

1023 Kavetski, D., Franks, S. W. and Kuczera, G.: Confronting Input Uncertainty in Environmental

1024   Modelling, Calibration Watershed Model., doi:doi:10.1029/WS006p0049, 2013.

1025 Keenan, T. F., Davidson, E., Moffat, A. M., Munger, W. and Richardson, A. D.: Using model-data

1026   fusion to interpret past trends, and quantify uncertainties in future projections, of terrestrial

1027   ecosystem carbon cycling, Glob. Chang. Biol., 18(8), 2555–2569, doi:10.1111/j.1365-

1028   2486.2012.02684.x, 2012.

1029 Kim, Y., Nishina, K., Chae, N., Park, S. J., Yoon, Y. J. and Lee, B. Y.: Constraint of soil moisture

1030   on CO 2 efflux from tundra lichen , moss , and tussock in Council , Alaska , using a

1031   hierarchical Bayesian model, Biogeosciences, 5567–5579, doi:10.5194/bg-11-5567-2014,

1032   2014.

1033 Klemedtsson, L., Jansson, P. E., Gustafsson, D., Karlberg, L., Weslien, P., Von Arnold, K.,

1034   Ernfors, M., Langvall, O. and Lindroth, A.: Bayesian calibration method used to elucidate

1035   carbon turnover in forest on drained organic soil, Biogeochemistry, 89(1), 61–79,

1036   doi:10.1007/s10533-007-9169-0, 2008.

1037 Laloy, E. and Vrugt, J. A.: High-dimensional posterior exploration of hydrologic models using

1038   multiple-try DREAM(ZS) and high-performance computing, Water Resour. Res., 48(1),

1039   doi:10.1029/2011WR010608, 2012.

1040 Li, J., Wang, G., Allison, S. D., Mayes, M. A. and Luo, Y.: Soil carbon sensitivity to temperature

1041 and carbon use efficiency compared across microbial-ecosystem models of varying

1042 complexity, Biogeochemistry, 119, 67–84 [online] Available from:

1043 http://www.jstor.org/stable/24716883, 2014.

1044 Li, M., Wang, Q. J., Bennett, J. C. and Robertson, D. E.: A strategy to overcome adverse effects

1045 of autoregressive updating of streamflow forecasts, Hydrol. Earth Syst. Sci., 19(1), 1–15,

1046 doi:10.5194/hess-19-1-2015, 2015.

1047 Li, M., Wang, Q. J., Bennett, J. C. and Robertson, D. E.: Error reduction and representation in

1048 stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting, Hydrol.

1049 Earth Syst. Sci., 20(9), 3561–3579, doi:10.5194/hess-20-3561-2016, 2016a.

1050 Li, Q., Xia, J., Shi, Z., Huang, K., Du, Z. and Lin, G.: Variation of parameters in a Flux-Based

1051 Ecosystem Model across 12 sites of terrestrial ecosystems in the conterminous USA, Ecol.

1052 Modell., 336, 57–69, doi:10.1016/j.ecolmodel.2016.05.016, 2016b.

1053 Lu, D., Ye, M., Meyer, P. D., Curtis, G. P., Shi, X., Niu, X.-F. and Yabusaki, S. B.: Effects of error

1054 covariance structure on estimation of model averaging weights and predictive performance,

1055 Water Resour. Res., 49(9), 6029–6047, doi:10.1002/wrcr.20441, 2013.

1056 Luo, Y., Ogle, K., Tucker, C., Fei, S., Gao, C., LaDeau, S., Clark, J. S. and Schimel, D. S.:

1057 Ecological forecasting and data assimilation in a data-rich era, Ecol. Appl., 21(5), 1429–

1058 1442, doi:10.1890/09-1275.1, 2011.

1059 Luo, Y., Keenan, T. F. and Smith, M.: Predictability of the terrestrial carbon cycle, Glob. Chang.

1060 Biol., 21(5), 1737–1751, doi:10.1111/gcb.12766, 2014.

1061 Manzoni, S., Taylor, P., Richter, A., Porporato, A. and Ågren, G. I.: Environmental and

1062 stoichiometric controls on microbial carbon-use efficiency in soils, New Phytol., 196(1),

79–91, doi:10.1111/j.1469-8137.2012.04225.x, 2012.

McInerney, D., Thyer, M., Kavetski, D., Lerat, J. and Kuczera, G.: Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors, Water Resour. Res., 53, 2199–2239, doi:10.1002/2016WR019168.Received, 2017.

Menichetti, L., Kätterer, T. and Leifeld, J.: Parametrization consequences of constraining soil organic matter models by total carbon and radiocarbon using long-term field data, Biogeosciences, 3003–3019, doi:10.5194/bg-13-3003-2016, 2016.

Nash, J. E. and Sutcliffe, J. V: River flow forecasting through conceptual models part I — A discussion of principles, J. Hydrol., 10(3), 282–290, doi:https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Ogle, K., Ryan, E., Dijkstra, F. A. and Pendall, E.: Journal of Geophysical Research: Biogeosciences, J. Geophys. Res. Biogeoscoences, 1–14, doi:10.1002/2016JG003385, 2016.

Pappenberger, F. and Beven, K. J.: Ignorance is bliss: Or seven reasons not to use uncertainty analysis, Water Resour. Res., 42(5), doi:10.1029/2005WR004820, 2006.

Peters, W., Jacobson, A. R., Sweeney, C., Andrews, A. E., Conway, T. J., Masarie, K., Miller, J. B., Bruhwiler, L. M. P., Pétron, G., Hirsch, A. I., Worthy, D. E. J., van der Werf, G. R., Randerson, J. T., Wennberg, P. O., Krol, M. C. and Tans, P. P.: An atmospheric perspective on North American carbon dioxide exchange: CarbonTracker., Proc. Natl. Acad. Sci. U. S. A., 104(48), 18925–30, doi:10.1073/pnas.0708986104, 2007.

Le Quéré, C., Peters, G. P., Andres, R. J., Andrew, R. M., Boden, T. A., Ciais, P., Friedlingstein, P., Houghton, R. A., Marland, G., Moriarty, R., Sitch, S., Tans, P., Arneth, A., Arvanitis,

A., Bakker, D. C. E., Bopp, L., Canadell, J. G., Chini, L. P., Doney, S. C., Harper, A., Harris, I., House, J. I., Jain, A. K., Jones, S. D., Kato, E., Keeling, R. F., Klein Goldewijk, K., Körtzinger, A., Koven, C., Lefèvre, N., Maignan, F., Omar, A., Ono, T., Park, G.-H., Pfeil, B., Poulter, B., Raupach, M. R., Regnier, P., Rödenbeck, C., Saito, S., Schwinger, J., Segschneider, J., Stocker, B. D., Takahashi, T., Tilbrook, B., van Heuven, S., Viovy, N., Wanninkhof, R., Wiltshire, A. and Zaehle, S.: Global carbon budget 2013, Earth Syst. Sci. Data, 6(1), 235–263, doi:10.5194/essd-6-235-2014, 2014.

Raich, J. W. J. J. W., Potter, C. S. C. and Bhagawati, D.: Interannual variability in global soil respiration, 1980-94, Glob. Chang. Biol., 8, 800–812, doi:10.1046/j.1365-2486.2002.00511.x, 2002.

Ren, X., He, H., Moore, D. J. P., Zhang, L., Liu, M., Li, F., Yu, G. and Wang, H.: Uncertainty analysis of modeled carbon and water fluxes in a subtropical coniferous plantation, J. Geophys. Res. Biogeosciences, 118(4), 1674–1688, doi:10.1002/2013JG002402, 2013.

Ricciuto, D. M., King, A. W., Dragoni, D. and Post, W. M.: Parameter and prediction uncertainty in an optimized terrestrial carbon cycle model: Effects of constraining variables and data record length, J. Geophys. Res. Biogeosciences, 116(1), 1–17, doi:10.1029/2010JG001400, 2011.

Richardson, A. D. and Hollinger, D. Y.: Statistical modeling of ecosystem respiration using eddy covariance data: Maximum likelihood parameter estimation, and Monte Carlo simulation of model and parameter uncertainty, applied to three simple models, Agric. For. Meteorol., 131(3–4), 191–208, doi:10.1016/j.agrformet.2005.05.008, 2005.

Sadegh, M. and Vrugt, J. A.: Bridging the gap between GLUE and formal statistical approaches: Approximate Bayesian computation, Hydrol. Earth Syst. Sci., 17(12), 4831–4850,

doi:10.5194/hess-17-4831-2013, 2013.

Scharnagl, B., Vrugt, J. A., Vereecken, H. and Herbst, M.: Inverse modelling of in situ soil water dynamics: Investigating the effect of different prior distributions of the soil hydraulic parameters, Hydrol. Earth Syst. Sci., 15(10), 3043–3059, doi:10.5194/hess-15-3043-2011, 2011.

Schimel, J. P. and Weintraub, M. N.: The implications of exoenzyme activity on microbial carbon and nitrogen limitation in soil: a theoretical model, Soil Biol. Biochem., 35(4), 549–563, doi:10.1016/S0038-0717(03)00015-4, 2003.

Schmidt, M. W. I., Torn, M. S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I. A., Kleber, M., Kögel-Knabner, I., Lehmann, J., Manning, D. A. C., Nannipieri, P., Rasse, D. P., Weiner, S. and Trumbore, S. E.: Persistence of soil organic matter as an ecosystem property, Nature, 478(7367), 49–56, doi:10.1038/nature10386, 2011.

Scholz, K., Hammerle, A., Hiltbrunner, E. and Wohlfahrt, G.: Analyzing the Effects of Growing Season Length on the Net Ecosystem Production of an Alpine Grassland Using Model – Data Fusion, Ecosystems, 21(5), 982–999, doi:10.1007/s10021-017-0201-5, 2018.

Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, Water Resour. Res., 46(10), 1–17, doi:10.1029/2009WR008933, 2010.

Scott, R. L., Jenerette, G. D., Potts, D. L. and Huxman, T. E.: Effects of seasonal drought on net carbon dioxide exchange from a woody-plant-encroached semiarid grassland, J. Geophys. Res. Biogeosciences, 114(4), doi:10.1029/2008JG000900, 2009.

Shi, X., Ye, M., Curtis, G. P., Miller, G. L., Meyer, P. D., Kohler, M., Yabusaki, S. and Wu, J.: Assessment of parametric uncertainty for groundwater reactive transport modeling, Water

1132    Resour. Res., 50(5), 4416–4439, doi:10.1002/2013WR013755, 2014.

1133 Sinsabaugh, R. L., Manzoni, S., Moorhead, D. L. and Richter, A.: Carbon use efficiency of

1134    microbial communities: stoichiometry, methodology and modelling, Ecol. Lett., 16(7),

1135    930–939, doi:10.1111/ele.12113, 2013.

1136 Smith, M. W., Bracken, L. J. and Cox, N. J.: Toward a dynamic representation of hydrological

1137    connectivity at the hillslope scale in semiarid areas, Water Resour. Res., 46(12),

1138    doi:10.1029/2009WR008496, 2010a.

1139 Smith, T., Sharma, A., Marshall, L., Mehrotra, R. and Sisson, S.: Development of a formal

1140    likelihood function for improved Bayesian inference of ephemeral catchments, Water

1141    Resour. Res., 46(12), 1–11, doi:10.1029/2010WR009514, 2010b.

1142 Smith, T., Marshall, L. and Sharma, A.: Modeling residual hydrologic errors with Bayesian

1143    inference, J. Hydrol., 528, 29–37, doi:10.1016/j.jhydrol.2015.05.051, 2015.

1144 Spaaks, J. H. and Bouten, W.: Resolving structural errors in a spatially distributed hydrologic

1145    model using ensemble Kalman filter state updates, Hydrol. Earth Syst. Sci., 17(9), 3455–

1146    3472, doi:10.5194/hess-17-3455-2013, 2013.

1147 Steinacher, M. and Joos, F.: Transient Earth system responses to cumulative carbon dioxide

1148    emissions: Linearities, uncertainties, and probabilities in an observation-constrained model

1149    ensemble, Biogeosciences, 13(4), 1071–1103, doi:10.5194/bg-13-1071-2016, 2016.

1150 Tang, J. and Riley, W. J.: Weaker soil carbon–climate feedbacks resulting from microbial and

1151    abiotic interactions, Nat. Clim. Chang., 5, 56 [online] Available from:

1152    http://dx.doi.org/10.1038/nclimate2438, 2014.

1153 Tang, J. and Zhuang, Q.: A global sensitivity analysis and Bayesian inference framework for

1154    improving the parameter estimation and prediction of a process-based Terrestrial

Ecosystem Model, J. Geophys. Res. Atmos., 114(D15), doi:10.1029/2009JD011724, 2009.

Thiemann, M., Trosset, M., Gupta, H. and Sorooshian, S.: Bayesian recursive parameter estimation for hydrologic models, Water Resour. Res., 37(10), 2521–2535, doi:10.1029/2000WR900405, 2001.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W. and Srikanthan, S.: Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, Water Resour. Res., 45(12), 1–22, doi:10.1029/2008WR006825, 2009.

Tiedeman, C. R. and Green, C. T.: Effect of correlated observation error on parameters, predictions, and uncertainty, Water Resour. Res., 49(10), 6339–6355, doi:10.1002/wrcr.20499, 2013.

Tsai, F. T.-C. and Elshall, A. S.: Hierarchical Bayesian model averaging for hydrostratigraphic modeling: Uncertainty segregation and comparative evaluation, Water Resour. Res., 49(9), doi:10.1002/wrcr.20428, 2013.

Tucker, C. L., Bell, J., Pendall, E. and Ogle, K.: Does declining carbon-use efficiency explain thermal acclimation of soil respiration with warming ?, Glob. Chang. Biol., 252–263, doi:10.1111/gcb.12036, 2013.

Tucker, C. L., Young, J. M., Williams, D. G. and Ogle, K.: Process-based isotope partitioning of winter soil respiration in a subalpine ecosystem reveals importance of rhizospheric respiration, Biogeochemistry, 121, 389–408 [online] Available from: http://www.jstor.org/stable/24717586, 2014.

Tuomi, M., Vanhala, P., Karhu, K., Fritze, H. and Liski, J.: Heterotrophic soil respiration-Comparison of different models describing its temperature dependence, Ecol. Modell.,

211(1–2), 182–190, doi:10.1016/j.ecolmodel.2007.09.003, 2008.

Vargas, R., Carbone, M. S., Reichstein, M. and Baldocchi, D. D.: Frontiers and challenges in soil respiration research: from measurements to model-data integration, Biogeochemistry, 102(1), 1–13, doi:10.1007/s10533-010-9462-1, 2011.

Vrugt, J. A. and Ter Braak, C. J. F.: DREAM(D): An adaptive Markov Chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems, Hydrol. Earth Syst. Sci., 15(12), 3701–3713, doi:10.5194/hess-15-3701-2011, 2011.

Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H. and Schoups, G.: Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications, Adv. Water Resour., 51, 457–478, doi:10.1016/j.advwatres.2012.04.002, 2013.

Wang, G., Post, W. M. and Mayes, M. A.: Development of microbial-enzyme-mediated decomposition model parameters through steady-state and dynamic analyses, Ecol. Appl., 23(1), 255–272, doi:10.1890/12-0681.1, 2013.

Weijs, S. V., Schoups, G. and Van De Giesen, N.: Why hydrological predictions should be evaluated using information theory, Hydrol. Earth Syst. Sci., 14(12), 2545–2558, doi:10.5194/hess-14-2545-2010, 2010.

Westerberg, I. K., Guerrero, J. L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E. and Xu, C. Y.: Calibration of hydrological models using flow-duration curves, Hydrol. Earth Syst. Sci., 15(7), 2205–2227, doi:10.5194/hess-15-2205-2011, 2011.

Wieder, W. R., Bonan, G. B. and Allison, S. D.: Global soil carbon projections are improved by modelling microbial processes, Nat. Clim. Chang., 3, 909 [online] Available from: http://dx.doi.org/10.1038/nclimate1951, 2013.

1201   Wieder, W. R., Allison, S. D., Davidson, E. A., Georgiou, K., Hararuk, O., He, Y., Hopkins, F.,

1202   Luo, Y., Smith, M. J., Sulman, B., Todd-Brown, K., Wang, Y.-P., Xia, J. and Xu, X.:

1203   Explicitly representing soil microbial processes in Earth system models, Global

1204   Biogeochem. Cycles, 29(10), 1782–1800, doi:10.1002/2015GB005188, 2015.

1205   Van Wijk, M. T., Van Putten, B., Hollinger, D. Y. and Richardson, A. D.: Comparison of different

1206   objective functions for parameterization of simple respiration models, J. Geophys. Res.

1207   Biogeosciences, 113(3), 1–11, doi:10.1029/2007JG000643, 2008.

1208   Xu, T., White, L., Hui, D. and Luo, Y.: Probabilistic inversion of a terrestrial ecosystem model:

1209   Analysis of uncertainty in parameter estimation and model prediction, Global Biogeochem.

1210   Cycles, 20(2), 1–15, doi:10.1029/2005GB002468, 2006.

1211   Xu, X., Schimel, J. P., Thornton, P. E., Song, X., Yuan, F. and Goswami, S.: Substrate and

1212   environmental controls on microbial assimilation of soil organic carbon: a framework for

1213   Earth system models, Ecol. Lett., 17(5), 547–555, doi:10.1111/ele.12254, 2014.

1214   Yeluripati, J. B., van Oijen, M., Wattenbach, M., Neftel, A., Ammann, A., Parton, W. J. and Smith,

1215   P.: Bayesian calibration as a tool for initialising the carbon pools of dynamic soil models,

1216   Soil Biol. Biochem., 41(12), 2579–2583, doi:10.1016/j.soilbio.2009.08.021, 2009.

1217   Yuan, W., Liang, S., Liu, S., Weng, E., Luo, Y. and Hollinger, D.: Improving model parameter

1218   estimation using coupling relationships between vegetation production and ecosystem

1219   respiration, Ecol. Modell., 240, 29–40, doi:10.1016/j.ecolmodel.2012.04.027, 2012.

1220   Yuan, W., Xu, W., Ma, M., Chen, S. and Liu, W.: Agricultural and Forest Meteorology Improved

1221   snow cover model in terrestrial ecosystem models over the Qinghai – Tibetan Plateau,

1222   Agric. For. Meteorol., 218–219, 161–170, doi:10.1016/j.agrformet.2015.12.004, 2016.

1223   Zhang, X., Niu, G.-Y., Elshall, A. S., Ye, M., Barron-Gafford, G. A. and Pavao-Zuckerman, M.:

1224   Assessing five evolving microbial enzyme models against field measurements from a

1225   semiarid savannah - What are the mechanisms of soil respiration pulses?, Geophys. Res.

1226   Lett., 41(18), doi:10.1002/2014GL061399, 2014.

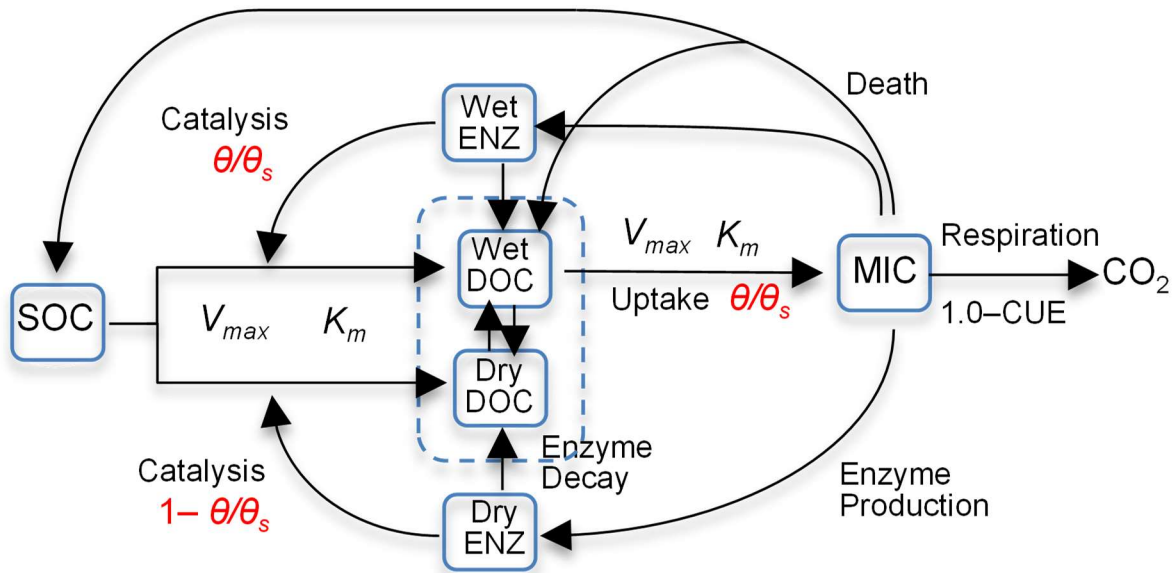
1227   Zhou, X., Luo, Y., Gao, C., Verburg, P. S. J., Arnone, J. A., Darrouzet-Nardi, A. and Schimel, D.

1228   S.: Concurrent and lagged impacts of an anomalously warm year on autotrophic and

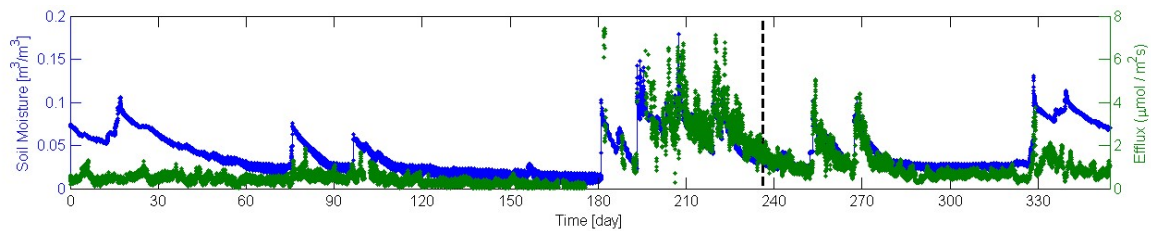
1229   heterotrophic components of soil respiration: A deconvolution analysis, New Phytol.,

1230   187(1), 184–198, doi:10.1111/j.1469-8137.2010.03256.x, 2010.

1231

Figure 1. Diagram of model 6C representing the processes of (1) degradation of soil organic carbon (SOC) to dissolved organic carbon (DOC) through catalysis of enzymes (ENZ) produced by microbes (MIC), (2) MIC uptake of DOC, and (3) microbial (MIC) respiration to produce CO_2 (CUE is the carbon use efficiency). SOC degradation and microbial uptake rates are controlled by water saturation (θ / θ_s). The DOC and ENZ pools are split into two subpools, one for the wet zone and the other for the dry zone of the soil pore space. Microbial uptake of DOC occurs only in the wet zone, and the uptake rate is linearly related to θ/θ_s. Catalysis through ENZ in the wet zone is proportional to θ/θ_s, while that in the dry zone is proportional to $1 - \theta/\theta_s$. V_{max} (s^{-1}) is the maximum rate, and K_m is the half-saturation concentration.

1244   Figure 2. Time series of soil moisture and efflux observations. The dashed line marks the divide

1245   of the dataset into calibration and validation periods.

1246

1247

1248   Figure 3. Residual analysis of the best realization (among multiple MCMC realizations) for model
1249   6C using data models (a-c) SLS and (d-f) WSEP-AC.

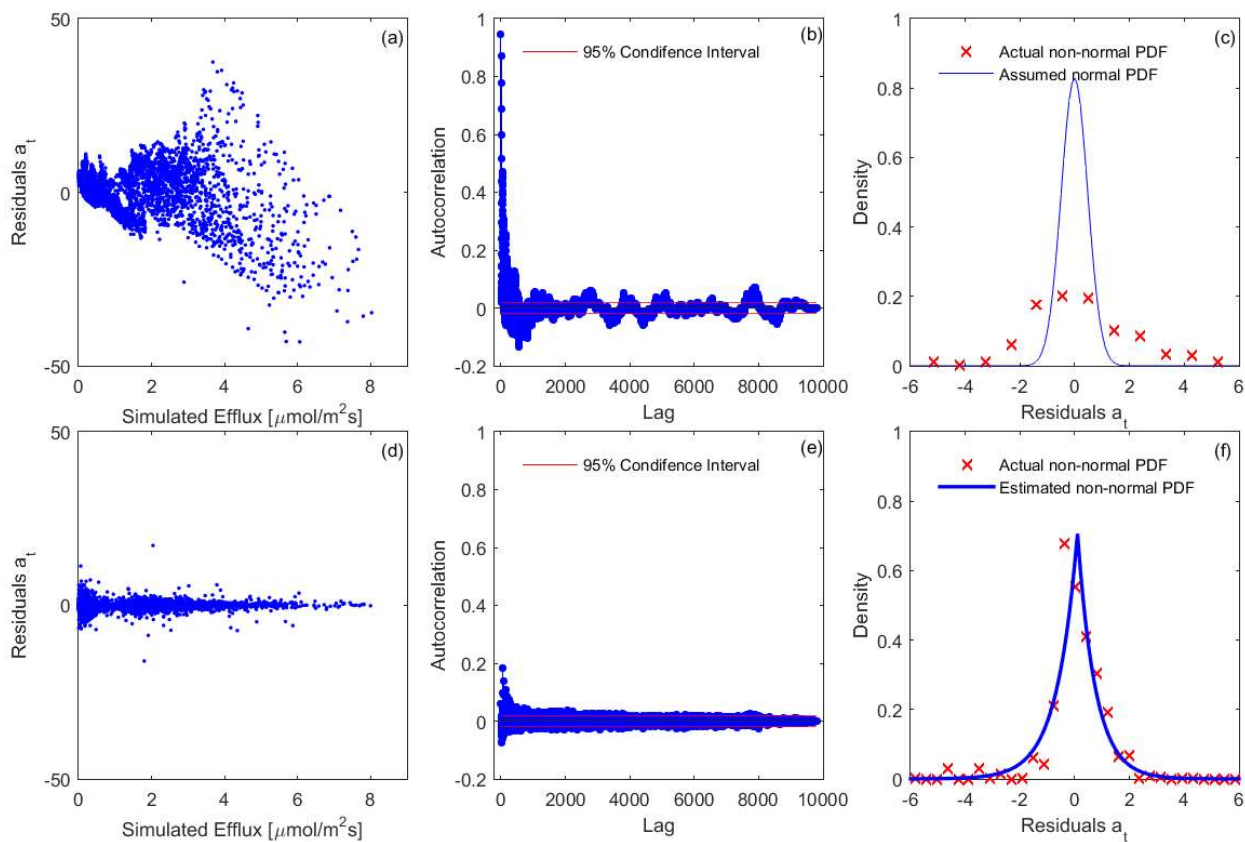
1250

1251

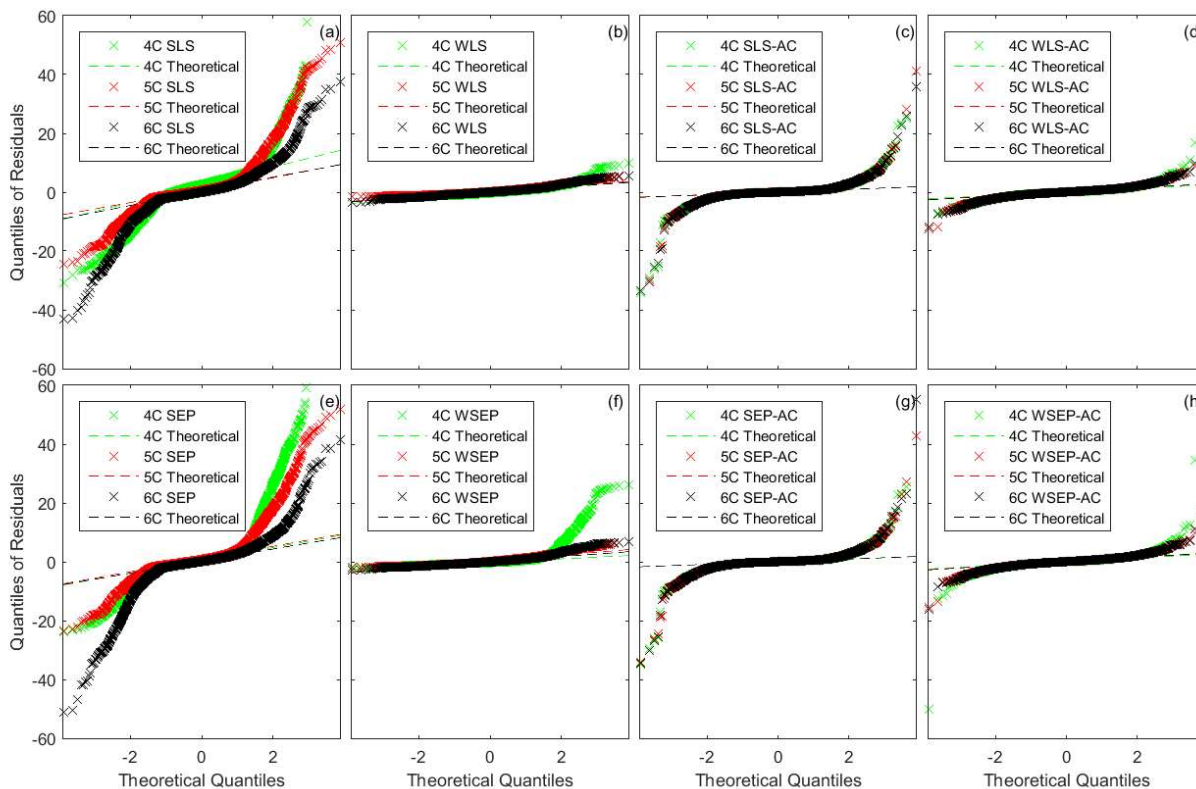
 Figure 4. Residual quantile-quantile (Q-Q) plots of the best realization (among multiple MCMC
1253 realizations) for the three soil respiration models and eight data models.
1254

1255

 Figure 5. Marginal posterior parameter density of carbon use efficiency (CUE) for the three soil
 respiration models and eight data models.

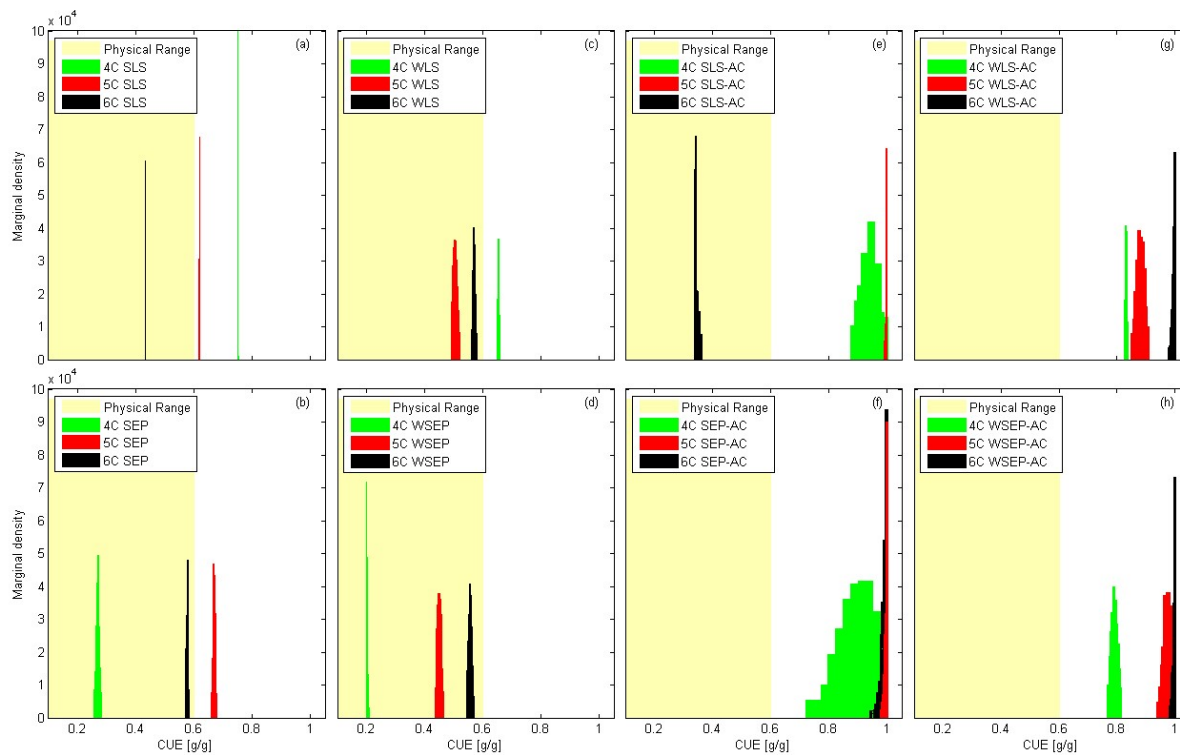
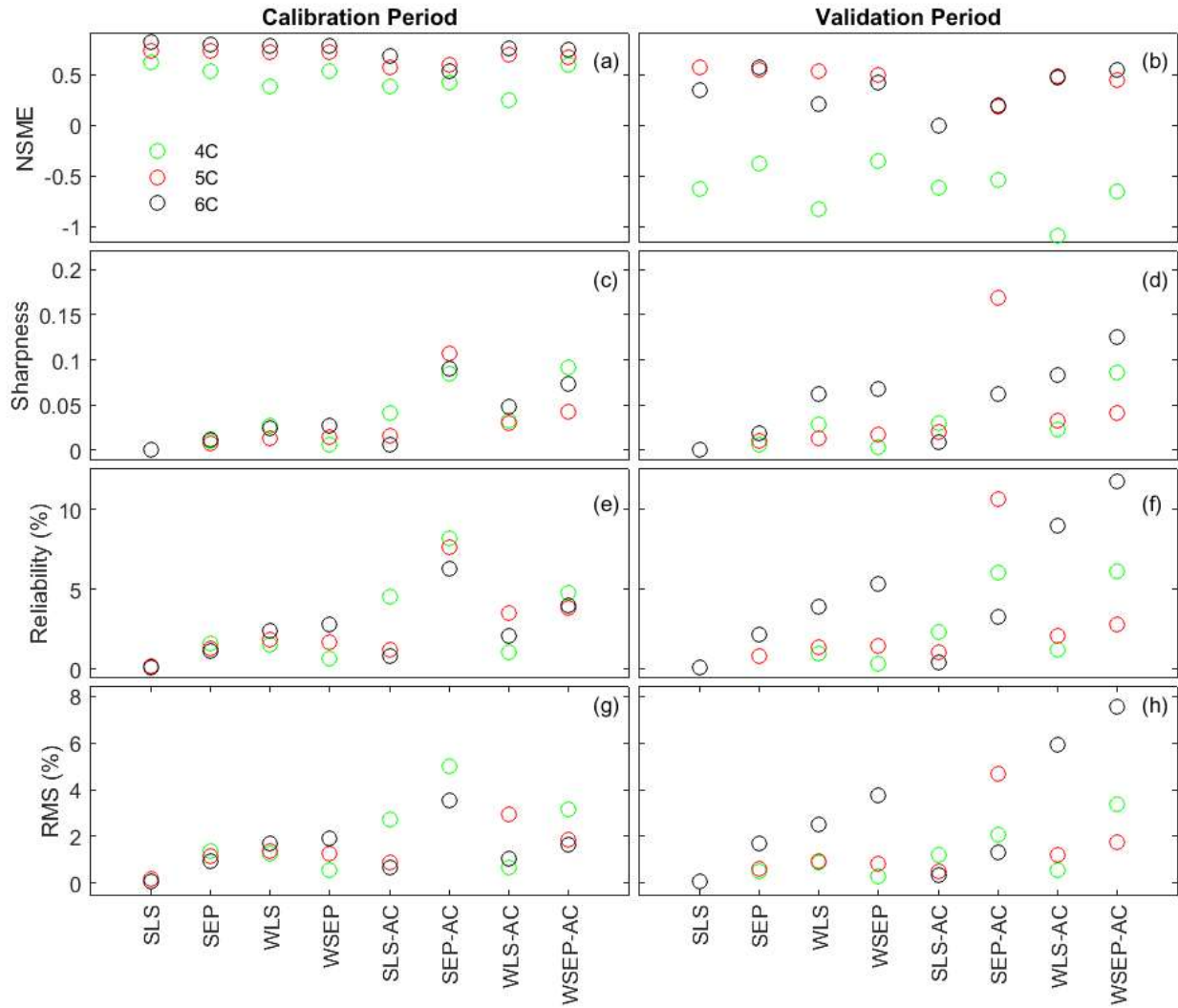



Figure 6. Observation data (blue dots) and mean prediction (green line) and 95% credible intervals (red line) of prediction ensembles for (a)-(f) the calibration period and (g)-(l) the validation period. The plots are for the three soil respiration models using data models SLS and WSEP-AC. *The prediction ensembles are generated to consider parametric uncertainty of the soil respiration models only.*

Figure 7. (a-b) Nash-Sutcliffe model efficiency (NSME), (c)-(d) sharpness, (e)-(f) predictive coverage, and (g)-(h) relative model score for measuring predictive performance of the three soil respiration models and the eight data models during the calibration and cross-validation periods. *The statistics are evaluated from the prediction ensembles generated to consider parametric uncertainty of the soil respiration models only.*

1275 Figure 8. Observation data (blue dots) and mean prediction (green line) and 95% credible intervals
1276 (red line) of prediction ensembles for (a)-(f) the calibration period and (g)-(l) the validation period.
1277 The plots are for the three soil respiration models using data models SLS and WSEP-AC. *The*
1278 *prediction ensembles are generated to consider parametric uncertainty of not only the soil*
1279 *respiration models but also the data models.*

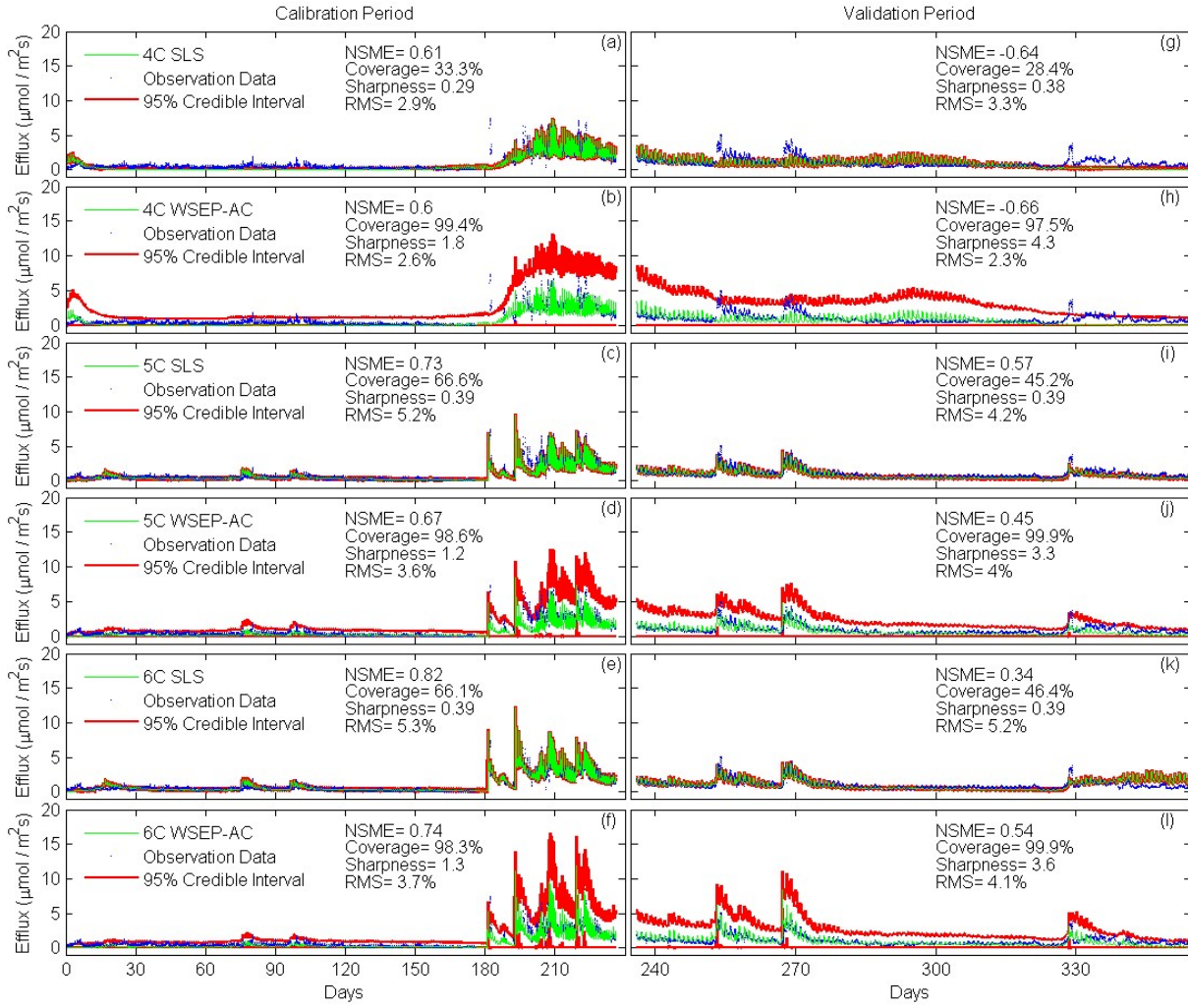
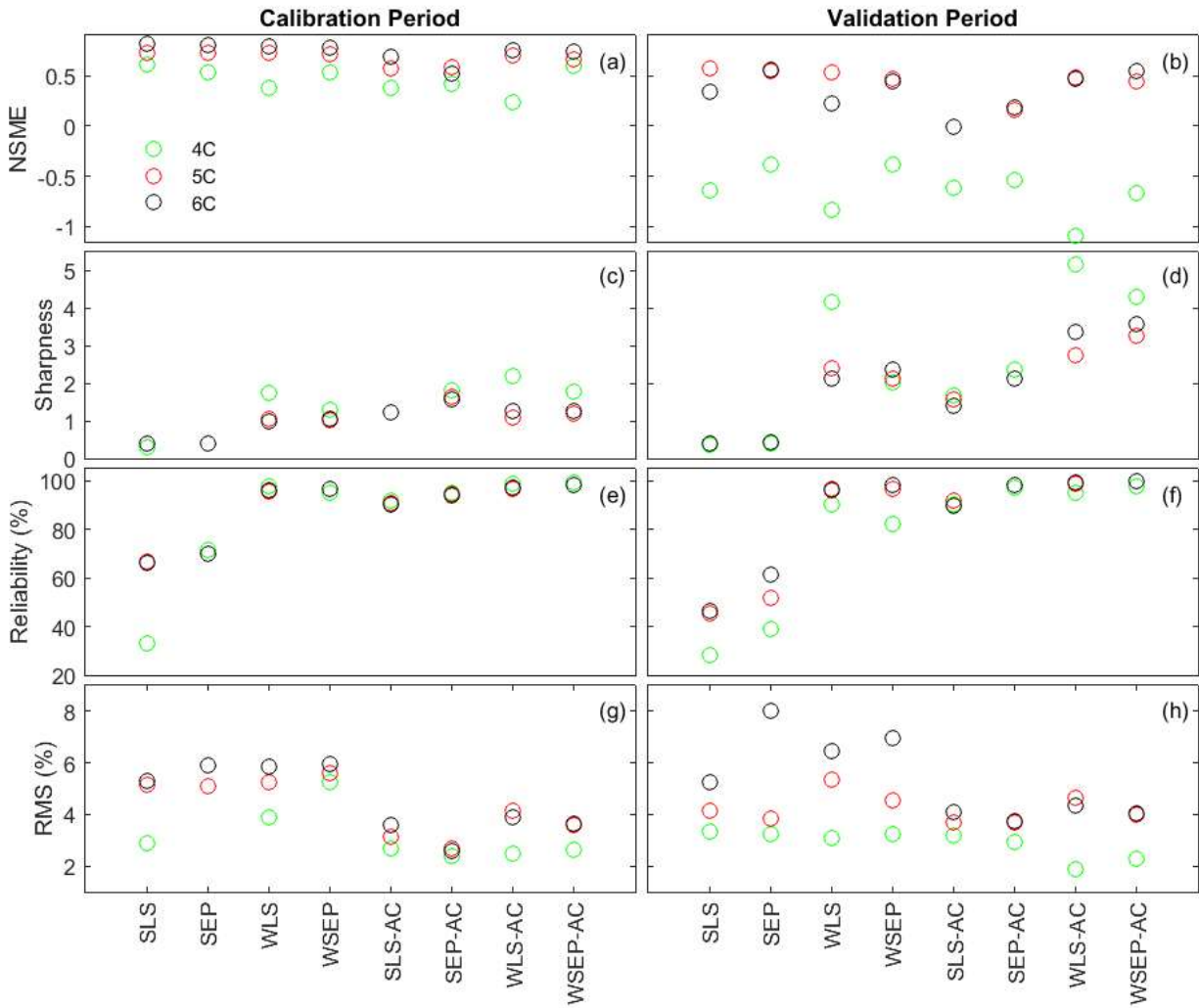
1280

1281

Figure 9. (a-b) Nash-Sutcliffe model efficiency (NSME), (c)-(d) sharpness, (e)-(f) predictive coverage, and (g)-(h) relative model score for measuring predictive performance of the three soil respiration models and the eight data models during the calibration and cross-validation periods. *The statistics are evaluated from the prediction ensembles generated to consider parametric uncertainty of not only the soil respiration models but also the data models.*

1289 Figure 10. Observation data (blue dots) and mean prediction (green line) and 95% credible
1290 intervals (red line) for 6C for the eight likelihood functions during the calibration period (a)-(h)
1291 and the validation period (i)-(p). *The prediction ensembles are generated to consider parametric*
1292 *uncertainty of not only the soil respiration models but also the data models.*
1293

1294